Spring 2017 – Epigenetics and Systems Biology
Discussion Session (Epigenetics)
Michael K. Skinner – Biol 476/576
Week 7 (February 23)

Epigenetics (History / Molecular Processes / Genomics)

Primary Papers

1. Haussmann, et al. (2016) Nature 540:301
2. Booth, et al. (2012) Science 336:934
3. Ernst, et al. (2011) Nature 473:43

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 chromatin marks were identified?
- What technology was used?
- How did the chromatin profiling correlate to gene expression?

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.

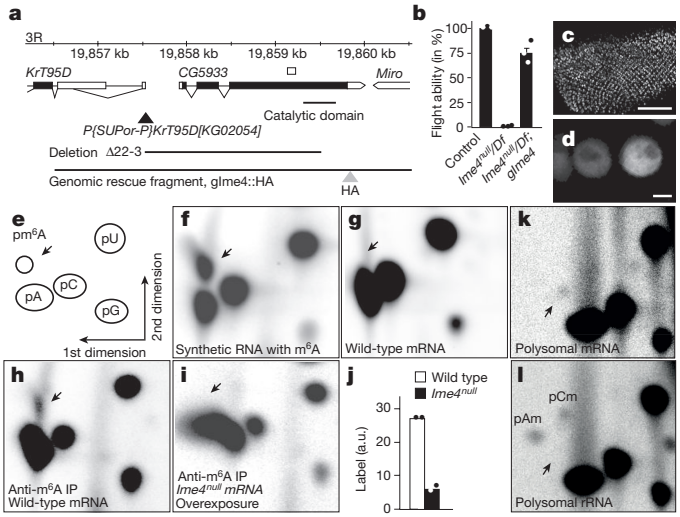


Figure 1 | Analysis of *Ime4* null mutants and m^{6}A methylation in *Drosophila*. **a**, Genomic organization of the *Ime4* locus depicting the transposon (black triangle) used to generate the deletion Δ22-3, which is a *Ime4* null allele and the hemaglutinin (HA)-tagged genomic rescue fragment. **b**, 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

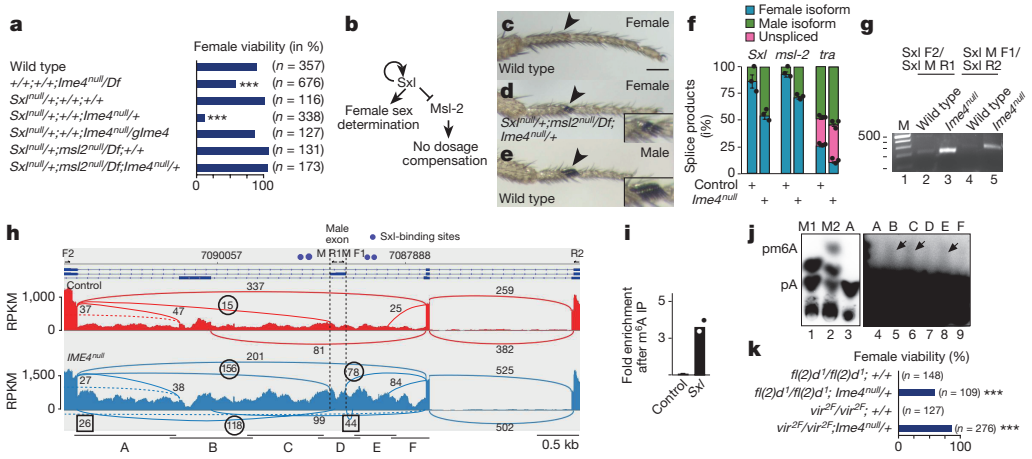


Figure 2 | m^{6}A methylation is required for *Sxl* alternative splicing in sex determination and dosage compensation. **a**, Female viability of indicated genotypes devoid of maternal m^{6}A (*n*, total number of flies, ***$P < 0.0001$). **b**, Schematic depicting Sxl control of female differentiation. **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*.

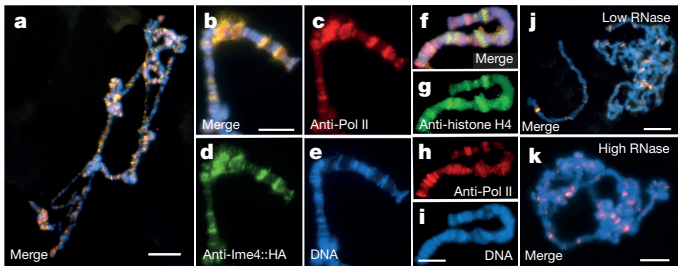


Figure 3 | Ime4 co-localizes to sites of transcription. **a**–**e**, Polytene chromosomes from salivary glands expressing IME::HA stained with anti-Pol II (red, **c**), anti-HA (green, **d**) and DAPI (DNA, blue, **e**), or merged (yellow, **a** and **b**). **f**–**i**, Polytene chromosomes stained with 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).

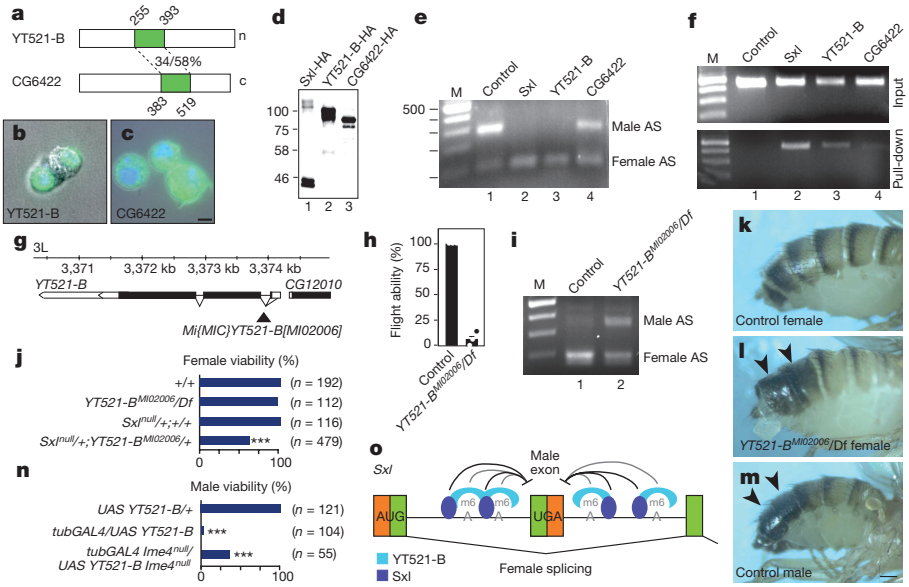


Figure 4 | YTH protein YT521-B decodes m^{6}A methylation in *Sxl*. **a**, Domain organization of *Drosophila* YTH proteins (YTH domain in green). n, nuclear; c, cytoplasmic. **b**–**d**, Cellular localization and size of HA-tagged YT521-B and CG6422 in S2 cells. Scale bar, 1 μm. **e**, Suppression of 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")
```

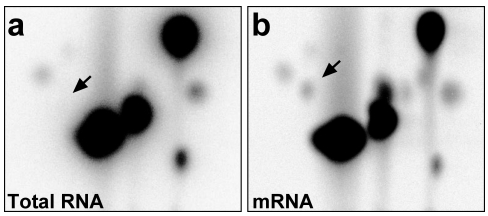
Polytene chromosome preparations and stainings. Ime4 or YT521-B were expressed in salivary glands with *C155-GAL4* from a *UAS* transgene. Larvae were grown at 18 °C under non-crowded conditions. Salivary glands were dissected in PBS containing 4% formaldehyde and 1% Triton X-100, and fixed for 5 min, 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.

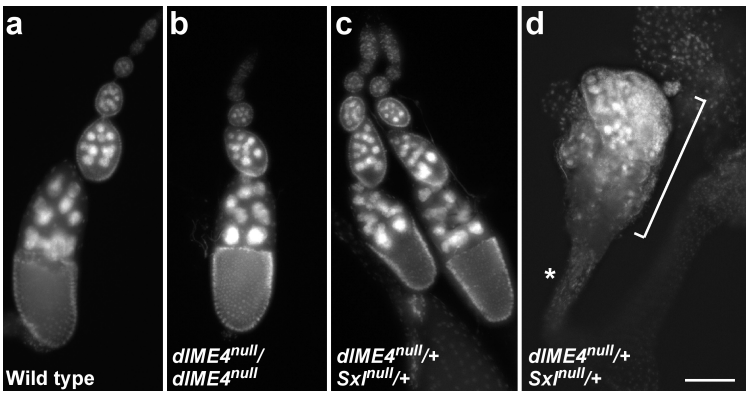
Data availability statement. RNA-seq data that support the findings of this study have been deposited at GEO under the accession number GSE79000, combining the single-end (GSE78999) and paired-end (GSE78992) experiments. All other data generated or analysed during this study are included in this published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

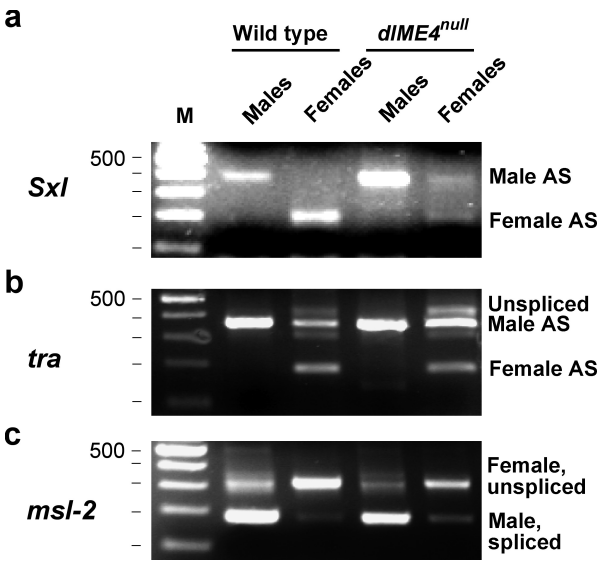
31. Haussmann, I. U., Li, M. & Soller, M. ELAV-mediated 3′-end processing of *ewg* transcripts is evolutionarily conserved despite sequence degeneration of the ELAV-binding site. *Genetics* **189,** 97–107 (2011).
32. Yan, D. & Perrimon, N. *spenito* is required for sex determination in *Drosophila melanogaster*. *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).



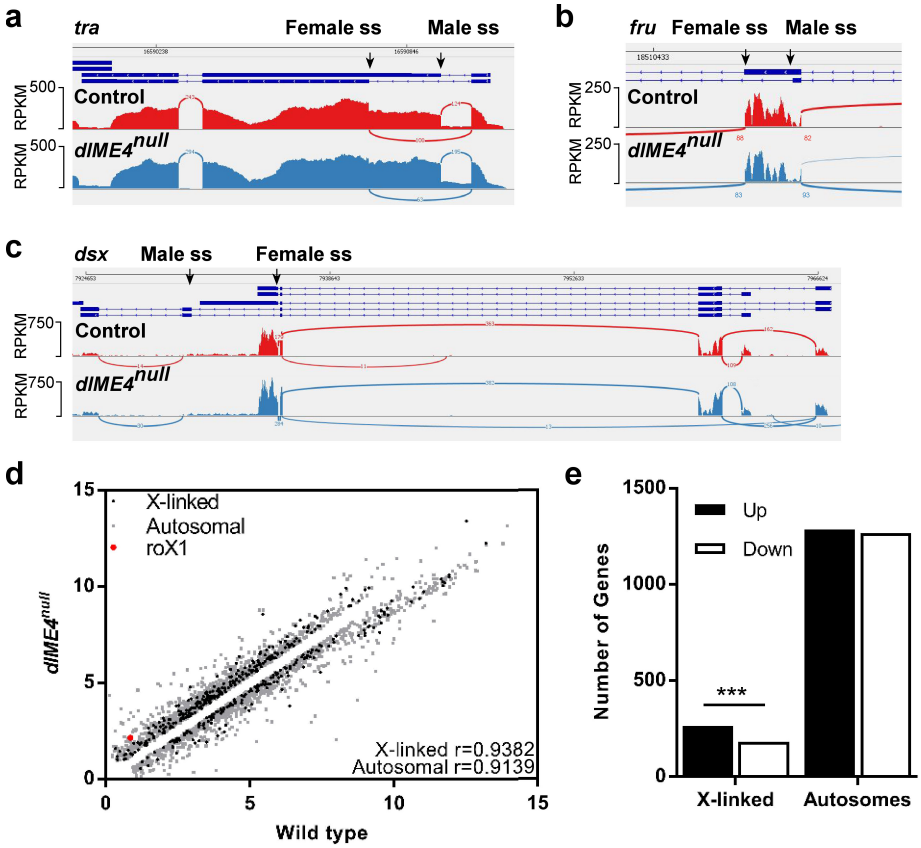
Extended Data Figure 1 | m^{6}A levels in unfertilized eggs. a, **b**, Thin-layer chromatography from maternal total RNA (**a**) and mRNA (**b**) present in unfertilized eggs. The arrow indicates m^{6}A.



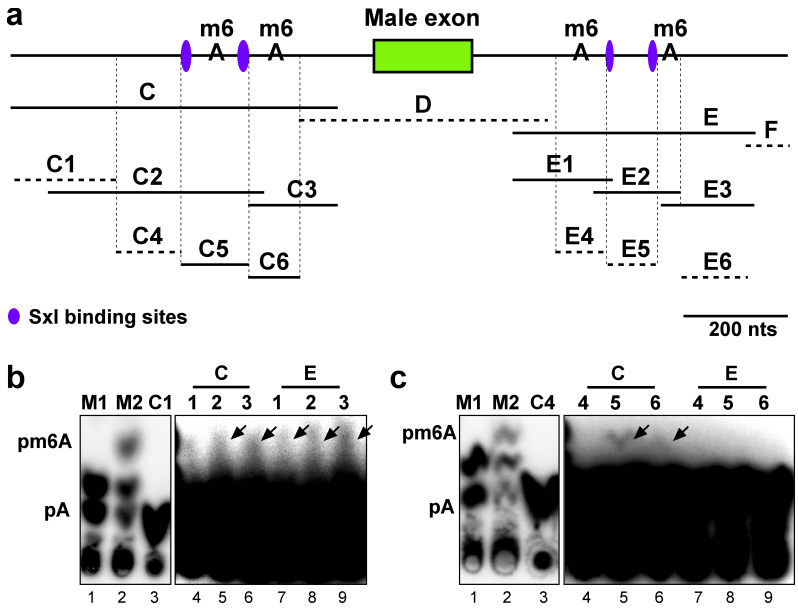
Extended Data Figure 2 | Ime4 supports Sxl in directing germline differentiation. a–**c**, Representative ovarioles of wild-type (**a**), *Ime4*^{null}/*Ime4*^{null} (**b**) and *Sxl*/+;*Ime4*^{null}/+ females (**c**), and a tumerous ovary of a *Sxl*/+;*Ime4*^{null}/+ female (**d**). The tumorous ovary 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).



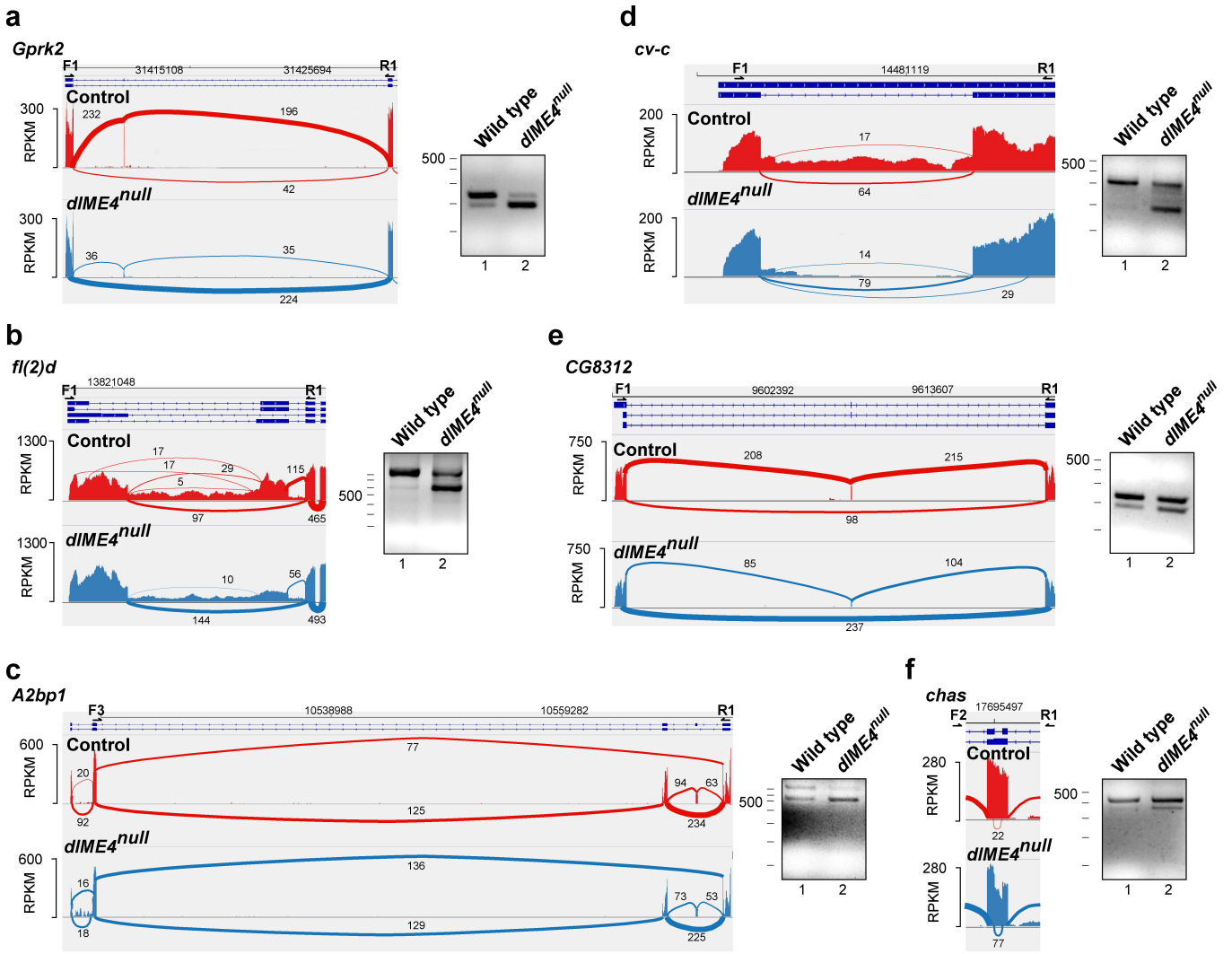
Extended Data Figure 3 | *Ime4* is required for female-specific splicing of *Sxl*, *tra* and *msl-2*. **a**–**c**, RT–PCR of *Sxl* (**a**), *tra* (**b**) and *msl-2* (**c**) sex-specific splicing in wild-type males and females, and *Ime4*^{null} males and females. 100-bp markers are shown on the left.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



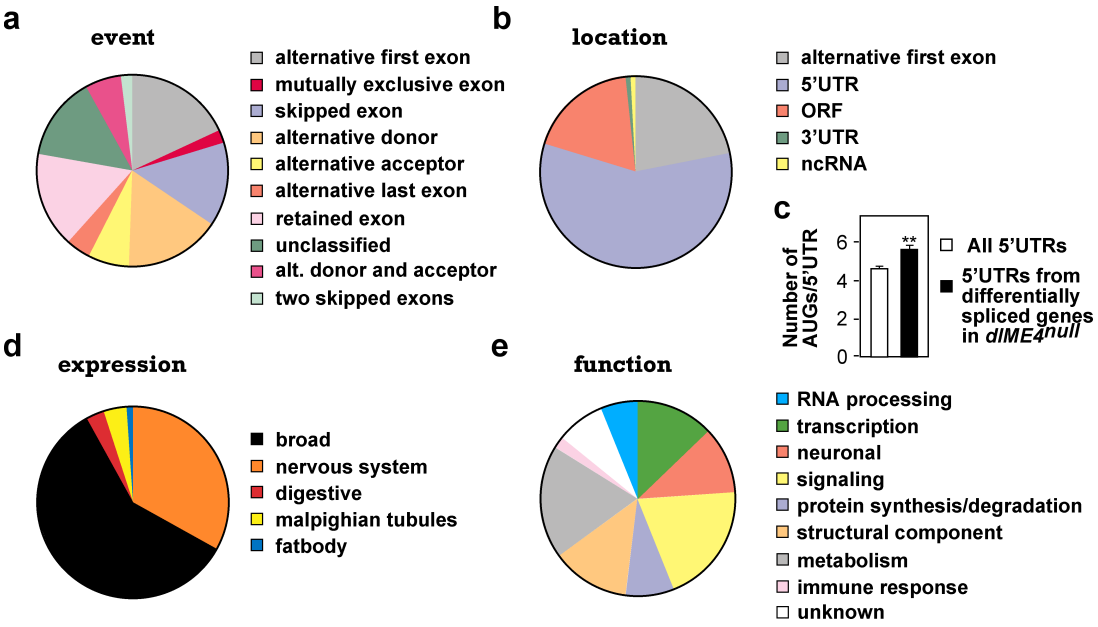
Extended Data Figure 4 | Alternative splicing of sex-determination genes and differential expression of X-linked genes in *Ime4*^{null} females. **a**–**c**, Sashimi plot depicting Tophat-mapped RNA sequencing reads and exon junction reads below the annotated gene model for sex-specific alternative splicing 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.



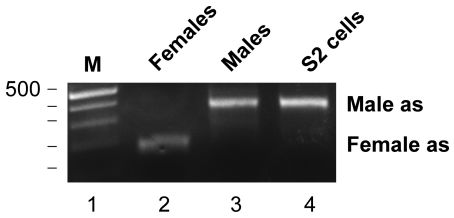
Extended Data Figure 5 | m^{6}A methylation sites map to the vicinity of *Sxl* binding sites. a, Schematic of the *Sxl* alternatively spliced intron around the male-specific exon depicting substrate RNAs used for *in vitro* m^{6}A methylation. Solid lines depict fragments containing m^{6}A methylation 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.



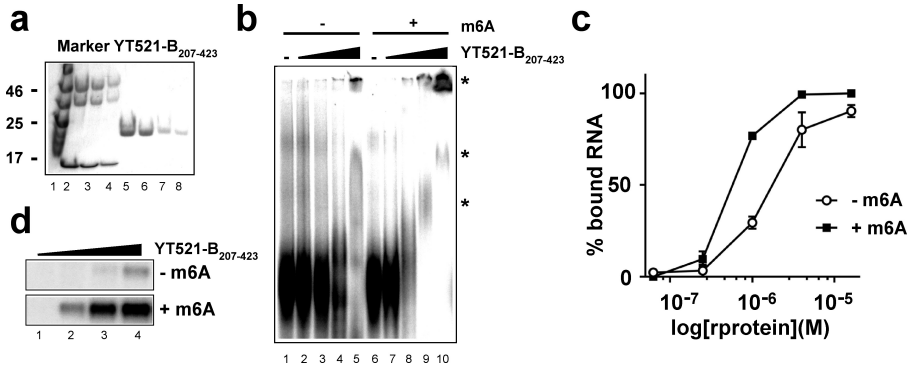
Extended Data Figure 6 | RT–PCR validation of differential alternative splicing in *Ime4*^{null} flies. a–f, Sashimi plots depicting Tophat-mapped RNA sequencing reads and exon junction reads below the annotated gene model of indicated genes on the left, and RT–PCR of alternative splicing shown on the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



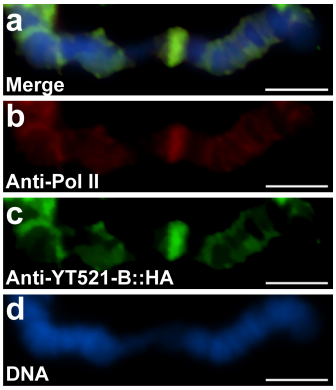
Extended Data Figure 7 | *Ime4* affects alternative splicing predominantly in 5′ UTRs in genes with a higher than average number of upstream start codons. a, **b**, Classification of differential alternative splicing in *Ime4*^{null} according to splicing event (**a**) and location of the event in the 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.)

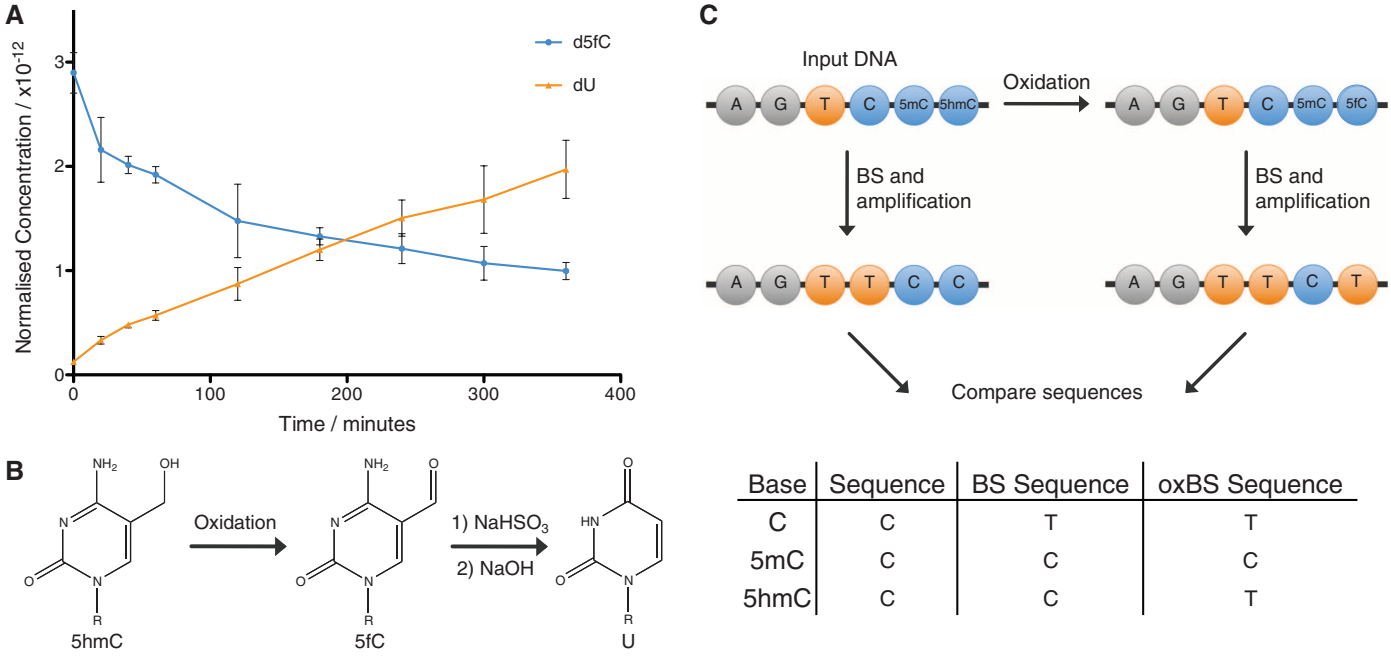


Fig. 1. A method for single-base resolution sequencing of 5hmC. (**A**) Reaction of 2′-deoxy-5-formylcytidine (d5fC) with $NaHSO_3$ (bisulfite) quenched by NaOH at different time points and then analyzed with high-performance liquid chromatography (HPLC). Data are mean ± SD of three replicates. (**B**) 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.

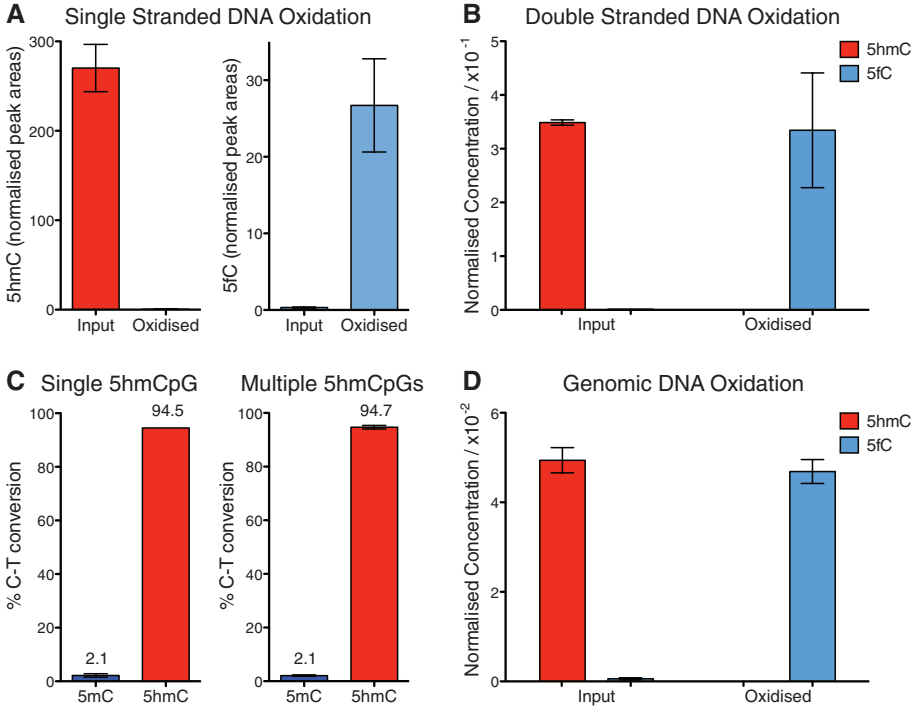


Fig. 2. Quantification of 5hmC oxidation. (**A**) Levels of 5hmC and 5fC (normalized to T) in a 15-nucleotide oligomer ssDNA oligonucleotide before and after $KRuO_4$ oxidation, measured with mass spectrometry. (**B**) Levels of 5hmC and 5fC (normalized to 5mC) in a 135-nucleotide oligomer dsDNA fragment 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

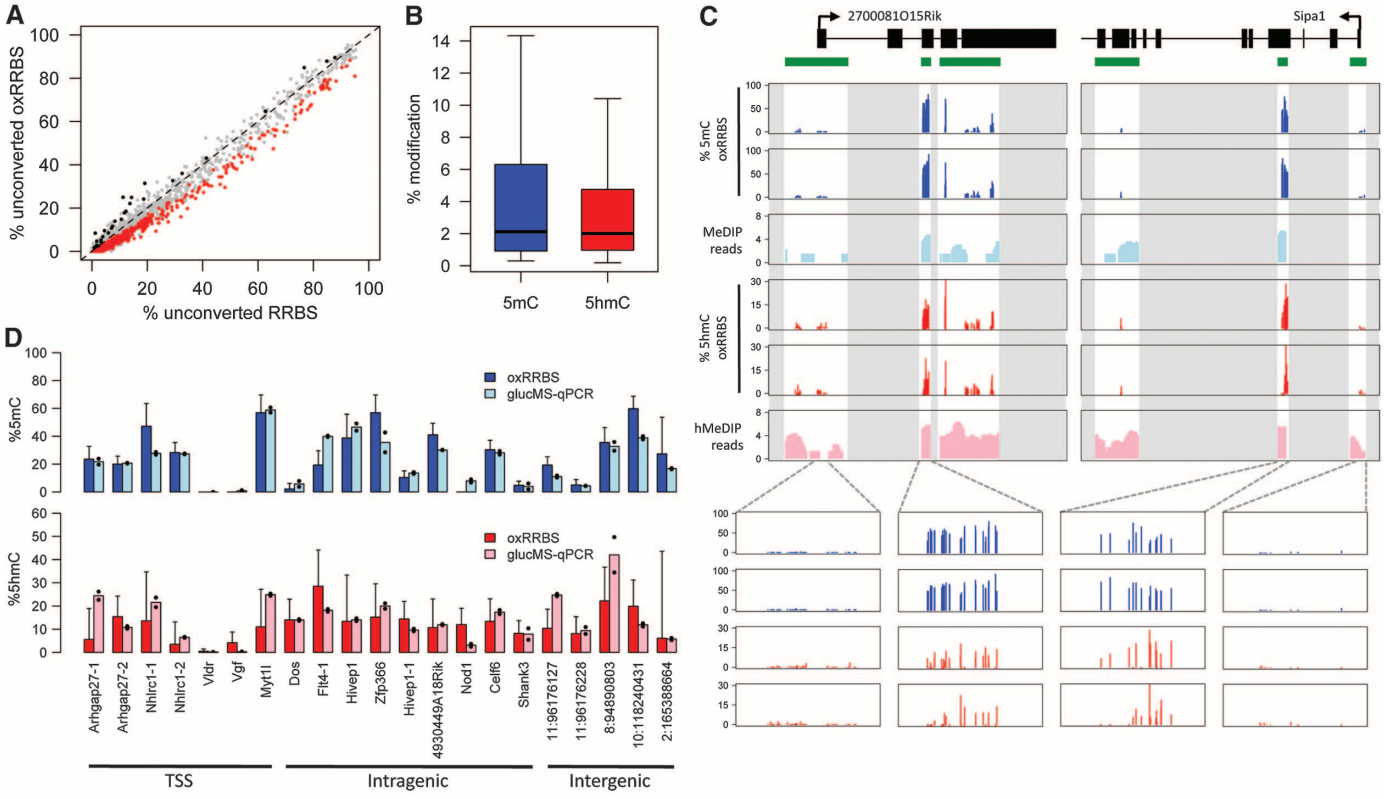


Fig. 3. Quantification of 5mC and 5hmC levels at CGIs by means of oxRRBS. (**A**) Fraction of unconverted cytosines per CGI; 5hmC-containing CGIs (red) have a statistically significant lower fraction in the oxRRBS data set; a false discovery rate of 3.7% was estimated from the CGIs with the opposite 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,

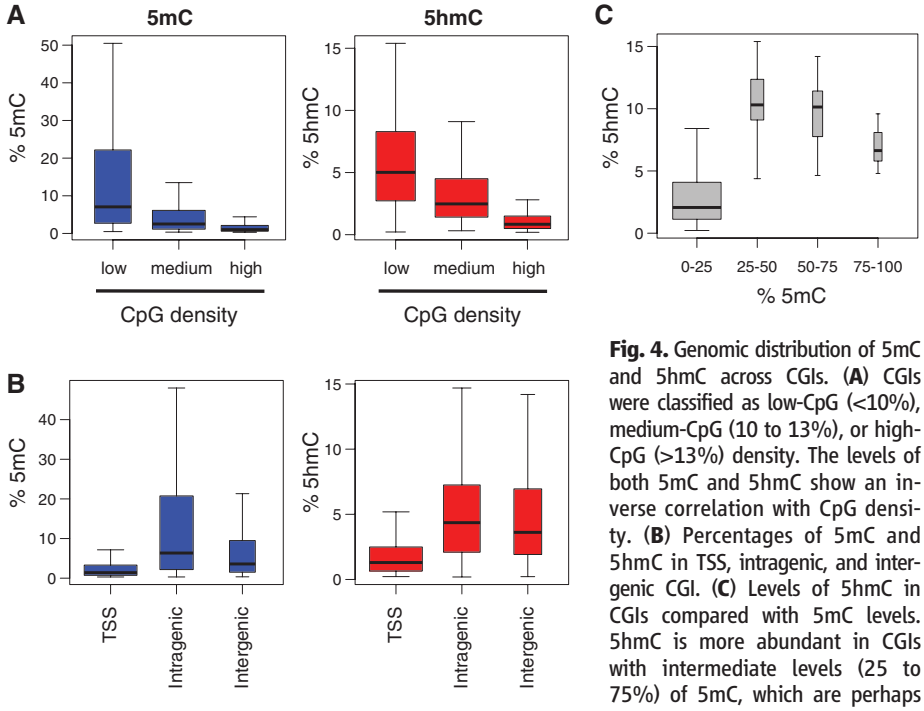


Fig. 4. Genomic distribution of 5mC and 5hmC across CGIs. (**A**) CGIs were classified as low-CpG (<10%), medium-CpG (10 to 13%), or high-CpG (>13%) density. The levels of both 5mC and 5hmC show an inverse correlation with CpG density. (**B**) Percentages of 5mC and 5hmC in TSS, intragenic, and intergenic 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

ARTICLE

doi:10.1038/nature09906

Mapping and analysis of chromatin state dynamics in nine human cell types

Jason Ernst[1,2], Pouya Kheradpour[1,2], Tarjei S. Mikkelsen[1], Noam Shoresh[1], Lucas D. Ward[1,2], Charles B. Epstein[1], Xiaolan Zhang[1], Li Wang[1], Robbyn Issner[1], Michael Coyne[1], Manching Ku[1,3,4], Timothy Durham[1], Manolis Kellis[1,2]* & Bradley E. Bernstein[1,3,4]*

Chromatin profiling has emerged as a powerful means of genome annotation and detection of regulatory activity. The approach is especially well suited to the characterization of non-coding portions of the genome, which critically contribute to cellular phenotypes yet remain largely uncharted. Here we map nine chromatin marks across nine cell types to systematically characterize regulatory elements, their cell-type specificities and their functional interactions. Focusing on cell-type-specific patterns of promoters and enhancers, we define multicell activity profiles for chromatin state, gene expression, regulatory motif enrichment and regulator expression. We use correlations between these profiles to link enhancers to putative target genes, and predict the cell-type-specific activators and repressors that modulate them. The resulting annotations and regulatory predictions have implications for the interpretation of genome-wide association studies. Top-scoring disease single nucleotide polymorphisms are frequently positioned within enhancer elements specifically active in relevant cell types, and in some cases affect a motif instance for a predicted regulator, thus suggesting a mechanism for the association. Our study presents a general framework for deciphering *cis*-regulatory connections and their roles in disease.

A major challenge in biology is understanding how a single genome can give rise to an organism comprising hundreds of distinct cell types. Much emphasis has been placed on the application of high-throughput tools to study interacting cellular components[1]. The field of systems biology has exploited dynamic gene expression patterns to reveal functional modules, pathways and networks[2]. Yet *cis*-regulatory elements, which may be equally dynamic, remain largely uncharted across cellular conditions.

Chromatin profiling provides a systematic means of detecting *cis*-regulatory elements, given the central role of chromatin in mediating regulatory signals and controlling DNA access, and the paucity of recognizable sequence signals. Specific histone modifications correlate with regulator binding, transcriptional initiation and elongation, enhancer activity and repression[1,3–6]. Combinations of modifications can provide even more precise insight into chromatin state[7,8].

Here we apply a high-throughput pipeline to map nine chromatin marks and input controls across nine cell types. We use recurrent combinations of marks to define 15 chromatin states corresponding to repressed, poised and active promoters, strong and weak enhancers, putative insulators, transcribed regions, and large-scale repressed and inactive domains. We use directed experiments to validate biochemical and functional distinctions between states.

The resulting chromatin state maps portray a highly dynamic landscape, with the specific patterns of change across cell types revealing strong correlations between interacting functional elements. We use correlated patterns of activity between chromatin state, gene expression and regulator activity to connect enhancers to likely target genes, to predict cell-type-specific activators and repressors, and to identify individual binding motifs responsible for these interactions.

Our results have implications for the interpretation of genome-wide association studies (GWASs). We find that disease variants frequently coincide with enhancer elements specific to a relevant cell type. In several cases, we can predict upstream regulators whose regulatory motif instances are affected or target genes whose expression may be altered, thereby suggesting specific mechanistic hypotheses for how disease-associated genotypes lead to the observed disease phenotypes.

Results

Systematic mapping of chromatin marks in multiple cell types

To explore chromatin state in a uniform way across multiple cell types, we applied a production pipeline for chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) to generate genome-wide chromatin data sets (Methods and Fig. 1a). We profiled nine human cell types, including common lines designated by the ENCODE consortium[1] and primary cell types. These consist of embryonic stem cells (H1 ES), erythrocytic leukaemia cells (K562), B-lymphoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblasts (NHLF), normal epidermal keratinocytes (NHEK) and mammary epithelial cells (HMEC).

We used antibodies for histone H3 lysine 4 trimethylation (H3K4me3), a modification associated with promoters[4,5,9]; H3K4me2 (dimethylation), associated with promoters and enhancers[1,3,6,9]; H3K4me1 (methylation), preferentially associated with enhancers[1,6]; lysine 9 acetylation (H3K9ac) and H3K27ac, associated with active regulatory regions[9,10]; H3K36me3 and H4K20me1, associated with transcribed regions[3–5]; H3K27me3, associated with Polycomb-repressed regions[3,4]; and CTCF, a sequence-specific insulator protein with diverse functions[11]. We validated each antibody by western blots and peptide competitions, and sequenced input controls for each cell type. We also collected data for H3K9me3, RNA polymerase II (RNAPII) and H2A.Z (also known as H2AFZ) in a subset of cells.

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [3]Howard Hughes Medical Institute, Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [4]Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
*These authors contributed equally to this work.

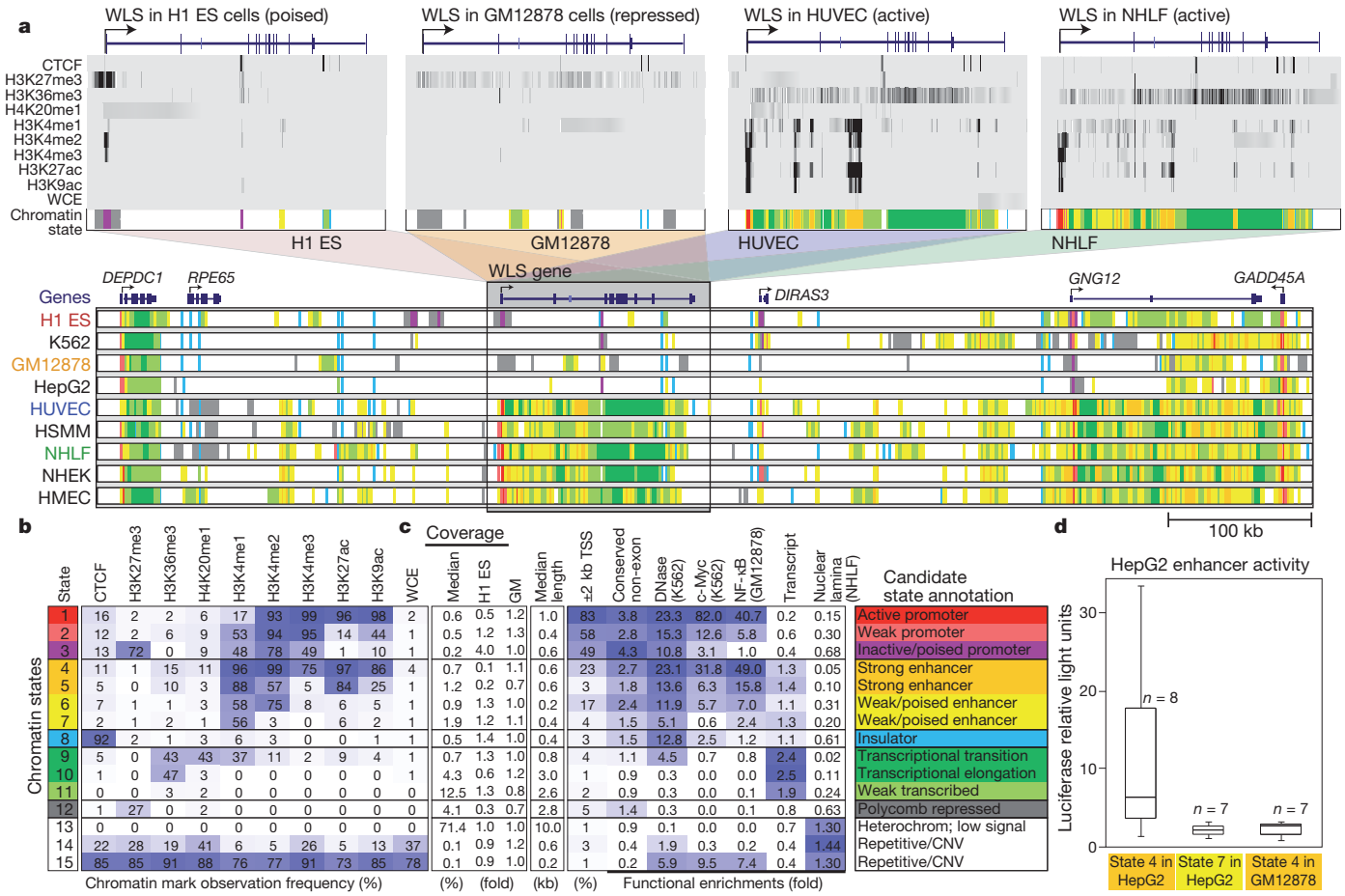


Figure 1 | Chromatin state discovery and characterization. **a**, Top: profiles for nine chromatin marks (greyscale) are shown across the WLS gene in four cell types, and summarized in a single chromatin state annotation track for each (coloured according to **b**). WLS is poised in ESCs, repressed in GM12878 and transcribed in HUVEC and NHLF. Its TSS switches accordingly between poised (purple), repressed (grey) and active (red) promoter states; enhancer regions within the gene body become activated (orange, yellow); and its gene body changes from low signal (white) to transcribed (green). These chromatin state changes summarize coordinated changes in many chromatin marks; for example, H3K27me3, H3K4me3 and H3K4me2 jointly mark a poised promoter, whereas loss of H3K27me3 and gain of H3K27ac and H3K9ac mark promoter activation. WCE, whole-cell extract. Bottom: nine chromatin state tracks, one per cell type, in a 900-kb region centred at WLS, summarizing 90 chromatin tracks in directly interpretable dynamic annotations and showing activation and repression patterns for six genes and hundreds of regulatory regions, including enhancer states. **b**, Chromatin states learned jointly across cell types by a multivariate hidden Markov model. The table shows emission parameters learned *de novo* on the basis of genome-wide recurrent combinations of chromatin marks. Each entry denotes the frequency with which a given mark is found at genomic positions corresponding to the chromatin state. **c**, Genome coverage, functional enrichments and candidate annotations for each chromatin state. Blue shading indicates intensity, scaled by column. CNV, copy number variation; GM, GM12878. **d**, Box plots depicting enhancer activity for predicted regulatory elements. Sequences 250 bp long corresponding either to strong or weak/poised HepG2 enhancer elements or to GM12878-specific strong enhancer elements were inserted upstream of a luciferase gene and transfected into HepG2. Reporter activity was measured in relative light units. Robust activity is seen for strong enhancers in the matched cell type, but not for weak/poised enhancers or for strong enhancers specific to a different cell type. Boxes indicate 25th, 50th and 75th percentiles, and whiskers indicate 5th and 95th percentiles.

This resulted in 90 chromatin maps corresponding to ~2,400,000,000 reads covering ~100,000,000,000 bases across nine cell types, which we set out to interpret computationally.

Learning a common set of chromatin states across cell types

To summarize these data sets into nine readily interpretable annotations, one per cell type, we applied a multivariate hidden Markov model that uses combinatorial patterns of chromatin marks to distinguish chromatin states[8]. The approach explicitly models mark combinations in a set of 'emission' parameters and spatial relationships between neighbouring genomic segments in a set of 'transition' parameters (Methods). It has the advantage of capturing regulatory elements with greater reliability, robustness and precision than is possible by studying individual marks[8].

We learned chromatin states jointly by creating a virtual concatenation of all chromosomes from all cell types. We selected 15 states that showed distinct biological enrichments and were consistently recovered (Fig. 1a, b and Supplementary Fig. 1). Even though states were learned *de novo* solely on the basis of the patterns of chromatin marks and their spatial relationships, they showed distinct associations with transcriptional start sites (TSSs), transcripts, evolutionarily conserved non-coding regions, DNase hypersensitive sites[12], binding sites for the regulators c-Myc[13] (MYC) and NF-κB[14], and inactive genomic regions associated with the nuclear lamina[15] (Fig. 1c).

We distinguished six broad classes of chromatin states, which we refer to as promoter, enhancer, insulator, transcribed, repressed and inactive states (Fig. 1c). Within them, active, weak and poised[4] promoters (states 1–3) differ in expression level, strong and weak candidate enhancers (states 4–7) differ in expression of proximal genes, and strongly and weakly transcribed regions (states 9–11) also differ in their positional enrichments along transcripts. Similarly, Polycomb-repressed regions (state 12) differ from heterochromatic and repetitive states (states 13–15), which are also enriched for H3K9me3 (Supplementary Figs 2–4).

The states vary widely in their average segment length (~500 base pairs (bp) for promoter and enhancer states versus 10 kb for inactive

regions) and in the portion of the genome covered (<1% for promoter and enhancer states versus >70% for inactive state 13). For each state, coverage was relatively stable across cell types (Supplementary Fig. 5), with the exception of embryonic stem cells (ESCs) in which the poised promoter state is more abundant but strong enhancer and Polycomb-repressed states are depleted, consistent with the unique biology of pluripotent cells[4,16].

We confirmed that promoter and enhancer states showed distinct biochemical properties (Supplementary Fig. 6). RNAPII was highly enriched at strong promoters, weakly enriched at strong enhancers and nearly undetectable at weak or poised enhancers, consistent with strong transcription at promoters and reports of weak transcription at active enhancers[17,18]. H2A.Z, a histone variant associated with nucleosome-free regions[19], was enriched in active promoters and strong enhancers, consistent with nucleosome displacement at TSSs and sites of abundant transcription factor binding in active enhancers.

We also used luciferase reporter assays to validate the functionality of predicted enhancers, the distinction between strong and weak enhancer states, and their predicted cell type specificity. We tested strong enhancers, weak enhancers and strong enhancers specific to an unmatched cell type by transfection in HepG2. We observed strong luciferase activity only for strong enhancer elements from the matched cell type (Fig. 1d).

These results and additional properties of the model (Supplementary Figs 7–10) suggest that chromatin states are an inherent, biologically informative feature of the genome. The framework enables us to reason about coordinated differences in marks by directly studying chromatin state changes between cell types (which we refer to as 'changes' or 'dynamics' without implying any temporal relationship).

Extent and significance of chromatin state changes across cell types

We next explored the extent to which chromatin states vary between pairs of cell types. The overall patterns of variability (Supplementary Figs 11 and 12) suggest that regulatory regions vary drastically in activity level across cell types. Enhancer states show frequent interchange between strong and weak, and promoter states vary between active, weak and poised. Promoter states seem more stable than enhancers; they are eight times more likely to remain promoter states, controlling for coverage. Switching was also observed among promoter, enhancer and transcriptional transition states, but no preferential changes to other groups were found. These general patterns suggest that despite varying activity levels, enhancer and promoter regions tend to preserve their chromatin identity as regions of regulatory potential.

Chromatin state differences between cell types relate to cell-type-specific gene functions. An unbiased clustering of chromatin state profiles across annotated TSSs in lymphoblastoid and skeletal muscle cells distinguished informative patterns predictive of downstream gene expression and functional gene classes (Supplementary Figs 13 and 14). Cell-type-specific patterns were also evident when TSSs were simply assigned to the most prevalent chromatin state. Promoters active in skeletal muscle were associated with extracellular structure genes (8.5-fold enrichment), those active in lymphoblastoid cells were associated with immune response genes (7.2-fold enrichment) and those active in both were associated with metabolic housekeeping genes.

Clustering of promoter and enhancer states on the basis of their activity patterns

Extending our pairwise promoter analysis, we clustered active promoter and strong enhancer regions across all cell types (Methods). This revealed clusters showing common activity and associated with highly coherent functions (Fig. 2). For promoter clusters, these include immune response (GM12878-specific clusters, $P < 10^{-18}$), cholesterol transport (HepG2 specific, $P < 10^{-4}$) and metabolic processes (all cells, $P < 10^{-131}$). Remarkably, genes assigned to enhancer clusters by proximity also showed strong functional enrichments, including immune

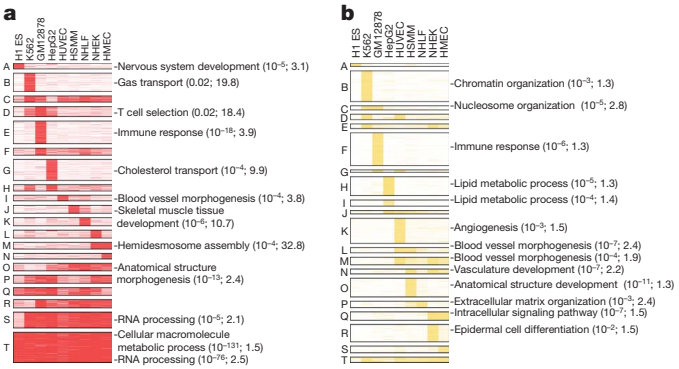


Figure 2 | Cell-type-specific promoter and enhancer states and associated functional enrichments. **a**, Clustering of genomic locations (rows) assigned to active promoter state 1 (red) across cell types (columns) reveals 20 common patterns of activity (A–T; Methods). For each cluster, enriched gene ontology terms are shown with hypergeometric *P* value and fold enrichment, based on the nearest TSS. For most clusters, several cell types show strong (dark red) or moderate (light red) activity. **b**, Analogous clustering and functional enrichments for strong enhancer state 4 (yellow). Enhancer states show greater cell type specificity, with most clusters active in only one cell type.

response (GM12878 specific, $P < 10^{-6}$), lipid metabolism (HepG2 specific, $P < 10^{-5}$) and angiogenesis (HUVEC specific, $P < 10^{-3}$).

Promoters and enhancers differed in their overall specificities. The majority of promoter clusters showed activity in multiple cell types, consistent with previous work[5,10] (Fig. 2a). Enhancer clusters are significantly more cell type specific, with few regions showing activity in more than two cell types and a majority being specific to a single cell type (Fig. 2b).

We also found differences in the relative contributions of enhancer-based and promoter-based regulation among gene classes. Developmental genes seem to be strongly regulated by both, showing the highest number of proximal enhancers and diverse promoter states, including poised and Polycomb repressed (Supplementary Fig. 15). Tissue-specific genes (for example immune genes and steroid metabolism genes) seem to be more dependent on enhancer regulation, showing multiple tissue-specific enhancers but less diverse promoter states. Lastly, housekeeping genes are primarily promoter regulated, with few enhancers in their vicinities.

Overall, this dynamic view of the chromatin landscape suggests that multicell chromatin profiles can be as productive for systems biology as expression analysis has traditionally been, and may hold additional information on genome regulatory programs, which we explore next.

Correlations in activity profiles link enhancers to target genes

We next investigated functional interconnections among enhancers, the factors that activate or repress them, and the genes whose expression they regulate, by defining 'activity profiles' for each across the cell types (Fig. 3). We complemented these enhancer activity profiles (Fig. 3a) with profiles for gene expression (Fig. 3b), sequence motif enrichment (Fig. 3d) and the expression of transcription factors recognizing each motif (Fig. 3e). We used correlations between these profiles to probabilistically link enhancers to their downstream targets and upstream regulators (Methods).

We found that patterns of enhancer activity (Figs 2b and 3a) correlated strongly with patterns of nearest-gene expression (Fig. 3b; correlation, >0.9 in 16 of 20 clusters). Because this correlation remained high even for large distances (>50 kb), we used activity correlation as a complement to genomic distance for linking enhancers to target genes (Methods). Activity-based linking yielded an increase in functional gene class enrichment for several clusters (Supplementary Fig. 16).

We validated our approach using quantitative trait locus mapping studies that use covariation between single nucleotide polymorphism (SNP) alleles and gene expression levels to link *cis*-regulatory regions

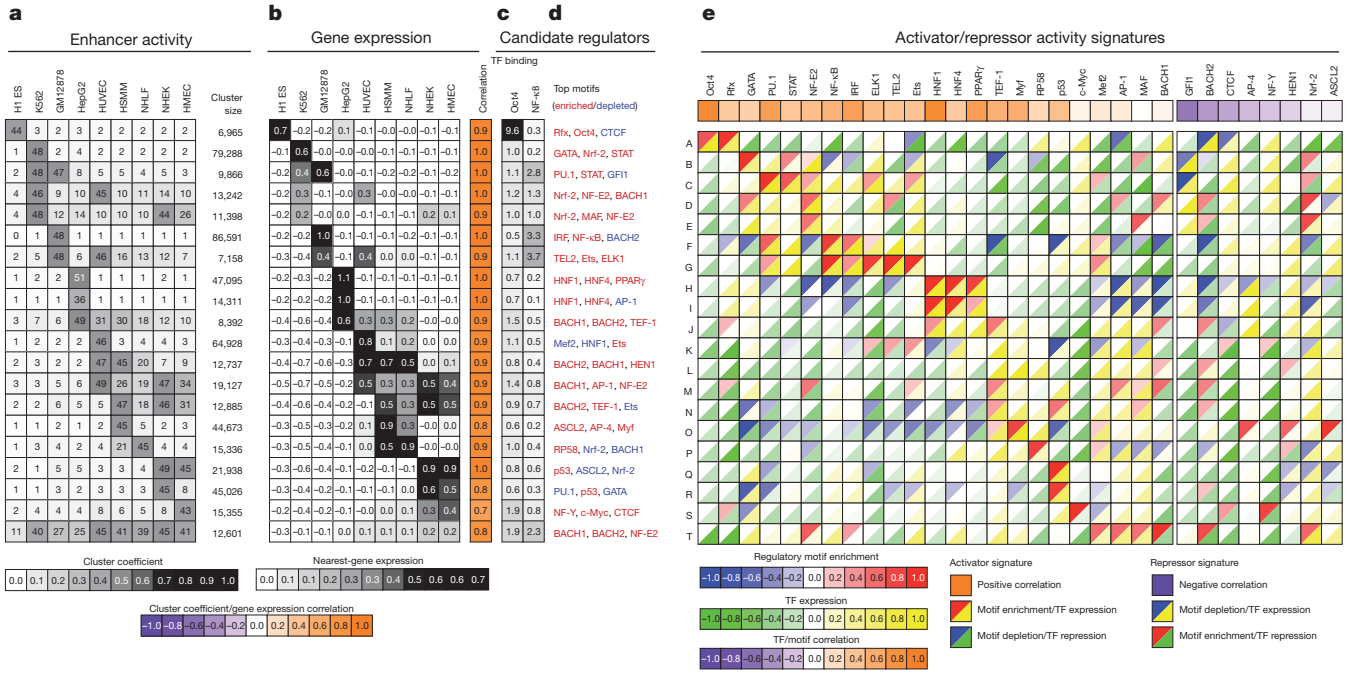


Figure 3 | Correlations in activity patterns link enhancers to gene targets and upstream regulators. **a**, Average enhancer activity across the cell types (columns) for each enhancer cluster (rows) defined in Fig. 2b (labelled A–T) and number of 200-bp windows in each cluster. **b**, Average messenger RNA expression of nearest gene across the cell types and correlation with enhancer activity profile from **a**. High correlations between enhancer activity and gene expression provide a means of linking enhancers to target genes. **c**, Enrichment for Oct4 binding in ESCs[24] and NF-κB binding in lymphoblastoid cells[14] for each cluster. TF, transcription factor. **d**, Strongly enriched (red) or depleted (blue) motifs for each cluster, from a catalogue of 323 consensus motifs. Rfx: Rfx family; Nrf-2: NFE2L2; STAT: STAT family; Ets: Ets family; Mef2: MEF2A and MYEF2; Myf: Myf family; NF-Y: NFYA, NFYB and NFYC. **e**, Predicted causal regulators for each cluster based on positive (activators) or negative (repressors) correlations between motif enrichment (top left triangles) and transcription factor expression (bottom right triangles). For example, the red–yellow combination indicates that Oct4 is a positive regulator of ESC-specific enhancers, as its motif-based predicted targets are enriched (red upper triangle) for enhancers active in ESCs (cluster A), and the Oct4 gene is expressed specifically in ESCs, resulting in a positive transcription factor expression correlation (yellow triangle). Overall correlations between motif enrichment and transcription factor expression across all clusters denote predicted activators (positive correlation, orange) and repressors (negative correlation, purple).

to target genes. Investigation of four recent quantitative trait locus studies in liver[20] and lymphoblastoid cells[21–23] revealed remarkable agreement with our enhancer predictions. Enhancers linked to a given target gene by our method were significantly enriched for SNPs correlated with the gene's expression level (Supplementary Fig. 17), thus confirming our enhancer–gene linkages with orthogonal data.

Correlations with transcription factor expression and motif enrichment predict upstream regulators

We next predicted, on the basis of regulatory motif enrichments, sequence-specific transcription factors likely to target enhancers in a given cluster. This implicated a number of transcription factors whose known biological roles matched the respective cell types (Fig. 3d and Supplementary Fig. 18). When ChIP-seq data on the relevant cell type was available, we confirmed that enriched motifs were preferentially bound by the cognate factor (Fig. 3c). Oct4 (POU5F1) motif instances in cluster A (ESC-specific enhancers) were preferentially bound by Oct4 in ESCs[24], and NF-κB motif instances in cluster F (lymphoblastoid-specific enhancers) were preferentially bound by NF-κB in lymphoblastoid cells[14]. In both cases, motif instances in cell-type-specific enhancers showed a ~5-fold increase in binding in comparison with other enhancers.

However, sequence-based motif enrichments do not distinguish causality. Enrichment could reflect a parallel binding event that does not affect the chromatin state, or the motif could actually be antagonistic to the enhancer state through specific repression in orthogonal cell types. To distinguish between these possibilities, we complemented the observed motif enrichments with cell-type-specific expression for the corresponding transcription factors (Fig. 3e). We then correlated a 'motif score' based on motif enrichment in a given cluster, and a 'transcription factor expression score' based on the agreement between the transcription factor expression pattern and the cluster activity profile (Methods). A positive correlation between the two scores implies that the transcription factor may be establishing or reinforcing the chromatin state. A negative correlation would instead imply that the transcription factor may act as a repressor. For example, in addition to the enrichment of the Oct4 motif in the ESC-specific cluster A, Oct4 is specifically expressed in ESCs, leading to the prediction that it is a causal regulator of ESCs (Fig. 3e), consistent with known biology[16].

For 18 of the 20 clusters, this analysis revealed one or more candidate regulators. Recovery of known roles for well-studied regulators validated our approach. For example, HNF1 (HNF1A), HNF4 (HNF4A) and PPARγ (PPARG) are predicted as activators of HepG2-specific enhancers (clusters H and I), PU.1 (SPI1) and NF-κB as activators of lymphoblastoid (GM12878) enhancers (clusters C, F and G), GATA1 as an activator of K562-specific enhancers (cluster B) and Myf family members as HSMM enhancers[14,25–27] (cluster O).

The analysis also revealed potentially novel regulatory interactions. ETS-related factors (ELK1, TEL2 (ETV7) and Ets family members) are predicted activators of enhancers active in both GM12878 and HUVEC (cluster G) but not of GM12878-specific or HUVEC-specific clusters, emphasizing the value of unbiased clustering. These connections are consistent with reported roles for ETS factors in lymphopoiesis and endothelium[28]. The prediction of p53 (TP53) as an activator in HSMM, NHLF, NHEK and HMEC (clusters N, Q and R) probably reflects its maintained activity in these primary cells, as opposed to cell models in which it may be suppressed by mutation (K562)[29], viral inactivation (GM12878)[30] or cytoplasmic localization (ESCs)[31]. A widespread role for p53 in regulating distal elements is consistent with its known binding to distal regions[32,33].

Our analysis also revealed several repressor signatures, including GFI1 in K562 and GM12878 (clusters B and C) and BACH2 in ESCs

(cluster A). Both regulators are known to repress transcription by recruiting histone deacetylases and methyltransferases to proximal promoters[34,35], and GFI1 has also been implicated in the silencing of satellite repeats[35]. Our regulatory inferences suggest that these regulators also modulate chromatin to inhibit enhancer activity, thus suggesting a new mechanism for distal gene regulation.

Validation of predicted binding events and regulatory outcomes

The regulatory inferences above imply transcription-factor-binding events at motif instances within enhancer regions in specific cellular contexts, and we sought to validate these inferences using a general molecular signature. Binding events are associated with nucleosome displacement, a structural change evident in ChIP-seq data for histones[36]. We thus studied local depletions in the chromatin intensity profiles ('dips') as these are indicative of transcription factor binding. We confirmed that dips were present in individual signal tracks for active enhancers and were associated with preferential sequence conservation and regulatory motif instances (Fig. 4a).

To test our specific predictions, we superimposed chromatin profiles of coordinately regulated enhancer regions, anchoring them on the implied motif instances. Striking dips precisely coincide with regulatory motifs, and are both cell type specific and region specific, exactly as predicted (Fig. 4b, c). Because dips only appear when the factor is expressed, they also support the identity of the *trans*-acting transcription factor.

To confirm that predicted causal motifs contribute to enhancer activity, we used luciferase reporters. Our model implicated HNF regulators as activators of HepG2-specific enhancers (Fig. 3), and context-specific dips supported binding interactions (Fig. 4c). We thus selected for functional analysis ten sites with HNF motifs showing dips in strong HepG2-specific enhancers, and evaluated them with and without the HNF motif. We found that permutation of the motif consistently led to a reduction in enhancer activity (Fig. 4d), supporting its predicted causal role.

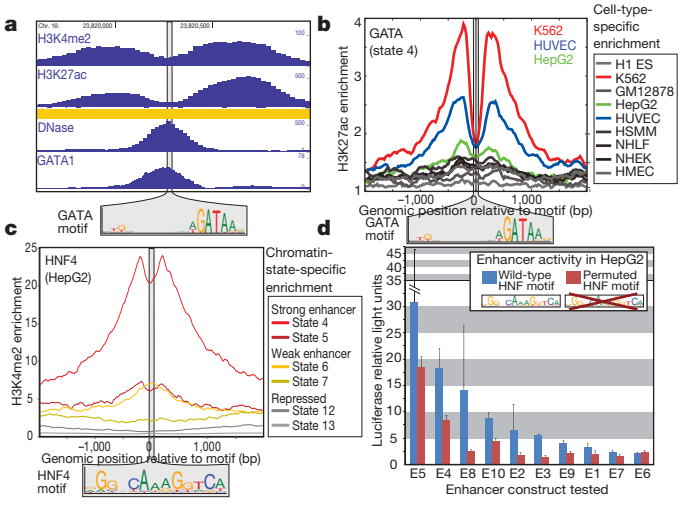


Figure 4 | Validation of regulatory predictions by nucleosome depletions and enhancer activity. **a**, Dips in chromatin intensity profiles in a K562-specific strong enhancer (orange) coincide with a predicted causal GATA motif instance (logo). The dips probably reflect nucleosome displacement associated with transcription factor binding, supported by DNase hypersensitivity[12] and GATA1 binding[25]. **b**, Superposition of H3K27ac signal across loci containing GATA motifs, centred on motif instances, shows dips in K562, as predicted. **c**, Superposition of H3K4me2 signal for HepG2 shows dips over HNF4 motifs in strong enhancer states, as predicted. **d**, HepG2-specific strong enhancers with predicted causal HNF motifs were tested in reporter assays. Constructs with permuted HNF motifs (red) led to significantly reduced luciferase activity in comparison with wild type (blue), with an average twofold reduction. Data shown are mean luciferase relative light units over three replicates and 95% confidence intervals.

Assigning candidate regulatory functions to disease-associated variants

Finally, we explored whether our chromatin annotations and regulatory predictions can provide insight into sequence variants associated with disease phenotypes. To that effect, we gathered a large set of non-coding SNPs from GWAS catalogues, an exceedingly small proportion of which are understood at present[37].

We found that disease-associated SNPs are significantly more likely to coincide with strong enhancers (states 4 and 5; twofold enrichment, $P < 10^{-10}$), despite the fact that no notable association with these states are seen for SNPs in general or for those SNPs tested in the studies. To test whether SNPs associated with a particular disease might have even more specific correspondences, we examined 426 GWAS data sets. We identified ten studies[38–47] whose variants showed significant correspondences to cell-type-specific strong enhancer states (Methods and Fig. 5a).

Individual variants from these studies were strongly enriched in enhancer states specifically active in relevant cell types (Fig. 5a, b). For example, SNPs associated with erythrocyte phenotypes[38] were found in erythrocytic leukaemia cell (K562) enhancers, SNPs associated with systemic lupus erythematosus[39] were found in lymphoblastoid cell (GM12878) enhancers and SNPs associated with triglyceride[40] phenotypes or blood lipid phenotypes[41] were found in hepatocellular carcinoma cell (HepG2) enhancers. We also applied our model to chromatin data for T cells[3] (Supplementary Fig. 19), for which strong enhancer states correlated to variants associated with risk of childhood acute lymphoblastic leukaemia[48], further validating our approach.

We also used our predicted enhancer/target gene associations to find candidate downstream genes whose expression might be affected by *cis* changes occurring in the enhancer region. Although most of the predicted target genes are proximal to the enhancer, a subset of more distal predicted targets could reflect novel candidates for the disease phenotypes (Fig. 5b).

In addition, we identified several instances in which a lead GWAS variant does not correspond to a particular chromatin element but a linked variant coincides with an enhancer with the predicted cell type specificity (Fig. 5c). Thus, chromatin profiles may provide a general means of triaging variants within a haplotype block, a common problem faced in GWASs.

Lastly, we identified several cases in which a disease-associated SNP created or disrupted a regulatory motif instance for a predicted causal transcription factor in the relevant cell type (Fig. 5d), suggesting a specific molecular mechanism by which the disease-associated genotype could lead to the observed disease phenotype consistent with our regulatory predictions.

Discussion

Our work demonstrates the power of multicell chromatin profiles as an additional and dynamic layer of genome annotation. We presented methods to distinguish different classes of functional elements, elucidate their cell type specificities and predict *cis*-regulatory interactions that drive gene expression programs. By intersecting our predictions with non-coding SNPs from GWAS data sets, we proposed potential mechanistic explanations for disease variants, either through their presence within cell-type-specific enhancer states or by their effect on binding motifs for predicted regulators.

Chromatin states drastically reduced the large combinatorial space of 90 chromatin data sets (2^{90} combinations) to a manageable set of biologically interpretable annotations, thus providing an efficient and robust way to track coordinated changes across cell types. This allowed the systematic identification and comparison of more than 100,000 promoter and enhancer elements. Both types of element are cell type specific, are associated with motif enrichments and assume

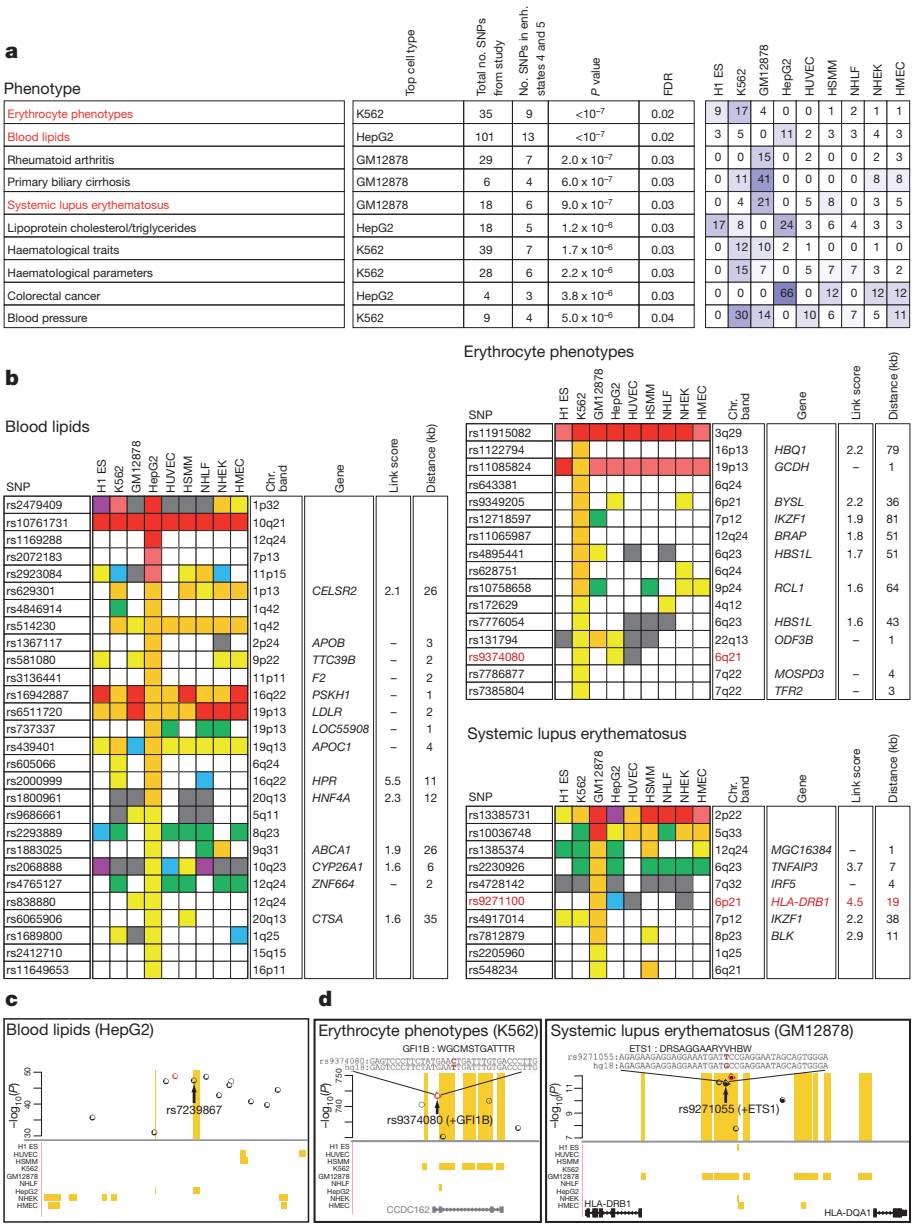


Figure 5 | Disease variants annotated by chromatin dynamics and regulatory predictions. **a**, Intersection of strong enhancer states (4 and 5) with disease-associated SNPs from GWASs shows significant enrichment (blue shading) in relevant cell types (Methods). Fold enrichments of the SNPs in strong enhancer states for each cell type are indicated. FDR, false-discovery rate. **b**, For three GWAS data sets[38–40], state annotations are shown for a subset of lead SNPs in the nine cell types (colours as in Fig. 1b, except state 11 is white). The strong enhancer state (orange) is most prevalent in cell types related to the phenotype. For SNPs overlapping strong enhancers, proximal genes with correlated expression are indicated, with linking score and distance. **c**, Example GWAS locus with blood lipid trait[41] association, where the lead variant (red circle) has no functional annotation but a linked SNP (arrow) coincides with a HepG2-specific strong enhancer (orange) and may represent a causal variant. Strong enhancer annotations are shown for all cell types. **d**, Example GWAS loci where a disease SNP affects a conserved instance of a predicted causal motif. Left: lead SNP rs9374080 in the erythrocyte phenotype GWAS[38] is <100 bp from a strong enhancer in K562 and strengthens a motif for GFI1B, a predicted repressor in K562 (Fig. 3e). Right: SNP rs9271055 associated with lupus[39] coincides with a lymphoblastoid (GM12878) strong enhancer and strengthens a motif for ETS1, a predicted activator of lymphoblastoid enhancers (Fig. 3e). This factor is further implicated by lupus-associated variants that directly affect the *ETS1* locus[39].

strong, weak and poised states that correlate with neighbouring gene expression and function. Enhancers showed very high tissue specificity, enrichment in the vicinity of developmental and cell-type-specific genes, and predictive power for proximal gene expression, reinforcing their roles as sentinels of tissue-specific gene expression[49]. By elucidating enhancers systematically, and linking them to upstream regulators and downstream genes, our analysis can help provide a missing link between regulators and target genes. The power of the approach should increase considerably as additional phenotypically distinct cell types are surveyed, and should enable a greater proportion of enhancer elements to be incorporated into the connectivity network.

The inferred *cis*-regulatory interactions make specific testable predictions, many of which were confirmed through additional experiments and analyses. Our enhancer/target gene linkages are supported by *cis*-regulatory inferences from quantitative trait locus mapping studies. Predicted transcription factor/motif interactions within cell-type-specific enhancers were confirmed in specific cases by transcription factor binding and more generally by depletions in the chromatin profiles at causal motifs in appropriate cellular contexts. Motifs predicted as causal regulators of cell-type-specific enhancers were also confirmed in enhancer assays.

The regulatory inferences afforded by multicell chromatin profiles are unique and highly complementary to data sets for transcription factor binding, expression, chromatin accessibility, nucleosome positioning and chromosome conformation[50]. For example, our regulatory predictions can help focus the spectrum of transcription-factor-binding events to a smaller number of functional interactions. The 'chromatin-centric' approach also complements the extensive body of work on biological network inference from expression data with the potential to introduce enhancers and other genomic elements into connectivity networks.

Our study has important implications for the understanding of disease. Our detailed and dynamic functional annotations of the relatively uncharted non-coding genome can facilitate the interpretation of GWAS data sets by predicting specific cell types and regulators related to specific diseases and phenotypes. Furthermore, the connections derived for enhancer regions, to upstream regulators and downstream genes, suggest *cis*- and *trans*-acting interactions that may be modulated by the sequence variants. Although the present study represents only a first, small step in this direction, we expect that future iterations with a greater diversity of cell types and improved methodologies will help define the molecular underpinnings of human disease.

METHODS SUMMARY

We performed ChIP-seq analysis in biological replicate as previously described[4], using antibodies validated by western blots and peptide competitions. ChIP DNA and input controls were sequenced using the Illumina Genome Analyser. Expression profiles were acquired using Affymetrix GeneChip arrays. Chromatin states were learned jointly by applying a hidden Markov model[8] to ten data tracks for each of the nine cell types. We focused on a 15-state model that provides sufficient resolution to resolve biologically meaningful patterns yet is reproducible across cell types when independently processed. We used this model to produce nine genome-wide chromatin state annotations, which were validated by additional ChIP experiments and reporter assays. Multicell type clustering was conducted on locations assigned to strong promoter state 1 (or strong enhancer state 4) in at least one cell type using the *k*-means algorithm. We predicted enhancer/target gene linkages by correlating normalized signal intensities of H3K27ac, H3K4me1 and H3K4me2 with gene expression across cell types as a function of distance to the TSS. Upstream regulators were predicted using a set of known transcription factor motifs assembled from multiple sources. Motif instances were identified by sequence match and evolutionary conservation. We based *P* values for GWAS studies on randomizing the location of SNPs, and based the false-discovery rate on randomizing the assignment of SNPs across studies. Data sets are available from the ENCODE website (http://genome.ucsc.edu/ENCODE), the supporting website for this paper (http://compbio.mit.edu/ENCODE_chromatin_states) and the Gene Expression Omnibus (GSE26386).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 August 2010; accepted 4 February 2011.
Published online 23 March 2011.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
2. Kim, H. D., Shay, T., O'Shea, E. K. & Regev, A. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* **325,** 429–432 (2009).
3. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
4. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
5. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130,** 77–88 (2007).
6. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39,** 311–318 (2007).
7. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLOS Comput. Biol.* **5,** e1000566 (2009).
8. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnol.* **28,** 817–825 (2010).
9. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120,** 169–181 (2005).
10. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
11. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (2009).
12. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107,** 139–144 (2010).
13. Raha, D. *et al.* Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc. Natl Acad. Sci. USA* **107,** 3639–3644 (2010).
14. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328,** 232–235 (2010).
15. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
16. Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132,** 567–582 (2008).
17. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8,** e1000384 (2010).
18. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
19. Talbert, P. B. & Henikoff, S. Histone variants — ancient wrap artists of the epigenome. *Nature Rev. Mol. Cell Biol.* **11,** 264–275 (2010).
20. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
21. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
22. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).
23. Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4,** e1000214 (2008).
24. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42,** 631–634 (2010).
25. Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol. Cell* **36,** 667–681 (2009).
26. Lemaigre, F. & Zaret, K. S. Liver development update: new embryo models, cell lineage control, and morphogenesis. *Curr. Opin. Genet. Dev.* **14,** 582–590 (2004).
27. Sabourin, L. A. & Rudnicki, M. A. The molecular regulation of myogenesis. *Clin. Genet.* **57,** 16–25 (2000).
28. Bartel, F. O., Higuchi, T. & Spyropoulos, D. D. Mouse models in the study of the Ets family of transcription factors. *Oncogene* **19,** 6443–6454 (2000).
29. Law, J. C., Ritke, M. K., Yalowich, J. C., Leder, G. H. & Ferrell, R. E. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. *Leuk. Res.* **17,** 1045–1050 (1993).
30. Forte, E. & Luftig, M. A. MDM2-dependent inhibition of p53 is required for Epstein-Barr virus B-cell growth transformation and infected-cell survival. *J. Virol.* **83,** 2491–2499 (2009).
31. Solozobova, V., Rolletschek, A. & Blattner, C. Nuclear accumulation and activation of p53 in embryonic stem cells after DNA damage. *BMC Cell Biol.* **10,** 46 (2009).
32. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116,** 499–509 (2004).
33. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124,** 207–219 (2006).
34. Hoshino, H. *et al.* Co-repressor SMRT and class II histone deacetylases promote Bach2 nuclear retention and formation of nuclear foci that are responsible for local transcriptional repression. *J. Biochem.* **141,** 719–727 (2007).
35. Vassen, L., Fiolka, K. & Moroy, T. Gfi1b alters histone methylation at target gene promoters and sites of gamma-satellite containing heterochromatin. *EMBO J.* **25,** 2409–2419 (2006).
36. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nature Genet.* **42,** 343–347 (2010).
37. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
38. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nature Genet.* **41,** 1191–1198 (2009).
39. Han, J. W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genet.* **41,** 1234–1237 (2009).
40. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genet.* **40,** 189–197 (2008).
41. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
42. Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genet.* **40,** 1426–1435 (2008).
43. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genet.* **41,** 666–676 (2009).
44. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genet.* **42,** 508–514 (2010).
45. Liu, X. *et al.* Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nature Genet.* **42,** 658–660 (2010).
46. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Genet.* **42,** 210–215 (2010).
47. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genet.* **41,** 1182–1190 (2009).
48. Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nature Genet.* **41,** 1006–1010 (2009).
49. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461,** 199–205 (2009).
50. Naumova, N. & Dekker, J. Integrating one-dimensional and three-dimensional maps of genomes. *J. Cell Sci.* **123,** 1979–1988 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the epigenomics community at the Broad Institute and the Bernstein and Kellis laboratories, and M. Daly, D. Altshuler and E. Lander for discussions and criticisms. We also thank M. Suva, E. Mendenhall and S. Gillespie for assistance with experiments, and L. Goff and A. Chess for critical reading of the manuscript. We acknowledge the Broad Institute Genome Sequencing Platform for their expertise and assistance with data production. This research was supported by the National Human Genome Research Institute under an ENCODE grant (U54 HG004570; B.E.B.), R01 HG004037 (M. Kellis), RC1 HG005334 (M. Kellis), the Howard Hughes Medical Institute (B.E.B.), the National Science Foundation (awards 0644282 (M. Kellis) and 0905968 (J.E.)) and the Sloan Foundation (M. Kellis).

Author Contributions J.E. conducted chromatin state analysis. J.E. and P.K. conducted regulatory motif analysis. J.E. and L.W. conducted GWAS SNP analysis. T.S.M., N.S. and T.D. implemented the ChIP-seq data processing pipeline. C.B.E., X.Z., L.W., R.I., M.C. and M. Ku developed the experimental pipeline and conducted experiments. M. Kellis designed and directed the computational analysis. B.E.B. designed the experimental approach and oversaw the work. J.E., M. Kellis and B.E.B. wrote the paper.

Author Information Sequencing and expression data has been deposited into the Gene Expression Omnibus under accession number GSE26386. 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 this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M. Kellis (manoli@mit.edu).

METHODS

Cell culture. Human H1 ES cells were cultured in TeSR media[51] on Matrigel by Cellular Dynamics International. Cells were split with dispase and collected at a passage number between 30 and 40. Before collection, cells were karyotyped and stained for Oct4 to confirm pluripotency. K562 erythrocytic leukaemia cells (ATCC CCL-243, lot no. 4607240) were grown in suspension in RPMI medium (HyClone SH30022.02) with 10% fetal bovine serum (FBS) and 1% Antibiotic-Antimycotic (GIBCO 15240-062). Cell density was maintained at between 3×10^5 and 7×10^5 cells ml^{-1}. GM12878 B-lymphoblastoid cells (Coriell Cell Repositories, 'expansion A') were grown in suspension in RPMI 1640 medium with 15% FBS (not heat inactivated), 2 mM L-glutamine and 1% penicillin/streptomycin. Cells were seeded at a concentration of $\sim 2 \times 10^5$ viable cells ml^{-1} with minimal disruption, and maintained at between 3×10^5 and 7×10^5 cells ml^{-1}. HepG2 hepatocellular carcinoma cells (ATCC HB-8065, lot no. 4968519) were cultured in DMEM (HyClone SH30022.02) with 10% FBS and 1% penicillin/streptomycin. Cells were trypsinized, resuspended to single-cell suspension, split to a confluence of between 15 and 20% and then collected at ~75% confluence. NHEK normal human epidermal keratinocytes isolated from skin (Lonza CC-2501, lot no. 4F1155J, passage 1) were grown in keratinocyte basal medium 2 (KGM-2 BulletKit, Lonza) supplemented with BPE, hEGF, hydrocortisone, GA-1000, transferrin, epinephrine and insulin. Cells were seeded at the recommended density (3,500 cells cm^{-2}), subjected to two or three passages on polystyrene tissue culture plates and collected at a confluence of 70 to 80%. HSMM primary human skeletal muscle myoblasts (Lonza CC-2580, lot no. 6F4444, passage 2) were cultured in Smooth Muscle Growth Medium 2 (SkGM-2 BulletKit, Lonza) supplemented with rhEGF, dexamethasone, L-glutamine, FBS and GA-1000. Cells were seeded at the recommended density (3,500 cells cm^{-2}), subjected to two or three passages and collected at a confluence of 50 to 70%. NHLF primary normal human lung fibroblasts (Lonza CC-2512, lot no. 4F0758, passage 2) were grown in Fibroblast Cell Basal Medium 2 (FGM-2 BulletKit, Lonza) supplemented with hFGF-β, insulin, FBS and GA-100. Cells were seeded at the recommended density (2,500 cells cm^{-2}), subjected to two or three passages and collected at an approximate confluence of 80%. HUVEC primary human umbilical vein endothelial cells (Lonza CC-2517, lot no. 7F3239, passage 1) were grown in endothelial basal medium 2 (EGM-2 BulletKit, Lonza) supplemented with hFGF-β, hydrocortisone, VEGF, R3-IGF-1, ascorbic acid, heparin, FBS, hEGF and GA-1000. Cells were seeded at the recommended density (2,500–5,000 cells cm^{-2}), subjected to two or three passages and collected at a confluence of 70 to 80%. HMEC primary human mammary epithelial cells from mammary reduction tissue (Lonza CC-2551, passage 7) were grown in mammary epithelia basal medium (MEGM BulletKit, Lonza) supplemented with hEGF-β, hydrocortisone, BPE, GA-1000 and insulin. Cells were seeded at the recommended density (2,500 cells cm^{-2}), subjected to two or three passages and collected at 60 to 80% confluence.

Antibodies. ChIP assays were performed using the following antibody reagents: H3K4me1 (Abcam ab8895, lot 38311/659352), H3K4me2 (Abcam ab7766, lot 56293), H3K4me3 (Abcam ab8580, lot 331024; Milipore 04-473, lot DAM1623866), H3K9ac (Abcam ab44441, lot 455103/550799), H3K27ac (Abcam ab4729, lot 31456), H3K36me3 (Abcam ab9050, lot 136353), H4K20me1 (Abcam ab9051, lot 104513/519198), H3K27me3 (Millipore 07-449, lot DAM1387952/DAM1514011), CTCF (Millipore 07-729, lot 1350637), H3K9me3 (Abcam ab8898, lot 484088), H2A.Z (Millipore 07-594, lot DAM1504736) and RNAPII N terminus (Santa Cruz sc-899X, lot H0510). All antibody lots were extensively validated for specificity and efficacy in ChIP-seq. Western blots were used to confirm specific recognition of histone protein (or CTCF). Dot plots performed using arrayed histone tail peptides representing various modification states were used to confirm specificity for the appropriate modification. ChIP-seq assays performed on a common cell reagent were used to confirm consistency between different lots of the same antibody.

Chromatin immunoprecipitation. Cells were harvested by crosslinking with 1% formaldehyde in cell culture medium for 10 min at 37 °C. After quenching with the addition of 125 mM glycine for 5 min at 37 °C, the cells were washed twice with cold PBS containing protease inhibitor (Roche). After aspiration of all liquid, pellets consisting of $\sim 10^7$ cells were flash frozen and stored at −80 °C. Fixed cells were thawed and sonicated to obtain chromatin fragments of ~200 to 700 bp using a Bioruptor (Diagenode). Immunoprecipitation was performed as previously described, retaining a fraction of input 'whole-cell extract' as a control[4]. Briefly, sonicated chromatin was diluted tenfold and incubated with ~5 μg antibody overnight. Antibody–chromatin complexes were pulled-down using protein A sepharose, washed and then eluted. After crosslink reversal and proteinase K treatment, immunoprecipitated DNA was extracted with phenol, precipitated in ethanol and treated with RNase. ChIP DNA was quantified by fluorometry using the Qubit assay (Invitrogen).

Next-generation sequencing. For each ChIP or control sample, ~5 ng of DNA was used to generate a standard Illumina sequencing library. Briefly, DNA fragments were end-repaired using the End-It DNA End-Repair Kit (Epicentre), extended with a 3′ 'A' base using Klenow (3′ → 5′ exo-, 0.3 U μl^{-1}, NEB), ligated to standard Illumina adapters (75 bp with a 'T' overhang) using DNA ligase (0.05 U μl^{-1}, NEB), gel-purified on 2% agarose, retaining products between 275 and 700 bp, and subjected to 18 PCR cycles. These libraries were quantified by fluorometry and evaluated by quantitative PCR or a multiplexed-digital-hybridization-based analysis[52] (NanoString nCounter) to confirm representation and specific enrichment of DNA species. Libraries were sequenced in one or two lanes on the Illumina Genome Analyser using standard procedures for cluster amplification and sequencing by synthesis.

Expression profiling. Cytosolic RNA was isolated using RNeasy Columns (Qiagen) from the same cell lots as above. Gene expression profiles were acquired using Affymetrix GeneChip arrays. The data were normalized using the GenePattern expression data analysis package[53]. CEL files were processed by RMA, quantile normalization and background correction. Two replicate expression data sets for each cell type were averaged and $\log_2$-transformed. Gene-level normalization across cell types was computed by mean normalization.

Primary processing of sequencing reads. ChIP-seq reads were aligned to human genome build HG18 with MAQ (http://maq.sourceforge.net/maq-man.shtml) using default parameters. All reads were truncated to 36 bases before alignment. Signal density maps for visualization were derived by extending sequencing reads by 200 bp in the 3′ direction (the estimated median size of ChIP fragments), and then counting the total number of overlapping reads at 25-bp intervals. Replicate ChIP-seq experiments were verified by comparing enriched intervals as previously described[4], and were then combined into a single data set. For the hidden Markov model (HMM), density maps were derived by extending sequencing reads by 200 bp in the 3′ direction and then assigning them to a single 200-bp window based on the midpoint of the extended read. These maps were then binarized at 200-bp resolution on the basis of a Poisson background model using a threshold of 10^{-4}.

Joint learning of HMM states across cell types. To handle data from the nine cell types, we concatenated their genomes to create an extended virtual genome that we used to train the HMM. We applied the model to ten tracks corresponding to the different chromatin marks and input using a multivariate HMM as previously described[8]. Here we used a Euclidean distance for determining initial parameters for the nested initialization step. After the HMM had learned and evaluated a set of roughly nested models, considering up to 25 states, we focused on a 15-state model that provides sufficient resolution to resolve biologically meaningful chromatin patterns and yet is highly reproducible across cell types when independently processed (Supplementary Fig. 7). We used this model to compute the probability that each location is in a given state, and then assigned each 200-bp interval to its most likely state for each cell type. Even though our model focuses on presence/absence frequencies of marks, we found that our states also capture signal intensity differences between high-frequency and low-frequency marks (Supplementary Fig. 9).

Enrichment analysis. For each state, enrichments for different annotations were computed at 200-bp resolution with the exception of conservation, which was computed at nucleotide resolution. We used annotations obtained through the UCSC Genome Browser[54] for RefSeq TSSs and transcribed regions[55], PhastCons[56], DNase-seq for K562 cells[12], c-Myc ChIP-seq for K562 cells[13], NF-κB ChIP-seq for GM12878[14], Oct4 in ESCs[24] and nuclear lamina[15]. Gene functional group enrichments were determined using STEM[57] and biological process annotations in the Gene Ontology database[58]. *P* values were calculated on the basis of the hypergeometric distribution and corrected for multiple testing using Bonferroni correction.

Comparing chromatin state assignments between cell types. For each pair of cell types, the chromatin state assignments at each genomic position were compared. We calculated the frequency with which each pair of states occurred, and normalized this against the expected frequency based on the amount of genome covered by each state. The fold enrichments in Fig. 2a reflect an aggregation across all 72 possible pairs of cell types.

Pairwise promoter clustering. Promoters for RefSeq genes were clustered on the basis of the most likely chromatin state assignment across a 2-kb region centred on the TSS. Clustering was performed jointly across GM12878 and HSMM, and was restricted to genes with corresponding Affymetrix expression. Briefly, each promoter was treated as a 330-element binary vector in which each component corresponded to a position along the promoter, cell type and state. Clustering was performed on these vectors using the *k*-means algorithm in MATLAB. Gene expression values were calculated on the basis of the corresponding Affymetrix probe set closest to the TSS.

Multicell type promoter and enhancer clustering. Promoter state clustering was performed for all 200-bp intervals assigned to the strong promoter state (state 1)

in at least one cell type. Each interval was represented by a single vector whose components are the estimated probabilities that it be in the strong promoter state for each of the nine cell types, accounting for model assignment uncertainty and biological noise. These were determined from the model posterior probabilities of state assignments and a comparison of state assignments in replicate experimental data. Clustering was performed using the *k*-means algorithm in MATLAB. We found that 20 clusters provided sufficient resolution to distinguish major cell-type-specific patterns. Enhancer state clustering was performed for all 200-bp intervals assigned to strong enhancer state 4 in at least one cell type using identical procedures. For the purposes of display in Fig. 2, the locations were randomly down-sampled. For the purpose of identifying enriched functional gene categories in Fig. 2b, enhancers were linked to the nearest TSS up to 50 kb distant excluding those within 5 kb. Enhancer–gene correspondences based on the nearest gene were used for the expression analysis of distance-based linked genes in Fig. 3b.

Linking enhancer locations to correlated genes. To predict linkages between enhancer states and target genes, we combined distance-based information with multicell type correlations between gene expression levels and normalized signal intensities for histone modifications associated with enhancer states (H3K4me1, H3K4me2 and H3K27ac). For each enhancer state (4–7), cell type, and 200-bp interval between 5 kb and 125 kb from the TSS, we trained logistic regression classifiers. The classifiers were trained to use mark intensity/expression correlation values to distinguish real instances of pairs of enhancer states and gene expression values from control pairs based on randomly re-assigning expression values to different genes. So that the classifiers learned a smooth and robust function at each position, we included as part of the training all enhancer state assignments within a 10-kb window centred at the position. The link score for a specific enhancer–gene linkage was defined as the ratio of the corresponding logistic regression classifier probability score to that for the randomized data.

For the evaluation of the expression quantitative trait loci (QTL) analysis, we used a link score threshold of 2.5. The expression QTL data was obtained from the University of Chicago QTL browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). In the QTL evaluation, each SNP that overlapped a strong enhancer state (4 or 5), was within 125 kb of a TSS, excluding locations within 5 kb, and was associated with a gene for which we had expression data was considered eligible to be supported by our linked predictions. We computed the fraction we observed linked on the basis of our linked predictions relative to the fraction that would be expected to be linked conditioned on knowing the distance distributions of the SNPs relative to the gene TSS.

For the evaluation of linked predictions using the Gene Ontology database, we used the same link score threshold and compared gene assignments against the distance-based assignments defined above. The base set of genes in the enrichment analysis here were all genes that could be linked in at least one cluster.

Motif and transcription factor analysis. A database of known transcription factor motifs was collated by combining motifs from TRANSFAC[59] (version 11.3), JASPAR[60] (2010-05-07) and protein-binding microarray data sets[61–63]. Motif instances in non-coding and non-repetitive regions of the genome were identified using these motifs and sequence conservation using a 29-way alignment of eutherian mammal genomes (K. Lindblad-Toh *et al.*, submitted). These were filtered using a significance threshold of $P < 4^{-8}$ for the motifs[64], and a confidence level based on conservation. Motifs were linked to corresponding transcription factors using metadata provided by the source. Motif enrichments for chromatin state clusters were computed as ratios to the instances of shuffled motifs, to correct for non-specific conservation and composition. A confidence interval was calculated for each ratio using Wilson score intervals ($z = 1.5$), selecting the most conservative value within the confidence interval. In cases where multiple motif variants were available for the same transcription factor, the one that showed the most variance in enrichment across clusters was selected.

For predicting causal activators and repressors, motif scores and transcription factor expression scores were correlated as follows. Motif scores were calculated as described above. Transcription factor expression scores were calculated for each cluster by correlating the expression of the transcription factor across the cell types with the activity profile of the enhancers in that cluster (defined by the cluster means from the *k*-means clustering). The motif scores and the transcription factor expression scores were then correlated against each other to identify positively and negatively correlated transcription factors.

Transcription factor/motif interactions predicted for strong enhancer states in specific cell types were validated by using the raw ChIP-seq tag enrichments as proxy for nucleosome positioning. For this purpose, sequencing reads were processed as above, except that the middle 75 bp of inferred ChIP fragments were used to derive signal density informative of nucleosome depletion (dips), as previously described[36]. Superposition plots show tag enrichments relative to a uniform background computed on the basis of sequencing depth.

Quantitative real-time PCR. Enrichment ratios for RNAPII and H2A.Z ChIPs were determined relative to input chromatin by quantitative real-time PCR using an ABI 7900 detection system, in biological replicate as described previously[65]. Regions used for validation correspond to three different chromatin states, including 13 for state 1 (arbitrarily selected), 11 for state 4 (arbitrarily selected but excluding regions within 2 kb of a state-1 annotation) and 11 for state 7 (arbitrarily selected but excluding regions within 2 kb of a state-1 or state-4 annotation). PCR primers are listed in Supplementary Data 1.

Functional enhancer assays. The SV40 promoter was first inserted between the HindIII and NcoI sites of pGL4.10 (Promega). Next, 250-bp sequences from the reference genome (hg18) corresponding to different chromatin states (eight from HepG2 state 4, seven from HepG2 state 7 and seven from GM12878 state 4) were synthesized (GenScript) and then inserted between the two SfiI sites upstream of the SV40 promoter. HepG2 cells were seeded into 96-well plates at a density of 5×10^4 cells per well and expanded overnight to ~50% confluency. The cells were then transfected with 400 ng of a pGL4.10-derived plasmid and 100 ng of pGL4.73 (Promega) using Lipofectamine LTX. Firefly and Renilla luciferase activities were measured 24 h post-transfection using Dual-Glow (Promega) and an EnVision 2103 multilabel reader (PerkinElmer), from triplicate experiments. Data are reported as light units relative to a control plasmid. For validation of causal transcription factor motifs, ten sequences of 250 bp corresponding to HepG2-specific strong enhancers (state 4) with dips and HNF motifs were tested as above, and compared with identical sequences except with the HNF motif permuted. Tested enhancer elements are listed in Supplementary Data 1.

GWAS SNP analysis. The GWAS variants and SNP coordinates were obtained from the NHGRI catalogue and the UCSC browser[37,54] (October 30, 2010). This set was refined by extending the blood lipid GWAS[41] set to contain all reported SNPs, and by bifurcating the haematological and biochemical traits study[46] into a haematological traits set and a biochemical traits set. We limited our analysis to studies reporting two or more associated SNPs. The variants from each study were intersected with chromatin states from each of the cell types. The reported *P* values were based on the overlap of associated SNPs with strong enhancer states 4 and 5. We controlled for non-independence between proximal SNPs by using a randomization test where SNPs were randomly shifted while preserving relative distance. We then defined an estimated false-discovery rate based on permutations in which SNPs were randomly re-assigned to different studies, and recomputed *P* values. Estimates of false-discovery rates based on these permutations control for multiple testing of studies and cell types and for general non-specific enrichments for states 4 and 5 with GWAS hits. Candidate gene targets were predicted for a subset of variants associated with enhancer states on the basis of the lead cell type using the linking method described above.

Data access. Data sets are available from the ENCODE website (http://genome.ucsc.edu/ENCODE), the supporting website for this paper (http://compbio.mit.edu/ENCODE_chromatin_states) and the Gene Expression Omnibus (GSE26386).

51. Ludwig, T. E. *et al.* Feeder-independent culture of human embryonic stem cells. *Nature Methods* **3,** 637–646 (2006).
52. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnol.* **26,** 317–325 (2008).
53. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38,** 500–501 (2006).
54. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
55. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65 (2007).
56. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
57. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7,** 191 (2006).
58. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
59. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31,** 374–378 (2003).
60. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–D94 (2004).
61. Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133,** 1266–1276 (2008).
62. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324,** 1720–1723 (2009).
63. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnol.* **24,** 1429–1435 (2006).
64. Touzet, H. & Varre, J. S. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol. Biol.* **2,** 15 (2007).
65. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).