identified above) were tested via Tol2-mediated transgenesis in zebrafish embryos. We observed tissue-specific enhancer activity with 3 of 5 fragments, which corresponded to the human enhancer tissue expression (Fig. 4). None of three control fragments without CAGE signal activated the *GATA2* promoter (Supplementary Table 9). Although the sample size is not high enough to reliably estimate the validation rates in zebrafish, the correlation between the enhancer usage profiles in zebrafish to those defined in human by CAGE is notable.

We grouped the primary cell and tissue samples into larger, mutually exclusive cell type and organ/tissue groups (referred to as facets), respectively, with similar function or morphology (Supplementary Tables 10 and 11). Figure 5 summarizes how many enhancers were detected in each facet and the degree of facet-specific CAGE expression (see also Supplementary Fig. 21). From the data we can draw several conclusions:

First, the majority of detected enhancers within any facet are not restricted to that facet. Exceptions, where facets use a higher fraction of specific enhancers, include immune cells, neurons, neural stem cells and hepatocytes amongst the cell-type facets, and brain, blood, liver and testis amongst the organ/tissue facets.

Second, despite their apparent promiscuity, enhancers are more generally detected in a much smaller subset of samples than mRNA transcripts (Supplementary Figs 21 and 22a, b), consistent with cell-line studies[7] and the higher specificity of ncRNAs in general[13]. Facets in which we detect many enhancers typically also have a higher fraction of facet-specific enhancers (Supplementary Fig. 22c, d).

Third, the number of detected expressed enhancers and mRNA transcripts is correlated (Supplementary Fig. 21b), but the number of detected expressed gene transcripts (>1 tag per million mapped reads (TPM)) is 19–34 fold larger than the number of detected enhancers with the cut-offs used. Noteworthy exceptions include blood and immune cells, testis, thymus and spleen, which have high enhancer/gene ratios. Conversely, smooth and skeletal muscle and skin, bone and epithelia-related cells have low ratios. Differential exosome activity between cell types might affect these results, but there was no correlation between *SKIV2L2* mRNA expression and the number of enhancers detected (Supplementary Fig. 22e, f).

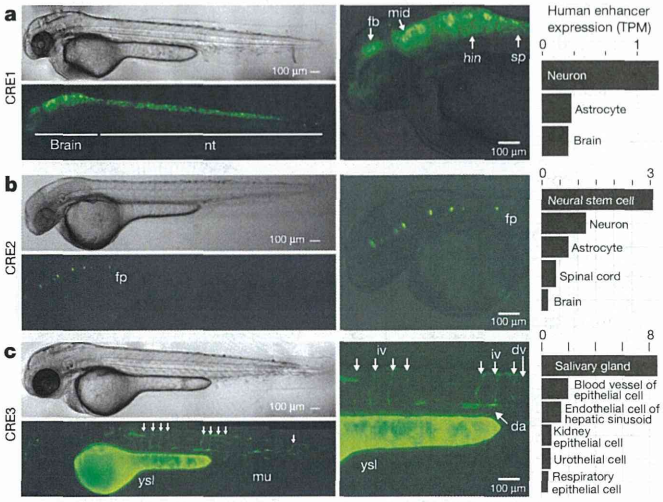

Figure 4 | *In vivo* validation in zebrafish of tissue-specific enhancers. Validations of *in vivo* activity of CAGE-defined human enhancers CRE1, CRE2 and CRE3 in zebrafish embryos at long-pec stage. Each panel shows, from left to right, representative yellow fluorescent protein (YFP) and bright field images of embryos injected with the human enhancer *gata2* promoter reporter gene construct (left), YFP zoom-ins (middle) and CAGE expression in TPM in human tissues/cell types for the enhancer (right). Muscle (mu) and yolk syncytial layer (ysl) activities are background expression coming from the *gata2* promoter-containing reporter construct. All images are lateral, head to the left. Note the correspondence between zebrafish and human enhancer usage/expression. Supplementary Fig. 20 shows UCSC browser images of each selected enhancer. **a**, CRE1, ~230 kb upstream of the *MEFC2* gene, drives highly robust expression in the brain (brain) and neural tube (nt). Right panel gives zoom-in overlay image showing expression in the forebrain (fb), midbrain (mid), hindbrain (hin) and spinal cord (sp). **b**, CRE2, 5 kb upstream of the *POU3F2* gene, is active in the floor plate (fp). **c**, CRE3, 10 kb upstream of the *SOX7* gene TSS, shows specific expression in the vasculature (including intersegmental vessels (iv), dorsal vein (dv) and dorsal aorta (da).

As expected, consensus motifs of known key regulators are over-represented in corresponding facet-specific enhancers, for instance ETS, C/EBP and NF-κB in monocyte-specific enhancers, RFX and SOX in neurons, and HNF1 and HNF4a in hepatocytes (Supplementary Fig. 23). Notably, the AP1 motif appears to be enriched across all facets, perhaps associated with a general role for AP1 in regulating open chromatin[19].

Expression clustering reveals ubiquitous enhancers

Hierarchical clustering of enhancers by facet expression revealed a small subset of enhancers (200 or 247, defined by primary cell or tissue facets, respectively) expressed in the large majority of facets (Supplementary Text, Supplementary Figs 24 and 25, and Supplementary Tables 12 and 13). Compared to other enhancers, these ubiquitous (u-) enhancers are 8 times more likely to overlap CGIs and they are twice as conserved (Supplementary Fig. 26a–c). U-enhancers overlap typical chromatin enhancer marks but have higher H3K4me3 signal (Supplementary Fig. 26d). Although they produce significantly longer ncRNAs than other enhancers (median 530 nucleotides, $P < 1.5 \times 10^{-8}$, Mann–Whitney *U* test), the transcripts remain predominantly (~78%) unspliced and significantly shorter ($P < 4.2 \times 10^{-18}$, Mann–Whitney *U* test) than mRNAs (Supplementary Figs 27 and 28), do not share exons with known genes, and are exosome-sensitive (Supplementary Fig. 14b). Therefore, it is unlikely that these are novel mRNA promoters. They are also highly enriched for P300 and cohesin ChIP-seq peaks[20] and RNAPII-mediated ChIA-PET signal[21] compared to other enhancers (Supplementary Fig. 26d). These results indicate that u-enhancers comprise a small but distinct subset of enhancers, which probably has specific regulatory functions used by virtually every human cell.

Linking enhancer usage with TSS expression

A major challenge is to link enhancers to their target genes[21,22]. Uniquely, FANTOM5 CAGE allows for direct comparison between transcriptional activity of the enhancer and of putative target gene TSSs across a diverse set of human cells. Based on pairwise expression correlation, nearly half (40%) of the inferred TSS-associated enhancers (Methods) were linked with the nearest TSS, and 64% of enhancers have at least one correlated TSS within 500 kilobases. Several associations (10,260; 15.3%) are supported by ChIA-PET (RNAPII-mediated) interaction data[21], and the supported fraction increases with the correlation threshold (Supplementary Fig. 29a). The fraction of supported associations is 4.8-fold higher than that of associations predicted from DNase I hypersensitivity correlations[10] (20.6% versus 4.3%, at the same correlation threshold), indicating that transcription is a better predictor of regulatory targets than chromatin accessibility. Conserved sequence motifs and ChIP-seq peaks also co-occurred significantly in associated enhancer-promoter pairs (Benjamini–Hochberg false discovery rate (FDR) < 0.05, binomial test), suggesting an additive or synergistic cooperation between enhancers and promoters at RNAPII foci.

On average, a RefSeq TSS was associated with 4.9 enhancers and an enhancer with 2.4 TSSs and we observed different regulatory architectures around genes (Supplementary Fig. 30). For example, at the beta-globin locus the CAGE expression patterns of four locus control region hypersensitive sites are highly correlated (Pearson's *r* between 0.88 and 0.98) with the expression of known target genes[23,24] *HBG2* and *HBD*, and to some extent *HBG1*.

©2014 Macmillan Publishers Limited. All rights reserved

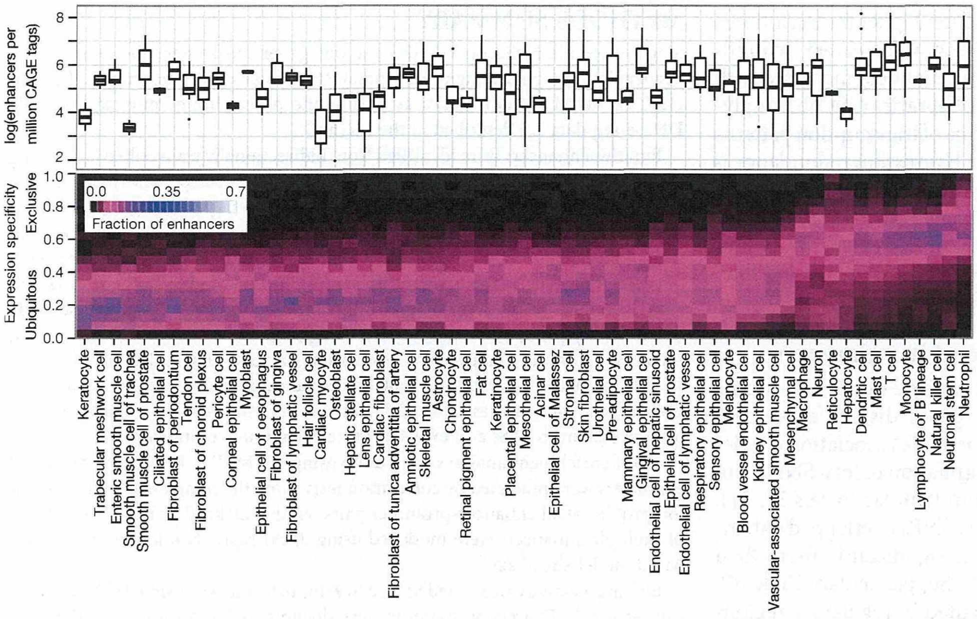

Figure 5 | Enhancer usage and specificity in groups of cells. The upper panel gives the number of detected enhancers per million CAGE tags within each group (facet) of related cell type libraries. The expression specificity of the enhancers is shown as a heat map in the panel below. Colours show the fraction of expressed enhancers in each facet (columns) that are in each specificity range (rows). For corresponding plots on organ/tissue facets and genes, see Supplementary Fig. 21.

These observations call for computational models of enhancer regulation, in which multiple enhancers may work in concert to enhance the expression of a gene. To this end, we focused on 2,206 RefSeq TSSs for which the joint expression of nearby enhancers (the closest ten enhancers within 500 kb) is highly predictive of the gene expression. Model shrinkage showed that in most cases, only one to three enhancers are necessary to explain the expression variance observed in the linked gene, and generally proximal enhancers are more predictive than distal ones (Fig. 6a, Supplementary Fig. 29b–d and Supplementary Text). One hypothesis explaining the function of multiple enhancers driving the

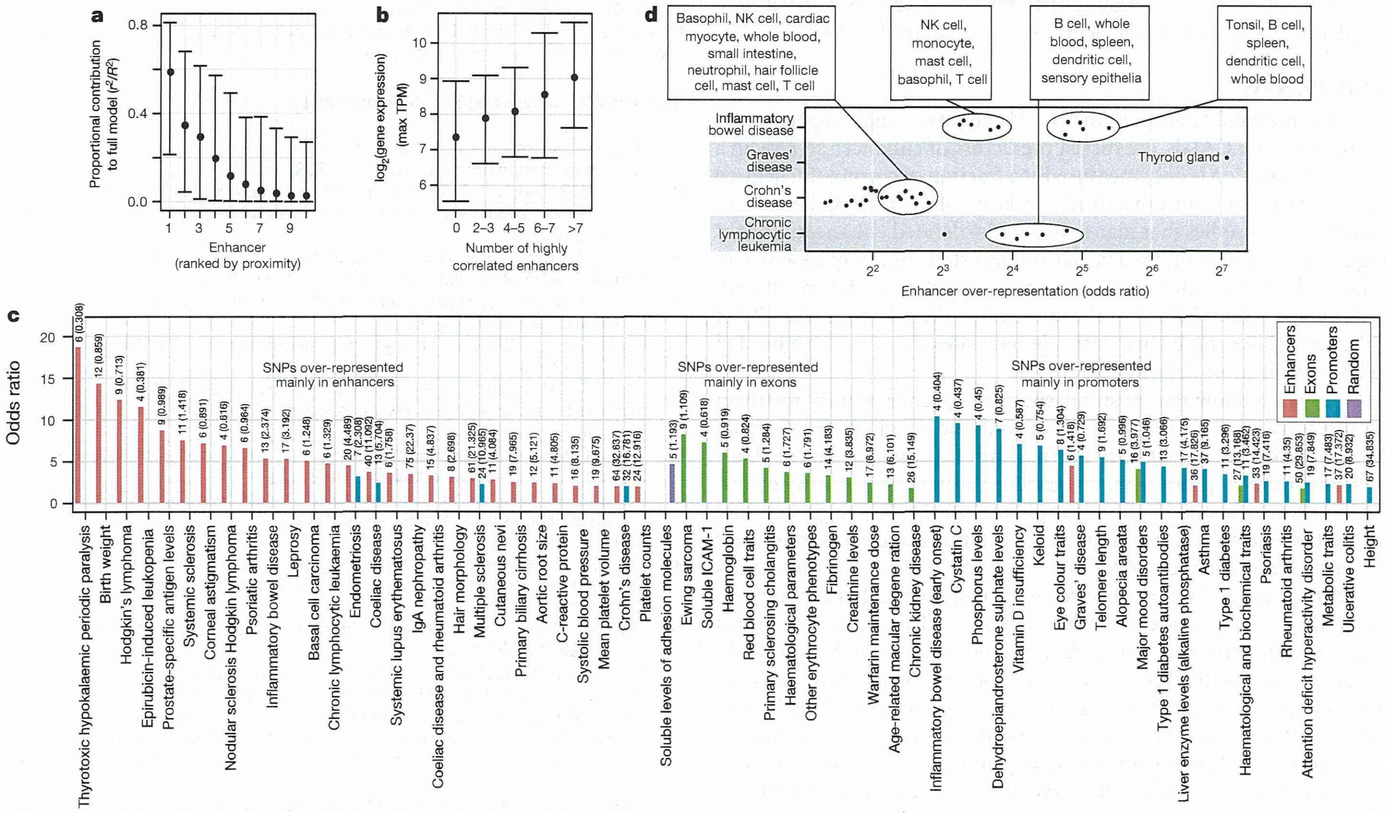

Figure 6 | Linking enhancers to TSSs and disease-associated SNPs. a, The proportional contribution (see Methods) of the 10 most proximal enhancers within 500 kb of a TSS in a model explaining gene expression variance (vertical axis) as a function of enhancer expression. *x* axis indicates the position of the enhancer relative to the TSS: 1 the closest, etc. Bars indicate interquartile ranges and dots medians. **b**, Relationship between the number of highly correlated ('redundant') enhancers per locus (horizontal axis) and the maximal expression (TPM) of the associated TSS in the same model over all CAGE libraries (vertical axis). Error bars as in **a**. **c**, GWAS SNP sets preferentially overrepresented within enhancers, exons and mRNA promoters. Observed and expected overlaps are shown above bars. The vertical axis gives enrichment odds ratios. The horizontal axis shows GWAS traits or diseases. **d**, Diseases with GWAS-associated SNPs over-represented in enhancers of certain expression facets. The horizontal axis gives the odds ratio as in panel **c**, broken up by expression facets: each point represents the odds ratio of GWAS SNP enrichment for a disease (vertical axis) in a specific expression facet. Summary annotations of point clouds are shown. See also Supplementary Fig. 31.

©2014 Macmillan Publishers Limited. All rights reserved

same expression pattern is that they might confer higher transcriptional output of a gene[25,26]. Indeed, the number of highly correlated (redundant) enhancers close to TSSs (Supplementary Methods) increased with the observed maximal TSS expression over all libraries (Fig. 6b), implying that these enhancers are redundant in terms of transcription patterns but additive in terms of expression strength. Expression redundancy is also common in genomic clusters of closely spaced enhancers (24% of 815 identified genomic clusters, Supplementary Table 15). These are associated with TSSs of genes involved in immune and defence responses and, as suggested by a previous study[27], have a higher expression than other enhancer-associated genes (eightfold increase on average).

Disease-associated SNPs are enriched in enhancers

Many disease-associated SNPs are located outside of protein-coding exons and a large proportion of human genes display expression polymorphism[28]. Using the NHGRI genome-wide association studies (GWAS) catalogue[29] and extending the compilation of lead SNPs with proxy SNPs in strong linkage disequilibrium (similar to refs 30, 31), we identified diseases/traits whose associated SNPs overlapped enhancers, promoters, exons and random regions significantly more than expected by chance (Fisher's exact test $P < 0.01$, Supplementary Table 16). Disease-associated SNPs were over-represented in regulatory regions to a greater extent than in exons (Fig. 6c). For many traits where enriched disease-associated SNPs were within enhancers, enhancer activity was detected in pathologically relevant cell types (Fig. 6d and Supplementary Figs 31 and 32). Examples include Graves' disease-associated SNPs enriched in enhancers that are expressed predominantly in thyroid tissue, and similarly lymphocytes for chronic lymphocytic leukaemia. As a proof of concept, we validated the impact of two disease-associated regulatory SNPs within enhancers (Supplementary Fig. 33).

Conclusions

The data presented here demonstrate that bidirectional capped RNAs, as measured by CAGE, are robust predictors of enhancer activity in a cell. Transcription is only measured at a fraction of chromatin-defined enhancers and few untranscribed enhancers show potential enhancer activity. This implies that many chromatin-defined enhancers are not regulatory active in that particular cellular state, but may be active in other cells of the same lineage[32] or are pre-marked for fast regulatory activity upon stimulation[33]. Of course, given the relative instability of enhancer RNAs some chromatin-defined sites may be active but fall below the limits of detection of CAGE.

Our results show that position-specific sequence signals upstream of the transcription initiation sites and the production of small, uncapped RNAs immediately downstream is present at both enhancers and mRNA promoters, suggesting similar mechanisms of initiation. Previous studies (for example refs 10, 34, 35) suggested that promoters and enhancers differ in motif composition. This view is not supported by the larger FANTOM5 data set. Instead, the differences reflect the local G+C content because transcribed enhancers tend to harbour low G+C content motifs like non-CGI promoters. Features distinguishing enhancers from mRNA promoters are (1) enhancer RNAs are exosome-sensitive regardless of direction whereas (sense) mRNAs have a longer half-life than their antisense counterpart; (2) enhancer RNAs are short, unspliced, nuclear and non-polyadenylated and (3) enhancers have downstream polyadenylation and 5′ splice motif frequencies at genomic background level similar to antisense PROMPTs, whereas mRNAs are depleted of termination signals and enriched for 5′ splice sites[11,12].

The collection of active enhancers presented here provides a resource that complements the activity of the ENCODE consortium[7] across a much greater diversity of tissues and cellular states. It has clear applications in human genetics, to narrow the search windows for functional association, and for the definition of regulatory networks that underpin the processes of cellular differentiation and organogenesis in human development.

METHODS SUMMARY

Single-molecule HeliScopeCAGE data was generated as described elsewhere[6]. Sequencing and processing of ribosomal RNA-depleted RNAs, short RNAs and H3K27ac or H3K4me1 ChIPs as well as the processing of publicly available DNase-seq data are described in the Methods.

Putative enhancers were identified from bidirectionally transcribed loci having divergent CAGE tag clusters separated by at most 400 bp (described in Supplementary Fig. 6a). We required loci to be divergently transcribed in at least one FANTOM5 sample, defined by CAGE tag 5′ ends within 200 bp divergent strand-specific windows immediately flanking the loci midpoints. The expression of each enhancer in each FANTOM5 sample was quantified as the normalized sum of strand-specific sums of CAGE tags in these windows. A sample-set wide directionality score, D, for each locus over aggregated normalized reverse, R, and forward, F, strand window-expression values across all samples, $D = (F - R)/(F + R)$, were then used to filter putative enhancers to have low, non-promoter-like, directionality scores ($|D| < 0.8$). Further filtering ensured enhancers to be located distant to TSSs and exons of protein- and noncoding genes.

Motif enrichment analyses were done using HOMER[36]. Regulatory targets of enhancers were predicted by correlation tests using the sample-set wide expression profiles of all enhancer-promoter pairs within 500 kb. The regulatory effects of multiple enhancers were modelled using linear regression followed by lasso-based model-shrinkage[37].

Enhancer activity was tested *in vivo* in zebrafish embryos using Tol2-mediated transgenesis[38]. Expression patterns were documented at 48 h post fertilization using >200 eggs per construct. Large-scale *in vitro* validations on randomly selected enhancers were performed using firefly/*Renilla* luciferase reporter plasmids with enhancer sequences cloned upstream of an EF1α basal promoter separated by a synthetic polyA signal/transcriptional pause site in a modified pGL4.10 (Promega) vector (Supplementary Fig. 9d). Full details are provided in the Methods.

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

Received 4 January; accepted 16 October 2013.

1. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339,** 250–257 (2010).
2. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Rev. Genet.* **13,** 233–245 (2012).
3. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
4. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
5. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature Methods* **3,** 211–222 (2006).
6. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* http://dx.doi.org/10.1038/nature13182 (this issue).
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
8. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).
9. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genet.* (in the press).
10. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
11. Ntini, E. *et al.* Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nature Struct. Mol. Biol.* **20,** 923–928 (2013).
12. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499,** 360–363 (2013).
13. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
14. Kowalczyk, M. S. *et al.* Intragenic enhancers act as alternative promoters. *Mol. Cell* **45,** 447–458 (2012).
15. Valen, E. *et al.* Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature Struct. Mol. Biol.* **18,** 1075–1082 (2011).
16. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41,** 572–578 (2009).
17. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
18. Rönnerblad, M. *et al.* Analysis of the DNA methylome and transcriptome in granulopoiesis reveal timed changes and dynamic enhancer methylation. *Blood* http://dx.doi.org/10.1182/blood-2013-02-482893 (in the press).

©2014 Macmillan Publishers Limited. All rights reserved

19. Biddie, S. C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol. Cell* **43,** 145–155 (2011).
20. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20,** 578–588 (2010).
21. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
22. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22,** 490–503 (2012).
23. Fraser, P., Pruzina, S., Antoniou, M. & Grosveld, F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes Dev.* **7,** 106–113 (1993).
24. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
25. Barolo, S. Shadow enhancers: frequently asked questions about distributed *cis*-regulatory information and enhancer redundancy. *Bioessays* **34,** 135–141 (2012).
26. Schaffner, G., Schirm, S., Müller-Baden, B., Weber, F. & Schaffner, W. Redundancy of information in enhancers as a principle of mammalian transcription control. *J. Mol. Biol.* **201,** 81–90 (1988).
27. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
28. Göring, H. H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39,** 1208–1216 (2007).
29. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
30. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930–D934 (2012).
31. Maurano, M. T., Wang, H., Kutyavin, T. & Stamatoyannopoulos, J. A. Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.* **8,** *e1002599 (2012).*
32. Mercer, E. M. *et al.* Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity* **35,** 413–425 (2011).
33. Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* **152,** 157–171 (2013).
34. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
35. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
36. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
37. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).
38. Gehrig, J. *et al.* Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nature Methods* **6,** 911–916 (2009).

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

Acknowledgements FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Y.H. and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan, to Y.H. The A.S. group was supported by funds from the European Research Council FP7/2007-2013/ERC no. 204135, the Novo Nordisk and Lundbeck foundations. Work in the M.R. group was funded by grants from the Deutsche Forschungsgemeinschaft (RE 1310/7, 11, 13) and Rudolf Bartling Stiftung. F.M. and I.M.E. were supported by "BOLD" Marie Curie ITN and "ZF- Health" Integrated project of the European Commission. We thank S. Noma, M. Sakai and H. Tarui for RNA-seq and sRNA-seq preparation, RIKEN GeNAS for generation and sequencing of the Heliscope CAGE libraries, Illumina RNA-seq and sRNA-seq, the Copenhagen National High-throughput DNA Sequencing Center for Illumina CAGE-seq, M. Edinger, P. Hoffmann and R. Eder for cell sorting, A. Albrechtsen, I. Moltke, W. Wasserman for advice, and the Netherlands Brain Bank for post-mortem human brain material.

Author Contributions R.A., I.H., E.A., E.V., K.L., Y.C., B.L., X.Z., M.J., H.K., T.F.M., T.L., N.B., O.R., A.M.B. , J.K.B, C.J.M, N.R., F.O.B., M.R., A.S. made the computational analysis. J.B., M.B., T.L., H.K., N.K., J.K., H.S., M.I., C.O.D, A.R.R.F., P.C., Y.H. prepared and pre-processed CAGE and/or RNA-seq libraries. E.N., P.R.A., T.H.J., J.B., M.B. made the knockdown experiments followed by CAGE. C.G., C.S., L.S., J.R., D.G., M.R. made the blood cell ChIP experiments, methylation assays and *in vitro* blood cell validations. T.S., C.G., Y.I., Y.S., E.F., S.M., Y.N., A.R.R.F., P.C. and H.S. made the HeLa/HepG2 *in vitro* validations. I.M.E., R.A., A.S., F.M. designed and carried out zebrafish *in vivo* tests. R.A., C.G., I.H., C.S., E.A., E.V., F.M., I.M.E., P.C., A.R.R.F, M.B., J.B., A.L., C.D., D.A.H., P.H., M.R., A.S. interpreted results. R.A., C.G., I.H., E.V., I.M.E., J.B., F.M., D.A.H., M.R., A.S. wrote the paper with input from all authors. M.R. and A.S. coordinated and supervised the project.

Author Information The FANTOM5 atlas is accessible *from http://fantom.gsc.riken.jp/5*. FANTOM5 CAGE, RNA-seq and sRNA data have been deposited in DDBJ/EMBL/GenBank (accession codes DRA000991, DRA001101). Genome browser tracks for enhancers with user-definable expression specificity-constraints can be generated at http://enhancer.binf.ku.dk. Here, processed enhancer expression data, predefined enhancer tracks and motif finding results are also deposited. Blood-cell ChIP-seq data and CAGE data on exosome-depleted HeLa cells have been deposited in the NCBI GEO database (accession codes GSE40668, GSE49834). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R.R.F. (alistair.forrest@gmail.com), P.C. (carninci@riken.jp), M.R. (michael.rehli@ukr.de) or A.S. (albin@binf.ku.dk).

©2014 Macmillan Publishers Limited. All rights reserved

Molecular Cell
Article

Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli

Katsutoshi Imamura,[1] Naoto Imamachi,[2] Gen Akizuki,[2] Michiko Kumakura,[3] Atsushi Kawaguchi,[3] Kyosuke Nagata,[3] Akihisa Kato,[4] Yasushi Kawaguchi,[4] Hiroki Sato,[5] Misako Yoneda,[5] Chieko Kai,[5] Tetsushi Yada,[6] Yutaka Suzuki,[7] Toshimichi Yamada,[8] Takeaki Ozawa,[8] Kiyomi Kaneki,[9] Tsuyoshi Inoue,[9] Mika Kobayashi,[9] Tatsuhiko Kodama,[9] Youichiro Wada,[2,9] Kazuhisa Sekimizu,[1] and Nobuyoshi Akimitsu[2,*]

[1]Graduate School of Pharmaceutical Sciences, The University of Tokyo, Tokyo 113-0033, Japan
[2]Radioisotope Centre, The University of Tokyo, Tokyo 113-0032, Japan
[3]Department of Infection Biology, Faculty of Medicine & Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba 305-8575, Japan
[4]Division of Molecular Virology, Department of Microbiology and Immunology, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan
[5]Laboratory Animal Research Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan
[6]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology Fukuoka 820-8502, Japan
[7]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8562, Japan
[8]Department of Chemistry, School of Science, The University of Tokyo, Tokyo 113-0033, Japan
[9]Laboratory for Systems Biology and Medicine, Research Centre for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan
*Correspondence: akimitsu@ric.u-tokyo.ac.jp

http://dx.doi.org/10.1016/j.molcel.2014.01.009

SUMMARY

Although thousands of long noncoding RNAs (lncRNAs) are localized in the nucleus, only a few dozen have been functionally characterized. Here we show that nuclear enriched abundant transcript 1 (NEAT1), an essential lncRNA for the formation of nuclear body paraspeckles, is induced by influenza virus and herpes simplex virus infection as well as by Toll-like receptor3-p38 pathway-triggered poly I:C stimulation, resulting in excess formation of paraspeckles. We found that NEAT1 facilitates the expression of antiviral genes including cytokines such as interleukin-8 (IL8). We found that splicing factor proline/glutamine-rich (SFPQ), a NEAT1-binding paraspeckle protein, is a repressor of *IL8* transcription, and that NEAT1 induction relocates SFPQ from the *IL8* promoter to the paraspeckles, leading to transcriptional activation of *IL8*. Together, our data show that NEAT1 plays an important role in the innate immune response through the transcriptional regulation of antiviral genes by the stimulus-responsive cooperative action of NEAT1 and SFPQ.

INTRODUCTION

Whole-transcriptome analyses have revealed that a new class of non-protein-coding transcripts, designated as long noncoding RNAs (lncRNAs), is transcribed from a large proportion of the mammalian genome (Carninci et al., 2005; Guttman et al., 2009; Kapranov et al., 2007). There is increasing evidence of lncRNA involvement in diverse biological processes (Chen and Carmichael, 2010; Gupta et al., 2010; Ponting et al., 2009; Yoon et al., 2013). Moreover, a large number of lncRNAs is induced by extracellular stimuli, suggesting that lncRNAs participate in stress responses (Mizutani et al., 2012; Tani et al., 2012). In addition, because lncRNAs are also implicated in many human diseases (Huarte and Rinn, 2010; Wang and Chang, 2011), understanding the precise molecular mechanisms by which lncRNAs function could prove important for developing new strategies for early diagnosis and molecular therapy. In particular, there are several emerging hypotheses on lncRNA involvement in infectious diseases (Scaria and Pasha, 2012). However, a mechanistic understanding of the role of lncRNAs in infection is limited. Hence, the functions of lncRNAs in host antiviral response have remained unclear.

The mammalian nucleus is highly organized and contains distinct structural components comprising approximately ten types of nuclear bodies, including speckles and paraspeckles, which are thought to be involved in gene regulation (Mao et al., 2011). Some of these nuclear bodies contain specific lncRNAs that regulate nuclear body function (Kapranov et al., 2007; Prasanth and Spector, 2007). Recent reports have suggested that crosstalk between architectural features of nuclear bodies and lncRNAs contributes to the precise control of gene expression. For example, speckles contain the metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), a lncRNA involved in regulating the expression of several specific genes (Bernard et al., 2010; Miyagawa et al., 2012; Tano et al., 2010; Yang et al., 2011). Paraspeckles contain another lncRNA, NEAT1, which is an essential architectural component of paraspeckle

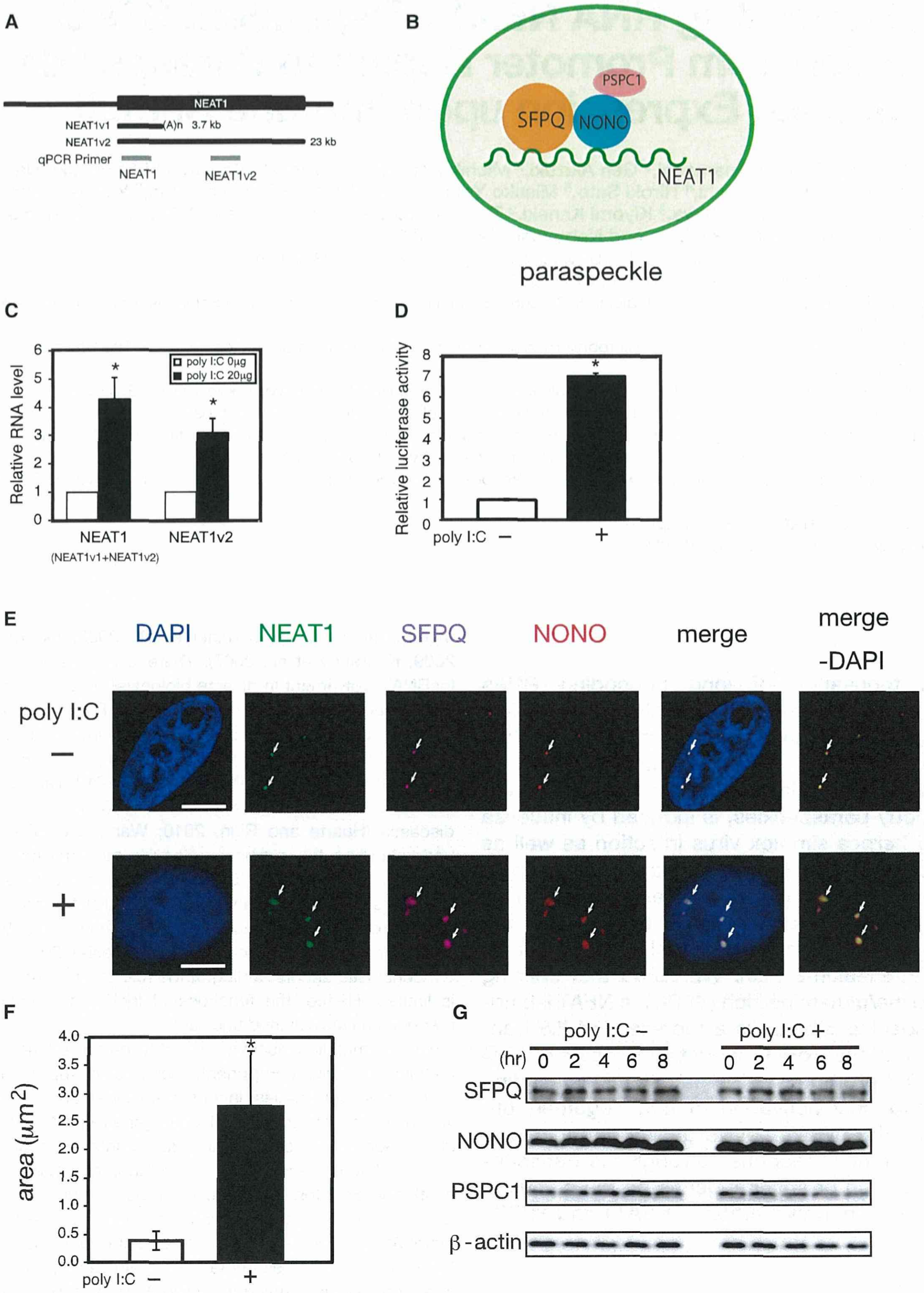

Figure 1. NEAT1 Induction and Excessive Formation of Paraspeckles by Poly I:C

(A) NEAT1 isoforms are shown schematically. The fragment positions amplified by the RT-qPCR primers are shown below.

(B) Entities of paraspeckles are shown schematically.

(legend continued on next page)

structure (Chen and Carmichael, 2009; Clemson et al., 2009; Sasaki et al., 2009; Sunwoo et al., 2009). The *NEAT1* gene (Figure 1A) produces two isoforms, 3.7 kb NEAT1v1 and 23 kb NEAT1v2 (Hutchinson et al., 2007). The effect of NEAT1v2 on the formation of paraspeckles is stronger than that of NEAT1v1 (Naganuma et al., 2012; Sasaki et al., 2009). Paraspeckles have been proposed to control several biological processes, including stress response and cellular differentiation, through control of the nuclear retention of mRNAs containing inverted repeats that form double-stranded RNA regions subject to adenosine-to-inosine editing (Fox and Lamond, 2010; Nakagawa and Hirose, 2012). Paraspeckles contain several protein factors: NONO/p54nrb, SFPQ/PSF, PSPC1, RBM14, and CPSF6 (Fox and Lamond, 2010). Among these, SFPQ and NONO form the heterodimer (Peng et al., 2002), which binds directly to NEAT1 (Sasaki et al., 2009) (Figure 1B). Several studies have demonstrated that SFPQ represses the transcription of several genes through direct promoter binding (Iacobazzi et al., 2005; Song et al., 2004; Urban et al., 2000). Recently, 35 proteins were added into the list of paraspeckle proteins (Naganuma and Hirose, 2013). Several of paraspeckle proteins are likely to be the factors involved in transcriptional control, suggesting that paraspeckles may integrate tightly coupled transcription and posttranscriptional events.

The innate immune response is crucial in the host cellular response to viral infection. Several pathogen-associated molecular pattern recognition receptors, such as the Toll-like receptors, sense the presence of viral molecules and trigger a robust program of gene expression involving the production of antiviral inflammatory cytokines, chemokines, and interferons through numerous transcriptional and posttranscriptional strategies (Arpaia and Barton, 2011; Rathinam and Fitzgerald, 2011; Thompson et al., 2011). For example, poly I:C, a double-stranded RNA (dsRNA)-mimicking immunostimulant that simulates viral infections, activates the TLR3-mediated signaling pathway, and consequently induces a set of antiviral genes (Kawai and Akira, 2010). To achieve the proper immune response, the transcriptional induction of immune response genes is highly coordinated by activators and repressors. For instance, the interleukin-8 (*IL8*) promoter is repressed by the binding of three factors in unstimulated cells (Hoffmann et al., 2002): NF-κB-repressing factor (NRF), octamer-1 (OCT-1), and deacetylation of histone proteins by histone deacetylase-1. When the cells are stimulated, NF-κB and C/EBP bind to the *IL8* promoter; C/EBP displaces OCT-1, whereas NRF switches its function to act as a coactivator. Recruitment of CREB-binding protein/p300 hyperacetylates the histones and remodels the chromatin, resulting in transcriptional activation of *IL8* gene. Although nuclear lncRNAs represent a large class of transcriptional units, the interplay between transcription factors and nuclear lncRNA to control gene expression during immune response remains to be elucidated.

RESULTS

Poly I:C Induces NEAT1 and Large Paraspeckle Formation

A previous study showed that NEAT1 is an inducible lncRNA in mice brains infected with Japanese encephalitis or rabies viruses, although it is unclear whether NEAT1 induction is a consequence of direct effect of viral infection to neural cells (Saha et al., 2006). This observation provided the rationale for the current study, which examined the relevance of NEAT1 in cellular response to viral infection. We therefore initially examined the expression levels of NEAT1 in response to transfection with poly I:C, a double-stranded RNA (dsRNA). As shown in Figure 1A, one primer set recognizes both NEAT1v1 and NEAT1v2 (total NEAT1), while the other recognizes only NEAT1v2. Expression levels of total NEAT1 and NEAT1v2 in HeLa cells and A549 cells were increased by poly I:C, but not by poly I or poly C alone (Figure 1C and Figures S1A–S1C). Treatment of the cells with either IFN-α or IFN-β induced 2′5′-OAS, an interferon response gene, but not NEAT1v2 (Figure S1D), ruling out the possibility of an indirect effect by which IFNs induced by poly I:C lead the expression of NEAT1. To examine whether upregulation of NEAT1 RNA levels by poly I:C stimulation was controlled by transcriptional regulation, we analyzed luciferase reporter activity in HeLa TO cells transfected with a luciferase reporter gene linked to a *NEAT1* promoter and found that poly I:C treatment enhanced the luciferase reporter activity (Figure 1D and Figure S1E). Next, we investigated the signaling pathway that activates transcription of the *NEAT1* gene by poly I:C stimulation. Because TLR3 is known as an intracellular sensor for dsRNAs such as poly I:C (Kawai and Akira, 2010), we tested the involvement of TLR3 in the poly I:C-induced transcriptional activation of the *NEAT1* gene. Knockdown of TLR3 reduced the levels of poly I:C-mediated NEAT1 induction compared with control cells (Figure S1H). We further examined other dsRNA sensors. We found that depletion of MDA-5, but not RIG-I, affected poly I:C-induced NEAT1 expression (Figures S1I and S1J). The effect of MDA-5 depletion for the reduction in poly I:C-induced NEAT1 expression was weaker than that for the reduction in TLR3, suggesting that TLR3 is the major receptor for inducing NEAT1 in response to poly I:C. TLR3-mediated signaling is branched to either the

(C) Total NEAT1 and NEAT1v2 levels of HeLa TO cells with or without poly I:C stimulation were quantified by RT-qPCR. The GAPDH mRNA level was used as the normalizing control. Values represent the mean ± SD (*p < 0.01, Student's t test).

(D) The luciferase reporter activity of HeLa TO cells transfected with a luciferase reporter gene harboring the *NEAT1* promoter was measured in the presence or absence of poly I:C. The activity of cotransfected pCMV-RL (Promega) was used as normalizing control. Values represent the mean ± SD (*p < 0.01, Student's t test).

(E) HeLa TO cells were transfected with and without poly I:C, followed by FISH staining and immunostaining. NEAT1 (green), SFPQ (magenta), NONO (red), and nuclei stained with DAPI (blue) are shown.

(F) The mean size of NEAT1 control cell foci (white bar; n = 50) and that of cells transfected with poly I:C (black bar; n = 50) was determined by FISH. Values represent the mean ± SD (*p < 0.01, Student's t test).

(G) The protein levels of paraspeckle proteins SFPQ, NONO, and PSPC1 were analyzed by western blotting at the indicated time points posttransfection with poly I:C. β-actin was used as the loading control.

p38 or JNK pathways (Arpaia and Barton, 2011). Pretreatment with ML3403, a p38 inhibitor, but not SP600125, a JNK inhibitor, abolished poly I:C-induced NEAT1 induction (Figure S1K). As expected, poly I:C-induced phosphorylation of p38 and JNK was eliminated by ML3403 and SP600125, respectively (Figures S1L and S1M). In contrast, NF-κB was not required for poly I:C-induced NEAT1 induction (Figures S1N and S1O). These results suggest that poly I:C leads to the transcriptional activation of the *NEAT1* gene mainly through the TLR3-p38 pathway.

Previous reports have shown that NEAT1 is an essential core component for the formation of paraspeckles (Chen and Carmichael, 2009; Clemson et al., 2009; Sasaki et al., 2009; Sunwoo et al., 2009). Corresponding with previous observation, paraspeckle proteins were dispersed to the nucleoplasm in the absence of NEAT1 (Figure S1Q). Because overexpression of NEAT1 results in the excess formation of paraspeckles (Clemson et al., 2009), we hypothesized that poly I:C stimulation induces this process. Combination staining of NEAT1 and of paraspeckle proteins SFPQ, NONO, and PSPC1 showed that poly I:C treatment resulted in excess paraspeckle formation in HeLa cells (Figures 1E and 1F and Figure S1P). Western blot analysis revealed that expression levels of paraspeckle proteins SFPQ and NONO remained unaltered throughout poly I:C stimulation (Figure 1G). In the absence of NEAT1, poly I:C stimulation did not induce the formation of paraspeckles (Figure S1Q). Fluorescence recovery after photobleaching (FRAP) analysis showed that the kinetics of paraspeckle-associated SFPQ in poly I:C-stimulated cells ($t_{1/2}$ = 7.08 s) was similar to that in naive cells ($t_{1/2}$ = 6.75 s) (Figures S1R and S1S; Movies S1 and S2), suggesting that the molecular quality of SFPQ was not changed by poly I:C stimulation. These findings suggest that poly I:C stimulation relocates paraspeckle proteins from the nucleoplasm to NEAT1, consequently inducing the excess formation of paraspeckles.

Identification of NEAT1-Regulated Antiviral Genes

We investigated whether NEAT1 induction followed by excess formation of paraspeckles was involved in poly I:C-inducible gene expression (Figures 2A and 2B). Microarray analysis revealed 1,232 poly I:C-inducible genes in HeLa TO cells. The induction of 259 of these poly I:C-inducible genes was abolished by NEAT1 knockdown (Figure 2B). We also identified 113 genes that were upregulated by solo overexpression of mNeat1v2 (Figure 2B). To eliminate false positives, we selected the 85 genes that form the overlap between these two groups of genes (Figure 2B). Interestingly, many antiviral factors, such as IL8 and CCL5, and virus sensors, such as RIG-I and MDA5, were identified in this group of 85 NEAT1-regulated genes, suggesting that NEAT1 is involved in the regulation of antiviral gene expression response. A gene ontology analysis using these data supported this idea (Tables S1 and S2). RT-qPCR experiments confirmed the NEAT1-dependent expression of genes involved in antiviral function, such as *IL8* and *CCL5* (Figure 2C and Table S3). To clarify whether NEAT1v2 is necessary for IL8 mRNA induction and excess paraspeckle formation, we specifically silenced NEAT1v2 using specific siRNAs (Figures S2A, S2B, and S2E) and found that NEAT1v2 depletion eliminated the induction of IL8 mRNA and excess formation of paraspeckles in response to poly I:C treatment (Figures S2C and S2D). mNeat1v2 is more active than mNeat1v1 in the formation of paraspeckles (Figure S2F). Corresponding to these findings, the effect of mNeat1v2 on gene induction was greater than that of mNeat1v1 (Figure 2E). The induction of IL8 mRNA and the size of the paraspeckles were correlated with the levels of mNeat1v2 overexpression (Figures S2G–S2I). Interestingly, NEAT1 knockdown affected the time point at which peak levels of IL8 mRNA induction were observed following poly I:C treatment (Figure 2D; 5 hr poststimulation in control cells; 3 hr poststimulation in NEAT1 knockdown cells), suggesting that NEAT1 also affects the kinetics of IL8 mRNA induction. The expression of IFN-β, a non-NEAT1-regulated gene, was not affected by the expression level of NEAT1 (Figures 2C and 2E).

Cooperative Action of NEAT1 and SFPQ Regulates *IL8* Transcription

Since paraspeckles contain many transcriptional regulators, such as SFPQ and NONO, it is reasonable to assume that excess formation of paraspeckles would affect the expression of a subset of genes. From this viewpoint, we assumed that certain paraspeckle proteins should regulate the expression of NEAT1-regulated genes such as *IL8*. As expected, we found that the expression of IL8, but not IFN-β, was increased in both SFPQ and NONO knockdown cells, but not in PSPC1 knockdown cells (Figure 3A and Figures S3A and S3L). Both SFPQ and NONO knockdown increased the promoter activity of the *IL8* gene, but not that of the *IFNB1* gene (Figure 3B and Figure S3B). In contrast, promoter activities of the *IL8* and *IFNB1* genes were unchanged in PSPC1-depleted cells (Figure 3B). Bioinformatics analysis revealed that the SFPQ-binding motif is located just 3′ downstream of the TATA box of the human *IL8* promoter and is evolutionarily conserved in the corresponding position of primate *IL8* genes (Figure 3C, Figures S3C–S3E, and Table S3). These findings suggest that SFPQ represses the *IL8* promoter. Notably, the SFPQ-binding motif was predicted with statistical significance in the promoter region of the majority of NEAT1-regulated genes (p value: 0.0015) (Table S3). We then employed a chromatin immunoprecipitation (ChIP) experiment to examine whether SFPQ binds the *IL8* promoter in vivo and is released upon poly I:C stimulation. ChIP experiment showed that SFPQ bound the SFPQ-binding motif of *IL8* gene in naive cells (Figure 3D). Conversely, SFPQ did not bind the SFPQ-binding motif when stimulated by poly I:C (Figure 3D). We also showed that binding of SFPQ to this motif decreased in cells transfected with mNeat1v2 expression plasmid (Figure 3E). We detected concomitantly increased binding of NEAT1v2 to SFPQ in response to poly I:C (Figure 3F). Consistent with this observation, the concentrations of SFPQ and NONO within enlarged paraspeckles were increased upon poly I:C treatment (Figures S3F and S3G). We performed kinetic and dose-response analyses of SFPQ binding to NEAT1v2 after poly I:C exposure. The data showed correlations among poly I:C stimulation, SFPQ binding to NEAT1v2, and IL8 mRNA induction (Figures S3H and S3I). These results suggest that SFPQ binds the SFPQ-binding motif of the *IL8* gene, thereby repressing *IL8* transcription in naive cells, and that poly I:C treatment relocates SFPQ from the *IL8* gene to NEAT1, resulting in the formation of excess paraspeckles, which in turn leads to the transcriptional

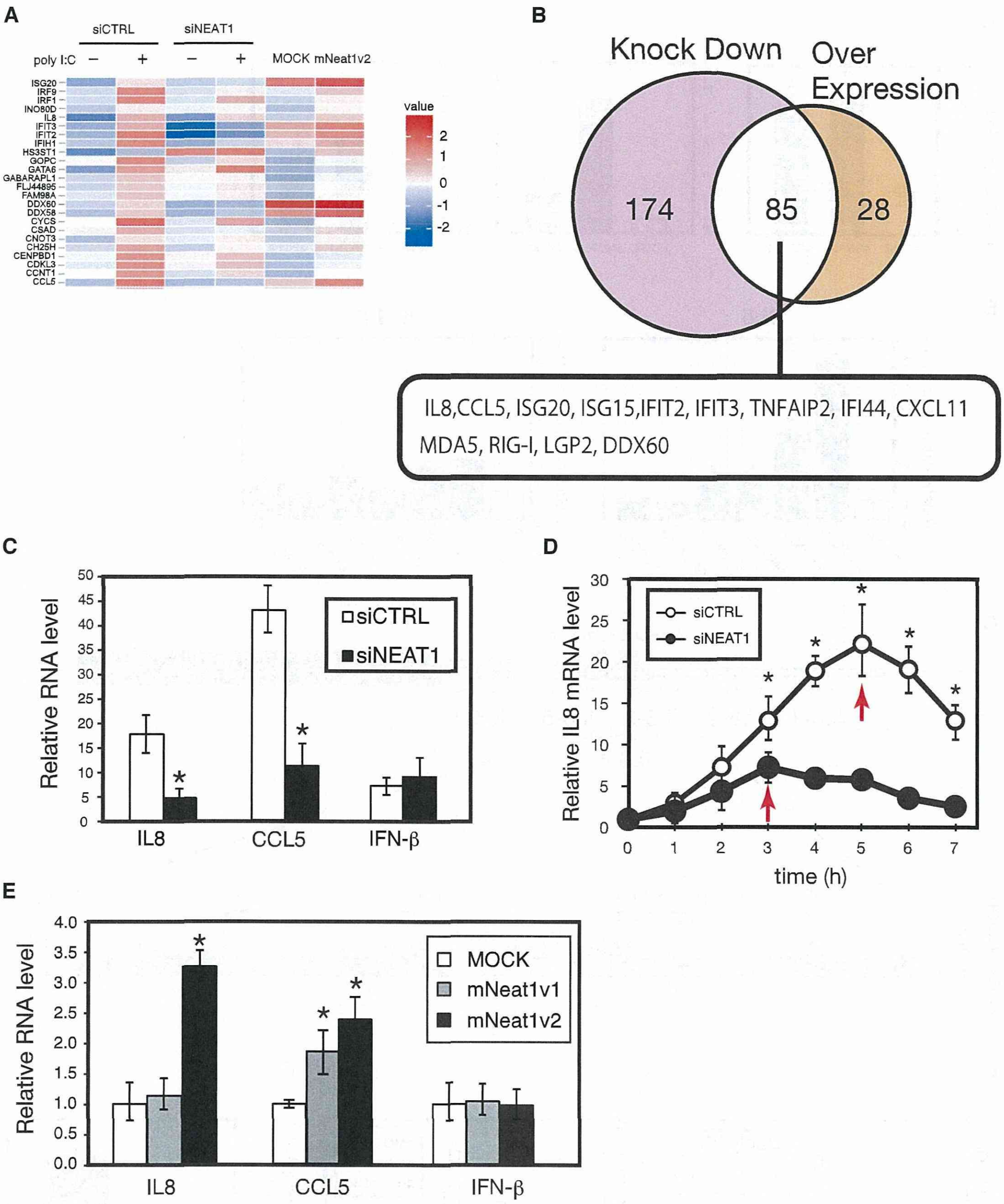

Figure 2. NEAT1-Regulated Genes

(A) Heat map image of microarray analysis of gene expression in the control cells with and without poly I:C stimulation, NEAT1 knockdown cells with and without poly I:C stimulation, and cells transfected with mock plasmid or mNeat1v2 expression plasmid alone.

(B) Venn diagram of genes with altered gene expression as identified by microarray analysis. Left circle contains the 259 poly I:C-induced genes whose induction was abolished by NEAT1 knockdown. Right circle contains the 113 genes induced by solo overexpression of mNeat1v2. Representative genes found in the overlap between these two groups are shown below.

(C) The relative mRNA levels of IL8, CCL5, and IFN-β in the poly I:C-treated cells and the nontreated cells as determined by RT-qPCR analysis. Values represent the mean ± SD (*p < 0.01, Student's t test).

(D) Induction kinetics of IL8 mRNA in control cells or NEAT1 knockdown cells after poly I:C stimulation.

(E) Relative mRNA levels of IL8, CCL5, and IFN-β in cells transfected with pCMV-mNeat1v1 or pCMV-mNeat1v2 compared with cells that have undergone mock transfection, as determined by RT-qPCR analysis. Values represent the mean ± SD (*p < 0.01, Student's t test).

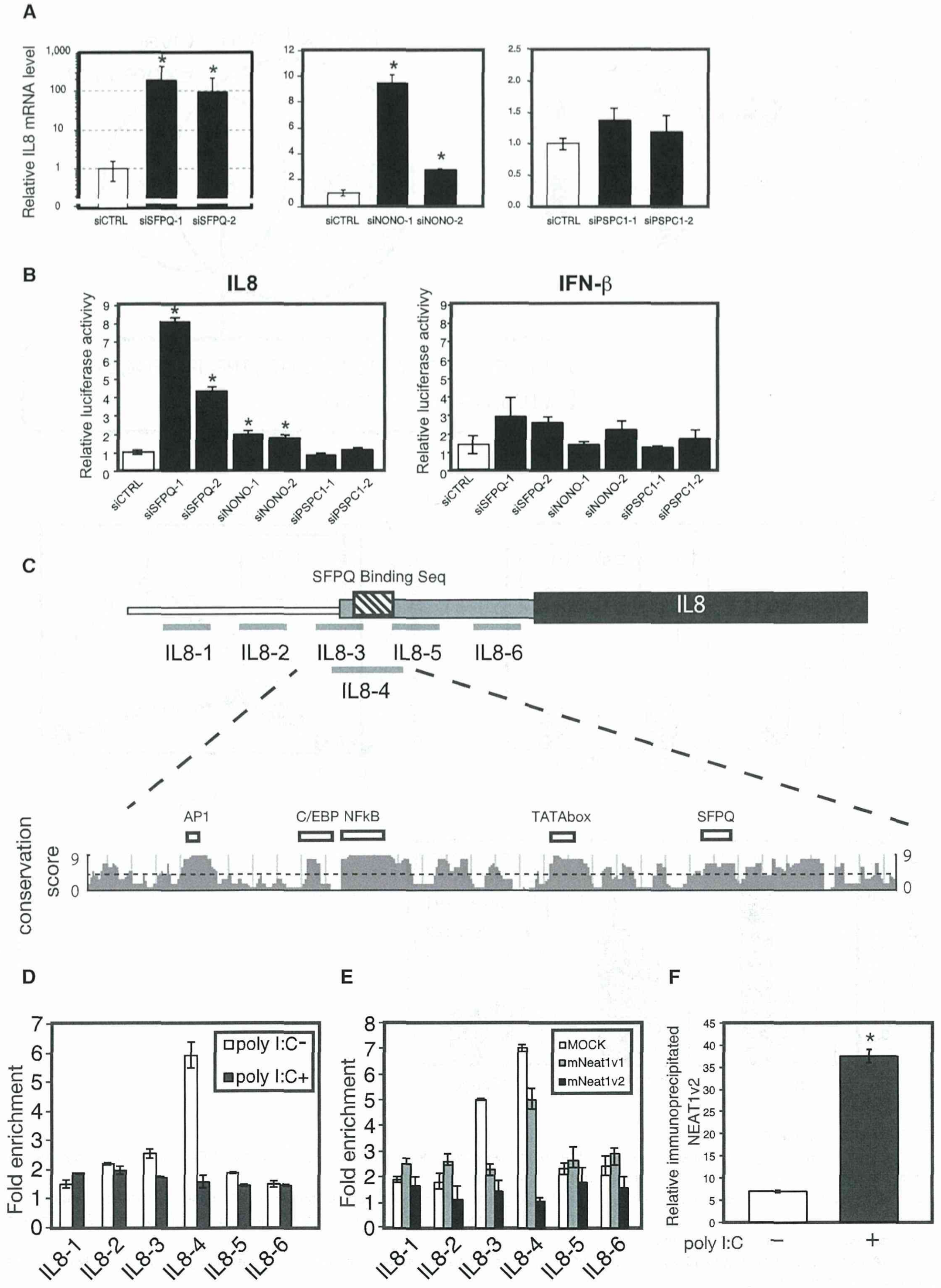

Figure 3. SFPQ-Mediated Transcriptional Repression of the *IL8* Gene and NEAT1-Mediated SFPQ Relocation

(A) IL8 mRNA levels of HeLa TO cells treated with various siRNAs as indicated. Values represent the mean ± SD (*p < 0.01, Student's t test).

(B) Luciferase reporter activities driven by the *IL8* promoter or the *INFB1* promoter were determined in cells transfected with the indicated siRNAs. Values represent the mean ± SD (*p < 0.01, Student's t test).

(legend continued on next page)