Spring 2023 – Epigenetics and Systems Biology
Discussion Session (Epigenetics)
Michael K. Skinner – Biol 476/576
Week 6 (February 16)

Epigenetics (History / Molecular Processes/ Genomics)

Primary Papers

1. Cuerda-Gil and Slotkin (2016) Nat Plants 2(11):16163. (PMID: 27808230)
2. Rao et al. (2014) Cell 159:1665. (PMID: 25497547)
3. Morris and Mattick (2014) Nat Rev Genet. 15(6):423-37. (PMID: 24776770)

Discussion

Student 13 – Ref #1 above

- What is RdDM and its structure?
- What are the effects of RdDM and different mechanisms?
- How are histone modifications linked to DNA methylation?

Student 14 – Ref #2 above

- What is mitotic technology used to map genome structure?
- What are the different loops identified and role CTCF?
- What is the hypothesis on the role of chromatin looping?

Student 15 – Ref #3 above

- What are non-coding RNAs?
- How does ncRNA influence gene expression?
- How do lncRNA influence chromatin structure?

nature plants

REVIEW ARTICLE

PUBLISHED ONLINE: 3 NOVEMBER 2016 | DOI: 10.1038/NPLANTS.2016.163

Non-canonical RNA-directed DNA methylation

Diego Cuerda-Gil[1] and R. Keith Slotkin[1,2]*

Small RNA-directed DNA methylation (RdDM) has been extensively studied in plants, resulting in a deep understanding of a major 'canonical RdDM' mechanism. However, current models of canonical RdDM cannot explain how this self-perpetuating mechanism is initiated. Recent investigations into the initiation of epigenetic silencing have determined that there are several alternative 'non-canonical RdDM' pathways that function through distinct mechanisms to modify chromatin. This Review aims to illustrate the diversity of non-canonical RdDM mechanisms described to date, recognize common themes within this dizzying array of interconnected pathways, and identify the key unanswered questions remaining in this field.

Identification of RNA-directed chromatin modification first occured in plants, where the accumulation of double-stranded RNA (dsRNA) resulted in sequence-specific targeting of cytosine DNA methylation[1]. Eight years later, small interfering RNAs (siRNAs) were identified as the causal molecule directing the deposition of DNA methylation and methylation of histone H3 at lysine 9 (H3K9me), resulting in the formation of stable heterochromatin[2,3]. Subsequently, siRNA-mediated heterochromatin formation (through methylation of cytosines or H3K9) was demonstrated in *Caenorhabditis elegans*, *Drosophila* and mice (reviewed in ref. 4), confirming that small RNAs are potent drivers of heterochromatic marks throughout the eukaryotic kingdoms. In plants, unlike animals, DNA methylation is not erased every generation, but rather epigenetically inherited (reviewed in ref. 5), explaining why transgenerational epigenetic inheritance patterns were first discovered and are more frequently observed in plants. This Review focuses on the different small RNA-guided pathways that establish heritable patterns of DNA methylation. Canonical and non-canonical forms of RdDM are compared and contrasted, with specific emphasis on the interconnection between post-transcriptional silencing and the establishment of chromatin marks.

Canonical RdDM

Soon after the discovery of siRNAs[6], it was recognized that they fall into two size categories with distinct molecular functions: 21–22-nucleotide (nt) siRNAs that function in post-transcriptional gene silencing (PTGS) and 24-nt siRNAs that are specifically associated with RdDM and transcriptional silencing[3]. The proteins that produce the 24-nt siRNAs are individual members from multi-protein families that have specialized for RdDM. The 24-nt siRNAs are produced by RNA-dependent RNA polymerase 2 (RDR2) and Dicer-like 3 (DCL3), and incorporated into Argonaute 4 (AGO4) and AGO6 (reviewed in ref. 7). In addition, two key plant-specific RNA polymerase II (Pol II) paralogs, Pol IV and Pol V, were identified, which originated from the duplication and sub-functionalization of Pol II genes. In the reference plant *Arabidopsis*, these protein complexes are not necessary for messenger RNA (mRNA) transcription or plant development, but rather function in gene silencing to target transposable elements (TEs), transgenes and viruses for RdDM. The identification of Pol IV and Pol V was instrumental for the molecular dissection of the canonical RdDM mechanism, as plant Pol IV and Pol V mutants are fertile, while in fission yeast (*Schizosaccharomyces pombe*), *C. elegans* and *Drosophila* the corresponding mutations in Pol II are lethal. This, coupled with the presence of DNA methylation in plant genomes, provided plant biologists with a competitive edge to dissect and functionally characterize small RNA-directed chromatin-modification mechanisms in *Arabidopsis*.

The canonical RdDM pathway begins with the transcription of heterochromatic loci by Pol IV, and the immediate conversion of this transcript into dsRNA via RDR2 (Fig. 1; for a more comprehensive model see ref. 7). Pol IV and RDR2 physically interact and produce short (26–45-nt) dsRNA transcripts[8–11]. Double-stranded RNA from Pol IV–RDR2 is specifically cleaved by DCL3 into 24-nt siRNAs, which are incorporated into AGO4 or the closely related protein AGO6[12]. In the downstream phase of RdDM, target loci are transcribed by Pol V, which generates a non-protein-coding transcript that is thought to remain attached to its locus of origin and function as a protein scaffold[13]. If the 24-nt siRNA directs AGO4/AGO6 to the Pol V transcript (via sequence complementarity), the methyltransferase protein DRM2 is recruited and results in DNA methylation of the Pol-V-transcribed region (Fig. 1). After DNA methylation is established, heterochromatin can be formed through the recruitment and activity of H3K9 methyltransferase proteins[14,15].

Canonical RdDM does not account for all RdDM

In a 2005 paper seminal to the RdDM field, Herr *et al.*[16] discovered that Pol IV functions in RdDM to silence endogenous TE expression. However, they found that Pol IV was not responsible for the RdDM of an infecting virus, and that the key protein for RdDM and silencing of viral DNA during infection was actually RDR6. RDR6 functions to produce dsRNA in PTGS, the mechanism responsible for the degradation of viral, transgene and TE Pol II mRNAs (Fig. 2; reviewed in ref. 17). Therefore, as far back as 2005, it was known that canonical RdDM was not the only RdDM mechanism, but the fact that alternative mechanisms of RdDM exist was largely overlooked over the next decade as the field focused on the molecular dissection of canonical RdDM. Focus did not return to alternative forms of RdDM until the finding that canonical RdDM did not function alone to initiate viral silencing was independently confirmed[18], and many proteins assumed to only function in PTGS were found to influence whole-genome DNA methylation patterns[19].

[1]Department of Molecular Genetics, The Ohio State University, 500 Aronoff Laboratory, 318 West Twelfth Avenue, Columbus, Ohio 43210, USA. [2]Center for RNA Biology, The Ohio State University, 105 Biological Sciences Building, 484 West Twelfth Avenue, Columbus, Ohio 43210, USA.
*e-mail: slotkin.2@osu.edu

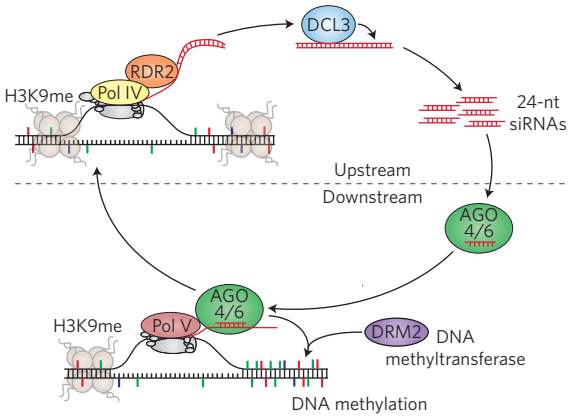


Figure 1 | The canonical RdDM pathway. In the upstream siRNA-generating phase, Pol IV and RDR2 function to produce dsRNA, which is cleaved into 24-nt siRNAs by DCL3. These siRNAs function to direct AGO4 or AGO6 to chromatin-bound transcripts produced by Pol V. In the downstream phase, the interaction of AGO4/6 with the Pol V transcript results in the recruitment and activity of the DNA methyltransferase DRM2. This pathway is reviewed in greater detail in ref. 7. Black strands represent DNA, red strands represent RNA. DNA methylation colours refer to the CG (red), CHG (blue) and CHH (green) sequence contexts, where H = A, C or T.

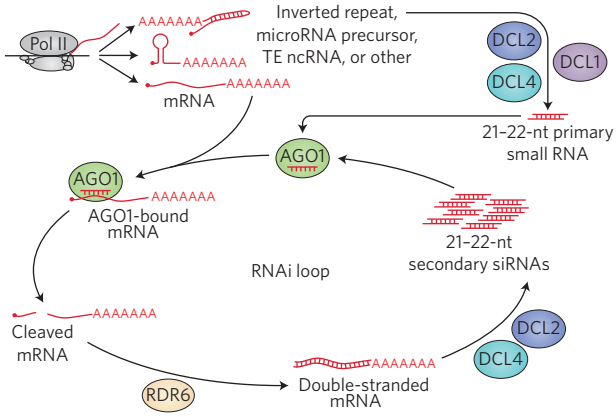


Figure 2 | PTGS in plants. Pol II transcription of regions such as palindromic TEs or microRNA precursors generates primary small RNAs[50,51], which are produced independently of RDR activity. Primary small RNAs target Pol II mRNAs for cleavage, and some of these cleaved mRNAs produce secondary siRNAs through the formation of dsRNA via RDR6 and cleavage by DCL2 and DCL4. Secondary siRNAs can target additional copies of the mRNA for cleavage and production of more siRNAs through the cycle of RNAi. ncRNA, non-coding RNA.

Mechanisms of non-canonical RdDM

Besides the production of 24-nt siRNAs by Pol IV–RDR2–DCL3, several additional small RNA pathways can direct RdDM and are referred to as non-canonical RdDM mechanisms. These varied mechanisms of small RNA production use different entry points (such as Pol-II-derived mRNAs) to feed into the canonical RdDM pathway (Figs 3,4). The various forms of non-canonical RdDM are described below. It is important to point out, however, that these mechanisms play minor roles compared with canonical RdDM, and are neither characterized nor understood to the same extent.

Inverted repeat and microRNA-directed DNA methylation. As with Pol IV–RDR2-derived transcripts, Pol II transcripts can also be cleaved by DCL3 into 24-nt small RNAs that participate in RdDM[20] (Fig. 3a). These small RNAs are produced independently of RDR activity[20], probably from transcripts that form imperfectly paired intramolecular dsRNA (hairpins). DCL proteins compete for these dsRNA substrates in a hierarchical fashion, which typically results in production of 21–22- and 24-nt small RNAs[21]. The 24-nt small RNAs are incorporated into AGO4 and mediate RdDM through the Pol V chromatin-bound downstream components of the canonical RdDM pathway (Fig. 3a). Intramolecular dsRNA is produced from transcription through inverted repeat (IR) sequences, and once cleaved, siRNAs from IRs can target RdDM in *cis* (the IR itself) or in *trans*[22]. A functional example of this Pol II–DCL3 RdDM pathway occurs in maize: a long Pol-II-derived IR of a Mutator family TE is cleaved into 24-nt siRNAs and directs *trans*-RdDM and epigenetic transcriptional silencing to the rest of this TE family[23]. Genome-wide analysis determined that the Pol II–DCL3 RdDM pathway targets many TEs when they are transcribed, contributing to their resilencing[20].

21–22-nt microRNAs are produced from short intramolecular stem-loop mRNA structures cleaved by DCL1, and function by the post-transcriptional or translational silencing of their target mRNAs. MicroRNAs are thought to evolve from longer IRs by reduction of the dsRNA length to only 20–30-nt of a stem-loop structure[24]. However, for some microRNA precursor transcripts, the small size of the dsRNA is sufficient for DCL3 cleavage into 24-nt small RNAs that participate in RdDM (Fig. 3b). In rice, moss and *Arabidopsis*, DCL3 competes with DCL1 for some microRNA stem-loops and cleaves these into 24-nt small RNAs[22,25,26]. MicroRNA-directed DNA methylation can occur in *cis* or *trans*, and can spread from its target site to repress adjacent regulatory features and affect gene expression[26].

The RDR6 RdDM pathway. 21–22-nt siRNAs produced through PTGS (involving Pol II–RDR6–DCL2–DCL4; Fig. 2) can also participate in RdDM. This pathway was first identified at *TAS* genes[27], which produce non-protein-coding Pol II mRNAs. These mRNAs are targeted for cleavage by microRNAs and subsequently converted into dsRNA by RDR6[28]. While RdDM is likely an off-target effect on *TAS* loci, where it does not alter TAS expression or siRNA production[27], this same mechanism was later discovered to target transcriptionally active TEs and play a critical role in the initiation and establishment of TE silencing[29]. This pathway, termed RDR6 RdDM, is not dependent on the canonical RdDM components Pol IV, RDR2 or DCL3, and further investigation uncovered the direct incorporation of 21–22-nt siRNAs produced from Pol II–RDR6-derived TE mRNAs into AGO6 (Fig. 3c)[30]. Once loaded with 21–22-nt siRNAs, AGO6 interacts with its target loci through a Pol V scaffolding transcript (which continues to associate with transcriptionally active TEs). RDR6 RdDM requires Pol V and DRM2, suggesting that the downstream targeting complex of RDR6 RdDM is the same as in canonical RdDM. On the genome-wide level, RDR6 RdDM acts on many long, full-length, structurally autonomous TEs when they are transcriptionally active[20].

RDR6–DCL3 RdDM pathway. Marí-Ordóñez *et al.*[31] also found that TE mRNAs converted into dsRNA by RDR6 can feed into RdDM, but in their example, RdDM occurred through dsRNA cleavage by DCL3, producing 24-nt siRNAs (Fig. 4a). The authors theorized that with an increasing TE copy number and high levels of RDR6-produced dsRNA, a threshold is crossed where DCL2/DCL4 become overwhelmed, and DCL3 compensates by producing 24-nt siRNAs from transcripts that are not typically DCL3 substrates. This mechanism relies on the known hierarchy of DCL activity, as in a *dcl2*/*dcl4* double mutant, some RDR6-generated dsRNAs are targeted by DCL3[32]. Thus, this pathway may represent an important mechanism that can detect high copy numbers or elevated expression of TEs and transgenes, and initiate RdDM when PTGS is saturated.

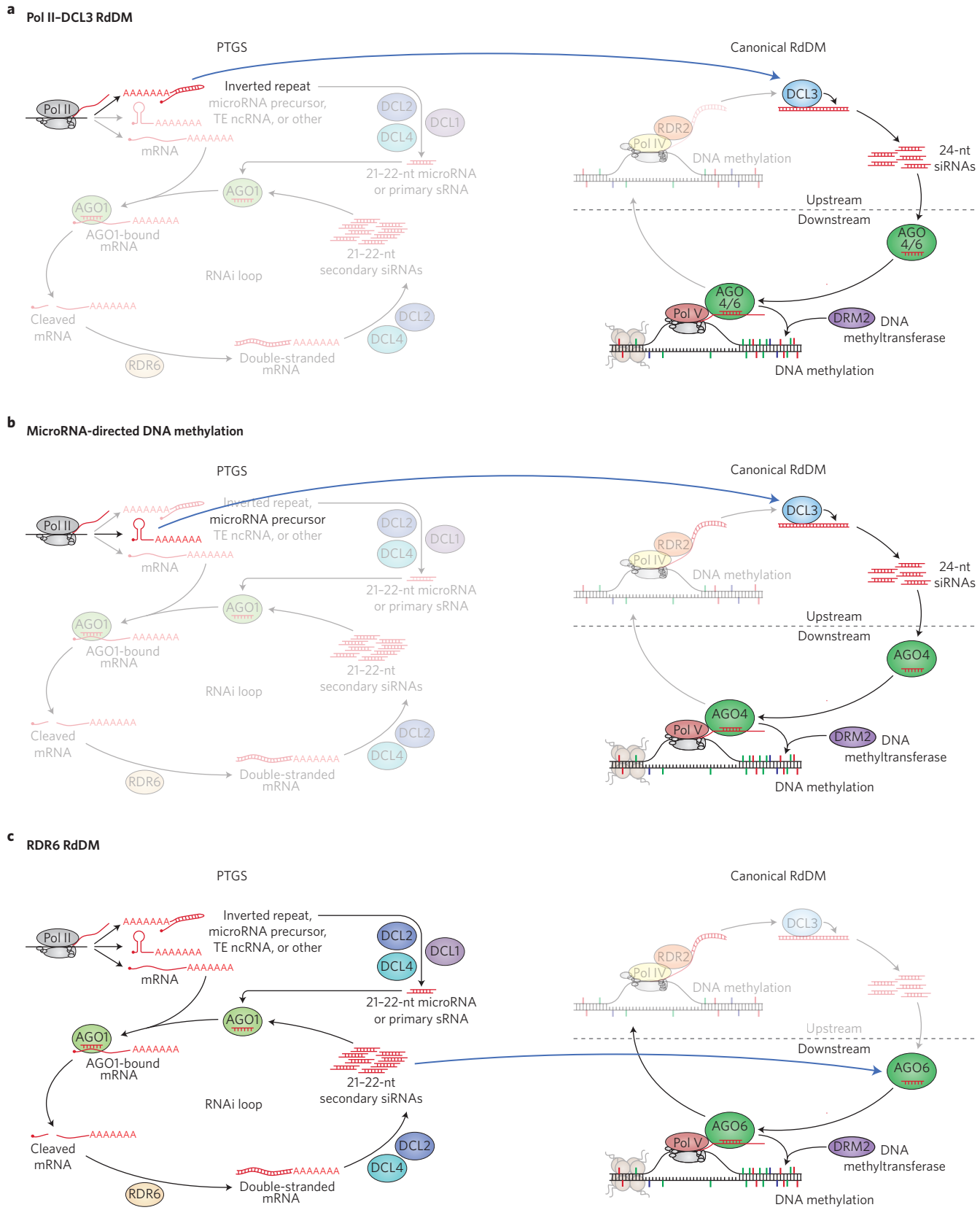


Figure 3 | Pol II transcripts target DNA methylation through the non-canonical RdDM pathways. a–c, Canonical RdDM (described in Fig. 1) is mechanistically connected to PTGS (described in Fig. 2) via at least six different non-canonical RdDM pathways (blue arrows), including: Pol II–DCL3 RdDM (**a**), microRNA-directed DNA methylation (**b**), and RDR6 RdDM (**c**). The remaining three non-canonical RdDM pathways linking PTGS to canonical RdDM are displayed in Fig. 4.

Pol IV–NERD RdDM pathway. Pontier *et al.*[33] identified an RdDM pathway that targets several intergenic loci and requires Pol IV and Pol V, yet is independent of RDR2 and AGO4 (Fig. 4b). Central to this pathway is a protein named NERD (Needed for RDR2-independent DNA methylation). GW (glycine–tryptophan) repeats in the NERD protein mediate the interaction of AGO2 with Pol IV and Pol V, and recruitment of NERD–AGO2 to chromatin results in RdDM and transcriptionally repressive histone tail modifications. Although the size and type of siRNAs that drive the Pol IV–NERD pathway remains to be determined, the target loci lose DNA methylation in *rdr1*, *rdr6*, *sde3*, *dcl2*, *dcl3* and *dcl4* mutants, suggesting that this pathway uses a combination of canonical RdDM and traditional PTGS proteins[33,34] (Fig. 4b).

Double-strand breaks recruit RdDM-related proteins. Upon DNA damage induced by a double-strand break, a RdDM-like complex is recruited and is necessary for efficient repair[35]. This repair pathway produces 21-nt siRNAs that are dependent on RDR6 and DCL2/DCL4, similar to PTGS or RDR6 RdDM, but differ in the fact that the primary transcripts responsible for this pathway are dependent on Pol IV, instead of Pol II. Double-strand break-induced 21-nt siRNAs participate in DNA break repair through their incorporation into AGO2 and presumed targeting of a Pol V scaffolding transcript[35]. Supporting the potential role of AGO2 directly interacting with the Pol V chromatin-bound complex, the closely related AGO3 protein was recently found to drive RdDM through its interaction with 24-nt siRNAs[36], suggesting that AGO proteins outside of the AGO4/AGO6/AGO9 clade can interact with the Pol V chromatin complex to direct RdDM.

Although this double-strand break repair pathway utilizes many of the proteins involved in canonical and non-canonical RdDM, DNA methylation has not been reported at the sites of DNA repair. Mechanistically, this pathway of siRNA production is most similar to the Pol IV–NERD pathway (Fig. 4b). However, NERD has not been specifically tested for a role in DNA damage repair, and a formal link between double-strand break repair and RdDM has not been established.

Dicer-independent RdDM. Recent publications have defined a Dicer-independent mechanism by which non-diced dsRNAs are incorporated into the AGO4 protein, and subsequently trimmed down at their 3' end to the appropriate siRNA size by exosome-core complex exonucleases[37,38]. The Dicer-independent mechanism produces an array of siRNA size classes including 21–24-nt siRNAs, and these Dicer-independent siRNAs (termed sidRNAs) contribute to RdDM of their targets (Fig. 4c). Thus far, it is unclear how significant the role of sidRNAs are outside the *dcl2*/*dcl3*/*dcl4* or *dcl1*/*dcl2*/*dcl3*/*dcl4* mutant contexts, but since sidRNAs can be generated from either Pol IV–RDR2 or Pol II–RDR6 dsRNA transcripts (two blue arrows leading to AGO4 in Fig. 4c), the authors of this study suggest that this pathway functions to initiate TE silencing[37].

Commonalities between non-canonical forms of RdDM

Five broad (and imperfect) commonalities between non-canonical RdDM pathways have been observed. First, like canonical RdDM, they all function through a similar, ancient mechanism of small RNA production followed by incorporation into an AGO complex and subsequent targeting of a Pol-V-derived chromatin-bound scaffolding transcript. Second, non-canonical RdDM targets few loci in wild-type cells and, overall, canonical RdDM targets more genome-wide regions than non-canonical RdDM[20]. Canonical RdDM perpetually targets thousands of small TEs near genes[39], presumably to maintain the sharp boundary between heterochromatic (TE) and euchromatic (gene) regions of the genome[40]. In wild-type cells, non-canonical RdDM targets only a few regions that produce Pol-II-generated siRNAs such as IRs or *TAS* loci[20,27]. For these regions, the Pol II promoters must be insulated from the spread of RdDM in order to continuously produce the mRNA substrates for non-canonical RdDM. Third, many of the non-canonical RdDM mechanisms require Pol II transcription of an mRNA, while in contrast canonical RdDM is independent of Pol II. Fourth, the non-canonical pathways may act transiently and, potentially, only in some cell types (such as in meristems for RDR6 RdDM[30]), to initiate silencing and transition to canonical RdDM. Fifth, the non-canonical RdDM pathways are not completely independent of the canonical pathway, and many loci are simultaneously targeted by canonical and non-canonical RdDM[20,29,41]. The non-canonical pathways represent alternative entry points from which small RNAs can feed into the canonical RdDM pathway, and are often dependent on canonical RdDM components such as DCL3 or AGO6 (Figs 3,4). Most notably, the downstream chromatin-bound phase of canonical RdDM (involving Pol V, DRM2 and others described in ref. 7) is required for all of the non-canonical RdDM mechanisms examined.

Functions of non-canonical RdDM

The current understanding is that both Pol IV and Pol V are recruited to participate in canonical RdDM of target loci through previously established heterochromatic marks. Pol IV is recruited to regions of the genome associated with GC (guanine–cytosine) context DNA methylation, histone deacetylation and H3K9me through the intermediary protein SHH1 (refs 42,43), while Pol V is recruited to regions of the genome associated with DNA methylation through SUVH2/SUVH9 (refs 44,45). Thus, with our current understanding, canonical RdDM cannot alone explain how heterochromatic marks are initiated at a target locus. Because many of the non-canonical RdDM mechanisms start with Pol II transcription and mRNA substrates, several of these mechanisms have been implicated in the initiation of virus, transgene and TE silencing[46,47]. For example, the initiation of TE silencing has been reported to function through a myriad of non-canonical RdDM mechanisms: RDR6–DCL3 RdDM[31], RDR6 RdDM[29], the Pol IV–NERD pathway[33] and the Dicer-independent pathway[37], and even the double-strand break repair pathway has been theorized to play a role in the initiation of TE silencing as new TE insertions must create a DNA break to insert into the genome. It is likely that these pathways act in a non-mutually exclusive manner to flood Pol II transcripts into various non-canonical RdDM mechanisms in any way possible, in order to efficiently target a TE, transgene or virus for chromatin-level transcriptional silencing. In addition, there are probably more interconnections and mechanisms to be discovered, since some silencing events are distinct from any of the known canonical or non-canonical RdDM pathways[46]. Therefore, the overall current dogma is that the function of non-canonical RdDM is to link PTGS to canonical RdDM (Figs 3,4). Once non-canonical RdDM initiates heterochromatic marks on a particular TE, virus or transgene, canonical RdDM is probably recruited through these marks to reinforce RdDM, form heterochromatin, and transcriptionally silence the locus (which many non-canonical RdDM mechanisms cannot perform due to their constant dependence on Pol II transcripts).

Key remaining questions for this field

- Is there an RdDM mechanism that is completely independent of both Pol IV and Pol V? Can Pol II alone produce both the siRNA-generating transcript and the downstream scaffold transcript required for RdDM? Multiple examples of Pol II production of siRNAs used in RdDM now exist (see above), and Pol II may be able to generate the scaffolding transcript as it interacts with AGO4 and is necessary for RdDM at some target loci[48]. Recent data suggests that Pol II can target a low but measurable level of RdDM without either Pol IV or Pol V to high-copy viral genomes, but cannot deposit H3K9me without Pol IV RdDM[41].

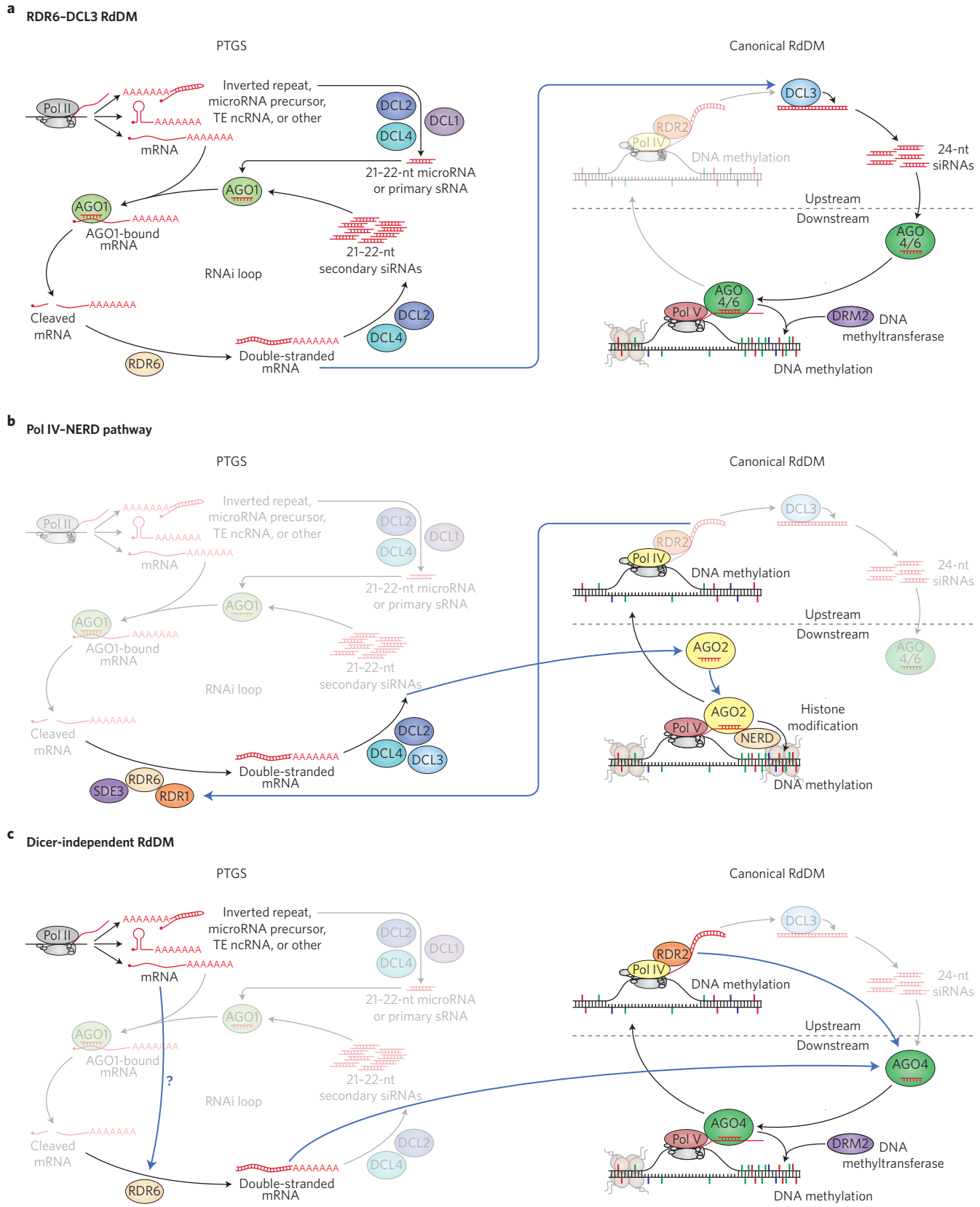


Figure 4 | Less understood mechanisms of non-canonical RdDM. a–c, Three additional non-canonical RdDM pathways link PTGS to canonical RdDM (blue arrows): RDR6–DCL3 RdDM (**a**), the Pol IV–NERD pathway (**b**), and Dicer-independent RdDM (**c**). For the Pol IV–NERD pathway (**b**), RNAi factors such as RDR6, RDR1, DCL2, DCL4 and the RNA helicase SDE3 influence DNA methylation, but the precise small RNAs that function in this pathway have not been identified. In addition, data suggests that Pol IV–NERD-induced DNA methylation is indirect, as chromatin modification probably occurs before DNA methylation in this pathway[33].

- Are Pol IV and Pol V recruited to regions of the genome that are not yet associated with heterochromatic marks? Does Pol V scan the whole genome surveying sequences for RdDM?
- Do small RNAs directly target H3K9me in plants? DNA methylaton and H3K9me are tightly interconnected and deposition of one mark is known to influence the second mark[49]. The fact that small RNAs direct H3K9me in fission yeast, *C. elegans* and *Drosophila* (which do not have DNA methylation) suggests that H3K9me modification can be directly targeted.
- Why are there so many small RNA-generating mechanisms that result in RdDM? Is biological redundancy required to generate small RNAs in any way possible for TEs, transgenes and viruses?
- What RNA quality control mechanisms drive transcripts into RNA interference (RNAi) and thus the non-canonical RdDM pathways?
- How do genes avoid non-canonical (and particularly Pol-II-dependent) RdDM mechanisms? Some post-transcriptionally degraded genes and transgenes produce abundant 21-nt siRNAs but are not methylated. Do these genes and transgenes completely avoid RdDM, or are they targeted and the chromatin modifications subsequently erased by DNA glycosylases and histone demethylases?
- In animals with Piwi-interacting RNAs, do endogenous siRNAs target chromatin modification as in fungi and plants?
- Why were non-canonical RdDM components generally not identified in screens for the initiation of transgene silencing? Are the various non-canonical RdDM mechanisms redundant, providing multiple routes to the initiation of silencing, while the canonical RdDM components identified in these screens are necessary/unique?

Received 27 April 2016; accepted 26 September 2016; published 3 November 2016; corrected 12 December 2016

References

1. Wassenegger, M., Heimes, S., Riedel, L. & Sänger, H. L. RNA-directed *de novo* methylation of genomic sequences in plants. *Cell* **76,** 567–576 (1994).
2. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297,** 1833–1837 (2002).
3. Hamilton, A., Voinnet, O., Chappell, L. & Baulcombe, D. Two classes of short interfering RNA in RNA silencing. *EMBO J.* **21,** 4671–4679 (2002).
4. Castel, S. E. & Martienssen, R. A. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.* **14,** 100–112 (2013).
5. Heard, E. & Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157,** 95–109 (2014).
6. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286,** 950–952 (1999).
7. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15,** 394–408 (2014).
8. Law, J. A., Vashisht, A. A., Wohlschlegel, J. A. & Jacobsen, S. E. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genet.* **7,** e1002195 (2011).
9. Haag, J. R. *et al. In vitro* transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol. Cell* **48,** 811–818 (2012).
10. Blevins, T. *et al.* Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding *de novo* DNA methylation in *Arabidopsis*. *Elife* **4,** e09591 (2015).
11. Zhai, J. *et al.* A one precursor one siRNA model for Pol IV-dependent siRNA biogenesis. *Cell* **163,** 445–455 (2015).
12. Havecker, E. R. *et al.* The *Arabidopsis* RNA-directed DNA methylation Argonautes functionally diverge based on their expression and interaction with target loci. *Plant Cell* **22,** 321–334 (2010).
13. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).
14. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416,** 556–560 (2002).
15. Ebbs, M. L. & Bender, J. Locus-specific control of DNA methylation by the *Arabidopsis* SUVH5 histone methyltransferase. *Plant Cell* **18,** 1166–1176 (2006).
16. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308,** 118–120 (2005).
17. Liu, L. & Chen, X. RNA quality control as a key to suppressing RNA silencing of endogenous genes in plants. *Mol. Plant* **9,** 826–836 (2016).
18. Raja, P., Sanville, B. C., Buchmann, R. C. & Bisaro, D. M. Viral genome methylation as an epigenetic defense against geminiviruses. *J. Virol.* **82,** 8997–9007 (2008).
19. Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V. & Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell* **152,** 352–364 (2013).
20. Panda, K. *et al.* Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol.* **17,** 170 (2016).
21. Henderson, I. R. *et al.* Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat. Genet.* **38,** 721–725 (2006).
22. Wu, L. *et al.* DNA methylation mediated by a microRNA pathway. *Mol. Cell* **38,** 465–475 (2010).
23. Slotkin, R. K., Freeling, M. & Lisch, D. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat. Genet.* **37,** 641–644 (2005).
24. Vazquez, F., Blevins, T., Ailhas, J., Boller, T. & Meins, F. Evolution of *Arabidopsis MIR* genes generates novel microRNA classes. *Nucleic Acids Res.* **36,** 6429–6438 (2008).
25. Chellappan, P. *et al.* siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic Acids Res.* **38,** 6883–6894 (2010).
26. Khraiwesh, B. *et al.* Transcriptional control of gene expression by microRNAs. *Cell* **140,** 111–122 (2010).
27. Wu, L., Mao, L. & Qi, Y. Roles of DICER-LIKE and ARGONAUTE proteins in TAS-derived small interfering RNA-triggered DNA methylation. *Plant Physiol.* **160,** 990–999 (2012).
28. Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-directed phasing during *trans*-acting siRNA biogenesis in plants. *Cell* **121,** 207–221 (2005).
29. Nuthikattu, S. *et al.* The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21–22 nucleotide small interfering RNAs. *Plant Physiol.* **162,** 116–131 (2013).
30. McCue, A. D. *et al.* ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *EMBO J.* **34,** 20–35 (2015).
31. Marí-Ordóñez, A. *et al.* Reconstructing *de novo* silencing of an active plant retrotransposon. *Nat. Genet.* **45,** 1029–1039 (2013).
32. Gasciolli, V., Mallory, A. C., Bartel, D. P. & Vaucheret, H. Partially redundant functions of *Arabidopsis* DICER-like enzymes and a role for DCL4 in producing *trans*-acting siRNAs. *Curr. Biol.* **15,** 1494–1500 (2005).
33. Pontier, D. *et al.* NERD, a plant-specific GW protein, defines an additional RNAi-dependent chromatin-based pathway in *Arabidopsis*. *Mol. Cell* **48,** 121–132 (2012).
34. Garcia, D. *et al.* Ago hook and RNA helicase motifs underpin dual roles for SDE3 in antiviral defense and silencing of nonconserved intergenic regions. *Mol. Cell* **48,** 109–120 (2012).
35. Wei, W. *et al.* A role for small RNAs in DNA double-strand break repair. *Cell* **149,** 101–112 (2012).
36. Zhang, Z., Liu, X., Guo, X., Wang, X.-J. & Zhang, X. *Arabidopsis* AGO3 predominantly recruits 24-nt small RNAs to regulate epigenetic silencing. *Nat. Plants* **2,** 16049 (2016).
37. Ye, R. *et al.* A Dicer-independent route for biogenesis of siRNAs that direct DNA methylation in *Arabidopsis*. *Mol. Cell* **61,** 222–235 (2016).
38. Yang, D.-L. *et al.* Dicer-independent RNA-directed DNA methylation in *Arabidopsis*. *Cell Res.* **26,** 66–82 (2016).
39. Zemach, A. *et al.* The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **153,** 193–205 (2013).
40. Li, Q. *et al.* RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc. Natl Acad. Sci. USA* **112,** 14728–14733 (2015).
41. Jackel, J. N., Storer, J. M., Coursey, T. & Bisaro, D. M. *Arabidopsis* RNA polymerases IV and V are required to establish H3K9 methylation, but not cytosine methylation, on geminivirus chromatin. *J. Virol.* **90,** 7529–7540 (2016).
42. Law, J. A. *et al.* Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498,** 385–389 (2013).
43. Blevins, T. *et al.* A two-step process for epigenetic inheritance in *Arabidopsis*. *Mol. Cell* **54,** 30–42 (2014).

44. Johnson, L. M. *et al.* SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature* **507,** 124–128 (2014).
45. Liu, Z.-W. *et al.* The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. *PLoS Genet.* **10,** e1003948 (2014).
46. Bond, D. M. & Baulcombe, D. C. Epigenetic transitions leading to heritable, RNA-mediated *de novo* silencing in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **112,** 917–922 (2015).
47. Fultz, D., Choudury, S. G. & Slotkin, R. K. Silencing of active transposable elements in plants. *Curr. Opin. Plant Biol.* **27,** 67–76 (2015).
48. Zheng, B. *et al.* Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in *Arabidopsis*. *Genes Dev.* **23,** 2850–2860 (2009).
49. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16,** 519–532 (2015).
50. Creasey, K. M. *et al.* miRNAs trigger widespread epigenetically activated siRNAs from transposons in *Arabidopsis*. *Nature* **508,** 411–415 (2014).
51. Bousios, A. *et al.* A role for palindromic structures in the *cis*-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response. *Genome Res.* **26,** 226–237 (2016).

Acknowledgements

D.C.-G. is supported by a Nuevo Leon state fellowship from the Mexico National Council of Science and Technology. The Slotkin lab is supported by U.S. National Science Foundation grants MCB-1252370 and MCB-1608392.

Additional information

Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to R.K.S.

Competing interests

The authors declare no competing interests.

Corrigendum: Non-canonical RNA-directed DNA methylation

Diego Cuerda-Gil and R. Keith Slotkin

Nature Plants **2,** 16163 (2016); published 3 November 2016; corrected 12 December 2016.

In the version of this Review originally published, the following NSF grants should have been acknowledged: MCB-1252370 and MCB-1608392. This has been corrected in all versions of the Review. In addition, a typographical error in the section 'Key remaining questions for this field' has been amended so that the sentence reads: "What RNA quality control mechanisms drive transcripts into RNA interference (RNAi) and thus the non-canonical RdDM pathways?"

Article

Cell

A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping

Suhas S.P. Rao,[1,2,3,4,10] Miriam H. Huntley,[1,2,3,4,5,10] Neva C. Durand,[1,2,3,4] Elena K. Stamenova,[1,2,3,4] Ivan D. Bochkov,[1,2,3] James T. Robinson,[1,4] Adrian L. Sanborn,[1,2,3,6] Ido Machol,[1,2,3] Arina D. Omer,[1,2,3] Eric S. Lander,[4,7,8,*] and Erez Lieberman Aiden[1,2,3,4,9,*]

[1]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[3]Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA
[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[6]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[7]Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA
[8]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[9]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA
[10]Co-first author
*Correspondence: lander@broadinstitute.org (E.S.L.), erez@erez.com (E.L.A.)
http://dx.doi.org/10.1016/j.cell.2014.11.021

SUMMARY

We use in situ Hi-C to probe the 3D architecture of genomes, constructing haploid and diploid maps of nine cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1 kb resolution. We find that genomes are partitioned into contact domains (median length, 185 kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs "facing" one another. The inactive X chromosome splits into two massive domains and contains large loops anchored at CTCF-binding repeats.

INTRODUCTION

The spatial organization of the human genome is known to play an important role in the transcriptional control of genes (Cremer and Cremer, 2001; Sexton et al., 2007; Bickmore, 2013). Yet important questions remain, like how distal regulatory elements, such as enhancers, affect promoters, and how insulators can abrogate these effects (Banerji et al., 1981; Blackwood and Kadonaga, 1998; Gaszner and Felsenfeld, 2006). Both phenomena are thought to involve the formation of protein-mediated "loops" that bring pairs of genomic sites that lie far apart along the linear genome into proximity (Schleif, 1992).

Various methods have emerged to assess the 3D architecture of the nucleus. In one seminal study, the binding of a protein to sites at opposite ends of a restriction fragment created a loop, which was detectable because it promoted the formation of DNA circles in the presence of ligase. Removal of the protein or either of its binding sites disrupted the loop, eliminating this "cyclization enhancement" (Mukherjee et al., 1988). Subsequent adaptations of cyclization enhancement made it possible to analyze chromatin folding in vivo, including nuclear ligation assay (Cullen et al., 1993) and chromosome conformation capture (Dekker et al., 2002), which analyze contacts made by a single locus, extensions such as 5C for examining several loci simultaneously (Dostie et al., 2006), and methods such as ChIA-PET for examining all loci bound by a specific protein (Fullwood et al., 2009).

To interrogate all loci at once, we developed Hi-C, which combines DNA proximity ligation with high-throughput sequencing in a genome-wide fashion (Lieberman-Aiden et al., 2009). We used Hi-C to demonstrate that the genome is partitioned into numerous domains that fall into two distinct compartments. Subsequent analyses have suggested the presence of smaller domains and have led to the important proposal that compartments are partitioned into condensed structures ~1 Mb in size, dubbed "topologically associated domains" (TADs) (Dixon et al., 2012; Nora et al., 2012). In principle, Hi-C could also be used to detect loops across the entire genome. To achieve this, however, extremely large data sets and rigorous computational methods are needed. Recent efforts have suggested that this is an increasingly plausible goal (Sexton et al., 2012; Jin et al., 2013).

Here, we report the results of an effort to comprehensively map chromatin contacts genome-wide, using in situ Hi-C, in which DNA-DNA proximity ligation is performed in intact nuclei. The protocol facilitates the generation of much denser Hi-C maps. The maps reported here comprise over 5 Tb of sequence

data recording over 15 billion distinct contacts, an order of magnitude larger than all published Hi-C data sets combined. Using these maps, we are able to clearly discern domain structure, compartmentalization, and thousands of chromatin loops. In addition to haploid maps, we were also able to create diploid maps analyzing each chromosomal homolog separately. The maps provide a picture of genomic architecture with resolution down to 1 kb.

RESULTS

In Situ Hi-C Methodology and Maps

Our in situ Hi-C protocol combines our original Hi-C protocol (here called dilution Hi-C) with nuclear ligation assay (Cullen et al., 1993), in which DNA is digested using a restriction enzyme, DNA-DNA proximity ligation is performed in intact nuclei, and the resulting ligation junctions are quantified. Our in situ Hi-C protocol involves crosslinking cells with formaldehyde, permeabilizing them with nuclei intact, digesting DNA with a suitable 4-cutter restriction enzyme (such as MboI), filling the 5′-overhangs while incorporating a biotinylated nucleotide, ligating the resulting blunt-end fragments, shearing the DNA, capturing the biotinylated ligation junctions with streptavidin beads, and analyzing the resulting fragments with paired-end sequencing (Figure 1A). This protocol resembles a recently published single-cell Hi-C protocol (Nagano et al., 2013), which also performed DNA-DNA proximity ligation inside nuclei to study nuclear architecture in individual cells. Our updated protocol has three major advantages over dilution Hi-C. First, in situ ligation reduces the frequency of spurious contacts due to random ligation in dilute solution—as evidenced by a lower frequency of junctions between mitochondrial and nuclear DNA in the captured fragments and by the higher frequency of random ligations observed when the supernatant is sequenced (Extended Experimental Procedures available online). This is consistent with a recent study showing that ligation junctions formed in solution are far less meaningful (Gavrilov et al., 2013). Second, the protocol is faster, requiring 3 days instead of 7 (Extended Experimental Procedures). Third, it enables higher resolution and more efficient cutting of chromatinized DNA, for instance, through the use of a 4-cutter rather than a 6-cutter (Data S1, I).

A Hi-C map is a list of DNA-DNA contacts produced by a Hi-C experiment. By partitioning the linear genome into "loci" of fixed size (e.g., bins of 1 Mb or 1 kb), the Hi-C map can be represented as a "contact matrix" M, where the entry $M_{i,j}$ is the number of contacts observed between locus L_i and locus L_j. (A "contact" is a read pair that remains after we exclude reads that are duplicates, that correspond to unligated fragments, or that do not align uniquely to the genome.) The contact matrix can be visualized as a heatmap, whose entries we call "pixels." An "interval" refers to a set of consecutive loci; the contacts between two intervals thus form a "rectangle" or "square" in the contact matrix. We define the "matrix resolution" of a Hi-C map as the locus size used to construct a particular contact matrix and the "map resolution" as the smallest locus size such that 80% of loci have at least 1,000 contacts. The map resolution is meant to reflect the finest scale at which one can reliably discern local features.

Contact Maps Spanning Nine Cell Lines Containing over 15 Billion Contacts

We constructed in situ Hi-C maps of nine cell lines in human and mouse (Table S1). Whereas our original Hi-C experiments had a map resolution of 1 Mb, these maps have a resolution of 1 kb or 5 kb. Our largest map, in human GM12878 B-lymphoblastoid cells, contains 4.9 billion pairwise contacts and has a map resolution of 950 bp ("kilobase resolution") (Table S2). We also generated eight in situ Hi-C maps at 5 kb resolution, using cell lines representing all human germ layers (IMR90, HMEC, NHEK, K562, HUVEC, HeLa, and KBM7) as well as mouse B-lymphoblasts (CH12-LX) (Table S1). Each map contains between 395 M and 1.1 B contacts.

When we used our original dilution Hi-C protocol to generate maps of GM12878, IMR90, HMEC, NHEK, HUVEC, and CH12-LX, we found that, as expected, in situ Hi-C maps were superior at high resolutions, but closely resembled dilution Hi-C at lower resolutions. For instance, our dilution map of GM12878 (3.2 billion contacts) correlated highly with our in situ map at 500, 50, and 25 kb resolutions ($R > 0.96$, 0.90, and 0.87, respectively) (Data S1, I; Figure S1).

We also performed 112 supplementary Hi-C experiments using three different protocols (in situ Hi-C, dilution Hi-C, and Tethered Conformation Capture) while varying a wide array of conditions such as extent of crosslinking, restriction enzyme, ligation volume/time, and biotinylated nucleotide. These include several in situ Hi-C experiments in which the formaldehyde crosslinking step was omitted, which demonstrate that the structural features we observe cannot be due to the crosslinking procedure. In total, 201 independent Hi-C experiments were successfully performed, many of which are presented in Data S1 and S2.

To account for nonuniformities in coverage due to the number of restriction sites at a locus or the accessibility of those sites to cutting (Lieberman-Aiden et al., 2009; Yaffe and Tanay, 2011) we use a matrix-balancing algorithm due to Knight and Ruiz (2012) (Extended Experimental Procedures).

Adequate tools for visualization of these large data sets are essential. We have therefore created the "Juicebox" visualization system that enables users to explore contact matrices, zoom in and out, compare Hi-C matrices to 1D tracks, superimpose all features reported in this paper onto the data, and contrast different Hi-C maps. All contact data and feature sets reported here can be explored interactively via Juicebox at http://www.aidenlab.org/juicebox/.

The Genome Is Partitioned into Small Domains Whose Median Length Is 185 kb

We began by probing the 3D partitioning of the genome. In our earlier experiments at 1 Mb map resolution (Lieberman-Aiden et al., 2009), we saw large squares of enhanced contact frequency tiling the diagonal of the contact matrices. These squares partitioned the genome into 5–20 Mb intervals, which we call "megadomains."

We also found that individual 1 Mb loci could be assigned to one of two long-range contact patterns, which we called compartments A and B, with loci in the same compartment showing more frequent interaction. Megadomains—and the associated squares along the diagonal—arise when all of the 1 Mb loci in

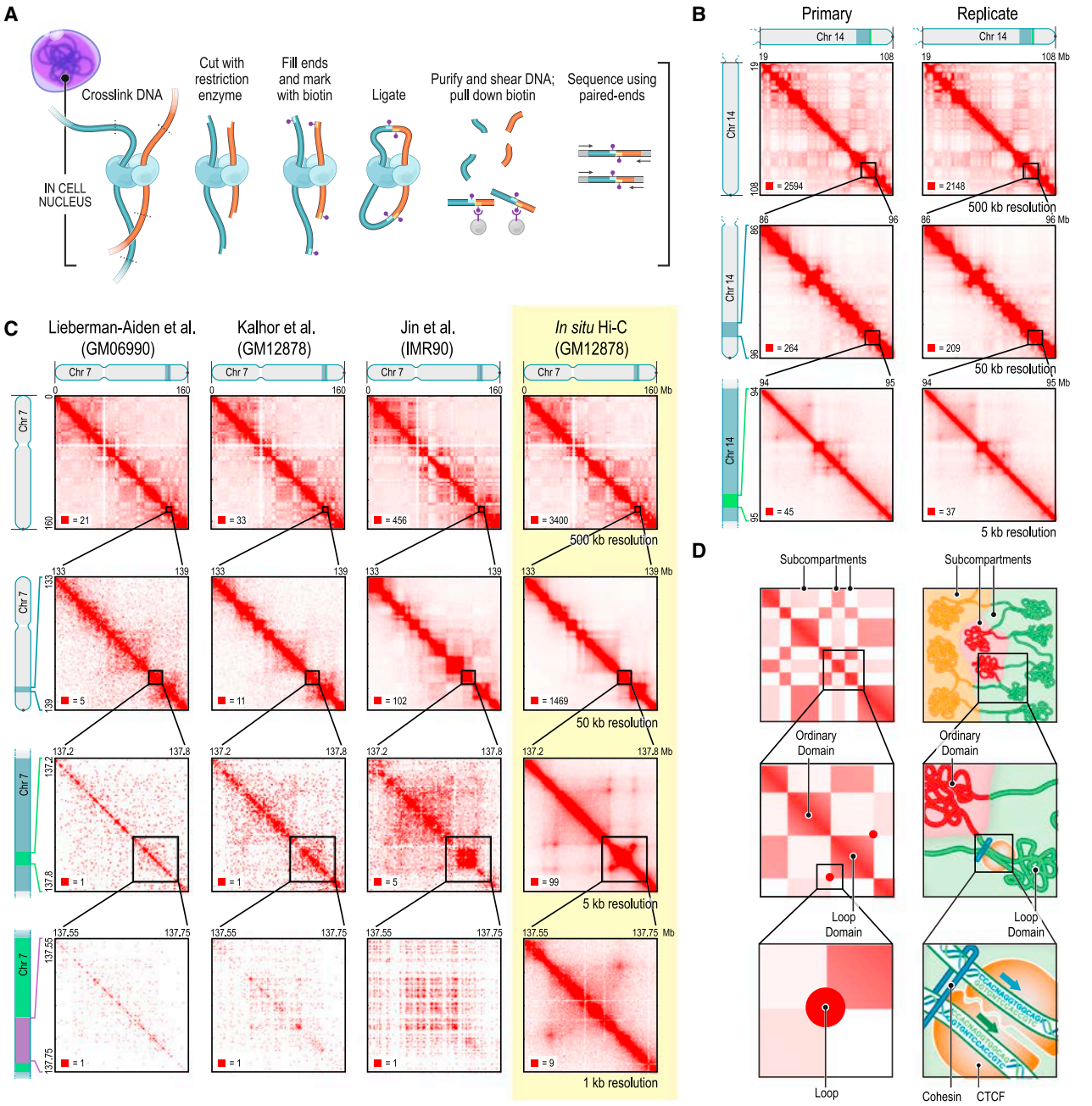


Figure 1. We Used In Situ Hi-C to Map over 15 Billion Chromatin Contacts across Nine Cell Types in Human and Mouse, Achieving 1 kb Resolution in Human Lymphoblastoid Cells

(A) During in situ Hi-C, DNA-DNA proximity ligation is performed in intact nuclei.

(B) Contact matrices from chromosome 14: the whole chromosome, at 500 kb resolution (top); 86–96 Mb/50 kb resolution (middle); 94–95 Mb/5 kb resolution (bottom). Left: GM12878, primary experiment; Right: biological replicate. The 1D regions corresponding to a contact matrix are indicated in the diagrams above and at left. The intensity of each pixel represents the normalized number of contacts between a pair of loci. Maximum intensity is indicated in the lower left of each panel.

(C) We compare our map of chromosome 7 in GM12878 (last column) to earlier Hi-C maps: Lieberman-Aiden et al. (2009), Kalhor et al. (2012), and Jin et al. (2013).

(D) Overview of features revealed by our Hi-C maps. Top: the long-range contact pattern of a locus (left) indicates its nuclear neighborhood (right). We detect at least six subcompartments, each bearing a distinctive pattern of epigenetic features. Middle: squares of enhanced contact frequency along the diagonal (left) indicate the presence of small domains of condensed chromatin, whose median length is 185 kb (right). Bottom: peaks in the contact map (left) indicate the presence of loops (right). These loops tend to lie at domain boundaries and bind CTCF in a convergent orientation.

See also Figure S1, Data S1, I–II, and Tables S1 and S2.

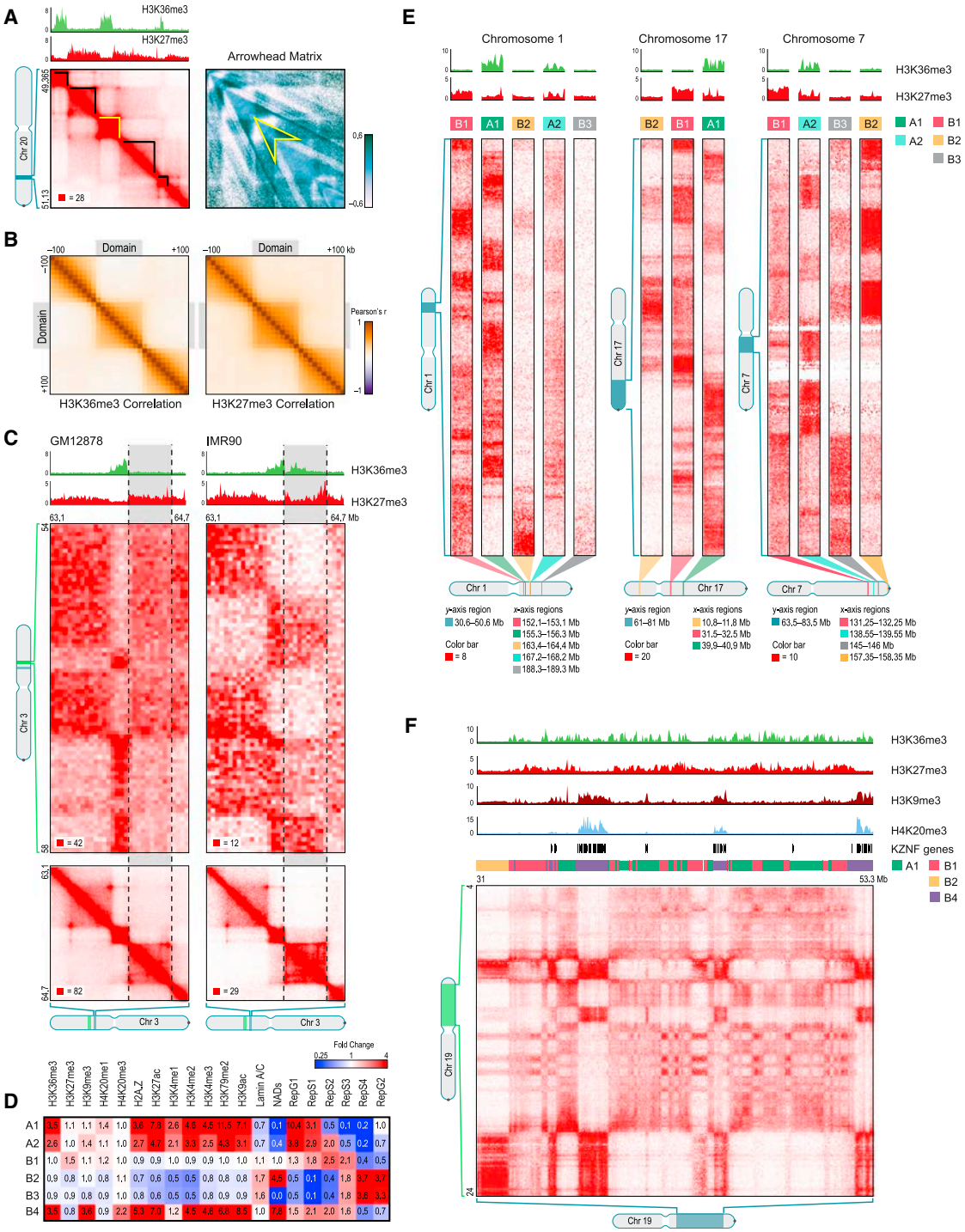


Figure 2. The Genome Is Partitioned into Contact Domains that Segregate into Nuclear Subcompartments Corresponding to Different Patterns of Histone Modifications

(A) We annotate thousands of domains across the genome (left, black highlight). To do so, we define an arrowhead matrix A (right) such that $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})$, where M^* is the normalized contact matrix. This transformation replaces domains with an arrowhead-shaped motif pointing toward the domain's upper-left corner (example in yellow); we identify these arrowheads using dynamic programming. See Experimental Procedures.

(B) Pearson correlation matrices of the histone mark signal between pairs of loci inside and within 100 kb of a domain. Left: H3K36me3; Right: H3K27me3.

(C) Conserved contact domains on chromosome 3 in GM12878 (left) and IMR90 (right). In GM12878, the highlighted domain (gray) is enriched for H3K27me3 and depleted for H3K36me3. In IMR90, the situation is reversed. Marks at flanking domains are the same in both: the domain to the left is enriched for H3K36me3 and the domain to the right is enriched for H3K27me3. The flanking domains have long-range contact patterns that differ from one another and are preserved in both

(legend continued on next page)

an interval exhibit the same genome-wide contact pattern. Compartment A is highly enriched for open chromatin; compartment B is enriched for closed chromatin (Lieberman-Aiden et al., 2009; Kalhor et al., 2012; Sexton et al., 2012).

In our new, higher resolution maps (200- to 1,000-fold more contacts), we observe many small squares of enhanced contact frequency that tile the diagonal of each contact matrix (Figure 2A). We used the Arrowhead algorithm (see Experimental Procedures) to annotate these contact domains genome-wide. The observed domains ranged in size from 40 kb to 3 Mb (median size 185 kb). As with megadomains, there is an abrupt drop in contact frequency (33%) for pairs of loci on opposite sides of the domain boundary (Figure S2G). Contact domains are often preserved across cell types (Figures S3A and S3B).

The presence of smaller domains in Hi-C maps is consistent with several other recent studies (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). We explore the relationship between the domains we annotate and those annotated in prior studies in the Discussion.

Contact Domains Exhibit Consistent Histone Marks Whose Changes Are Associated with Changes in Long-Range Contact Pattern

Loci within a contact domain show correlated histone modifications for eight different factors (H3K36me3, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K79me2, and H4K20me1) based on data from the ENCODE project in GM12878 cells (ENCODE Project Consortium, 2012). By contrast, loci at comparable distance but residing in different domains showed much less correlation in chromatin state (Figures 2B, S2I, and S2K; Extended Experimental Procedures). Strikingly, changes in a domain's chromatin state are often accompanied by changes in the long-range contact pattern of domain loci (i.e., the pattern of contacts between loci in the domain and other loci genome-wide), indicating that changes in chromatin pattern are accompanied by shifts in a domain's nuclear neighborhood (Figures 2C and S3C–S3E; Extended Experimental Procedures). This observation is consistent with microscopy studies associating changes in gene expression with changes in nuclear localization (Finlan et al., 2008).

There Are at Least Six Nuclear Subcompartments with Distinct Patterns of Histone Modifications

Next, we partitioned loci into categories based on long-range contact patterns alone, using four independent approaches: manual annotation and three unsupervised clustering algorithms (HMM, K-means, Hierarchical). All gave similar results (Figure S4B; Extended Experimental Procedures). We then investigated the biological meaning of these categories.

When we analyzed the data at low matrix resolution (1 Mb), we reproduced our earlier finding of two compartments (A and B). At high resolution (25 kb), we found evidence for at least five "subcompartments" defined by their long-range interaction patterns, both within and between chromosomes. These findings expand on earlier reports suggesting three compartments in human cells (Yaffe and Tanay, 2011). We found that the median length of an interval lying completely within a subcompartment is 300 kb. Although the subcompartments are defined solely based on their Hi-C interaction patterns, they exhibit distinct genomic and epigenomic content.

Two of the five interaction patterns are correlated with loci in compartment A (Figure S4E). We label the loci exhibiting these patterns as belonging to subcompartments A1 and A2. Both A1 and A2 are gene dense, have highly expressed genes, harbor activating chromatin marks such as H3K36me3, H3K79me2, H3K27ac, and H3K4me1 and are depleted at the nuclear lamina and at nucleolus-associated domains (NADs) (Figures 2D, 2E, and S4I; Table S3). While both A1 and A2 exhibit early replication times, A1 finishes replicating at the beginning of S phase, whereas A2 continues replicating into the middle of S phase. A2 is more strongly associated with the presence of H3K9me3 than A1, has lower GC content, and contains longer genes (2.4-fold).

The other three interaction patterns (labeled B1, B2, and B3) are correlated with loci in compartment B (Figure S4E) and show very different properties. Subcompartment B1 correlates positively with H3K27me3 and negatively with H3K36me3, suggestive of facultative heterochromatin (Figures 2D and 2E). Replication of this subcompartment peaks during the middle of S phase. Subcompartments B2 and B3 tend to lack all of the above-noted marks and do not replicate until the end of S phase (see Figure 2D). Subcompartment B2 includes 62% of pericentromeric heterochromatin (3.8-fold enrichment) and is enriched at the nuclear lamina (1.8-fold) and at NADs (4.6-fold). Subcompartment B3 is enriched at the nuclear lamina (1.6-fold), but strongly depleted at NADs (76-fold).

Upon closer visual examination, we noticed the presence of a sixth pattern on chromosome 19 (Figure 2F). Our genome-wide clustering algorithm missed this pattern because it spans only 11 Mb, or 0.3% of the genome. When we repeated the algorithm on chromosome 19 alone, the additional pattern was detected. Because this sixth pattern correlates with the Compartment B pattern, we labeled it B4. Subcompartment B4 comprises a handful of regions, each of which contains many KRAB-ZNF superfamily genes. (B4 contains 130 of the 278 KRAB-ZNF genes in the genome, a 65-fold enrichment). As noted in previous studies (Vogel et al., 2006; Hahn et al., 2011), these regions exhibit a highly distinctive chromatin pattern, with strong enrichment for

cell types. In IMR90, the highlighted domain is marked by H3K36me3 and its long-range contact pattern matches the similarly-marked domain on the left. In GM12878, it is decorated with H3K27me3, and the long-range pattern switches, matching the similarly-marked domain to the right. Diagonal submatrices, 10 kb resolution; long-range interaction matrices, 50 kb resolution.

(D) Each of the six long-range contact patterns we observe exhibits a distinct epigenetic profile (data sources are listed in Table S3). Each subcompartment also has a visually distinctive contact pattern.

(E) Each example shows part of the long-range contact patterns for several nearby genomic intervals lying in different subcompartments.

(F) A large contiguous region on chromosome 19 contains intervals in subcompartments A1, B1, B2, and B4.

See also Figures S2, S3, and S4 and Data S1, III–IV.

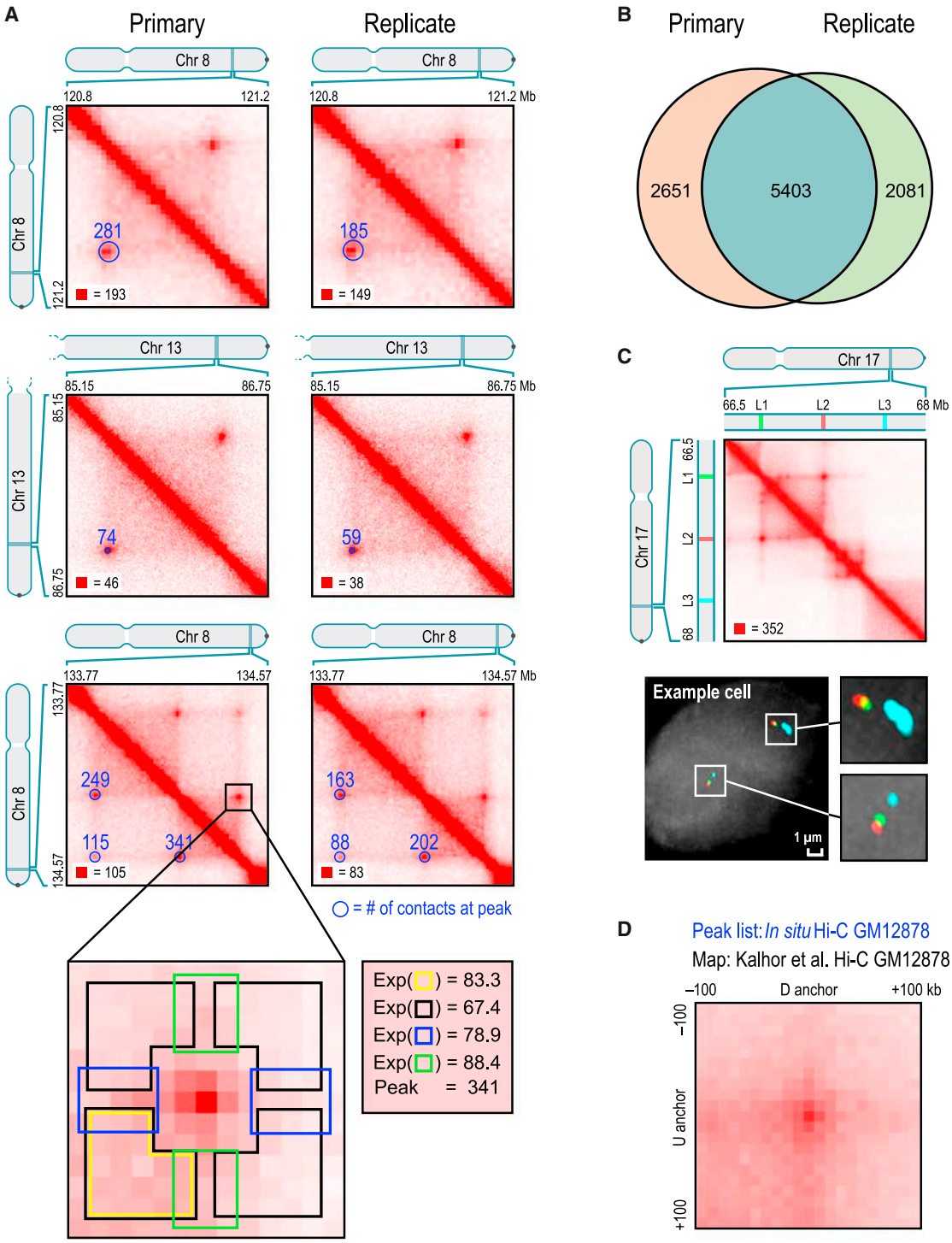


Figure 3. We Identify Thousands of Chromatin Loops Genome-wide Using a Local Background Model

(A) We identify peaks by detecting pixels that are enriched with respect to four local neighborhoods (blowout): horizontal (blue), vertical (green), lower-left (yellow), and donut (black). These "peak" pixels indicate the presence of a loop and are marked with blue circles (radius = 20 kb) in the lower-left of each heatmap. The number of raw contacts at each peak is indicated. Left: primary GM12878 map; Right: replicate; annotations are completely independent. All contact matrices in this and subsequent figures are 10 kb resolution unless noted.

(B) Overlap in peak annotations between replicates.

(C) Top: location of 3D-FISH probes used to verify a peak in the chromosome 17 contact map. Bottom: example cell.

(legend continued on next page)

both activating chromatin marks, such as H3K36me3, and heterochromatin-associated marks, such as H3K9me3 and H4K20me3.

Approximately 10,000 Peaks Mark the Position of Chromatin Loops

We next sought to identify the positions of chromatin loops by using an algorithm to search for pairs of loci that show significantly closer proximity with one another than with the loci lying between them (Figure 3A). Such pairs correspond to pixels with higher contact frequency than typical pixels in their neighborhood. We refer to these pixels as "peaks" in the Hi-C contact matrix and to the corresponding pair of loci as "peak loci." Peaks reflect the presence of chromatin loops, with the peak loci being the anchor points of the chromatin loop. (Because contact frequencies vary across the genome, we define peak pixels relative to the local background. We note that some papers [Sanyal et al., 2012; Jin et al., 2013] have sought to define peaks relative to a genome-wide average. This choice is problematic because, for example, many pixels within a domain may be reported as peaks despite showing no locally distinctive proximity; see Discussion.)

Our algorithm detected 9,448 peaks in the in situ Hi-C map for GM12878 at 5 kb matrix resolution. These peaks are associated with a total of 12,903 distinct peak loci (some peak loci are associated with more than one peak). The vast majority of peaks (98%) reflected loops between loci that are <2 Mb apart.

These findings were reproducible across all of our high-resolution Hi-C maps. Examining the primary and replicate maps separately, we found 8,054 peaks in the former and 7,484 peaks in the latter, with 5,403 in both lists (see Figures 3A and 3B; Data S1, V; Table S4). The differences were almost always the result of our conservative peak-calling criteria (Extended Experimental Procedures). We also called peaks using our GM12878 dilution Hi-C experiment. Because the map is sparser and thus noisier, we called only 3,073 peaks. Nonetheless, 65% of these peaks were also present in the list of peaks from our in situ Hi-C data set, again reflecting high interreplicate reproducibility.

To independently confirm that peak loci are closer than neighboring locus pairs, we performed 3D-FISH (Beliveau et al., 2012) on four loops (Table S5). In each case, we compared two peak loci, *L1* and *L2*, with a control locus, *L3*, that lies an equal genomic distance away from *L2* but on the opposite side (Figures 3C and S5B). In all cases, the 3D-distance between *L1* and *L2* was consistently shorter than the 3D-distance between *L2* and *L3* (Extended Experimental Procedures).

We also confirmed that our list of peaks was consistent with previously published Hi-C maps. Although earlier maps contained too few contacts to reliably call individual peaks, we developed a method called Aggregate Peak Analysis (APA) that compares the aggregate enrichment of our peak set in these low-resolution maps to the enrichment seen when our peak set is translated in any direction (Experimental Procedures). APA showed strong consistency between our loop calls and all six previously published Hi-C experiments in lymphoblastoid cell lines (Lieberman-Aiden et al., 2009; Kalhor et al., 2012) (Figure 3D; Data S2, I.E; Table S6).

Finally, we demonstrated that the peaks observed were robust to particular protocol conditions by performing APA on our GM12878 dilution Hi-C map and on our 112 supplemental Hi-C experiments exploring a wide range of protocol variants. Enrichment was seen in every experiment. Notably, these include five experiments (HIC043-HIC047; Table S1) in which the Hi-C protocol was performed without crosslinking, demonstrating that the peaks observed in our experiments cannot be byproducts of the formaldehyde-crosslinking procedure.

Conservation of Peaks among Human Cell Lines and across Evolution

We also identified peaks in the other seven human cell lines (Table S1). Because these maps contain fewer contacts, sensitivity is reduced, and fewer peaks are observed (ranging from 2,634 to 8,040). APA confirmed that these peak calls were consistent with the dilution Hi-C maps reported here (in IMR90, HMEC, HUVEC, and NHEK), as well as with all previously published Hi-C maps in these cell types (Lieberman-Aiden et al., 2009; Dixon et al., 2012; Jin et al., 2013) (Data S2, I.F).

We found that peaks were often conserved across cell types (Figure 4A): between 55% and 75% of the peaks found in any given cell type were also found in GM12878 (Figure S5D).

Next, we compared peaks across species. In CH12-LX mouse B-lymphoblasts, we identified 2,927 high-confidence contact domains and 3,331 peaks. When we examined orthologous regions in GM12878, we found that 50% of peaks and 45% of domains called in mouse were also called in humans. This suggests substantial conservation of 3D genome structure across the mammals (Figures 4B–4E).

Loops Anchored at a Promoter Are Associated with Enhancers and Increased Gene Activation

Various lines of evidence indicate that many of the observed loops are associated with gene regulation.

First, our peaks frequently have a known promoter at one peak locus (as annotated by ENCODE's ChromHMM) (Hoffman et al., 2013) and a known enhancer at the other (Figure 5A). For instance, 2,854 of the 9,448 peaks in our GM12878 map bring together known promoters and known enhancers (30% versus 7% expected by chance). The peaks include classic promoter-enhancer loops, such as at *MYC* (chr8:128.35–128.75 Mb, in HMEC) and alpha-globin (chr16:0.15–0.22 Mb, in K562). Second, genes whose promoters are associated with a loop are much more highly expressed than genes whose promoters are not associated with a loop (6-fold).

Third, the presence of cell type-specific peaks is associated with changes in expression. When we examined RNA sequencing (RNA-seq) data produced by ENCODE, we found

(D) APA plot shows the aggregate signal from the 9,448 GM12878 loops we report by summing submatrices surrounding each peak in a low-resolution GM12878 Hi-C map due to Kalhor et al. (2012). Although individual peaks cannot be seen in the Kalhor et al. (2012) data (that contains 42 M contacts), the peak at the center of the APA plot indicates that the aggregate signal from our peak set as a whole can be clearly discerned using their data set.
See also Figure S5, Data S1, V. and Data S2,I, and Tables S4, S5, and S6.

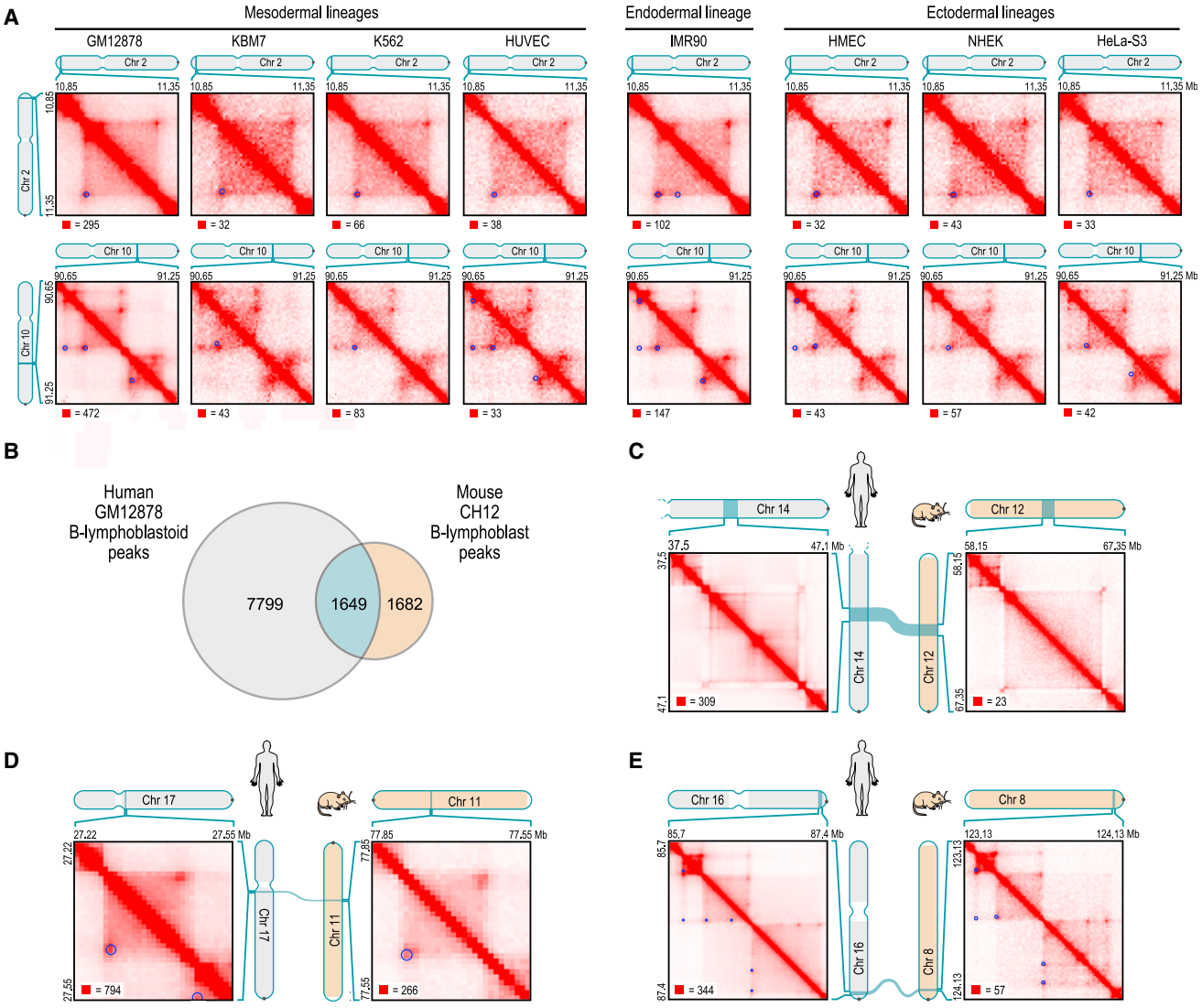


Figure 4. Loops Are Often Preserved across Cell Types and from Human to Mouse

(A) Examples of peak and domain preservation across cell types. Annotated peaks are circled in blue. All annotations are completely independent.

(B) Of the 3,331 loops we annotate in mouse CH12-LX, 1,649 (50%) are orthologous to loops in human GM12878.

(C–E) Conservation of 3D structure in synteny blocks. The contact matrices in (C) are shown at 25 kb resolution. (D) and (E) are shown at 10 kb resolution.

that the appearance of a loop in a cell type was frequently accompanied by the activation of a gene whose promoter overlapped one of the peak loci. For example, a cell-type-specific loop is anchored at the promoter of the gene encoding L-selectin (*SELL*), which is expressed in GM12878 (where the loop is present), but not in IMR90 (where the loop is absent, Figure 5B). Genome-wide, we observed 557 loops in GM12878 that were clearly absent in IMR90. The corresponding peak loci overlapped the promoters of 43 genes that were markedly upregulated (>50-fold) in GM12878, but of only one gene that was markedly upregulated in IMR90. Conversely, we found 510 loops in IMR90 that were clearly absent in GM12878. The corresponding peak loci overlapped the promoters of 94 genes that were markedly upregulated in IMR90, but of only three genes that were markedly upregulated in GM12878. When we compared GM12878 to the five other human cell types for which ENCODE RNA-seq data were available, the results were very similar (Figure 5C; Table S7).

Occasionally, gene activation is accompanied by the emergence of a cell-type-specific network of peaks. Figure 5D illustrates the case of *ADAMTS1*, which encodes a protein involved in fibroblast migration. The gene is expressed in IMR90, where its promoter is involved in six loops. In GM12878, it is not expressed, and the promoter is involved in only two loops. Many of the IMR90 peak loci form transitive peaks with one another (see discussion of "transitivity" below), suggesting that the *ADAMTS1* promoter and the six distal sites may all be located at a single spatial hub.

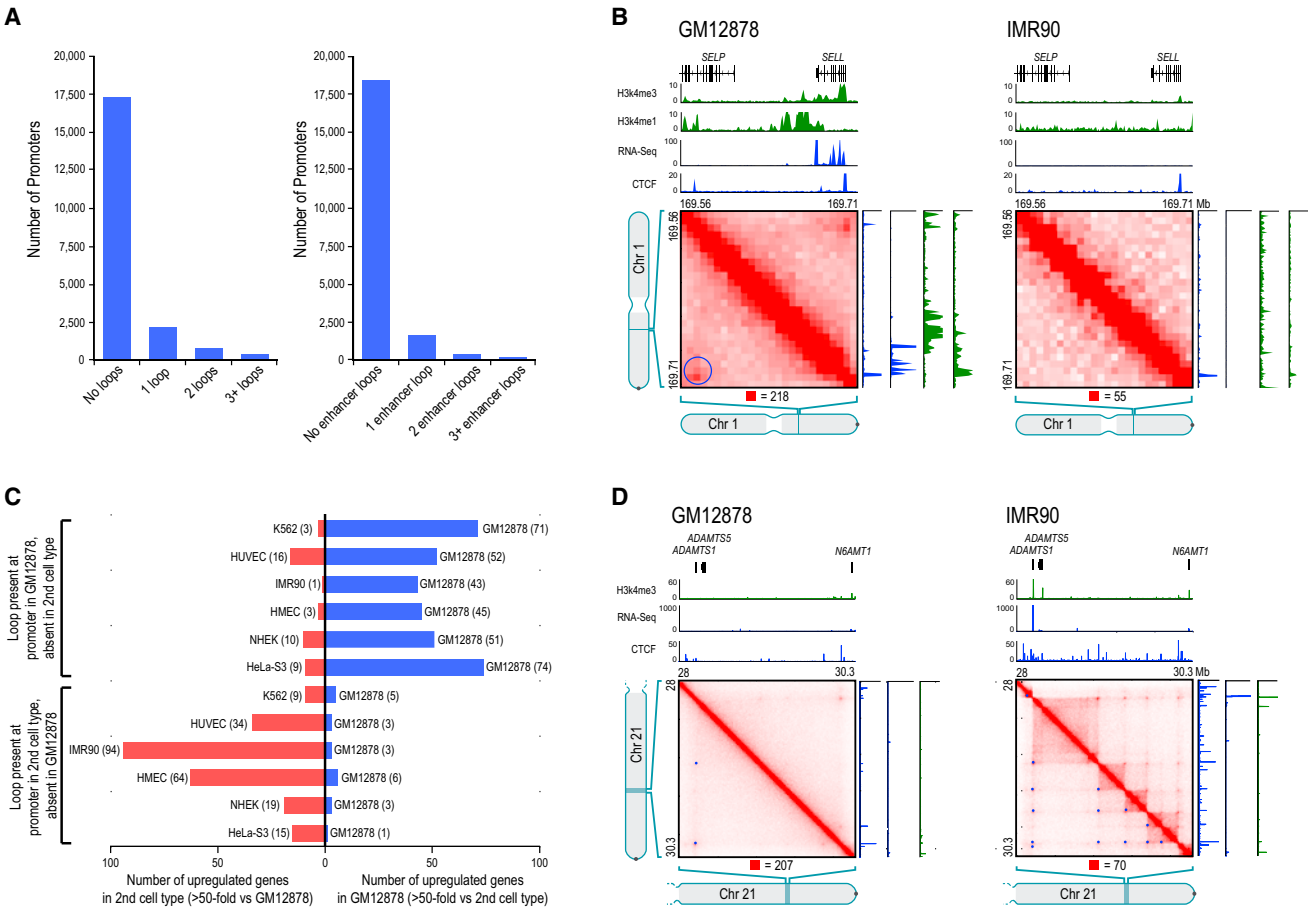


Figure 5. Loops between Promoters and Enhancers Are Strongly Associated with Gene Activation

(A) Histogram showing loop count at promoters (left); restricted to loops where the distal peak locus contains an enhancer (right).

(B) Left: a loop in GM12878, with one anchor at the *SELL* promoter and the other at a distal enhancer. The gene is on. Right: the loop is absent in IMR90, where the gene is off.

(C) Genes whose promoters participate in a loop in GM12878 but not in a second cell type are frequently upregulated in GM12878 and vice versa.

(D) Left: two loops in GM12878 are anchored at the promoter of the inactive *ADAMTS1* gene. Right: a series of loops and domains appear, along with transitive looping. *ADAMTS1* is on.

See also Data S1, VI and Table S7.

These observations are consistent with the classic model in which looping between a promoter and enhancer activates a target gene (Tolhuis et al., 2002; Amano et al., 2009; Ahmadiyeh et al., 2010).

Loops Frequently Demarcate the Boundaries of Contact Domains

A large fraction of peaks (38%) coincide with the corners of a contact domain—that is, the peak loci are located at domain boundaries (Figures 6A and S6). Conversely, a large fraction of domains (39%) had peaks in their corner. Moreover, the appearance of a loop is usually (in 65% of cases) associated with the appearance of a domain demarcated by the loop. Because this configuration is so common, we use the term "loop domain" to refer to contact domains whose endpoints form a chromatin loop.

In some cases, adjacent loop domains (bounded by peak loci *L1-L2* and *L2-L3*, respectively) exhibit transitivity—that is, *L1* and *L3* also correspond to a peak. This may indicate that the three loci simultaneously colocate at a single spatial position. However, many peaks do not exhibit transitivity, suggesting that the corresponding loci do not colocate. Figure 6B shows a region on chromosome 4 exhibiting both configurations.

We also found that overlapping loops are strongly disfavored: pairs of loops *L1-L3* and *L2-L4* (where *L1*, *L2*, *L3* and *L4* occur consecutively in the genome) are found 4-fold less often than expected under a random model (Extended Experimental Procedures).

The Vast Majority of Loops Are Associated with Pairs of CTCF Motifs in a Convergent Orientation

We next wondered whether peaks are associated with specific proteins. We examined the results of 86 chromatin immunoprecipitation sequencing (ChIP-seq) experiments performed by ENCODE in GM12878. We found that the vast majority of peak

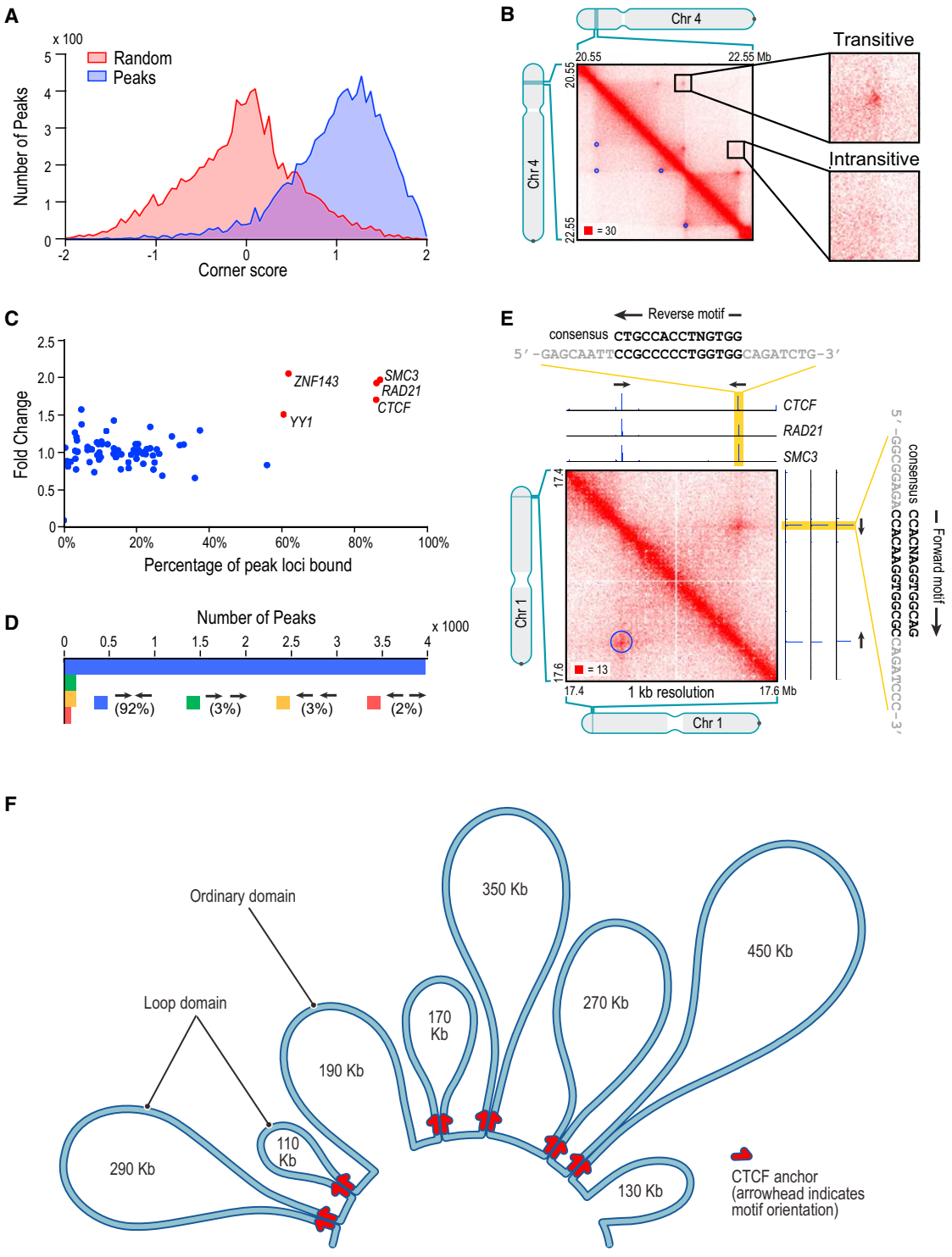


Figure 6. Many Loops Demarcate Contact Domains; The Vast Majority of Loops Are Anchored at a Pair of Convergent CTCF/RAD21/SMC3 Binding Sites

(A) Histograms of corner scores for peak pixels versus random pixels with an identical distance distribution.

(B) Contact matrix for chr4:20.55 Mb–22.55 Mb in GM12878, showing examples of transitive and intransitive looping behavior.

(C) Percent of peak loci bound versus fold enrichment for 76 DNA-binding proteins.

(D) The pairs of CTCF motifs that anchor a loop are nearly all found in the convergent orientation.

(legend continued on next page)

loci are bound by the insulator protein CTCF (86%) and the cohesin subunits RAD21 (86%) and SMC3 (87%) (Figure 6C). This is consistent with numerous reports, using a variety of experimental modalities, that suggest a role for CTCF and cohesin in mediating DNA loops (Splinter et al., 2006; Hou et al., 2008; Phillips and Corces, 2009). Because many of our loops demarcate domains, this observation is also consistent with studies suggesting that CTCF delimits structural and regulatory domains (Xie et al., 2007; Cuddapah et al., 2009; Dixon et al., 2012).

We found that most peak loci encompass a unique DNA site containing a CTCF-binding motif, to which all three proteins (CTCF, SMC3, and RAD21) were bound (5-fold enrichment). We were thus able to associate most of the peak loci (6,991 of 12,903, or 54%) with a specific CTCF-motif "anchor."

The consensus DNA sequence for CTCF-binding sites is typically written as 5′-CCACNAGGTGGCAG-3′. Because the sequence is not palindromic, each CTCF motif has an orientation; we designate the consensus motif above as the "forward" orientation. Thus, a pair of CTCF sites on the same chromosome can have four possible orientations: (1) same direction on one strand, (2) same direction on the other strand, (3) convergent on opposite strands, and (4) divergent on opposite strands.

If CTCF sites were randomly oriented, one would expect all four orientations to occur equally often. But when we examined the 4,322 peaks in GM12878 where the two corresponding peak loci each contained a single CTCF-binding motif, we found that the vast majority (92%) of motif pairs are convergent (Figures 6D and 6E). Overall, the presence, at pairs of peak loci, of bound CTCF sites in the convergent orientation was enriched 102-fold over random expectation (Extended Experimental Procedures). The convergent orientation was overwhelmingly more frequent than the divergent orientation, despite the fact that divergent motifs also lie on opposing strands: in GM12878, the counts were 3,971-78 (51-fold enrichment, convergent versus divergent); in IMR90, 1,456-5 (291-fold); in HMEC, 968-11 (88-fold); in K562, 723-2 (362-fold); in HUVEC, 671-4 (168-fold); in HeLa, 301-3 (100-fold); in NHEK, 556-9 (62-fold); and in CH12-LX, 625-8 (78-fold). This pattern suggests that a pair of CTCF sites in the convergent orientation is required for the formation of a loop.

The observation that looped CTCF sites occur in the convergent orientation also allows us to analyze peak loci containing multiple CTCF-bound motifs to predict which motif instance plays a role in a given loop. In this way, we can associate nearly two-thirds of peak loci (8,175 of 12,903, or 63.4%) with a single CTCF-binding motif.

The specific orientation of CTCF sites at observed peaks provides evidence that our peak calls are biologically correct. Because randomly chosen CTCF pairs would exhibit each of the four orientations with equal probability, the near-perfect association between our loop calls and the convergent orientation could not occur by chance ($p < 10^{-1,900}$, binomial distribution).

In addition, the presence of CTCF and RAD21 sites at many of our peaks provides an opportunity to compare our results to three recent ChIA-PET experiments reported by the ENCODE Consortium (in GM12878 and K562) in which ligation junctions bound to CTCF (or RAD21) were isolated and analyzed. We found strong concordance with our results in all three cases (Li et al., 2012; Heidari et al., 2014) (Extended Experimental Procedures).

The CTCF-Binding Exapted SINEB2 Repeat in Mouse Shows Preferential Orientation with Respect to Loops

In mouse, we found that 7% of peak anchors lie within SINEB2 repeat elements containing a CTCF motif, which has been exapted to be functional. (The spread of CTCF binding via retrotransposition of this element, which contains a CTCF motif in its consensus sequence, has been documented in prior studies [Bourque et al., 2008; Schmidt et al., 2012].) The CTCF motifs at peak anchors in SINEB2 elements show the same strong bias toward convergent orientation seen throughout the genome (89% are oriented toward the opposing loop anchor versus 94% genome-wide). The orientation of these CTCF motifs is aligned with the orientation of the SINEB2 consensus sequence in 97% of cases. This suggests that exaptation of a CTCF in a SINEB2 element is more likely when the orientation of the inserted SINEB2 is compatible with local loop structure.

Diploid Hi-C Maps Reveal Homolog-Specific Features, Including Imprinting-Specific Loops and Massive Domains and Loops on the Inactive X Chromosome

Because many of our reads overlap SNPs, it is possible to use GM12878 phasing data (McKenna et al., 2010; 1000 Genomes Project Consortium et al., 2012) to assign contacts to specific chromosomal homologs (Figure 7A; Table S8). Using these assignments, we constructed a "diploid" Hi-C map of GM12878 comprising both maternal (238 M contacts) and paternal (240 M) maps.

For autosomes, the maternal and paternal homologs exhibit very similar inter- and intrachromosomal contact profiles (Pearson's $R > 0.998$). One interchromosomal difference was notable: an elevated contact frequency between the paternal homologs of chromosome 6 and 11 that is consistent with an unbalanced translocation fusing chr11q:73.5 Mb and all distal loci (a stretch of over 60 Mb) to the telomere of chromosome 6p (Figures 7B and S7B). The signal intensity suggests that the translocation is present in between 1.2% and 5.6% of our cells (Extended Experimental Procedures). We tested this prediction by karyotyping 100 GM12878 cells using Giemsa staining and found three abnormal chromosomes, each showing the predicted

(E) A peak on chromosome 1 and corresponding ChIP-seq tracks. Both peak loci contain a single site bound by CTCF, RAD21, and SMC3. The CTCF motifs at the anchors exhibit a convergent orientation.

(F) A schematic rendering of a 2.1 Mb region on chromosome 20 (48.78–50.88 Mb). Eight domains tile the region, ranging in size from 110 kb to 450 kb; 95% of the region is contained inside a domain (contour lengths are shown to scale). Six of the eight domains are demarcated by loops between convergent CTCF-binding sites located at the domain boundaries. The other two domains are not demarcated by loops. The motif orientation is indicated by the direction of the arrow. Note that not every CTCF-binding site is shown.

See also Figure S6.

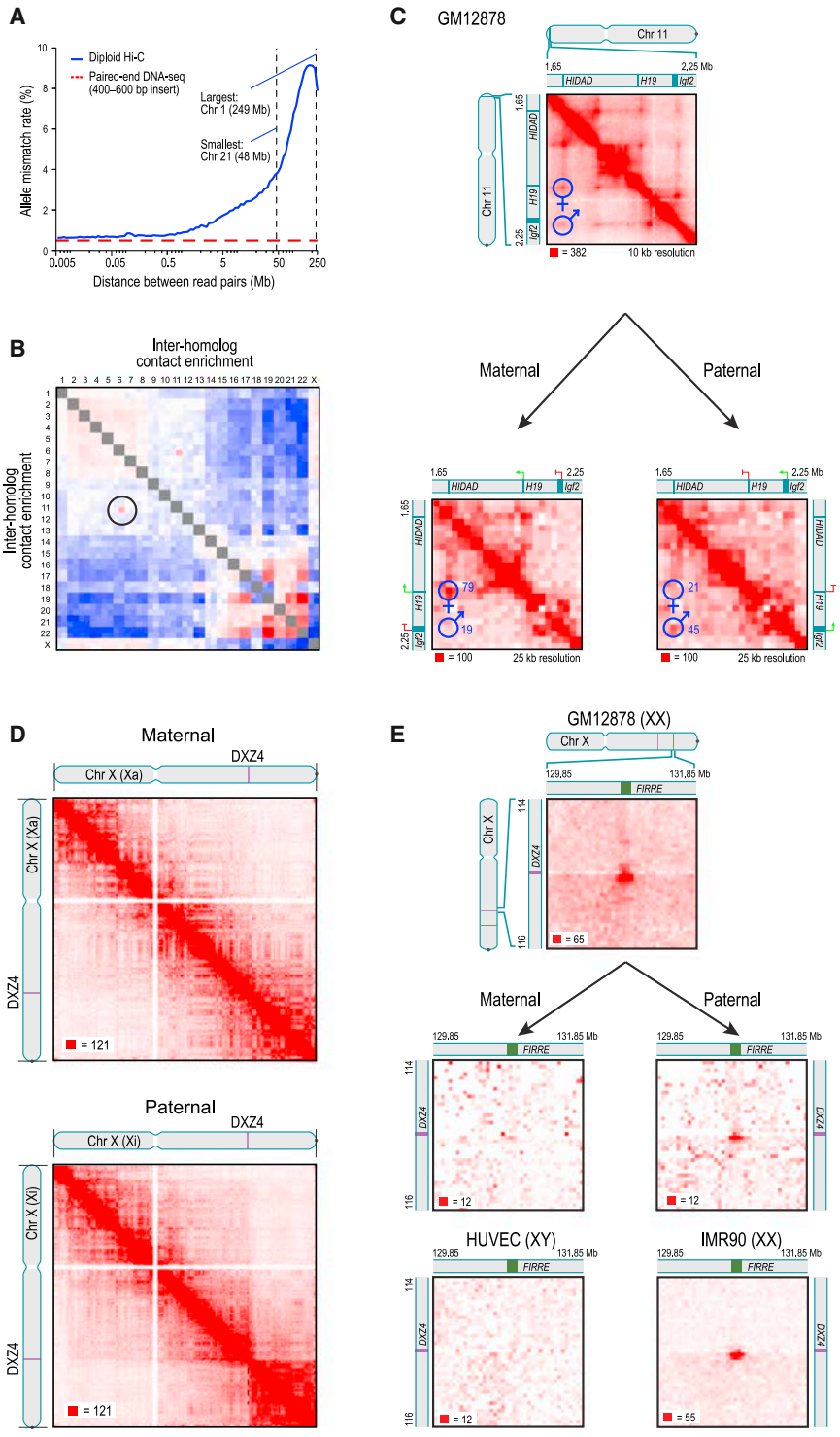


Figure 7. Diploid Hi-C Maps Reveal Superdomains and Superloops Anchored at CTCF-Binding Tandem Repeats on the Inactive X Chromosome

(A) The frequency of mismatch (maternal-paternal) in SNP allele assignment versus distance between two paired read alignments. Intrachromosomal read pairs are overwhelmingly intramolecular.

(B) Preferential interactions between homologs. Left/top is maternal; right/bottom is paternal. The aberrant contact frequency between 6/paternal and 11/paternal (circle) reveals a translocation.

(C) Top: in our unphased Hi-C map of GM12878, we observe two loops joining both the promoter of the maternally-expressed *H19* and the promoter of the paternally-expressed *Igf2* to a distal locus, HIDAD. Using diploid Hi-C maps, we phase these loops: the HIDAD-*H19* loop is present only on the maternal homolog (left) and the HIDAD-*Igf2* loop is present only on the paternal homolog (right).

(D) The inactive (paternal) copy of chromosome X (bottom) is partitioned into two massive "superdomains" not seen in the active (maternal) copy (top). *DXZ4* lies at the boundary. Contact matrices are shown at 500 kb resolution.

(E) The "superloop" between *FIRRE* and *DXZ4* is present in the unphased GM12878 map (top), in the paternal GM12878 map (middle right), and in the map of the female cell line IMR90 (bottom right); it is absent from the maternal GM12878 map (middle left) and the map of the male HUVEC cell line (bottom left). Contact matrices are shown at 50 kb resolution.

See also Figure S7 and Table S8.

translocation, der(6)t(6,11)(pter;q) (Figures S7C–S7F). The Hi-C data reveal that the translocation involves the paternal homologs, which cannot be determined with ordinary cytogenetic methods.

We also observed differences in loop structure between homologous autosomes at some imprinted loci. For instance, the *H19/Igf2* locus on chromosome 11 is a well-characterized case of genomic imprinting. In our unphased maps, we clearly see two loops from a single distal locus at 1.72 Mb (that binds CTCF in the forward orientation) to loci located near the promoters of both *H19* and *Igf2* (both of which bind CTCF in the reverse orientation, i.e., the above consensus motif lies on the opposite strand; see Figure 7C). We refer to this distal locus as the *H19/Igf2* Distal Anchor Domain (HIDAD). Our diploid maps reveal that the loop to the *H19* region is present on the maternal chromosome (from which *H19* is expressed), but the loop to the *Igf2* region is absent or greatly attenuated. The opposite pattern is found on the paternal chromosome (from which *Igf2* is expressed).

Pronounced differences were seen on the diploid intrachromosomal maps of chromosome X. The paternal X chromosome, which is usually inactive in GM12878, is partitioned into two massive domains (0–115 Mb and 115–155.3 Mb). These "superdomains" are not seen in the active, maternal X (Figure 7D). When we examined the unphased maps of chromosome X for the karyotypically normal female cell lines in our study (GM12878, IMR90, HMEC,

NHEK), the superdomains on X were evident, although the signal was attenuated due to the superposition of signals from active and inactive X chromosomes. When we examined the male HUVEC cell line and the haploid KBM7 cell line, we saw no evidence of superdomains (Figure S7G).

Interestingly, the boundary between the superdomains (ChrX: 115 Mb ± 500 kb) lies near the macrosatellite repeat *DXZ4* (ChrX: 114,867,433–114,919,088) near the middle of Xq. *DXZ4* is a CpG-rich tandem repeat that is conserved across primates and monkeys and encodes a long noncoding RNA. In males and on the active X, *DXZ4* is heterochromatic, hypermethylated and does not bind CTCF. On the inactive X, *DXZ4* is euchromatic, hypomethylated, and binds CTCF. *DXZ4* has been hypothesized to play a role in reorganizing chromatin during X inactivation (Chadwick, 2008).

There were also significant differences in loop structure between the chromosome X homologs. We observed 27 large "superloops," each spanning between 7 and 74 Mb, present only on the inactive X chromosome in the diploid map (Figure 7E). The superloops were also seen in all four unphased maps from karyotypically normal XX cells, but were absent in unphased maps from X0 and XY cells (Figure S7I). Two of the superloops (chrX:56.8 Mb-*DXZ4* and *DXZ4*-130.9 Mb) were reported previously in a locus-specific study (Horakova et al., 2012).

Like the peak loci of most other loops, nearly all the superloop anchors bind CTCF (23 of 24). The six anchor regions most frequently associated with superloops are large (up to 200 kb). Four of these anchor regions contain whole long noncoding RNA (lncRNA) genes: *loc550643*, *XIST*, *DXZ4*, and *FIRRE*. Three (*loc550643*, *DXZ4*, and *FIRRE*) contain CTCF-binding tandem repeats that only bind CTCF on the inactive homolog.

DISCUSSION

Using the in situ Hi-C protocol, we probed genomic architecture with high resolution; in the case of GM12878 lymphoblastoid cells, better than 1 kb. We observe the presence of contact domains that were too small (median length = 185 kb) to be seen in previous maps. Loci within a domain interact frequently with one another, have similar patterns of chromatin modifications, and exhibit similar long-range contact patterns. Domains tend to be conserved across cell types and between human and mouse. When the pattern of chromatin modifications associated with a domain changes, the domain's long-range contact pattern also changes. Domains exhibit at least six distinct patterns of long-range contacts (subcompartments), which subdivide the two compartments that we previously reported based on low resolution data. The subcompartments are each associated with distinct chromatin patterns. It is possible that the chromatin patterns play a role in bringing about the long-range contact patterns, or vice versa.

Our data also make it possible to create a genome-wide catalog of chromatin loops. We identified loops by looking for pairs of loci that have significantly more contacts with one another than they do with other nearby loci. In our densest map (GM12878), we observe 9,448 loops.

The loops reported here have many interesting properties. Most loops are short (<2 Mb) and strongly conserved across cell types and between human and mouse. Promoter-enhancer loops are common and associated with gene activation. Loops tend not to overlap; they often demarcate contact domains, and may establish them. CTCF and the cohesin subunits RAD21 and SMC3 associate with loops; each of these proteins is found at over 86% of loop anchors.

The most striking property of loops is that the pair of CTCF motifs present at the loop anchors occurs in a convergent orientation in >90% of cases (versus 25% expected by chance). The importance of motif orientation between loci that are separated by, on average, 360 kb is surprising and must bear on the mechanism by which CTCF and cohesin form loops, which seems likely to involve CTCF dimerization. Experiments in which the presence or orientation of CTCF sites is altered may enable the engineering of loops, domains, and other chromatin structures.

It is interesting to compare our results to those seen in previous reports. The contact domains we observe are similar in size to the "physical domains" that have been reported in Hi-C maps of *Drosophila* (Sexton et al., 2012) and to the "topologically constrained domains" (mean length: 220 kb) whose existence was demonstrated in the 1970s and 1980s in structural studies of human chromatin (Cook and Brazell, 1975; Vogelstein et al., 1980; Zehnbauer and Vogelstein, 1985). On the other hand, the domains we observe are much smaller than the TADs (1 Mb) (Dixon et al., 2012) that have been reported in humans and mice on the basis of lower-resolution contact maps. This is because detecting TADs involves detection of domain boundaries. With higher resolution data, it is possible to detect additional boundaries beyond those seen in previous maps. Interestingly, nearly all the boundaries we observe are associated with either a subcompartment transition (that occur approximately every 300 kb), or a loop (that occur approximately every 200 kb); and many are associated with both.

Our annotation identifies many fewer loops than were reported in several recent high-throughput studies, despite the fact that we have more data. The key reason is that we call peaks only when a pair of loci shows elevated contact frequency relative to the local background—that is, when the peak pixel is enriched as compared to other pixels in its neighborhood. In contrast, prior studies have defined peaks by comparing the contact frequency at a pixel to the genome-wide average (Sanyal et al., 2012; Jin et al., 2013). This latter definition is problematic because many pixels within a domain can be annotated as peaks despite showing no local increase in contact frequency. Papers using the latter definition imply the existence of more than 100,000 loops (1,187 loops were reported in 1% of the genome [Sanyal et al., 2012]) or even more than 1 million loops (reported in a genome-wide Hi-C study [Jin et al., 2013]). The vast majority of the loops annotated by these papers show no enrichment relative to the local background when examined one-by-one and no enrichment with respect to any published Hi-C data set when analyzed using APA (see Extended Experimental Procedures; Figure S8; Data S2). This suggests that these peak annotations may correspond to pairs of loci that lie in the same domain or compartment, but rarely correspond to loops.

We created diploid Hi-C maps by using polymorphisms to assign contacts to distinct chromosomal homologs. We found that the inactive X chromosome is partitioned into two large

superdomains whose boundary lies near the locus of the lncRNA *DXZ4*. We also detect a network of long-range superloops, the strongest of which are anchored at locations containing lncRNA genes (*loc550643*, *XIST*, *DXZ4*, and *FIRRE*). With the exception of *XIST*, all of these lncRNAs contain CTCF-binding tandem repeats that bind CTCF only on the inactive X.

In our original report on Hi-C, we observed that Hi-C maps can be used to study physical models of genome folding, and we proposed a fractal globule model for genome folding at the megabase scale. The kilobase-scale maps reported here allow the physical properties of genome folding to be probed at much higher resolution. We will report such studies elsewhere.

Just as loops bring distant DNA loci into close spatial proximity, we find that they bring disparate aspects of DNA biology—domains, compartments, chromatin marks, and genetic regulation—into close conceptual proximity. As our understanding of the physical connections between DNA loci continues to improve, our understanding of the relationships between these broader phenomena will deepen.

EXPERIMENTAL PROCEDURES

In Situ Hi-C Protocol

All cell lines were cultured following the manufacturer's recommendations. Two to five million cells were crosslinked with 1% formaldehyde for 10 min at room temperature. Nuclei were permeabilized. DNA was digested with 100 units of MboI, and the ends of restriction fragments were labeled using biotinylated nucleotides and ligated in a small volume. After reversal of crosslinks, ligated DNA was purified and sheared to a length of ~400 bp, at which point ligation junctions were pulled down with streptavidin beads and prepped for Illumina sequencing. Dilution Hi-C was performed as in Lieberman-Aiden et al. (2009).

3D-FISH

3D DNA-FISH was performed as in Beliveau et al. (2012) with minor modifications.

Hi-C Data Pipeline

All sequence data were produced using Illumina paired-end sequencing. We processed data using a custom pipeline that was optimized for parallel computation on a cluster. The pipeline uses BWA (Li and Durbin, 2010) to map each read end separately to the b37 or mm9 reference genomes; removes duplicate and near-duplicate reads; removes reads that map to the same fragment; and filters the remaining reads based on mapping quality score. Contact matrices were generated at base pair delimited resolutions of 2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, and 5 kb, as well as fragment-delimited resolutions of 500 f, 200 f, 100 f, 50 f, 20 f, 5 f, 2 f, and 1 f. For our largest maps, we also generated a 1 kb contact matrix. Normalized contact matrices are produced at all resolutions using Knight and Ruiz (2012).

Annotation of Domains: Arrowhead

To annotate domains, we apply an "arrowhead" transformation, defined as $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})$. M^* denotes the normalized contact matrix (see Figures S2A–S2F). This is equivalent to calculating a matrix equal to −1*(observed/expected − 1), where the expected model controls for local background and distance from the diagonal in the simplest possible way: the "expected" value at $i,i + d$ is simply the mean of the observed values at $i,i - d$ and $i,i + d$. $A_{i,i+d}$ will be strongly positive if locus $i - d$ is inside a domain and locus $i + d$ is not. If the reverse is true, $A_{i,i+d}$ will be strongly negative. If the loci are both inside or both outside a domain, $A_{i,i+d}$ will be close to zero. Consequently, if there is a domain at $[a,b]$, we find that A takes on very negative values inside a triangle whose vertices lie at $[a,a]$, $[a,b]$, and $[(a + b)/2,b]$ and very positive values inside a triangle whose vertices lie at $[(a + b)/2,b]$, $[b,b]$, and $[b,2b - a]$. The size and positioning of these triangles creates the arrowhead-shaped feature that replaces each domain in M^*. A "corner score" matrix, indicating each pixel's likelihood of lying at the corner of a domain, is efficiently calculated from the arrowhead matrix using dynamic programming.

Assigning Loci to Subcompartments

To cluster loci based on long-range contact patterns, we constructed a 100 kb resolution interchromosomal contact matrix such that loci from odd chromosomes appeared on the rows, and loci from even chromosomes appeared on the columns. (Intrachromosomal data and data involving chromosome X were excluded.) We cluster this matrix using the Python package *scikit*. For subcompartment B4, the 100 kb interchromosomal matrix for chromosome 19 was constructed and clustered separately, using the same procedure.

Annotation of Peaks: HiCCUPS

Our peak-calling algorithm examines each pixel in a Hi-C contact matrix and compares the number of contacts in the pixel to the number of contacts in a series of regions surrounding the pixel. The algorithm thus identifies "enriched pixels" $M^*_{i,j}$ where the contact frequency is higher than expected and where this enrichment is not the result of a larger structural feature. For instance, we rule out the possibility that the enrichment of pixel $M^*_{i,j}$ is the result of L_i and L_j lying in the same domain by comparing the pixel's contact count to an expected model derived by examining the "lower-left" neighborhood. (The "lower-left" neighborhood samples pixels $M_{i',j'}$ where $i \leq i' \leq j' \leq j$; if a pixel is in a domain, these pixels will necessarily be in the same domain.) We require that the pixel being tested contain at least 50% more contacts than expected based on the lower-left neighborhood and the enrichment be statistically significant after correcting for multiple hypothesis testing (False Discovery Rate < 10%). The same criteria are applied to three other neighborhoods. Thus, to be labeled an enriched pixel, a pixel must be significantly enriched relative to four neighborhoods: (1) pixels to its lower-left, (2) pixels to its left and right, (3) pixels above and below, and (4) a donut surrounding the pixel of interest (Figure 3A). The resulting enriched pixels tend to form contiguous interaction regions comprising 5–20 pixels each. We define the "peak pixel" (or simply the "peak") to be the pixel in an interaction region with the most contacts.

Because of the enormous number of pixels that must be examined, this calculation requires weeks of central processing unit (CPU) time to execute. (For instance, at a matrix resolution of 5 kb, the algorithm must be run on 20 billion pixels.) To accelerate it, we created a highly parallelized implementation using general-purpose graphical processing units resulting in a 200-fold speedup.

Aggregate Peak Analysis

We perform APA on 10 kb resolution contact matrices. To measure the aggregate enrichment of a set of putative peaks in a contact matrix, we plot the sum of a series of submatrices derived from that contact matrix. Each of these submatrices is a 210 kb × 210 kb square centered at a single putative peak in the upper triangle of the contact matrix. The resulting APA plot displays the total number of contacts that lie within the entire putative peak set at the center of the matrix; the entry immediately to the right of center corresponds to the total number of contacts in the pixel set obtained by shifting the peak set 10 kb to the right; the entry two positions above center corresponds to an upward shift of 20 kb and so on. Focal enrichment across the peak set in aggregate manifests as larger values at the center of the APA plot. The APA plots shown only include peaks whose loci are at least 300 kb apart.

ACCESSION NUMBERS

The Gene Expression Omnibus (GEO) accession number for the data sets reported in this paper is GSE63525. The dbGaP accession number for the HeLa data reported in this paper is phs000640.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, eight figures, two data files, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.11.021.

AUTHOR CONTRIBUTIONS

E.L.A. conceived this project. S.S.P.R., M.H.H., E.K.S., and E.L.A. designed experiments. S.S.P.R., E.K.S., I.D.B., A.D.O., and M.H.H. performed Hi-C experiments. E.K.S. and I.D.B. performed 3D-FISH experiments. N.C.D. built the computational pipeline for Hi-C data. N.C.D. and J.T.R. built the visualization system for Hi-C data. S.S.P.R., M.H.H., N.C.D., A.L.S., I.M., E.S.L., and E.L.A. analyzed data. S.S.P.R., M.H.H., N.C.D., E.S.L., and E.L.A. prepared the manuscript.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Aharon Lieberman. Our work was supported by an NSF Graduate Research Fellowship (DGE0946799 and DGE1144152) to M.H.H., an NIH New Innovator Award (OD008540-01), an NSF Physics Frontier Center (PHY-1427654, Center for Theoretical Biological Physics), an NHGRI CEGS (HG006193), NVIDIA, IBM, Google, a CPRIT Scholar Award (R1304), a McNair Medical Institute Scholar Award, and the President's Early Career Award in Science and Engineering to E.L.A., and an NHGRI grant (HG003067) to E.S.L. We thank Leslie Gaffney, Lauren Solomon, and Bang Wong for assistance with figures; BCM's Integrated Microscopy Core, Michael Mancini, Justin Demmerle, Wendy Salmon, Fabio Stossi, Radhika Dandekar, Sanjay Krishna, and especially Asha Multani for microscopy assistance; Aharon Lieberman, Aviva Presser Aiden, Nicholas Christakis, James Lupski, José Onuchic, Mitchell Guttman, Andreas Gnirke, Louise Williams, Chad Nusbaum, John Bohannon, Olga Dudchenko, and the Aiden laboratory for discussions; and Robbyn Issner and Broad's ENCODE group for several cell lines. The Center for Genome Architecture is grateful to Janice, Robert, and Cary McNair for support. A provisional patent covering in situ Hi-C and related methods has been filed. All sequence data reported in this paper that were not derived from HeLa cells have been deposited at GEO (http://www.ncbi.nlm.nih.gov/geo/) (GSE63525). Some of the genome sequences described in this research were derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. The HeLa data generated from this research were submitted to the database of Genotypes and Phenotypes (dbGaP) as a substudy under accession number phs000640.

Received: October 12, 2014
Revised: November 5, 2014
Accepted: November 13, 2014
Published: December 11, 2014

REFERENCES

Ahmadiyeh, N., Pomerantz, M.M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., He, H.H., Brown, M., Liu, X.S., Davis, M., et al. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc. Natl. Acad. Sci. USA *107*, 9742–9746.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev. Cell *16*, 47–57.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., et al. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Natl. Acad. Sci. USA *109*, 21301–21306.

Bickmore, W.A. (2013). The spatial organization of the human genome. Annu. Rev. Genomics Hum. Genet. *14*, 67–84.

Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60–63.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L.L., Ruan, Y., Wei, C.-L.L., Ng, H.H., and Liu, E.T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. *18*, 1752–1762.

Chadwick, B.P. (2008). DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Genome Res. *18*, 1259–1269.

Cook, P.R., and Brazell, I.A. (1975). Supercoils in human DNA. J. Cell Sci. *19*, 261–279.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.-Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24–32.

Cullen, K.E., Kladde, M.P., and Seyfred, M.A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. Science *261*, 203–206.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet. *4*, e1000039.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet. *7*, 703–713.

Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, I.I., Iarovaia, O.V., and Razin, S.V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res. *41*, 3563–3575.

Hahn, M.A., Wu, X., Li, A.X., Hahn, T., and Pfeifer, G.P. (2011). Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. PLoS ONE *6*, e18844.

Heidari, N., Phanstiel, D.H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M.Q., and Snyder, M.P. (2014). Genome-wide map of regulatory interactions in the human genome. Genome Res. Published online September 16, 2014. http://dx.doi.org/10.1101/gr.176586.114.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.

Horakova, A.H., Moseley, S.C., McLaughlin, C.R., Tremblay, D.C., and Chadwick, B.P. (2012). The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. Hum. Mol. Genet. *21*, 4367–4377.

Hou, C., Zhao, H., Tanimoto, K., and Dean, A. (2008). CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc. Natl. Acad. Sci. USA *105*, 20398–20403.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Knight, P., and Ruiz, D. (2012). A fast algorithm for matrix balancing. IMA J. Numer. Anal. Published online October 26, 2012. http://dx.doi.org/10.1093/imanum/drs019.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mukherjee, S., Erickson, H., and Bastia, D. (1988). Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein. Cell *52*, 375–383.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schleif, R. (1992). DNA looping. Annu. Rev. Biochem. *61*, 199–223.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348.

Sexton, T., Schober, H., Fraser, P., and Gasser, S.M. (2007). Gene regulation through nuclear organization. Nat. Struct. Mol. Biol. *14*, 1049–1055.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. *20*, 2349–2354.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell *10*, 1453–1465.

Vogel, M.J., Guelen, L., de Wit, E., Peric-Hupkes, D., Lodén, M., Talhout, W., Feenstra, M., Abbas, B., Classen, A.K., and van Steensel, B. (2006). Human heterochromatin proteins form large domains containing KRAB-ZNF genes. Genome Res. *16*, 1493–1504.

Vogelstein, B., Pardoll, D.M., and Coffey, D.S. (1980). Supercoiled loops and eucaryotic DNA replicaton. Cell *22*, 79–85.

Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E.S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc. Natl. Acad. Sci. USA *104*, 7145–7150.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

Zehnbauer, B.A., and Vogelstein, B. (1985). Supercoiled loops and the organization of replication and transcription in eukaryotes. BioEssays *2*, 52–54.

TIMELINE

The rise of regulatory RNA

Kevin V. Morris and John S. Mattick

Abstract | Discoveries over the past decade portend a paradigm shift in molecular biology. Evidence suggests that RNA is not only functional as a messenger between DNA and protein but also involved in the regulation of genome organization and gene expression, which is increasingly elaborate in complex organisms. Regulatory RNA seems to operate at many levels; in particular, it plays an important part in the epigenetic processes that control differentiation and development. These discoveries suggest a central role for RNA in human evolution and ontogeny. Here, we review the emergence of the previously unsuspected world of regulatory RNA from a historical perspective.

RNA has long been at the centre of molecular biology and was likely the primordial molecule of life, encompassing both informational and catalytic functions. Its informational functions are thought to have subsequently devolved to the more stable and easily replicable DNA, and its catalytic functions to the more chemically versatile polypeptides[1]. The idea that the contemporary role of RNA is to function as the intermediary between the two had its roots in the early 1940s with the entry of chemists into the study of biology, notably Beadle and Tatum[2], whose work underpinned the one gene–one enzyme hypothesis (FIG. 1 (TIMELINE)). This idea later matured into the more familiar one gene–one protein concept and became widely accepted despite the prescient misgivings of experienced geneticists, notably McClintock[3]. The concept that genes encode only the functional components of cells (that is, the 'enzymes') itself had deeper roots in the mechanical zeitgeist of the era, which was decades before the widespread understanding of the use of digital information for systems control.

Although the one gene–one protein hypothesis has long been abandoned owing to the discovery of alternative splicing in the 1970s, the protein-centric view of molecular biology has persisted. Such persistence was aided by phenotypic and ascertainment bias towards protein-coding mutations in genetic studies and by the assumption that these mutations affected *cis*-acting regulatory protein-binding sites[4]. However, this view was challenged by the discovery of nuclear introns and RNA interference (RNAi), as well as by the advent of high-throughput sequencing, which led to the identification of large numbers and different types of large and small RNAs, the functions of which are still under investigation.

> "emerging evidence suggests … that the amount and type of gene regulation in complex organisms have been substantially misunderstood"

In this Timeline article, we examine the history of, and report the shift in thinking that is still underway about, the role of RNA in cell and developmental biology, especially in animals. The emerging evidence suggests that there are more genes encoding regulatory RNAs than those encoding proteins in the human genome, and that the amount and type of gene regulation in complex organisms have been substantially misunderstood for most of the past 50 years.

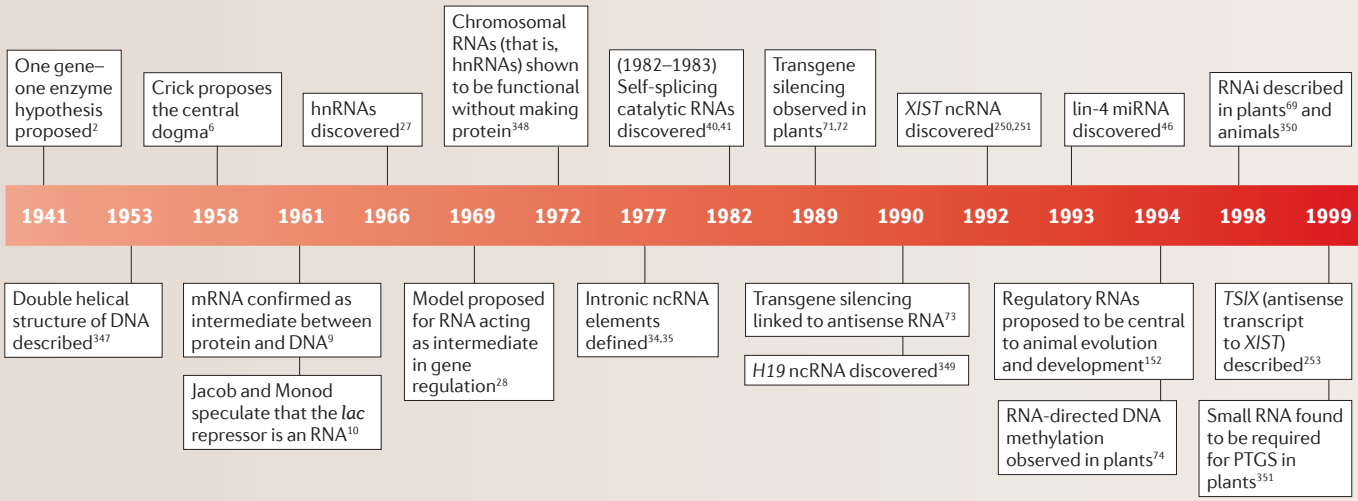
Early ideas for the role of RNA

RNA — the central dogma and gene regulation. After the elucidation of the double helical structure of DNA in 1953 (REF. 5), the following years were preoccupied with deciphering the 'genetic code' and establishing the mechanistic pathway between genes and proteins: the identification of a transitory template (mRNA), an adaptor (tRNA) and the ribosome 'factory' comprised of ribosomal RNAs and proteins for translating the code into a polypeptide. In 1958, Crick published the celebrated central dogma to describe the flow of genetic information from DNA to RNA to protein, which has proved remarkably accurate and durable, including the prediction of reverse transcription[6]. Nonetheless, in conceptual terms, RNA was tacitly consigned to be the template and an infrastructural platform (with regard to rRNAs and tRNAs) for protein synthesis or has at least been interpreted in this way by most people since that time.

In the mid-1950s, the link was established between rRNA (which is highly expressed in essentially all cells) and the structures termed ribosomes as the platform for protein synthesis[7]. The roles of tRNA and mRNA were experimentally confirmed in 1958 (REF. 8) and 1961 (REF. 9), respectively. The latter occurred in the same year that Jacob and Monod published their classic paper on the *lac* operon of *Escherichia coli*[10], which was the first locus to be characterized at the molecular genetic level. These studies confirmed that at least some, but presumably most, genes encoded proteins and supported the emerging idea that gene expression is controlled by regulating the transcription of the gene, as indicated by the locus encoding the *lac* repressor in the repressor–operator model. At the time, Jacob and Monod did not know the chemical identity of the repressor and speculated in passing that it "may be a polyribonucleotide" (that is, RNA)[10]. However, Gilbert later showed that the repressor is a polypeptide that allosterically binds to the lactose substrate, and the brief idea faded[11].

These studies reinforced and extended the concept that proteins are not only enzymes but also the primary analogue components and control factors that constitute the cellular machinery. This, in turn, has led to the prevailing transcription factor

Timeline | **The rise of regulatory RNA**



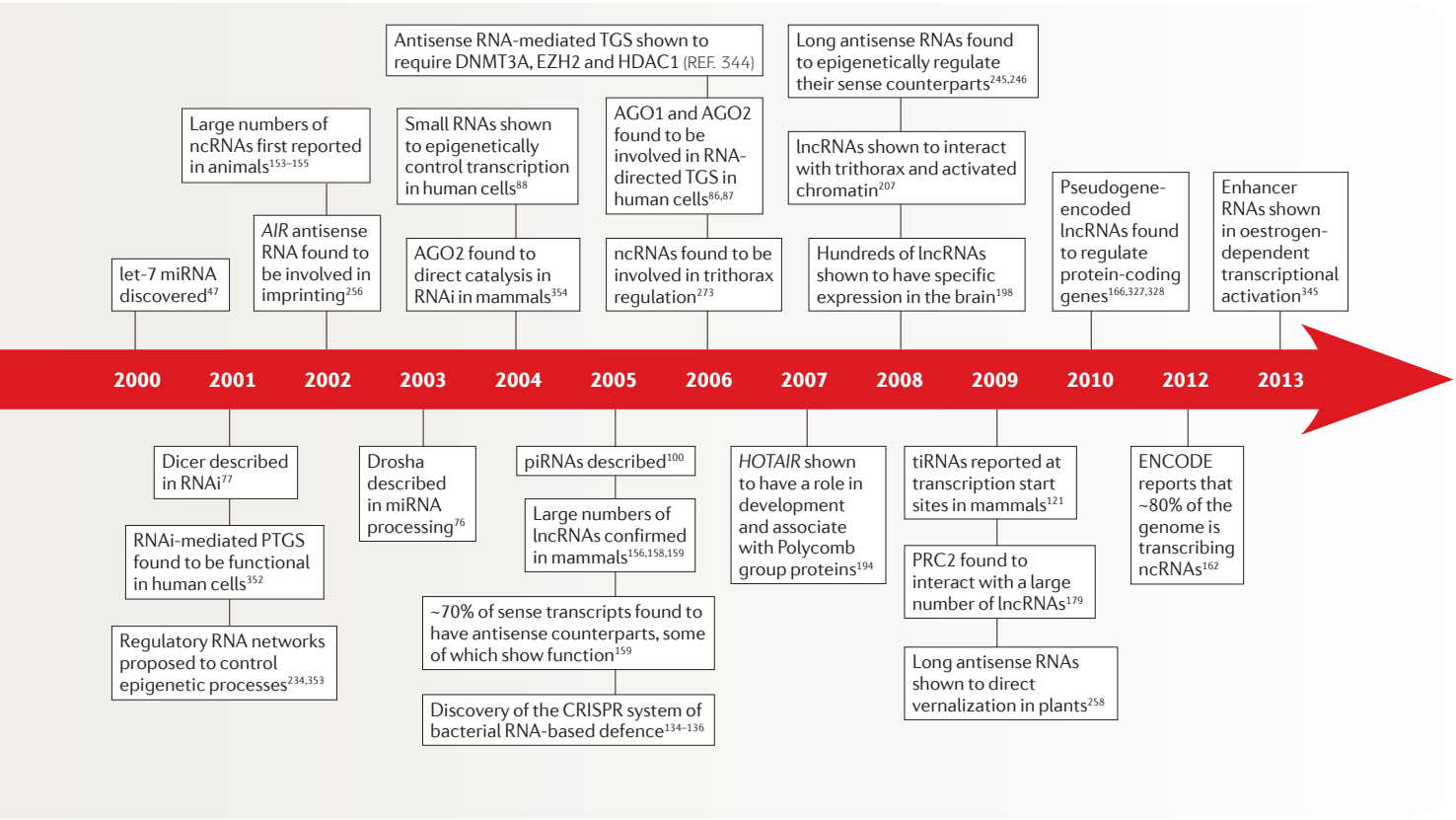
AGO, Argonaute; *AIR*, also known as *AIRN* (antisense of *IGF2R* non-protein coding RNA); CRISPR, clustered regularly interspaced short palindromic repeat; DNMT3A, DNA (cytosine-5)-methyltransferase 3A; ENCODE, Encyclopedia of DNA Elements; EZH2, enhancer of Zeste 2; *H19*, H19 imprinted maternally expressed transcript; HDAC1, histone deacetylase 1; hnRNA, heterogeneous nuclear RNA; *HOTAIR*, HOX transcript antisense RNA; lncRNA, long non-coding RNA; miRNA, microRNA; ncRNA, non-coding RNA; piRNA, PIWI-interacting RNA; PRC2, Polycomb repressive complex 2; PTGS, post-transcriptional gene silencing; RNAi, RNA interference; TGS, transcriptional gene silencing; tiRNA, transcription initiation RNA; *XIST*, X inactive specific transcript.

paradigm of gene regulation, including the derived assumption that combinatorial interactions would provide an enormous range of regulatory possibilities[12] that are more than enough to control human ontogeny. However, this assumption has not been substantiated theoretically or mechanistically, and both the observed scaling of regulatory genes and the extent of the regulatory challenge in programming human developmental architecture seem to be different from these expectations[13]. In this context, it is noteworthy that genome-wide association studies have shown that most haplotype blocks influencing complex diseases are outside the known boundaries of protein-coding genes[14].

Small nuclear RNAs and small nucleolar RNAs. Following the discovery and functional description of tRNAs and rRNAs, new classes of common small RNAs in the nucleus were identified by biochemical fractionation[15]. Many of these small RNAs were found to be part of ribonucleoprotein (RNP) complexes (reviewed in REF. 16). One class — the small nuclear RNAs (snRNAs) (FIG. 2) — was later found to be a central cofactor in RNA splicing[17] (see below) and was therefore given the newer designation as spliceosomal RNAs. The snRNAs U1, U2, U4, U5 and U6 participate in various RNA–RNA and RNA–protein interactions in the assembly and function of canonical spliceosomes: U1 and U2 recognize the 5′ splice site and the branch point, respectively, followed by the recruitment of U4, U5 and U6, which displace U1 and interact with U2 (through U6) as well as the 5′ and 3′ splice sites (through U5)[18]. A set of less abundant snRNAs (U11, U12, U4atac and U6atac) and U5 are found in a variant 'minor' spliceosome termed U12-type[19].

Other small RNAs were found to be localized to the nucleolus and to guide the methylation (the box C/D subclass) and pseudouridylation (the box H/ACA subclass) of rRNAs, tRNAs and snRNAs[20–22] (FIG. 2). The chemical modifications of rRNAs, tRNAs and snRNAs proved to be essential in ribosomal and cellular function, particularly in tRNA and mRNA maturation, and in pre-mRNA splicing (which requires modification of the U2 snRNA). Notably, the disruption of small nucleolar RNAs (snoRNAs) was found to cause a loss of processing of the 5.8S, 18S and 28S (or 25S in plants) rRNAs[20]. Early studies found that some snoRNAs are subject to parental imprinting and/or differentially expressed (for example, in the brain[23,24]), and that they seem to target a wide range of RNAs (including mRNAs[25]), which suggests a regulatory role. Related small RNAs have also been identified in subnuclear structures called Cajal bodies (which process telomerase RNA), and these were termed small Cajal body-specific RNAs (scaRNAs)[26]. However, none of these studies suggested anything other than that the role of RNA was limited to protein synthesis.

The emergence of heterogeneous nuclear RNAs. The first hint that RNA may have additional roles in complex organisms was the discovery of heterogeneous nuclear RNA (hnRNA)[27] and the observation that the complexity of this population, as determined by denaturation–renaturation hybridization kinetics, was much greater in the nucleus than in the cytoplasm. The existence of hnRNA and the concomitant discovery of the large amount of repetitive sequences (that is, different classes of retrotransposon sequences with similar composition that occupy large portions of plant and animal genomes) led Britten and Davidson to speculate in 1969 that animal cells contain extensive RNA-based regulatory networks[28–30]. Although this hypothesis attracted a great deal of interest at the time, it also quickly



lapsed. Its proponents did not revisit the hypothesis even after the subsequent discovery of introns (see below) and instead focused on regulatory networks controlled by transcription factors[31,32] or on the importance of transposons in protein evolution[33].

The discovery of introns. The discovery of introns in 1977 (REFS 34,35) was perhaps the biggest surprise in the history of molecular biology[36] (FIG. 1 (TIMELINE)), as no one expected that the genes of higher organisms would be mosaics of coding and non-coding sequences, all of which are transcribed. However, the prevailing concept of the flow of genetic information was not overly disturbed, as the removal of the intervening sequences (that is, introns) and the reconstruction of a mature mRNA by splicing preserved the conceptual status quo; that is, genes still made proteins. In parallel, it was assumed that the excised intronic RNAs were simply degraded, although the technology of the time was too primitive to confirm this. In any case, introns were immediately and universally dismissed as genomic debris, and their presence was rationalized as evolutionary remnants involved in the prebiotic modular assembly of protein-coding RNAs that have remained (and been expanded by transposition) in complex organisms[37]. This notion was consistent, at least superficially, with the implication of the C-value enigma that eukaryotes contained varying amounts of DNA 'baggage'. It is also in agreement with the accompanying conclusion that retrotransposon sequences are mainly 'selfish', parasitic DNA[38,39].

RNA as a catalyst. A few years later, Cech, Altman and colleagues demonstrated that RNA itself was capable of enzymatic catalysis (that is, they are ribozymes)[40,41], which provided evidence in support of the RNA early hypothesis. They also showed that RNA catalysis exists and has persisted in particular contexts, notably at the core of RNA splicing[42] and mRNA translation[43]. This finding reinforced both the mechanical concept of molecular biology and the role of RNA as the platform for protein synthesis, but did not give any hint of RNA as a widespread regulatory factor, although that possibility is perfectly feasible. Indeed, there is increasing evidence that catalytic RNA exists in animal and plant cells, in introns, untranslated regions (UTRs) and elsewhere, and that these RNAs may have various roles, for example, in the regulation of post-transcriptional cleavage reactions[44,45].

The small RNA revolution

The discovery of microRNAs. In 1993, Ambros and colleagues showed the first evidence for small (~22-nucleotide) regulatory RNAs with the discovery of the genetic loci *lin-4* and *let-7*, which regulate the timing of *Caenorhabditis elegans* development[46,47] (FIG. 1 (TIMELINE)). Although *let-7* is highly conserved from nematodes to humans[48], very few microRNAs (miRNAs) were discovered genetically[49,50], and these RNAs remained interesting idiosyncrasies until the discovery of RNAi (see below). This discovery led to the targeted cloning after size selection of many more miRNAs[51–53] and the demonstration that these miRNAs act, at least partly, by imperfect base-pairing — typically with the 3′UTRs of target mRNAs — to inhibit their translation and to accelerate their degradation[54].

Current databases list large numbers of evolutionarily widespread miRNAs[55], almost all of which had evaded prior detection by genetic screens but many were subsequently validated by reverse genetics. Although many miRNAs can be identified by conservation, it is also evident that many are tissue and lineage specific[56,57], and that there may be many more to be discovered.

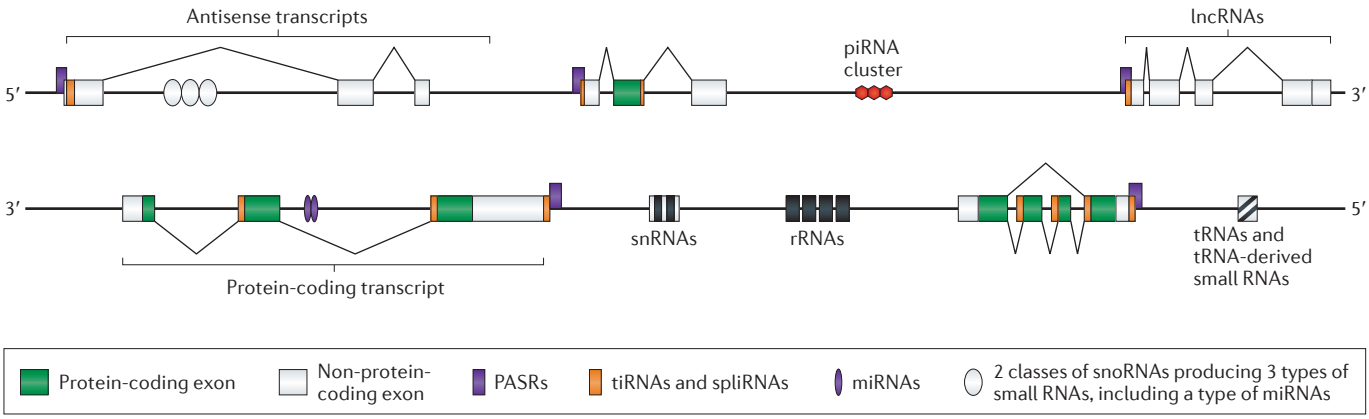


Figure 2 | **Complex expression of the genome and examples of non-coding RNA expression.** The mammalian transcriptional landscape is represented graphically with genes expressing ribosomal RNAs, tRNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), various protein-coding and non-coding genes (which encode mRNAs and long non-coding RNAs (lncRNAs), respectively), as well as genes expressing small regulatory RNAs such as microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), promoter-associated short RNAs (PASRs), transcription initiation RNAs (tiRNAs) and splice site RNAs (spliRNAs), snoRNA-derived small RNAs and tRNA-derived small RNAs. The transcriptional units are not depicted to scale.

There is also evidence that many, if not most, protein-coding transcripts are targets for miRNA regulation[58,59]. In some cases, miRNAs can regulate large numbers of target mRNAs[60] and, reciprocally, many mRNAs contain target sites for a large number of miRNAs[61], although the implied regulatory logic of this complex multiplex arrangement has not been explained. The targets of miRNAs are usually thought to be mRNAs but may also include other types of RNAs[62]. Biologically, miRNAs have been shown to regulate many physiological, developmental and disease processes, including pluripotency[63], epithelial–mesenchymal transition and metastasis[64], testis differentiation[65], diabetes[66], and neural plasticity and memory[67], among others[68].

The RNA interference pathway. miRNAs are only one aspect of the phenomenon of RNAi, which silences gene expression after the introduction of sense–antisense RNA pairs. This process was discovered in 1998 in plants[69] and *C. elegans*[70] (FIG. 1 (TIMELINE)). These discoveries were presaged by the curious phenomenon of transgene silencing, which is mainly found in plants[71,72] and linked to both antisense RNA and small RNA-directed DNA methylation, thus indicating transcriptional and post-transcriptional silencing[73,74]. Mechanistic analyses of these silencing mechanisms showed that exogenous double-stranded RNA (dsRNA) is processed into short fragments (known as small interfering RNAs (siRNAs)) with similar sizes to miRNAs, which implies that miRNAs may represent a similar endogenous system.

This hypothesis was confirmed and led to the elucidation of natural dsRNA precursors in stem–loop structures[75], as well as the identification of key genes and enzymes involved in their biogenesis and function, notably Drosha[76], Dicer[77] and several Argonaute (AGO) proteins[78]. AGO proteins were already known to have central roles in differentiation and development[79] but are now known to also be involved in defence against RNA viruses in many organisms[80]. Drosha and exportin 5 are involved in the cleavage and export of dsRNA precursors from the nucleus to the cytoplasm[76], where they are further processed by Dicer to small (21–24-nucleotide) dsRNA moieties. One strand of the dsRNA is loaded into the AGO component of the RNA-induced silencing complex (RISC), which also comprises other proteins[77]. The RISC is guided by the small RNA strand to complementary RNA targets, which are subsequently silenced by translational repression and/or RNA destabilization[81,82] (FIG. 3).

Although still under discussion, the current view is that siRNAs (and short hairpin RNAs (shRNAs)) — which seem to naturally occur more commonly in plants — act primarily by perfect base-pairing and by AGO-mediated cleavage of complementary target RNAs; hence, they are used widely as experimental tools and potential therapeutic agents[83]. By contrast, miRNAs have incomplete homology with their target sequences and act primarily at the translational level[81,82] (FIG. 3).

Both miRNAs and siRNAs are thought to act post-transcriptionally in the cytoplasm, but the existence of AGO in the nucleus[84–87] and the role of the RNAi pathway in epigenetic modulation[88] suggest that the system is more complex and multifaceted than expected. For example, it has been shown that miRNA isoforms are developmentally regulated[89], that the target 'seed' sequence is only one factor in target recognition[90,91] and that miRNAs can also impose transcriptional gene silencing[92] (FIG. 3). There is also growing evidence of intersecting pathways, such as RNA editing and modification, in these networks[93–96].

PIWI-associated small RNAs. Although most AGO proteins are expressed ubiquitously and associate with both miRNAs and siRNAs, there is a subclade of AGO proteins termed PIWI that are required for germ cell development[97–100]. PIWI and PIWI-like proteins associate with a distinctive class of small (26–30-nucleotide) RNAs termed PIWI-interacting RNAs (piRNAs), which epigenetically and post-transcriptionally silence transposons in germ cells[101–110]. PIWI is found predominantly in the nucleus[111], colocalizes in an RNA-dependent manner with Polycomb group proteins[112] and seems to be expressed in other tissues (including the brain[113]), which suggests a role beyond genome protection in epigenetic processes[114,115].

Other classes of small RNAs in eukaryotes. The molecular genetics, biochemistry and structural biology of the RNAi system are still being unravelled but indicate an ancient, widespread and multilaterally adapted system that controls many cellular processes, the dimensions of which are still being explored. These include potentially lineage-specific

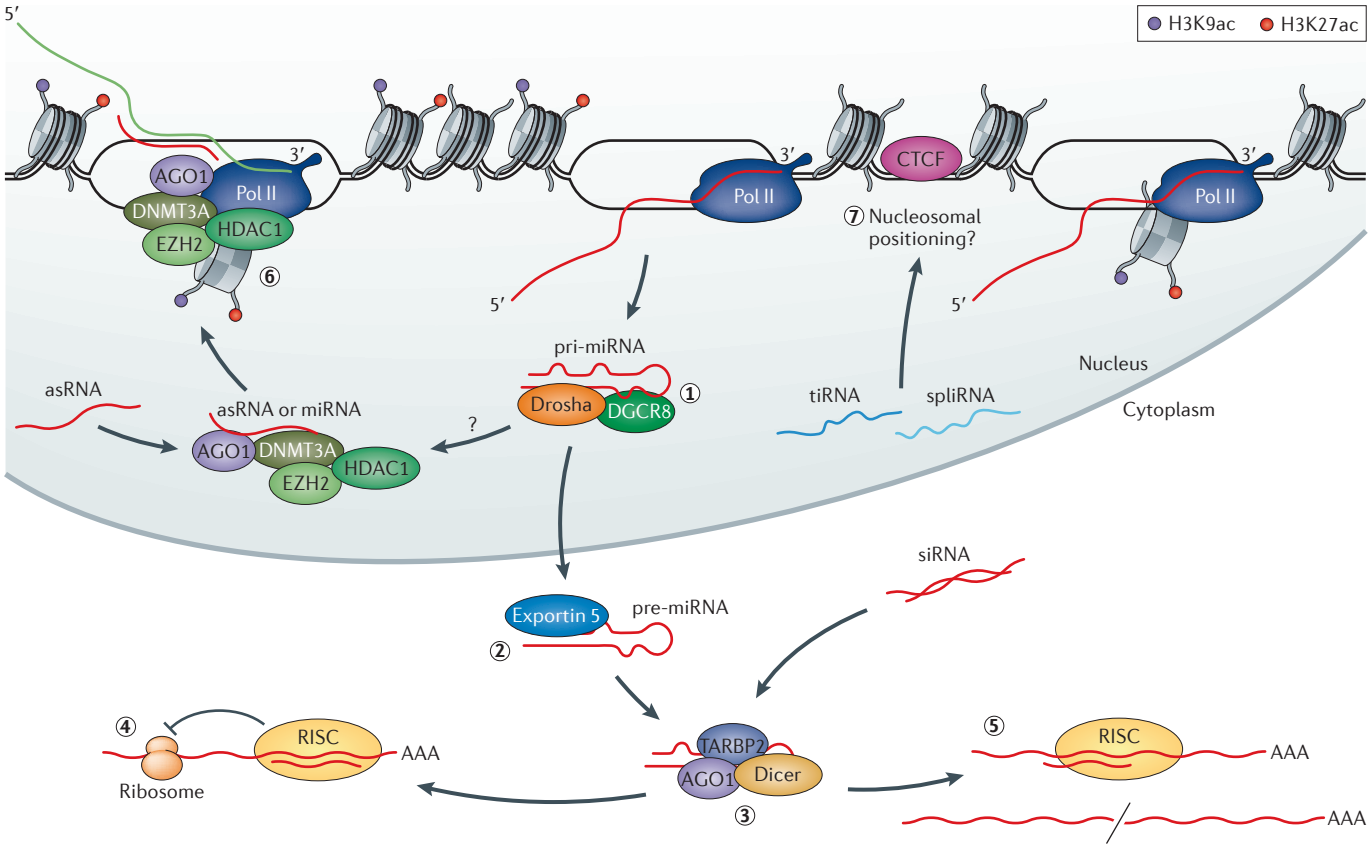


Figure 3 | **Functional pathways of small regulatory RNAs.** MicroRNA (miRNA) precursors (that is, pri-miRNAs) are expressed as stem–loop structures[75], which interact with Drosha[76] and DGCR8 (also known as Pasha) (step 1). They are then processed into pre-miRNAs and exported from the nucleus by exportin 5 (step 2). These transcripts are further processed by Dicer to small (21–23-nucleotide) double-stranded RNAs, one strand of which is loaded into the Argonaute (AGO) component of the RNA-induced silencing complex (RISC) (step 3). Exogenously introduced small interfering RNAs (siRNAs) can also be processed by RISC. The endogenous miRNA or siRNA, or exogenously added siRNA, can then target the repression of translation (step 4) and/or cleavage of homology-containing transcripts[81,82] (step 5). Some small RNAs are functional in the nucleus. Exogenously introduced small antisense RNAs (asRNAs) can induce epigenetic silencing of targeted loci[88,342,343] — a pathway that miRNAs may also use in the nucleus[92] (step 6). Transcription initiation RNAs (tiRNAs) and splice site RNAs (spliRNAs)[121,122] are expressed through an unknown pathway that may involve RNA polymerase II (Pol II) backtracking and TFIIS cleavage[123] (not shown); tiRNAs and spliRNAs are shown to modulate CCCTC-binding factor (CTCF) chromatin localization and to be associated with nucleosome positioning[124] (step 7). DNMT3A, DNA (cytosine-5)-methyltransferase 3A; EZH2, enhancer of Zeste 2; H3K9ac, histone H3 lysine 9 acetylation; HDAC1, histone deacetylase 1; TARBP2, RISC-loading complex subunit TARBP2 (also known as TRBP).

variations such as the 21U RNAs in *C. elegans*[116]. Surprisingly, it seems that all snoRNAs from fission yeast to humans produce at least three different subclasses of small RNAs[117], one of which has the same size and functions as miRNAs[118], and another that is similar in size to piRNAs[117]. There are also intriguing and recurring reports of tRNA fragments that are produced in tissue-specific patterns[119] and that are associated with AGO proteins[120].

More recently, deep sequencing of small RNA populations has revealed the existence of two other classes of small RNAs in animals but not in plants, which are 17–18 nucleotides in length and associated with transcription initiation[121] and splice sites[122] (termed transcription initiation RNAs (tiRNAs) and splice site RNAs (spliRNAs), respectively) (FIG. 3). The origin and function of these RNAs are uncertain, but preliminary evidence suggests that they play a part in nucleosome positioning[123] and/or in other levels of chromatin organization[124]. There are also other reports of less distinct classes of promoter-associated RNAs called promoter-associated short RNAs (PASRs)[125], transcription start site-associated RNAs (TSSa-RNAs)[126] and promoter upstream transcripts (PROMPTS)[127], some of which may have a role in RNA-directed transcriptional gene silencing[128].

Regulatory RNAs in bacteria and archaea. Many small regulatory RNAs have been identified in bacteria, in which they regulate a wide variety of adaptive responses. Bacterial small regulatory RNAs generally function by simple antisense mechanisms to regulate translation or stability of target mRNAs through altering their secondary structure to expose or sequester *cis*-acting sites[129,130]. Studies in bacteria have also identified *cis*-acting regulatory RNA sequences known as riboswitches, which act allosterically by binding metabolites to regulate gene expression[131,132] and almost certainly exist as part of the RNA regulatory landscape in all kingdoms of life.

Very recently, the bacterial and archaeal kingdoms have once again surprised us with the sophistication of their molecular machinery. Many bacterial and most

archaeal genomes have loci comprised of regularly spaced repeats that are interspersed by other virus-derived DNA sequences[133–136] (termed clustered regularly interspaced short palindromic repeats (CRISPRs)). These loci act as an innate immune system by incorporating fragments of viral DNA between the repeats, which are then transcribed and processed to produce small guide RNAs that are linked to their effector complexes through the repeat sequence and that target and destroy viral DNA[137–140] or RNA[141]. This system has recently been adapted for RNA programmable sequence-specific genome manipulation in eukaryotes (including mammals[142–145]) with extraordinary versatility, including targeted gene excision and fusion, as well as engineered CRISPRs that can recruit silencing and activating proteins to target loci[146–150]. Moreover, the biological 'arms race' continues, as bacteriophages encode their own CRISPR system to evade host innate immunity[151].

Long non-coding RNAs

The eukaryotic transcriptome. Noting that the density and size of introns (and, as it turned out later, intergenic sequences) increased with developmental complexity, Mattick posited in 1994 that introns had evolved to express an expanding range of *trans*-acting regulatory RNAs (FIG. 1 (TIMELINE)). He postulated that some genes subsequently evolved to express only intronic or exonic regulatory RNAs, and that this RNA-based regulatory system was the essential prerequisite for the emergence of developmentally complex organisms[152]. Subsequently, the application of genome tiling array technology and deep sequencing to the characterization of the transcriptome showed that tens of thousands of loci in mammals express long transcripts that do not encode proteins, which are located intergenic, intronic and antisense to protein-coding genes. The initial findings[153–155] were confirmed in 2005 (REFS 156–159) and extended by the Encyclopedia of DNA Elements (ENCODE) project[160–162], all of which showed that the vast majority (at least 80%) of the human and mouse genomes are differentially transcribed in one context or another; other studies also reported similar findings in all organisms examined. Indeed, it seems that most intergenic and, by definition, intronic sequences are differentially transcribed, and that the extent of the transcriptome therefore expands with developmental complexity[163].

Using more focused deep sequencing methodologies, it has become evident that the full range of the protein-coding and non-protein-coding transcriptome is still vastly under-sampled[164]. In addition, many transcripts are not polyadenylated and represent a largely different sequence class[156,165], some of which seem to be relevant to development (for example, the *POU5F1* (also known as *OCT4*) transcript[166,167]). Moreover, 95% of human transcription initiation sites are not associated with mRNA transcription but rather mainly with transcription of non-polyadenylated non-coding RNAs (ncRNAs)[168]. These non-polyadenylated transcripts are so far mostly uncharacterized because of the historical use of poly(A) tails to remove the overwhelming rRNA contamination in RNA preparations. This issue is being alleviated by more sophisticated approaches such as cap trapping[169], oligonucleotide subtraction[170] and array capture[164,171].

Defining long non-coding RNAs. Long ncRNAs (lncRNAs) are operationally defined as non-protein-coding RNAs that are >200 nucleotides in length, which corresponds to a convenient cutoff in biochemical fractionation and excludes all known classes of small RNAs[172]. Transcripts are classified as non-coding if they lack long open reading frames (traditionally >100 codons) and/or do not show codon conservation, although there was considerable uncertainty, as genomic and transcriptomic data were initially limited for comparison. However, recent analyses provide strong evidence that most annotated lncRNAs do not encode proteins; nonetheless, some specify small proteins that had not been identified previously using bioinformatic approaches[173–175].

These ncRNAs can be parsed into intronic, antisense or intergenic (that is, large intergenic non-coding RNA (lincRNA)) subsets in experimental studies and databases[159,176,177], partly because of mechanistic expectations[178] and because of a desire to reduce ambiguity and overlap with protein-coding loci in functional analyses[179–181]. However, there is no evidence of any intrinsic difference between RNAs that are intronic, intergenic or antisense, or that overlap with protein-coding transcripts (FIG. 2), for example, in their interaction with chromatin-activating or chromatin-repressive complexes (see below). Nonetheless, ncRNA subclasses will inevitably exist and be defined, some of which may be biased in relation to genomic origin.

Long non-coding RNAs: transcriptional noise or functional? The unexpected discovery of large numbers of non-coding transcripts in eukaryotes, some of which span tens or hundreds of kilobases[182], led to debates about their functionality[183,184]. In particular, as many lncRNAs were shown to have fairly low evolutionary conservation and low levels of expression, some have posited that they represent transcriptional noise and/or redundant transcripts with no biological importance. This hypothesis remains, at least partly, a possibility. Nevertheless, lncRNAs show a wide range of evolutionary conservation, from ultraconserved ones[185] to primate-specific ones[186–188], which can be explained as the result of different structure–function constraints and lineage-specific adaptive radiation[189]. Indeed, there is now considerable evidence that lack of primary sequence conservation in lncRNAs does not indicate lack of function[190,191], and many lncRNAs show evidence of structural conservation[192,193].

Loci that express lncRNAs show all of the hallmarks of bona fide genes[4], including conservation of promoters[169], indicative chromatin structure[194], and regulation by conventional morphogens and transcription factors[195]. Moreover, lncRNAs were found to have a similar range of cellular half-lives as mRNAs[196] and to be differentially expressed in a tissue-specific manner[158,197], especially in the brain[198]. The study in the brain showed that, although the expression levels of many lncRNAs seem to be lower than those of mRNAs in whole tissues, lncRNAs are highly expressed and easily detectable in particular cell types[198]. In addition, lncRNAs were found to have, on average, higher cell specificity than proteins[165,199]; this is consistent with their proposed role in architectural (as opposed to 'cell-type') regulation, in which each cell has a unique positional identity in precisely sculpted organs, bones and muscles[200].

Many lncRNAs are alternatively spliced[201], which is further evidence of the precision of their expression and is hard to reconcile with the suggestion that they are simply transcriptional noise. It should also be noted that some functionally validated lncRNAs can have isoforms that encode proteins[202] and that, reciprocally, some (perhaps many) mRNAs have intrinsic functions as *trans*-acting regulatory RNAs[203–205]. In some contexts, 3′UTRs can be separately expressed and convey genetic functions in *trans*[204], and both lncRNAs and mRNAs may be further processed to produce subsidiary species[206].

lncRNAs have been shown to be dynamically expressed in a range of differentiating systems, including embryonic stem cells[207], muscles[208], T cells[209], breast tissues[210,211], the erythroid system[211] and neurons[212–214], as well as in cancer and other diseases[210,215–222]. Such dynamic expression of lncRNAs is at least partly controlled by conventional transcription factors[195,213].

Emerging roles of non-coding RNAs

The validation of ncRNA functions has so far mainly relied on knockdown of candidate ncRNAs. Knockdown of ncRNA expression has proved to be surprisingly easy using chemically engineered antisense oligonucleotides, or using siRNA- or shRNA-mediated approaches, frequently resulting in phenotypic changes in cultured cells, in which most studies have been carried out.

Development and differentiation. Many small ncRNAs[63–65] and most functionally analysed lncRNAs[223] seem to have a role in the regulation of differentiation and development. On the basis of studies in cell culture, these include the regulation of apoptosis and metastatic processes[211,218,220,221,224], retinal and erythroid development[211,225], breast development[210,226] and epidermal differentiation[227], among many others. Antisense knockdown of some lncRNAs in zebrafish and deletion of sequences that specify lncRNAs in mice have resulted in visible developmental defects[181,191,228,229]. However, knockouts of the widely expressed nuclear paraspeckle assembly transcript 1 (*Neat1*)[230] or of some of the most highly conserved sequences in the mammalian genome[231] have not shown any detrimental effect on development. These results suggest that more sophisticated phenotypic screens are required to delineate functions, especially cognitive ones, because most mammalian lncRNAs are expressed in the brain[198] and many are specific to mammals or primates[188,232]. A good example is brain cytoplasmic RNA 1 (*BC1*) — a retrotransposon-derived lncRNA that is widely expressed in the brain — the knockout of which causes no visible anatomical abnormality but leads to behavioural changes that would be lethal in the wild[233].

Epigenetic roles. Consistent with their roles in differentiation and development, a range of genetic and biochemical evidence suggests that a major function of many small RNAs and lncRNAs is the regulation of epigenetic processes[234,235], probably by guiding chromatin-modifying enzymes to their sites of action and/or by acting as scaffolds for chromosomal organization[179,235–238] (FIG. 4).

RNAs were shown to induce transcriptional gene silencing first in plants[74,239], then in fungi[240] and human cells[88], and both small RNAs and the RNAi machinery were implicated in the underlying epigenetic processes[240–242]. These studies were consistent with the observations that small RNAs interact with Polycomb group proteins[243] and that AGO proteins are found in the nucleus[86,87] (FIG. 3). In parallel, dating back to 1990, antisense RNAs were shown to affect gene expression, again initially in plants[73] and later in animals[159,166,244–246]. Similar to small ncRNAs[247], some lncRNAs have been shown to control alternative splicing[248,249]. Other naturally occurring lncRNAs were shown to control epigenetic processes *in vivo*, notably in X chromosome dosage compensation[250–254] and parental imprinting in mammals[255–257], and vernalization in plants[258]. Subsequent studies showed that intergenic and antisense RNAs bind to Polycomb repressive complexes (PRCs)[194,259–261], to trithorax chromatin-activating complexes and activated forms of histones[207], and to DNA methyltransferases[201,262,263]. These observations were writ large in 2009 when it was shown that ~20% of ~3,300 lncRNAs examined were bound by PRC2 and that others were bound by different chromatin-modifying complexes. siRNA-mediated knockdown of PRC2-associated lncRNAs was found to result in gene expression changes, and the upregulated genes were enriched for those normally silenced by PRC2 (REF. 179). Polycomb group proteins were also discovered to bind to RNA with high affinity but low specificity[264], which is consistent with the idea that many RNAs interact with these proteins.

One of the notable lncRNAs to emerge — HOX transcript antisense RNA (*HOTAIR*) — is derived from the *HOXC* locus and regulates *HOXD* in *trans*[194]. It is involved in cancer metastasis[220] and, when inactivated, results in homeotic transformation *in vivo*[229]. lncRNAs have also been shown to act as scaffolds for the assembly of histone modification complexes[265], and the widespread alternative splicing of these RNAs suggests that the cargo and/or target specificity can be varied in a context-dependent and differentiation-specific manner.

Glossary

Antisense RNA
A single-stranded RNA that is complementary to an mRNA or a gene.

Encyclopedia of DNA Elements
(ENCODE). An international consortium involved in building a comprehensive list of functional elements in the human genome.

Heterogeneous nuclear RNA
(hnRNA). A type of RNA that is similar to mRNA or pre-mRNA but that is retained predominantly in the nucleus.

Introns
A term first coined by Gilbert to describe the RNA regions that are removed, by being spliced out, to produce mRNAs.

PIWI-interacting RNAs
(piRNAs). Small RNAs that are associated with the PIWI protein complex and that emanated from transposon-like elements

RNA CaptureSeq
A method that combines the ability to capture RNA (that is, to isolate and enrich for certain types of RNA) with deep sequencing technology to mine the human transcriptome.

RNA-directed DNA methylation
An epigenetic process whereby processed double-stranded small (21–24-nucleotide) RNAs guide the methylation of homologous DNA loci.

Small interfering RNAs
(siRNAs). Small interfering, double-stranded RNAs that can be used to suppress homology-containing transcripts in a transcriptional and post-transcriptional manner.

Splice site RNAs
(spliRNAs). Small RNAs that are derived from the 3′ ends of exons adjacent to splice sites and that are similar to transcription initiation RNAs (tiRNAs).

Transcriptional gene silencing
The regulation of a gene at the transcriptional level, in contrast to post-transcriptional gene silencing, in which silencing of gene expression occurs at the mRNA or translational level, after transcription has occurred.

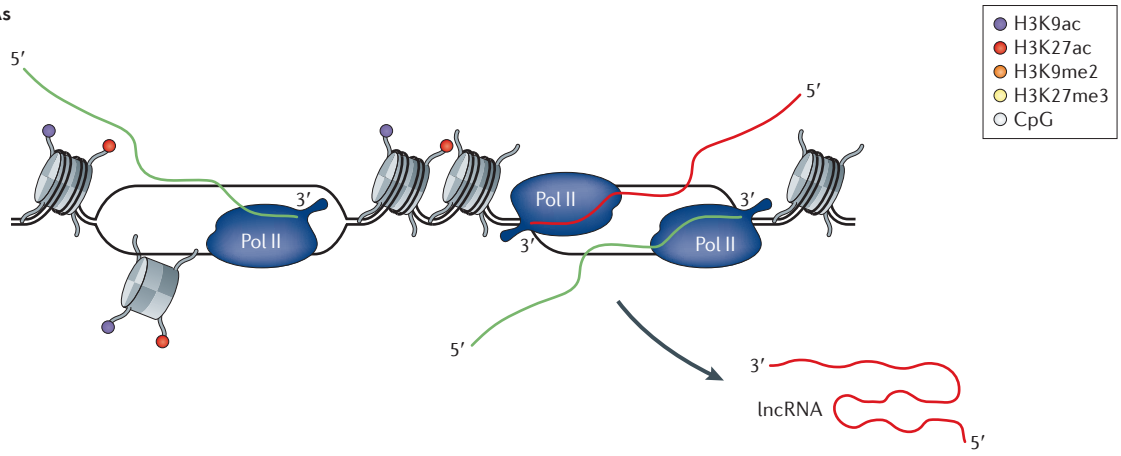
Transcription initiation RNAs
(tiRNAs). Small RNAs associated with promoters with peak density at ~15–35 nucleotides downstream of transcription start sites.

Transinduction
A genetic phenomenon whereby mRNA transcription induces transcription of nearby enhancers and intergenic non-coding RNAs.

Transposons
Mobile genetic elements with evolutionary links to retroviruses.

Transvection
A genetic phenomenon whereby non-coding regions can induce transcription of coding regions on other chromosomes.

Untranslated regions
(UTRs). Sequences either side of a coding sequence on a strand of mRNA; these can be 5′ leader sequences or 3′ trailer sequences.



◀ Figure 4 | **Various roles for long non-coding RNAs in cellular regulation. A** | Long non-coding RNAs (lncRNAs) are expressed from many loci in the genome — sense and antisense, intronic, overlapping and intergenic with respect to nearby protein-coding loci — and function in both *cis* and *trans*. **B** | Nuclear functional lncRNAs can modulate gene expression both transcriptionally and epigenetically. Some lncRNAs interact with proteins to control the access of chromatin to cellular components and/or guide epigenetic regulatory complexes to target loci, which results in both transcriptional suppression[201] (part **Ba**) and activation or suppression (that is, bimodal control)[194] (part **Bb**). Proteins involved in chromatin modification — such as DNA (cytosine-5)-methyltransferase 3A (DNMT3A), enhancer of Zeste 2 (EZH2), euchromatic histone-lysine *N*-methyltransferase 2 (EHMT2; also known as G9a), chromodomain Y-like protein (CDYL), repressor element 1-silencing transcription factor (REST), co-repressor of REST (coREST), trithorax-activating complex MLL1 (REF. 207) (not shown) and Polycomb repressive complex 2 (PRC2) — have been associated with lncRNA-mediated epigenetic silencing[194,201,265]; the histone demethylase LSD1 (also known as KDM1A) has been associated with activation of silent loci. Enhancer functional lncRNAs tether distal enhancer elements with their promoters[344,345], presumably in concert with a protein component that has yet to be determined (shown as 'unknown') (part **Bc**). Decoy functional lncRNAs affect transcription by binding to proteins such as DNMT1 to sequester them from their sites of action, which leads to a loss of maintenance of DNA methylation and gene activation[263] (part **Bd**). **C** | Some lncRNAs can function in both nuclear and cytoplasmic compartments of the cell to affect gene expression and translation of mRNAs. Decoy functional lncRNA complexes affect microRNA (miRNA) targeting of mRNAs (part **Ca**). Some lncRNAs can interact with each other or with mRNAs to sequester small regulatory RNAs, such as miRNAs and therefore RNA-induced silencing complex (RISC), from protein-coding mRNAs[201,337,338]. Translational regulatory lncRNAs have been observed to recruit protein complexes that consist of heterogeneous nuclear ribonucloprotein K (hnRNPK), fragile X mental retardation syndrome-related protein 1 (FXR1), FXR2 and Poly(U)-binding splicing factor (PUF60) to homology-containing protein-coding mRNAs, where they bind to and sequester the mRNAs from the translational machinery[346] and regulate translation (part **Cb**). lncRNAs can also bind to homology-containing mRNAs and recruit proteins such as QKI and serine/arginine-rich splicing factor 1 (SRSF1), both of which modulate the splicing of the targeted mRNA[341] (part **Cc**). H3K9ac, histone H3 lysine 9 acetylation; me, methylation; Pol II, RNA polymerase II.

lncRNAs may also be involved in orchestrating the highly dynamic spatial structure of chromatin during differentiation and development[164,266], which would explain their often highly cell-specific expression patterns[200]. Developmental enhancers, as well as Polycomb- and trithorax-response elements, are transcribed in the cells in which they are active[203,267–272]. These elements may not only be scaffolds for the recruitment of epigenetic regulators[273] but also be the physical mediators of the complex phenomena of transvection and transinduction[234].

Moreover, many lncRNAs show the properties of enhancers[180]. These RNAs might guide the physical looping that occurs between enhancers, target promoters and exons with precise positioning of nucleosomes[274–278] to control transcription and alternative splicing[237,279,280]. Indeed, the emerging picture is of a chromatin and transcriptional landscape that is exquisitely and precisely controlled in four dimensions by a range of regulatory RNAs that assemble fairly generic (albeit often cell- or differentiation state-specific) enzyme complexes and isoforms to their sites of action in a context-dependent manner[238].

A substantial proportion of lncRNAs reside within, or are dynamically shuttled to, the cytoplasm, which indicates roles in other cellular processes (BOX 1), including the regulation of protein localization[281], mRNA translation[282] and mRNA stability[283].

RNA modification, evolution and inheritance. Regulatory RNAs may also be influenced by environmental signals and transmitted between cells and generations, which has important implications for understanding gene–environment interactions and evolution. There is evidence that plasticity has been superimposed on RNA-directed epigenetic networks by the expansion of RNA editing, especially during cognitive evolution[284,285], and by the use and mobility of retrotransposons[114,286–289], which is consistent with the insights of McClintock and of Britten and Davidson. The 'raw material' for evolution is gene duplication and transposition; the latter has the advantage of being able to mobilize functional cassettes in regulatory networks[290], which seems to be the main 'driver' of adaptive radiation[234,291]. Indeed, many lncRNAs may have originated from retrotransposons, and the evolution of mRNAs and lncRNAs may have been accelerated by retrotransposition of functional modules[292–296].

Moreover, apart from snoRNA-directed modifications, there are more than 100 other documented modifications of RNA[297,298], including cytosine and adenosine methylation that have known physiological and cognitive effects[299–302]. This indicates an additional layer of RNA informational code and epitranscriptomics — an exciting field that is just beginning to emerge[303,304].

There is evidence for systemic transmission of RNA[305,306] and RNA-mediated epigenetic inheritance in plants and animals[307–311]. There is also the intriguing possibility of RNA-directed DNA recoding, which may place RNA at the centre not only of gene regulation in the developmental ontogeny of higher organisms but also of both 'hard-wired' and 'soft-wired' somatic and germline evolution[312–314].

Conclusions and outlook

Our understanding of the previously hidden and unanticipated world of ncRNAs has greatly expanded in the past two decades. Indeed, in retrospect, it seems that we may have fundamentally misunderstood the nature of the genetic programming in complex organisms because of the assumption that most genetic information is transacted by proteins. This may be true to a large extent in simpler organisms but is turning out not to be the case in more complex organisms, the genomes of which seem to be progressively dominated by regulatory RNAs that orchestrate the epigenetic trajectories of differentiation and development.

The emerging picture is one of an extraordinarily complex transcriptional landscape in mammals and other multicellular organisms. Such a landscape is comprised of overlapping, intergenic and intronic, sense and antisense, small and large RNAs with interlaced exons[315,316], which have varying promoters, splicing patterns, polyadenylation sites and localization in different cells and developmental contexts (see below). As there seem to be few distinct boundaries to genes in humans, it might be better to change the focus of analysis to the transcript and to redefine genetic loci as 'fuzzy' transcription clusters[165,316,317] that are nonetheless semantically anchored or related to an enclosed or nearby protein-coding locus. However, this can only be stretched to a certain extent, and non-protein-coding loci raise problems for existing schema of human genome nomenclature.

Indeed, even the notion of a simple protein-coding sequence needs to be reassessed. It is becoming evident not only that mRNAs can have multiple functions[205] but also that

protein-coding sequences themselves can have other embedded functions, as suggested by constraints on synonymous codon usage[318,319], including regulatory functions as epigenetic modulators[203], tissue-specific enhancers[319,320] and transcription factor binding sites[321]. The possibility, if not likelihood, is that there is a very complex functional and evolutionary interplay between the protein-coding and regulatory functions of RNAs[200], and that some lncRNAs may have evolved, at least partly, from protein-coding genes — as in the case of X inactive specific transcript (*XIST*) — by duplication or pseudogenization and the subsequent emergence of paralogous regulatory and/or coding functions[201,322]. Conversely, new protein-coding capacity may also appear in lncRNAs[174].

The sheer number and diversity of RNAs juxtaposed with their extraordinarily complex molecular functions (FIG. 4) — for example, in regulating epigenetic processes, subcellular organelles, protein-coding and non-coding gene transcription, translation, RNA turnover, chromosomal organization and integrity, and genome defence — suggests that we have a long way to go to understand the structure and functions of what is surely a highly interconnected system. Tens of thousands (if not more) of individual non-coding RNAs exist, and their roles in cell and developmental biology, as well as in brain function, remain to be determined. Moreover, many (if not most) regulatory RNAs have yet to be identified, especially in complex organisms. These include new classes such as the circular RNAs and others that may function as miRNA 'sponges' (REFS 62,323–328), the identification of which will require targeted deep sequencing of small and large RNAs that are derived from different genomic locations in various cell types, using targeted techniques such as RNA CaptureSeq[164,171].

RNA is not a linear molecule but can fold into complex and allosterically responsive three-dimensional structures that can both recruit generic effector proteins and guide the resulting complexes in a sequence-specific manner to other RNAs and DNA through duplex or triplex formation. Important issues that remain include the identification of functional domains in RNA and their interacting partners, so that we can predict and explain RNA functional interactions in the same way that has already been done by recognition of well-characterized motifs and domains in proteins. One way to do this, which is already underway in many laboratories, is to combine immunoprecipitation of different types of RNA-binding proteins (for example, chromatin-modifying proteins, transcription factors and RNA transport proteins) with deep sequencing of the associated RNAs, followed by analysis of primary and predicted secondary structures, and ultimately by biochemical validation and characterization.

Determination of the structure of RNA species, RNA–RNA, RNA–DNA and RNP complexes will be a rapidly growing field that requires the development of new technologies, such as RNA footprinting using high-throughput sequencing[329] and *in vivo* studies using RNA-based genetic techniques, for example, CRISPR-mediated mutation[143]. Other objectives include determination of whether small RNA pathways are used in viral defence in humans[80]; the functions of tiRNAs, spliRNAs and snoRNA-derived small RNAs; the roles of piRNAs in retrotransposon dynamics and genome remodelling by retrotransposons in the brain[114]; the mechanisms and extent of RNA-mediated transgenerational epigenetic inheritance[330]; the locations of RNA-binding sites (that is, RNA–DNA duplexes and RNA–DNA:DNA triplexes) in the genome; the crosstalk between different types of regulatory RNAs; the logic and hierarchy of RNA- and protein-mediated regulation of gene expression; and finally, the extent, mechanisms and information content of RNA-mediated communication between cells both within[306] and between organisms (that is, 'social RNA')[331].

Indeed, it seems that RNA is the computational engine of cell biology, developmental biology, brain function and perhaps even evolution itself[313]. The complexity and interconnectedness of these systems should not be cause for concern but rather the motivation for exploring the vast unknown universe of RNA regulation, without which we will not understand biology.

Kevin V. Morris is at the School of Biotechnology and Biomedical Sciences, University of New South Wales, Sydney, NSW 2052, Australia; and Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California 92037, USA.

John S. Mattick is at the Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010, Australia; the School of Biotechnology and Biomedical Sciences, and St. Vincent's Clinical School, University of New South Wales, Sydney, NSW 2052, Australia.

Correspondence to K.V.M.
e-mail: kmorris@scripps.edu

doi:10.1038/nrg3722
Published online 29 April 2014

Box 1 | **Examples of specific long non-coding RNAs and their functions**

Long non-coding RNAs have a role in a wide range of biological processes in the cell, for example:

- Template RNAs guide chromosomal rearrangements in ciliates[332]
- Telomeric repeat-containing RNA (TERRA) is involved in telomere biology[333]
- 7S RNA is an essential component of the signal recognition particle, which is involved in protein export[334]
- 7SK is a highly expressed structured RNA that acts as a scaffold to assemble a multimeric protein complex containing SR splicing proteins and positive transcription elongation factor b (P-TEFb, which is a cyclin-dependent kinase required for transcriptional elongation by RNA polymerase II and other factors)[335]
- Nuclear paraspeckle assembly transcript 1 (*NEAT1*) is an essential component of paraspeckles, which are enigmatic subnuclear organelles that appear in mammalian differentiated cells but not stem cells[336,337]
- Metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*) is localized to the nucleus and regulates alternative splicing[338] and cell cycle progression[339]
- Myocardial infarction-associated transcript (*MIAT*; also known as gomafu) is expressed in an unknown subnuclear structure, possibly a specialized spliceosome, in a subset of neurons[340] and has recently been implicated in schizophrenia[341]

1. Gilbert, W. Origin of life: the RNA world. *Nature* **319**, 618 (1986).
2. Beadle, G. W. & Tatum, E. L. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl Acad. Sci. USA* **27**, 499–506 (1941).
3. Comfort, N. C. *The Tangled Field: Barbara McClintock's Search for the Patterns of Genetic Control* (Harvard Univ. Press, 2003).
4. Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet.* **5**, e1000459 (2009).
5. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
6. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
7. Palade, G. E. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.* **1**, 59–68 (1955).
8. Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**, 241–257 (1958).
9. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576–581 (1961).
10. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).

11. Gilbert, W. & Muller-Hill, B. Isolation of the *lac* repressor. *Proc. Natl Acad. Sci. USA* **56**, 1891–1898 (1966).
12. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
13. Mattick, J. S. & Gagen, M. J. Accelerating networks. *Science* **307**, 856–858 (2005).
14. Freedman, M. L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genet.* **43**, 513–518 (2011).
15. Weinberg, R. A. & Penman, S. Small molecular weight monodisperse nuclear RNA. *J. Mol. Biol.* **38**, 289–304 (1968).
16. Dreyfuss, G., Philipson, L. & Mattaj, I. W. Ribonucleoprotein particles in cellular processes. *J. Cell Biol.* **106**, 1419–1425 (1988).
17. Butcher, S. E. & Brow, D. A. Towards understanding the catalytic core structure of the spliceosome. *Biochem. Soc. Trans.* **33**, 447–449 (2005).
18. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).
19. Pessa, H. K. *et al.* Minor spliceosome components are predominantly localized in the nucleus. *Proc. Natl Acad. Sci. USA* **105**, 8655–8660 (2008).
20. Maxwell, E. S. & Fournier, M. J. The small nucleolar RNAs. *Annu. Rev. Biochem.* **64**, 897–934 (1995).
21. Henras, A. K., Dez, C. & Henry, Y. RNA structure and function in C/D and H/ACA s(no)RNPs. *Curr. Opin. Struct. Biol.* **14**, 335–343 (2004).
22. Meier, U. T. The many facets of H/ACA ribonucleoproteins. *Chromosoma* **114**, 1–14 (2005).
23. Cavaille, J., Seitz, H., Paulsen, M., Ferguson-Smith, A. C. & Bachellerie, J. P. Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. *Hum. Mol. Genet.* **11**, 1527–1538 (2002).
24. Rogelj, B., Hartmann, C. E., Yeo, C. H., Hunt, S. P. & Giese, K. P. Contextual fear conditioning regulates the expression of brain-specific small nucleolar RNAs in hippocampus. *Eur. J. Neurosci.* **18**, 3089–3096 (2003).
25. Bachellerie, J. P., Cavaille, J. & Huttenhofer, A. The expanding snoRNA world. *Biochimie* **84**, 775–790 (2002).
26. Jady, B. E., Bertrand, E. & Kiss, T. Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body-specific localization signal. *J. Cell Biol.* **164**, 647–652 (2004).
27. Warner, J. R., Soeiro, R., Birnboim, H. C., Girard, M. & Darnell, J. E. Rapidly labeled HeLa cell nuclear RNA. I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA. *J. Mol. Biol.* **19**, 349–361 (1966).
28. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
29. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **66**, 111–138 (1971).
30. Davidson, E. H., Klein, W. H. & Britten, R. J. Sequence organization in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript. *Dev. Biol.* **55**, 69–84 (1977).
31. Howard, M. L. & Davidson, E. H. *Cis*-regulatory control circuits in development. *Dev. Biol.* **271**, 109–118 (2004).
32. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic Press, 2006).
33. Britten, R. Transposable elements have contributed to thousands of human proteins. *Proc. Natl Acad. Sci. USA* **103**, 1798–1803 (2006).
34. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl Acad. Sci. USA* **74**, 3171–3175 (1977).
35. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* **12**, 1–8 (1977).
36. Williamson, B. DNA insertions and gene structure. *Nature* **270**, 295–297 (1977).
37. Gilbert, W., Marchionni, M. & McKnight, G. On the antiquity of introns. *Cell* **46**, 151–154 (1986).
38. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).
39. Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).
40. Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **31**, 147–157 (1982).
41. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).
42. Fica, S. M. *et al.* RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**, 229–234 (2013).
43. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28**, 411–418 (2003).
44. Webb, C. H., Riccitelli, N. J., Ruminski, D. J. & Luptak, A. Widespread occurrence of self-cleaving ribozymes. *Science* **326**, 953 (2009).
45. de la Pena, M. & Garcia-Robles, I. Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep.* **11**, 711–716 (2010).
46. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
47. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
48. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).
49. Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. & Cohen, S. M. *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell* **113**, 25–36 (2003).
50. Johnston, R. J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845–849 (2003).
51. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
52. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).
53. Lee, R. C. & Ambros, V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294**, 862–864 (2001).
54. Williams, T. M. *et al.* The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *Drosophila*. *Cell* **134**, 610–623 (2008).
55. Kaya, K. D., Karakulah, G., Yakicier, C. M. & Acar, A. C. & Konu, O. mESAdb: microRNA expression and sequence analysis database. *Nucleic Acids Res.* **39**, D170–D180 (2011).
56. Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nature Genet.* **38**, 1375–1377 (2006).
57. Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. & Peterson, K. J. MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl Acad. Sci. USA* **105**, 2946–2950 (2008).
58. John, B. *et al.* Human microRNA targets. *PLoS Biol.* **2**, e363 (2004).
59. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
60. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79**, 351–379 (2010).
61. Schnall-Levin, M. *et al.* Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. *Genome Res.* **21**, 1395–1403 (2011).
62. Hansen, T. B. *et al.* miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J.* **30**, 4414–4422 (2011).
63. Leonardo, T. R., Schultheisz, H. L., Loring, J. F. & Laurent, L. C. The functions of microRNAs in pluripotency and reprogramming. *Nature Cell Biol.* **14**, 1114–1121 (2012).
64. Bracken, C. P., Gregory, P. A., Khew-Goodall, Y. & Goodall, G. J. The role of microRNAs in metastasis and epithelial–mesenchymal transition. *Cell. Mol. Life Sci.* **66**, 1682–1699 (2009).
65. Rakoczy, J. *et al. MicroRNAs-140-5p/140-3p* modulate Leydig cell numbers in the developing mouse testis. *Biol. Reprod.* **88**, 143 (2013).
66. Fernandez-Valverde, S. L., Taft, R. J. & Mattick, J. S. MicroRNAs in β-cell biology, insulin resistance, diabetes and its complications. *Diabetes* **60**, 1825–1831 (2011).
67. Bredy, T. W., Lin, Q., Wei, W., Baker-Andresen, D. & Mattick, J. S. MicroRNA regulation of neural plasticity and memory. *Neurobiol. Learn. Mem.* **96**, 89–94 (2011).
68. Park, C. Y., Choi, Y. S. & McManus, M. T. Analysis of microRNA knockouts in mice. *Hum. Mol. Genet.* **19**, R169–R175 (2010).
69. Waterhouse, P. M., Graham, M. W. & Wang, M. B. Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proc. Natl Acad. Sci. USA* **95**, 13959–13964 (1998).
70. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
71. Napoli, C., Lemieux, C. & Jorgensen, R. Introduction of a chimeric chalcone synthase gene into Petunia results in reversible co-suppression of homologous genes in *trans*. *Plant Cell* **2**, 279–289 (1990).
72. Matzke, M. A., Primig, M., Trnovsky, J. & Matzke, A. J. M. Reversible methylation and inactivation of marker genes in sequentially transformed tobacco plants. *EMBO J.* **8**, 643–649 (1989).
73. van der Krol, A. R., Mur, L. A., de Lange, P., Mol, J. N. & Stuitje, A. R. Inhibition of flower pigmentation by antisense *CHS* genes: promoter and minimal sequence requirements for the antisense effect. *Plant Mol. Biol.* **14**, 457–466 (1990).
74. Wassenegger, M., Heimes, S., Riedel, L. & Sanger, H. L. RNA-directed *de novo* methylation of genomic sequences in plants. *Cell* **76**, 567–576 (1994).
75. Basyuk, E., Suavet, F., Doglio, A., Bordonne, R. & Bertrand, E. Human *let-7* stem–loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res.* **31**, 6593–6597 (2003).
76. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
77. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
78. Doi, N. *et al.* Short-interfering-RNA-mediated gene silencing in mammalian cells requires Dicer and eIF2C translation initiation factors. *Curr. Biol.* **13**, 41–46 (2003).
79. Peters, L. & Meister, G. Argonaute proteins: mediators of RNA silencing. *Mol. Cell* **26**, 611–623 (2007).
80. Maillard, P. V. *et al.* Antiviral RNA interference in mammalian cells. *Science* **342**, 235–238 (2013).
81. Zeng, Y., Yi, R. & Cullen, B. R. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc. Natl Acad. Sci. USA* **100**, 9779–9784 (2003).
82. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).
83. Ramachandran, P. V. & Ignacimuthu, S. RNA interference — a silent but an efficient therapeutic tool. *Appl. Biochem. Biotechnol.* **169**, 1774–1789 (2013).
84. Ahlenstiel, C. L. *et al.* Direct evidence of nuclear Argonaute distribution during transcriptional silencing links the actin cytoskeleton to nuclear RNAi machinery in human cells. *Nucleic Acids Res.* **40**, 1579–1595 (2012).
85. Ameyar-Zazoua, M. *et al.* Argonaute proteins couple chromatin silencing to alternative splicing. *Nature Struct. Mol. Biol.* **19**, 998–1004 (2012).
86. Rudel, S., Flatley, A., Weinmann, L., Kremmer, E. & Meister, G. A multifunctional human Argonaute2-specific monoclonal antibody. *RNA* **14**, 1244–1253 (2008).
87. Kim, D. H., Villeneuve, L. M., Morris, K. V. & Rossi, J. J. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature Struct. Mol. Biol.* **13**, 793–797 (2006).
88. Morris, K. V., Chan, S. W., Jacobsen, S. E. & Looney, D. J. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305**, 1289–1292 (2004).
89. Fernandez-Valverde, S. L., Taft, R. J. & Mattick, J. S. Dynamic isomiR regulation in *Drosophila* development. *RNA* **16**, 1881–1888 (2010).
90. Didiano, D. & Hobert, O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Struct. Mol. Biol.* **13**, 849–851 (2006).
91. Shin, C. *et al.* Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell* **38**, 789–802 (2010).

92. Kim, D. H., Saetrom, P., Snove, O. Jr & Rossi, J. J. MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proc. Natl Acad. Sci. USA* **105**, 16230–16235 (2008).
93. Blow, M. J. *et al.* RNA editing of human microRNAs. *Genome Biol.* **7**, R27 (2006).
94. Hundley, H. A. & Bass, B. L. ADAR editing in double-stranded UTRs and other noncoding RNA sequences. *Trends Biochem. Sci.* **35**, 377–383 (2010).
95. Kawahara, Y., Zinshteyn, B., Chendrimada, T. P., Shiekhattar, R. & Nishikura, K. RNA editing of the *microRNA-151* precursor blocks cleavage by the Dicer–TRBP complex. *EMBO Rep.* **8**, 763–769 (2007).
96. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genet.* **36**, 40–45 (2004).
97. Lin, H. & Spradling, A. C. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* **124**, 2463–2476 (1997).
98. Cox, D. N. *et al.* A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev.* **12**, 3715–3727 (1998).
99. Kuramochi-Miyagawa, S. *et al. Mili*, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development* **131**, 839–849 (2004).
100. Kim, J. K. *et al.* Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308**, 1164–1167 (2005).
101. Pal-Bhadra, M., Bhadra, U. & Birchler, J. A. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Mol. Cell* **9**, 315–327 (2002).
102. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320–324 (2006).
103. Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203–207 (2006).
104. Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199–202 (2006).
105. Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* **20**, 1709–1714 (2006).
106. Watanabe, T. *et al.* Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.* **20**, 1732–1743 (2006).
107. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761–764 (2007).
108. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
109. Brennecke, J. *et al.* An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**, 1387–1392 (2008).
110. Kuramochi-Miyagawa, S. *et al.* DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev.* **22**, 908–917 (2008).
111. Cox, D. N., Chao, A. & Lin, H. *piwi* encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127**, 503–514 (2000).
112. Grimaud, C. *et al.* RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* **124**, 957–971 (2006).
113. Rajasethupathy, P. *et al.* A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* **149**, 693–707 (2012).
114. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).
115. Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature* **505**, 353–359 (2014).
116. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
117. Taft, R. J. *et al.* Small RNAs derived from snoRNAs. *RNA* **15**, 1233–1240 (2009).
118. Ender, C. *et al.* A human snoRNA with microRNA-like functions. *Mol. Cell* **32**, 519–528 (2008).
119. Kawaji, H. *et al.* Hidden layers of human small RNAs. *BMC Genomics* **9**, 157 (2008).
120. Haussecker, D. *et al.* Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA* **16**, 673–695 (2010).
121. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41**, 572–578 (2009).
122. Taft, R. J. *et al.* Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nature Struct. Mol. Biol.* **17**, 1030–1034 (2010).
123. Taft, R. J., Kaplan, C. D., Simons, C. & Mattick, J. S. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* **8**, 2332–2338 (2009).
124. Taft, R. J., Hawkins, P. G., Mattick, J. S. & Morris, K. V. The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization. *Epigenetics Chromatin* **4**, 13 (2011).
125. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
126. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
127. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–1854 (2008).
128. Han, J., Kim, D. & Morris, K. V. Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. *Proc. Natl Acad. Sci. USA* **104**, 12422–12427 (2007).
129. Wassarman, K. M., Zhang, A. & Storz, G. Small RNAs in *Escherichia coli*. *Trends Microbiol.* **7**, 37–45 (1999).
130. Gottesman, S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.* **21**, 399–404 (2005).
131. Tucker, B. J. & Breaker, R. R. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* **15**, 342–348 (2005).
132. Winkler, W. C. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr. Opin. Chem. Biol.* **9**, 594–602 (2005).
133. Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of archaea, bacteria and mitochondria. *Mol. Microbiol.* **36**, 244–246 (2000).
134. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
135. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
136. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
137. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
138. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
139. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 (2008).
140. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
141. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA–Cas protein complex. *Cell* **139**, 945–956 (2009).
142. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
143. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
144. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
145. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
146. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
147. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
148. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR–Cas9-based transcription factors. *Nature Methods* **10**, 973–976 (2013).
149. Cheng, A. W. *et al.* Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res.* **23**, 1163–1171 (2013).
150. Hu, J. *et al.* Direct activation of human and mouse *Oct4* genes using engineered TALE and Cas9 transcription factors. *Nucleic Acids Res.* http://dx.doi.org/10.1093/nar/gku109 (2014).
151. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–491 (2013).
152. Mattick, J. S. Introns: evolution and function. *Curr. Opin. Genet. Dev.* **4**, 823–831 (1994).
153. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
154. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
155. Rinn, J. L. *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
156. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
157. Kapranov, P. *et al.* Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15**, 987–997 (2005).
158. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
159. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
160. 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).
161. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
162. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* **40**, D912–D917 (2012).
163. Taft, R. J., Pheasant, M. & Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**, 288–299 (2007).
164. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature Biotech.* **30**, 99–104 (2012).
165. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
166. Hawkins, P. G. & Morris, K. V. Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. *Transcription* **1**, 165–175 (2010).
167. Sheik Mohamed, J., Gaughwin, P. M., Lim, B., Robson, P. & Lipovich, L. Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* **16**, 324–337 (2010).
168. Venters, B. J. & Pugh, B. F. Genomic organization of human transcription initiation complexes. *Nature* **502**, 53–58 (2013).
169. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38**, 626–635 (2006).
170. Huang, R. *et al.* An RNA-seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS ONE* **6**, e27288 (2011).
171. Roberts, A. & Pachter, L. RNA-seq and find: entering the RNA deep field. *Genome Med.* **3**, 74 (2011).
172. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* **10**, 155–159 (2009).
173. Frith, M. C. *et al.* The abundance of short proteins in the mammalian proteome. *PLoS Genet.* **2**, e52 (2006).
174. Gascoigne, D. K. *et al.* Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28**, 3042–3050 (2012).
175. Banfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657 (2012).

176. Dinger, M. E. *et al.* NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.* **37**, D122–D126 (2009).
177. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
178. Wahlestedt, C. Natural antisense and noncoding RNA transcripts as potential drug targets. *Drug Discov. Today* **11**, 503–508 (2006).
179. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106**, 11667–11672 (2009).
180. Ørom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).
181. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**, e01749 (2013).
182. Furuno, M. *et al.* Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet.* **2**, e37 (2006).
183. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
184. Clark, M. B. *et al.* The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625 (2011).
185. Calin, G. A. *et al.* Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**, 215–229 (2007).
186. Tay, S. K., Blythe, J. & Lipovich, L. Global discovery of primate-specific genes in the human genome. *Proc. Natl Acad. Sci. USA* **106**, 12019–12024 (2009).
187. Lipovich, L. *et al.* Developmental changes in the transcriptome of human cerebral cortex tissue: long noncoding RNA transcripts. *Cereb. Cortex* http://dx.doi.org/10.1093/cercor/bhs414 (2013).
188. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
189. Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. *Genome Res.* **17**, 1245–1253 (2007).
190. Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1–5 (2006).
191. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).
192. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41**, 8220–8236 (2013).
193. Johnsson, P., Lipovich, L., Grander, D. & Morris, K. V. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta* **1840**, 1063–1071 (2014).
194. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
195. 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).
196. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22**, 885–898 (2012).
197. Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**, 11–19 (2006).
198. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci. USA* **105**, 716–721 (2008).
199. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
200. Mattick, J. S., Taft, R. J. & Faulkner, G. J. A global view of genomic information — moving beyond the gene and the master regulator. *Trends Genet.* **26**, 21–28 (2010).
201. Johnsson, P. *et al.* A pseudogene long-noncoding-RNA network regulates *PTEN* transcription and translation in human cells. *Nature Struct. Mol. Biol.* **20**, 440–446 (2013).
202. Chooniedass-Kothari, S. *et al.* The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett.* **566**, 43–47 (2004).
203. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
204. Mercer, T. R. *et al.* Expression of distinct RNAs from 3′ untranslated regions. *Nucleic Acids Res.* **39**, 2393–2403 (2011).
205. Dinger, M. E., Gascoigne, D. K. & Mattick, J. S. The evolution of RNAs with multiple functions. *Biochimie* **93**, 2013–2018 (2011).
206. Mercer, T. R. *et al.* Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.* **20**, 1639–1650 (2010).
207. Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18**, 1433–1445 (2008).
208. Sunwoo, H. *et al.* MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* **19**, 347–359 (2009).
209. Pang, K. C. *et al.* Genome-wide identification of long noncoding RNAs in $CD8^+$ T cells. *J. Immunol.* **182**, 7738–7748 (2009).
210. Askarian-Amiri, M. E. *et al.* SNORD-host RNA *Zfas1* is a regulator of mammary development and a potential marker for breast cancer. *RNA* **17**, 878–891 (2011).
211. Hu, W., Yuan, B., Flygare, J. & Lodish, H. F. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev.* **25**, 2573–2578 (2011).
212. Mercer, T. R. *et al.* Long noncoding RNAs in neuronal–glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci.* **11**, 14 (2010).
213. Johnson, R. *et al.* Regulation of neural macroRNAs by the transcriptional repressor REST. *RNA* **15**, 85–96 (2009).
214. Ng, S.-Y., Johnson, R. & Stanton, L. W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* **31**, 522–533 (2012).
215. Takeda, K. *et al.* Identification of a novel bone morphogenetic protein-responsive gene that may function as a noncoding RNA. *J. Biol. Chem.* **273**, 17079–17085 (1998).
216. Bussemakers, M. J. *et al. DD3*: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* **59**, 5975–5979 (1999).
217. Pasmant, E. *et al.* Characterization of a germ-line deletion, including the entire *INK4/ARF* locus, in a melanoma-neural system tumor family: identification of *ANRIL*, an antisense noncoding RNA whose expression coclusters with *ARF*. *Cancer Res.* **67**, 3963–3969 (2007).
218. Wang, F., Li, X., Xie, X., Zhao, L. & Chen, W. *UCA1*, a non-protein-coding RNA up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. *FEBS Lett.* **582**, 1919–1927 (2008).
219. Mourtada-Maarabouni, M., Pickard, M. R., Hedge, V. L., Farzaneh, F. & Williams, G. T. *GAS5*, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* **28**, 195–208 (2009).
220. Gupta, R. A. *et al.* Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
221. Khaitan, D. *et al.* The melanoma-upregulated long noncoding RNA *SPRY4-IT1* modulates apoptosis and invasion. *Cancer Res.* **71**, 3852–3862 (2011).
222. Kerin, T. *et al.* A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci. Transl Med.* **4**, 128ra40 (2012).
223. Amaral, P. P. & Mattick, J. S. Noncoding RNA in development. *Mamm. Genome* **19**, 454–492 (2008).
224. Mourtada-Maarabouni, M., Hedge, V. L., Kirkham, L., Farzaneh, F. & Williams, G. T. Growth arrest in human T-cells is controlled by the non-coding RNA growth-arrest-specific transcript 5 (*GAS5*). *J. Cell Sci.* **121**, 939–946 (2008).
225. Young, T. L., Matsuda, T. & Cepko, C. L. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr. Biol.* **15**, 501–512 (2005).
226. Ginger, M. R. *et al.* A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc. Natl Acad. Sci. USA* **103**, 5781–5786 (2006).
227. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA *TINCR*. *Nature* **493**, 231–235 (2013).
228. Gutschner, T. *et al.* The noncoding RNA *MALAT1* is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* **73**, 1180–1189 (2013).
229. Li, L. *et al.* Targeted disruption of *Hotair* leads to homeotic transformation and gene derepression. *Cell Rep.* **5**, 3–12 (2013).
230. Nakagawa, S., Naganuma, T., Shioi, G. & Hirose, T. Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J. Cell Biol.* **193**, 31–39 (2011).
231. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**, e234 (2007).
232. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
233. Lewejohann, L. *et al.* Role of a neuronal small non-messenger RNA: behavioural alterations in *BC1* RNA-deleted mice. *Behav. Brain Res.* **154**, 273–289 (2004).
234. Mattick, J. S. & Gagen, M. J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **18**, 1611–1630 (2001).
235. Mattick, J. S., Amaral, P. P., Dinger, M. E., Mercer, T. R. & Mehler, M. F. RNA regulation of epigenetic processes. *Bioessays* **31**, 51–59 (2009).
236. Koziol, M. J. & Rinn, J. L. RNA traffic control of chromatin complexes. *Curr. Opin. Genet. Dev.* **20**, 142–148 (2010).
237. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Struct. Mol. Biol.* **20**, 300–307 (2013).
238. Mercer, T. R. & Mattick, J. S. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Res.* **23**, 1081–1088 (2013).
239. Wassenegger, M. RNA-directed DNA methylation. *Plant Mol. Biol.* **43**, 203–220 (2000).
240. Hall, I. M. *et al.* Establishment and maintenance of a heterochromatin domain. *Science* **297**, 2232–2237 (2002).
241. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
242. Verdel, A. *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303**, 672–676 (2004).
243. Kanhere, A. *et al.* Short RNAs are transcribed from repressed Polycomb target genes and interact with Polycomb repressive complex-2. *Mol. Cell* **38**, 675–688 (2010).
244. Imamura, T. *et al.* Non-coding RNA directed DNA demethylation of *Sphk1* CpG island. *Biochem. Biophys. Res. Commun.* **322**, 593–600 (2004).
245. Morris, K. V., Santoso, S., Turner, A. M., Pastori, C. & Hawkins, P. G. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.* **4**, e1000258 (2008).
246. Yu, W. *et al.* Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**, 202–206 (2008).
247. Allo, M. *et al.* Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nature Struct. Mol. Biol.* **16**, 717–724 (2009).
248. Beltran, M. *et al.* A natural antisense transcript regulates *Zeb2/Sip1* gene expression during *Snail1*-induced epithelial–mesenchymal transition. *Genes Dev.* **22**, 756–769 (2008).
249. Morrissy, A. S., Griffith, M. & Marra, M. A. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res.* **21**, 1203–1212 (2011).
250. Brockdorff, N. *et al.* The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515–526 (1992).
251. Brown, C. J. *et al.* The human *XIST* gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992).
252. Meller, V. H., Wu, K. H., Roman, G., Kuroda, M. I. & Davis, R. L. *roX1* RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system. *Cell* **88**, 445–457 (1997).

253. Lee, J. T., Davidow, L. S. & Warshawsky, D. *Tsix*, a gene antisense to *Xist* at the X-inactivation centre. *Nature Genet.* **21**, 400–404 (1999).
254. Sado, T., Wang, Z., Sasaki, H. & Li, E. Regulation of imprinted X-chromosome inactivation in mice by *Tsix*. *Development* **128**, 1275–1286 (2001).
255. Ripoche, M. A., Kress, C., Poirier, F. & Dandolo, L. Deletion of the *H19* transcription unit reveals the existence of a putative imprinting control element. *Genes Dev.* **11**, 1596–1604 (1997).
256. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002).
257. Thakur, N. *et al.* An antisense RNA regulates the bidirectional silencing property of the *Kcnq1* imprinting control region. *Mol. Cell. Biol.* **24**, 7855–7862 (2004).
258. Swiezewski, S., Liu, F., Magusin, A. & Dean, C. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. *Nature* **462**, 799–802 (2009).
259. Nagano, T. *et al.* The *Air* noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717–1720 (2008).
260. Mohammad, F. *et al. Kcnq1ot1/Lit1* noncoding RNA mediates transcriptional silencing by targeting to the perinucleolar region. *Mol. Cell. Biol.* **28**, 3713–3728 (2008).
261. Kotake, Y. *et al.* Long non-coding RNA *ANRIL* is required for the PRC2 recruitment to and silencing of *p15*(*INK4B*) tumor suppressor gene. *Oncogene* **30**, 1956–1962 (2011).
262. Mohammad, F., Mondal, T., Guseva, N., Pandey, G. K. & Kanduri, C. *Kcnq1ot1* noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development* **137**, 2493–2499 (2010).
263. Di Ruscio, A. *et al.* DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503**, 371–376 (2013).
264. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nature Struct. Mol. Biol.* **20**, 1250–1257 (2013).
265. Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
266. Zhang, H. *et al.* Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the *Kcnq1* locus. *J. Cell Biol.* **204**, 61–75 (2014).
267. Sanchez-Herrero, E. & Akam, M. Spatially ordered transcription of regulatory DNA in the bithorax complex of *Drosophila*. *Development* **107**, 321–329 (1989).
268. Bae, E., Calhoun, V. C., Levine, M., Lewis, E. B. & Drewell, R. A. Characterization of the intergenic RNA profile at *abdominal-A* and *Abdominal-B* in the *Drosophila* bithorax complex. *Proc. Natl Acad. Sci. USA* **99**, 16847–16852 (2002).
269. Jones, E. A. & Flavell, R. A. Distal enhancer elements transcribe intergenic RNA in the IL-10 family gene cluster. *J. Immunol.* **175**, 7437–7446 (2005).
270. Petruk, S. *et al.* Transcription of *bxd* noncoding RNAs promoted by trithorax represses *Ubx* in *cis* by transcriptional interference. *Cell* **127**, 1209–1221 (2006).
271. Feng, J. *et al.* The *Evf-2* noncoding RNA is transcribed from the *Dlx-5/6* ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20**, 1470–1484 (2006).
272. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
273. Sanchez-Elsner, T., Gou, D., Kremmer, E. & Sauer, F. Noncoding RNAs of trithorax response elements recruit *Drosophila* Ash1 to Ultrabithorax. *Science* **311**, 1118–1123 (2006).
274. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19**, 1732–1741 (2009).
275. Nahkuri, S., Taft, R. J. & Mattick, J. S. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* **8**, 3420–3424 (2009).
276. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon–intron structure. *Nature Struct. Mol. Biol.* **16**, 990–995 (2009).
277. Spies, N., Nielsen, C. B., Padgett, R. A. & Burge, C. B. Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* **36**, 245–254 (2009).
278. Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nature Struct. Mol. Biol.* **16**, 996–1001 (2009).
279. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).
280. Mercer, T. R. *et al.* DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genet.* **45**, 852–859 (2013).
281. Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570–1573 (2005).
282. Carrieri, C. *et al.* Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature* **491**, 454–457 (2012).
283. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470**, 284–288 (2011).
284. Mattick, J. S. RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays* **32**, 548–552 (2010).
285. Mattick, J. S. The central role of RNA in human development and cognition. *FEBS Lett.* **585**, 1600–1616 (2011).
286. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
287. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).
288. Muotri, A. R. *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443–446 (2010).
289. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nature Genet.* **41**, 563–571 (2009).
290. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Rev. Genet.* **9**, 397–405 (2008).
291. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
292. Brosius, J. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118**, 99–116 (2003).
293. Krull, M., Brosius, J. & Schmitz, J. Alu–SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* **22**, 1702–1711 (2005).
294. Cordaux, R., Udit, S., Batzer, M. A. & Feschotte, C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl Acad. Sci. USA* **103**, 8101–8106 (2006).
295. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).
296. Kapusta, A. *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
297. Czerwoniec, A. *et al.* MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res.* **37**, D118–D121 (2009).
298. Cantara, W. A. *et al.* The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–D201 (2011).
299. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res.* **38**, 1415–1430 (2010).
300. Abbasi-Moheb, L. *et al.* Mutations in *NSUN2* cause autosomal-recessive intellectual disability. *Am. J. Hum. Genet.* **90**, 847–855 (2012).
301. Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* **149**, 1635–1646 (2012).
302. Jia, G., Fu, Y. & He, C. Reversible RNA adenosine methylation in biological regulation. *Trends Genet.* **29**, 108–115 (2013).
303. Saletore, Y. *et al.* The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* **13**, 175 (2012).
304. Saletore, Y., Chen-Kiang, S. & Mason, C. E. Novel RNA regulatory mechanisms revealed in the epitranscriptome. *RNA Biol.* **10**, 342–346 (2013).
305. Brosnan, C. A. *et al.* Nuclear gene silencing directs reception of long-distance mRNA silencing in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **104**, 14741–14746 (2007).
306. Dinger, M. E., Mercer, T. R. & Mattick, J. S. RNAs as extracellular signaling molecules. *J. Mol. Endocrinol.* **40**, 151–159 (2008).
307. Alleman, M. *et al.* An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature* **442**, 295–298 (2006).
308. Rassoulzadegan, M. *et al.* RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* **441**, 469–474 (2006).
309. Chandler, V. L. Paramutation: from maize to mice. *Cell* **128**, 641–645 (2007).
310. Nadeau, J. H. Transgenerational genetic effects on phenotypic variation and disease risk. *Hum. Mol. Genet.* **18**, R202–R210 (2009).
311. Buckley, B. A. *et al.* A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature* **489**, 447–451 (2012).
312. Herbert, A. & Rich, A. RNA processing in evolution: the logic of soft-wired genomes. *Ann. NY Acad. Sci.* **870**, 119–132 (1999).
313. Herbert, A. & Rich, A. RNA processing and the evolution of eukaryotes. *Nature Genet.* **21**, 265–269 (1999).
314. Mattick, J. S. Has evolution learnt how to learn? *EMBO Rep.* **10**, 665 (2009).
315. Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **15**, R17–29 (2006).
316. Gingeras, T. R. Origin of phenotypes: genes and transcripts. *Genome Res.* **17**, 682–690 (2007).
317. Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**, 930–939 (2003).
318. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Rev. Genet.* **7**, 98–108 (2006).
319. Lin, M. F. *et al.* Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res.* **21**, 1916–1928 (2011).
320. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* **22**, 1059–1068 (2012).
321. Stergachis, A. B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367–1372 (2013).
322. Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653–1655 (2006).
323. Capel, B. *et al.* Circular transcripts of the testis-determining gene *Sry* in adult mouse testis. *Cell* **73**, 1019–1030 (1993).
324. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).
325. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
326. Zhang, Y. *et al.* Circular intronic long noncoding RNAs. *Mol. Cell* **51**, 792–806 (2013).
327. Wang, J. *et al.* CREB up-regulates long non-coding RNA, *HULC* expression through interaction with *microRNA-372* in liver cancer. *Nucleic Acids Res.* **38**, 5366–5383 (2010).
328. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
329. Wan, Y. *et al.* Genome-wide measurement of RNA folding energies. *Mol. Cell* **48**, 169–181 (2012).
330. Arteaga-Vazquez, M. A. & Chandler, V. L. Paramutation in maize: RNA mediated *trans*-generational gene silencing. *Curr. Opin. Genet. Dev.* **20**, 156–163 (2010).
331. Sarkies, P. & Miska, E. A. Is there social RNA? *Science* **341**, 467–468 (2013).
332. Nowacki, M. *et al.* RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* **451**, 153–158 (2008).
333. Lopez de Silanes, I., Stagno d'Alcontres, M. & Blasco, M. A. *TERRA* transcripts are bound by a complex array of RNA-binding proteins. *Nature Commun.* **1**, 33 (2010).
334. Walter, P. & Blobel, G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**, 691–698 (1982).
335. Ji, X. *et al.* SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell* **153**, 855–868 (2013).
336. Fox, A. H., Bond, C. S. & Lamond, A. I. P54nrb forms a heterodimer with PSP1 that localizes to paraspeckles in an RNA-dependent manner. *Mol. Biol. Cell* **16**, 5304–5315 (2005).

337. Mao, Y. S., Sunwoo, H., Zhang, B. & Spector, D. L. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nature Cell Biol.* **13**, 95–101 (2011).
338. Tripathi, V. *et al.* The nuclear-retained noncoding RNA *MALAT1* regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39**, 925–938 (2010).
339. Tripathi, V. *et al.* Long noncoding RNA *MALAT1* controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet.* **9**, e1003368 (2013).
340. Sone, M. *et al.* The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. *J. Cell Sci.* **120**, 2498–2506 (2007).
341. Barry, G. *et al.* The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Mol. Psychiatry* **19**, 486–494 (2014).
342. Weinberg, M. S. *et al.* The antisense strand of small interfering RNAs directs histone methylation and transcriptional gene silencing in human cells. *RNA* **12**, 256–262 (2006).
343. Janowski, B. A. *et al.* Involvement of AGO1 and AGO2 in mammalian transcriptional silencing. *Nature Struct. Mol. Biol.* **13**, 787–792 (2006).
344. Ling, J., Baibakov, B., Pi, W., Emerson, B. M. & Tuan, D. The HS2 enhancer of the beta-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a *cis*-linked globin promoter. *J. Mol. Biol.* **350**, 883–896 (2005).
345. Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013).
346. Gumireddy, K. *et al.* Identification of a long non-coding RNA-associated RNP complex regulating metastasis at the translational step. *EMBO J.* **32**, 2672–2684 (2013).
347. Watson, J. D. & Crick, F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).
348. Holmes, D. S., Mayfield, J. E., Sander, G. & Bonner, J. Chromosomal RNA: its properties. *Science* **177**, 72–74 (1972).
349. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the *H19* gene may function as an RNA. *Mol. Cell. Biol.* **10**, 28–36 (1990).
350. Fire, A. *et al.* Potent and specific genetic interference by double stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
351. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952 (1999).
352. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
353. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2**, 986–991 (2001).
354. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437–1441 (2004).

Acknowledgements
This work was supported by the US National Institutes of Health grant PO1 AI099783-01 and the Australian Research Council Future Fellowship FT130100572 to K.V.M., and by the National Health and Medical Research Council Australia Fellowship 631688 to J.S.M.

Competing interests statement
The authors declare no competing interests.