JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Communication

pubs.acs.org/JACS

Chemoproteomic Discovery of Cysteine-Containing Human Short Open Reading Frames

Adam G. Schwaid,[†,⊥] D. Alexander Shannon,[‡,⊥] Jiao Ma,[†] Sarah A. Slavoff,[†] Joshua Z. Levin,[§] Eranthie Weerapana,*[,‡] and Alan Saghatelian*[,†]

[†]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States
[‡]Department of Chemistry, Merkert Chemistry Center, Boston College, Chestnut Hill, Massachusetts 02467, United States
[§]Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States

Supporting Information

ABSTRACT: The application of ribosome profiling and mass spectrometry technologies has recently revealed that the human proteome is larger than previously appreciated. Short open reading frames (sORFs), which are difficult to identify using traditional gene-finding algorithms, constitute a significant fraction of unknown protein-coding genes. Thus, experimental approaches to identify sORFs provide invaluable insight into the protein-coding potential of genomes. Here, we report an affinity-based approach to enrich and identify cysteine-containing human sORF-encoded polypeptides (ccSEPs) from cells. This approach revealed 16 novel ccSEPs, each derived from an uncharacterized sORF, demonstrating its potential for discovering new genes. We validated expression of a SEP from its endogenous RNA, and demonstrated the specificity of our labeling approach using synthetic SEP. The discovery of additional human SEPs and their conservation indicate the potential importance of these molecules in biology.

Short open reading frame (sORF)-encoded polypeptides (SEPs) are an emerging class of biomolecules that are comprised of peptides and small proteins from sORFs (defined here as <150 codons).[1] The existence of these molecules is of interest because they appear to be present in a variety of different cells[1,2] and organisms[3,4] but are missed by traditional gene-finding algorithms.[5] The discovery of these molecules has already revealed a great deal about protein translation in cells.[1,2,6,7] Ribosome profiling[2] and mass spectrometry discovery of sORFs,[1,2,7] for example, revealed the prevalent use of non-ATG start codons.

Genetic screens have also identified several bioactive protein-producing sORFs.[4] The search for genes that prevent cell death, for instance, led to the discovery of a 75-base pair sORF that inhibits apoptosis of neuronal cells. It was shown that this sORF produces a 24-amino acid (aa) peptide[4] called humanin that binds and inhibits BAX,[8] revealing a new endogenous molecule with a role in cell death. The complete extent of SEPs in the human genome is unknown—there may be additional bioactive peptides and small proteins awaiting discovery.

SEPs are difficult to predict with traditional gene annotation algorithms due to their small size.[3] Additionally, SEPs have been shown to violate several canonical rules of protein translation. They often initiate with non-ATG start codons, and some have been shown to be bicistronic.[1,2] The recent discovery of this hidden proteome by ribosome profiling[2] and mass spectrometry[1] has generated intense interest toward identifying additional SEPs.

In order to identify additional SEPs, and also to discover SEPs that have properties similar to those of functional proteins, making them more likely to be functional, we applied a cysteine affinity enrichment approach to identify novel cysteine-containing SEPs (ccSEPs). Reactive cysteines play a variety of critical roles in protein structure and function. In particular, cysteines are important catalytic residues in the active site of many enzymes.[9] Furthermore, cysteine oxidation to sulfenic, solfinic, and sulfonic acid in addition to *S*-nitrosylation are important post translational modifications.[10] For example, *S*-nitrosylation on histone deacetylase 2 (HDAC2) was found to induce chromatin remodeling in neurons.[11] Lastly, cysteines are important metal chelators and are found in the metal binding site of many metalloproteins. The incorporation of metal ions in metalloproteins is important for metalloprotein folding and also stabilizes metalloprotein secondary structure.[12−14] The ability of metal binding cysteines to stabilize the secondary structure of proteins is particularly interesting in the case of SEPs. Short proteins are intrinsically more disordered so SEPs that contain metal binding cysteines are more likely to be structured and consequently more likely to be functional.[15,16] In addition to selecting for cysteines that may be amenable to further functional characterization, by using a different strategy to enrich the peptidome, we anticipate the discovery of novel ccSEPs.

Our strategy began with isolating the peptidome from K-562 cells, a human leukemia cell line, by lysis of these cells followed by a molecular weight cutoff (MWCO) filter to remove proteins larger than 30 kDa (Figures 1 and S1).[1] We incubated the peptidome with a previously described iodoacetamide (IA)-alkyne) probe[17,18] that reacts with the sulfhydryl side chain of cysteine to form a covalent bond to the peptide. Notably, when used at 100 μM concentrations the IA-alkyne probe will only label reactive cysteines.[18] After cysteine capture by IA-alkyne, the probe is conjugated to a biotin-labeled tobacco etch virus

Received: June 28, 2013
Published: October 23, 2013

   dx.doi.org/10.1021/ja406606j | J. Am. Chem. Soc. 2013, 135, 16750−16753

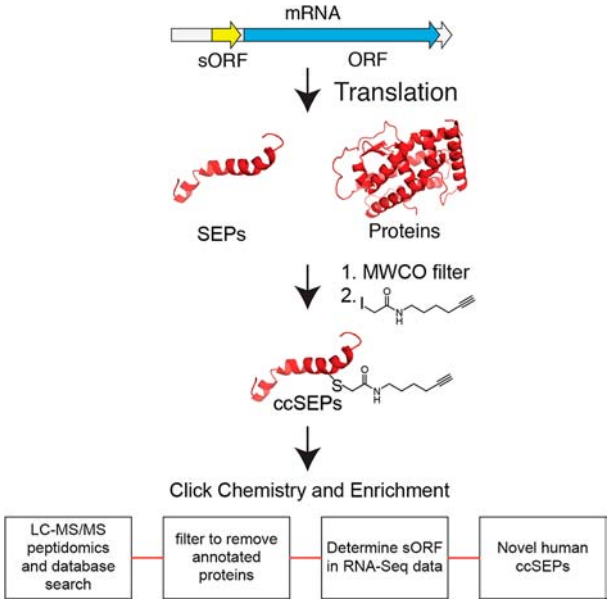

Figure 1. Workflow for identifying cysteine-containing SEPs. The proteome and peptidome are separated by a MWCO filter, and the peptidome fraction is carried forward to identify ccSEPs. Incubation of the peptidome with an iodoacetamide-alkyne probe leads to alkylation of cysteine-containing peptides including ccSEPs. Labeled peptides are then selectively enriched by conjugation to an azide-TEV-biotin tag using copper-activated click chemistry followed by affinity chromatography with streptavidin-coated beads. This sample is then analyzed by LC-MS/MS peptidomics and filtered to remove annotated proteins, which leads to the identification of novel protein-generating sORFs that produce ccSEPs.

(TEV) recognition peptide through copper-activated click chemistry (CuACC).[17−19] Probe-labeled peptides are then separated from unlabeled peptides via streptavidin affinity chromatography to afford an enriched peptidome sample. On-bead trypsin digestion was performed, and unlabeled peptides were eluted and analyzed by offline electrostatic hydrophilic repulsion liquid chromatography (ERLIC) fractionation followed by LC-MS/MS.[1,20] The remaining bead-bound labeled peptides were subsequently released from the beads by the addition of TEV protease and analyzed by MudPIT-LC-MS/MS.[21]

The data from this peptidomics analysis contain known as well as novel (i.e. non-annotated) peptides, including ccSEPs. In order to identify peptides originating from non-annotated RNAs, we used a custom database using K-562 RNA-Seq data,[1,22] which contains information on the vast majority of mRNAs in K-562 cells. Since these RNAs must be the source of any polypeptide produced we can include non-annotated genes in our peptidomics search by translating this database in three frames to generate a protein database that contains all possible peptide products.

We then matched our peptide spectra against this RNA-Seq database to reveal candidate SEPs. This approach yielded 175 hits that surpassed our preliminary cross-correlation score requirements.[17] After removing annotated peptides, we were left with 109 candidate SEPs. Our K-562 RNA-Seq database was too large to perform a reverse database search directly. To overcome this, we constructed a forward and reversed database by appending our candidate SEPs to the Uniprot database. We used this database to filter our candidate SEP spectra using a reversed database search, and we only accepted peptides with a false discovery rate <0.05. Subsequently, we validated that detected peptides could only originate from a single sORF (i.e., there are not two different ORFs in the RNA-Seq data that could account for the peptide). Additionally, SEPs with more than two missed cleavages were removed, along with SEPs detected from peptides <7 aa in length. Furthermore, spectra were visually inspected to ensure good sequence coverage and confirm that peptides detected from the TEV fraction contained an IA-modified cysteine residue (Figure S2). After this, we were left with 16 novel human ccSEPs (Tables 1 and S1), with the majority having <6 ppm mass error (Table S2).

In cases where a detected peptide contained multiple cysteines, the labeled cysteine could be determined from the MS/MS data (Figure 2A). To verify that our labeling and enrichment are specific to the cysteine on a ccSEP, we performed an *in vitro* assay in cell lysates. We first synthesized TCT-SEP (named for the detected peptide; Figure 2B) by solid-phase peptide synthesis, along with a mutant of this TCT-SEP where the cysteine is replaced by a serine, TST-SEP. We

Table 1. Newly Discovered ccSEPs

detected peptide[a]	start codon	length (aa)	transcript origin	conserved?
C*GFFSYCSSESVSCSTS	ATC	34	non-annotated	no
STSLYCHSTILC*	AAG	24	CDS	no
TC*DGNSNEGGGTR	AAG	19	non-annotated	no
NFPLASSPERC*FFVPK	AAG	48	3′UTR	yes
VEKLELLYIAGGNVNWYSPC*	GTG	22	non-annotated	yes
YPAC*SPSPALI	CTG	29	non-annotated	no
GRGCC*RGFSAVGQGPSST	ATG	84	non-annotated	no
CPSINFQHFCHFVLCAFPIHC*	CTG	35	non-annotated	no
TC*TIPVPAGGRPR	CTG	32	non-annotated	no
IC*DIKGLIDNV	TTG	41	non-annotated	no
TSPADAVC*PGLGRDLCGSSRCCLRP	ATG	79	5′UTR	yes
RGPGEAGMSWEEAGGLAPHLLC*CR	GTG	86	CDS	yes
QIVLGGC*GEMV	alternate	16	non-annotated	no
GASFSEDGC*LLVG	CTG	37	non-annotated	no
GSSDIISVPC*	ATG	40	3′UTR	yes
SSMPLIC*FLILEGLGR	ATG	29	3′UTR	yes

[a]Asterisk denotes labeled cysteine.

dx.doi.org/10.1021/ja406606j | J. Am. Chem. Soc. 2013, 135, 16750−16753

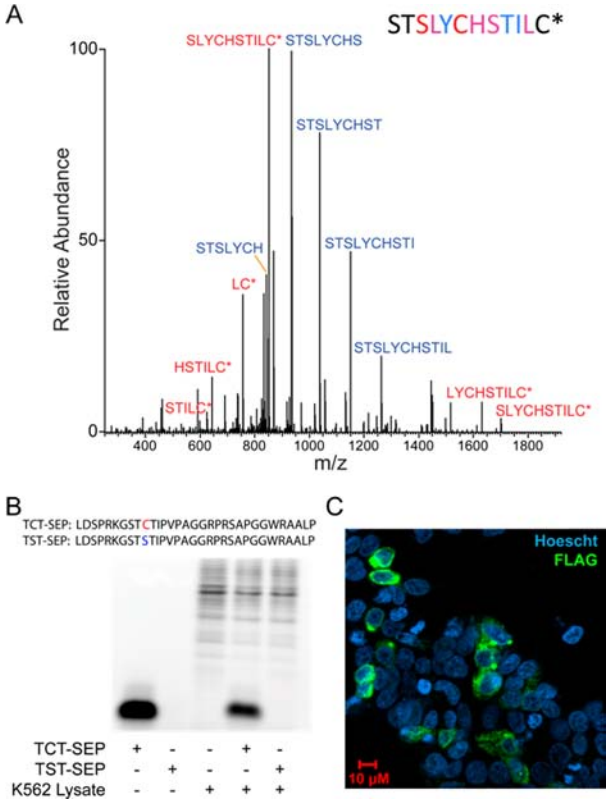

Figure 2. Validation of site of labeling and cellular expression of newly discovered ccSEPs. (A) In the case of ccSEPs with multiple cysteines, examination of the tandem MS spectra reveals the site of labeling. In this case, STS-ccSEP labels at the C terminal cysteine. Red indicates fragments detected by y ions, blue indicates fragments detected by b ions, and purple indicates fragments detected by both. (B) We tested labeling of one of the ccSEPs in a complex mixture by spiking the purified ccSEP into lysate and then performing a labeling reaction with rhodamine azide. If the ccSEP reacted it would be fluorescently labeled. Mutation of the cysteine on the ccSEP to a serine abrogates labeling. (C) A C-terminal Flag tag appended to the sORF coding for TSP-ccSEP validated that this sORF does indeed produce protein. Staining of the protein product with an anti-Flag antibody confirmed expression and cellular stability of the ccSEP.

incubated TCT-SEP in K-562 cell lysates and then added the IA-alkyne probe. After labeling, the lysate was mixed with a fluorescent azide in the presence of copper(II) sulfate and TCEP to promote CuACC. This fluorescently labeled lysate was then resolved on an SDS-PAGE gel to assess labeling of the TCT-SEP. Labeling of TCT-SEP was specific and robust and could be easily observed within total K-562 lysate (Figure 2B). The control TST-SEP was not labeled when probe-treated alone or in K-562 lysate, demonstrating that labeling is occurring on the cysteine residue (Figure S3).

To validate the production of ccSEPs from their endogenous RNA, we transfected cells with a vector containing the sORF TSP-ccSEP, which is found on the same transcript as MRS2L. This construct contained the entire endogenous 5′UTR, which includes the sORF, and a FLAG tag was appended to the sORF to enable easy detection of protein production (Figures 2C and S4). Stable ccSEP expression was then observed by immunofluorescence using an anti-FLAG antibody (green) (Figures 2C and S5) and Western blot (Figure S6). This sORF was not annotated previously, thereby highlighting the ability of this workflow to discover novel protein-coding genes. More generally, this affinity strategy successfully identified a new pool of SEPs with characteristic hallmarks of this emerging class of peptides.[1]

An overview of these newly identified ccSEPs revealed many similarities with previously identified SEPs. First, the length of their sORFs ranged between 16 and 86 codons (Figure 3A).

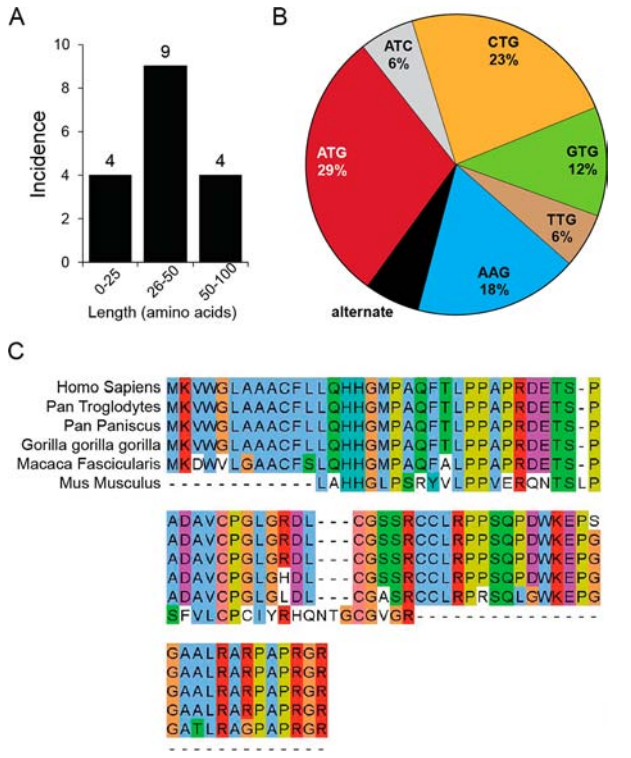

Figure 3. ccSEP overview. (A) Distribution of ccSEPs by their length in amino acids. SEP length was determined using the distance from an upstream in-frame AUG start codon to a downstream in-frame stop codon; when no in-frame AUG was present, a near-cognate start codon or stop codon was used instead. (B) While AUG is the predominant start codon for the production of ccSEPs, near-cognate start codons (i.e., one base different from AUG) are also common. (C) TSP-SEP is strongly conserved among several species of primates, suggesting this SEP may be functional.

SEP length was determined by measuring the number of codons between the stop codon of the sORF and the first start codon on the 5′ side of this stop codon. In the case where a start codon could not be identified, the number of codons reflects the distance between the stop codon of the sORF and the 5′ end of the transcript. Second, these SEPs had both AUG start codons or non-canonical near-cognate start codons (Figure 3B), similar to previously discovered SEPs. Moreover, SEPs could be found in the 3′UTR, frame-shifted within known genes or within the 5′UTR, in non-annotated RNAs, or in antisense transcripts (Supporting Information). As expected, we did not detect any previously observed SEPs, since our workflow was optimized toward the detection of SEPs with reactive cysteines. These identified SEPs are very small relative to the average length of a human protein, which is 335 aa.[23] The small size of these SEPs contributes to the difficulties associated with computationally predicting the sORFs that encode them.

While specific functions for these ccSEPs await future studies, we examined these ccSEPs for sequence conservation, which is

dx.doi.org/10.1021/ja406606j | J. Am. Chem. Soc. 2013, 135, 16750−16753

an important and well-documented signifier of biological function.[24] We examined the conservation of our SEPs in several species by alignment of the translated RNA to *in silico* translated RNA and DNA databases comprising the GenBank, EMBL, DDBJ, PDB, and Refseq sequences. Of the ccSEPs we discovered, over one-third (6/16) are conserved among several species of primates, indicating that they have been maintained throughout evolution and highlighting these ccSEPs as likely having functions. Notably, the cysteine residue that we find labeled by the IA probe is also conserved between species, including mice, despite the low overall sequence conservation across the entire SEP. This implies that this residue may be important for the SEP's biological function (Figures 3C and S7). The conservation of these SEPs makes them good leads for further functional characterization and demonstrates that this platform allows for the identification of peptides that are of significant biological interest.

In summary, we have utilized a chemoproteomics approach to identify new human ccSEPs. These results demonstrate the value of chemoproteomics to promote the discovery of additional sORFs. In this case, we identified 16 novel ccSEPs, indicating the presence of even more of these molecules than had been predicted, and representing a 15% increase in the number of known SEPs. Moreover, conservation indicates that some of these ccSEPs may be functional. Furthermore, cysteine reactivity is governed by secondary structure and local environment, suggesting that enriching ccSEPs with highly reactive cysteines may identify proteins with distinct secondary structures. Additionally, certain biologically important post-translational modifications, such as protein *S*-nitrosylation, occur at, and can be regulated by, redox-active cysteines.[25] Some of these ccSEPs are likely targeted by these oxidative modifications, which could serve to further regulate SEP function. The struggle to identify the whole range of SEPs in human cells as well as their functional role remains a key question in biology. The development of mass spectrometry methods focused on the identification of SEPs, such as chemoproteomic approaches, is a critical step toward answering these questions.

■ ASSOCIATED CONTENT

Supporting Information

Experimental details. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

eranthie.weerapana@bc.edu
saghatelian@chemistry.harvard.edu

Author Contributions

⊥A.G.S. and D.A.S. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Xian Adiconis and Lin Fan for constructing the cDNA libraries used in this study. S.A.S. is supported by an NRSA postdoctoral fellowship (1F32GM099408-01). This work was supported by a Eli Lilly graduate fellowship (A.G.S), an NIH grant R01GM102491 (A.S.), a US National Human Genome Research Institute grant HG03067 (J.Z.L), the Damon Runyon Cancer Research Foundation (grant DRR-18-12, E.W.), and the Smith Family Foundation (E.W.).

■ REFERENCES

(1) Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M.; Ma, J.; Levin, J. Z.; Budnik, B.; Rinn, J. L. S. *Nat. Chem. Biol.* **2012**, *9*, 59.
(2) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. *Cell* **2011**, *147*, 789.
(3) Galindo, I. G.; Pueyo, J. I.; Fouix, S.; Bishop, S. A.; Couso, J. P. *PLOS Biol* **2007**, *5*, 1052.
(4) Hashimoto, Y.; Niikura, T.; Tajima, H.; Yasukawa, T.; Sudo, H.; Ito, Y.; Kita, Y.; Kawasumi, M.; Kouyama, K.; Doyu, M.; Sobue, G.; Koide, T.; Tsuji, S.; Lang, J.; Kurokawa, K.; Nishimoto, I. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 6336.
(5) Frith, M. C.; Forrest, A. R.; Nourbakhsh, E.; Pang, K. C.; Kai, C.; Kawai, J.; Carninci, P.; Hayashizaki, Y.; Bailey, T. L.; Grimmond, S. M. *Proteins* **2006**, 2.
(6) Lee, S.; Liu, B.; Lee, S.; Huang, S.; Shen, B.; Qian, S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, 109.
(7) Stern-Ginossar, N.; Weisburd, B.; Michalski, A.; Le, V. T.; Hein, M. Y.; Huang, S. X.; Ma, M.; Shen, B.; Qian, S. B.; Hengel, H.; Mann, M.; Ingolia, N. T.; Weissman, J. S. *Science* **2013**, *338*, 1088.
(8) Guo, B.; Zhai, D.; Cabezas, E.; Welsh, K.; Nouraini, S.; Satterthwait, A. C.; Reed, J. C. *Nature* **2003**, *423*, 456.
(9) Chapman, H. A.; Riese, R. J.; Shi, G.-P. *Annu. Rev. Physiol.* **1997**, *59*, 63.
(10) Reddie, K. G.; Carroll, K. S. *Curr. Opin. Chem. Biol.* **2008**, *12*, 746.
(11) Nott, A.; Nitarska, J.; Veenvliet, J. V.; Schacke, S.; Derijck, A. A. H. A.; Sirko, P.; Muchardt, C.; Pasterkamp, R. J.; Smidt, M. P.; Riccio, A. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 3113.
(12) Ikeguchi, M.; Kuwajima, K.; Sugai, S. *J. Biochem.* **1986**, *99*, 1191.
(13) Coyne, H. J., III; Ciofi-Baffoni, S.; Banci, L.; Bertini, I.; Zhang, L.; George, G. N.; Winge, D. R. *J. Biol. Chem.* **2007**, *282*, 8926.
(14) Morleo, A.; Bonomi, F.; Iamentti, S.; Huang, V. W.; Kurtz, D., Jr *Biochemistry* **2010**, *49*, 6627.
(15) Scholtz, J. M.; Baldwin, R. L. *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 95.
(16) Kozlowski, H.; Bal, W.; Dyba, M.; Kowalik-Jankowska, T. *Coord. Chem. Rev.* **1999**, *184*, 319.
(17) Weerapana, E.; Speers, A. E.; Cravatt, B. F. *Nat. Protoc.* **2007**, *2*, 1414.
(18) Weerapana, E.; Wang, C.; Simon, G.; Richter, F.; Khare, S.; Dillon, M. B. D.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F. *Nature* **2010**, *468*, 790.
(19) Wu, P. F., A. K.; Nugent, A. K.; Hawker, C. J.; Scheel, A.; Voit, B.; Pyun, J.; Fréchet, J. M. J.; Sharpless, B. K.; Fokin, V. V. *Angew. Chem., Int. Ed* **2004**, *43*, 3928.
(20) Alpert, A. *Anal. Chem.* **2008**, *80*, 62.
(21) Washburn, M.; Wolters, D.; Yates, J. *Nat. Biotechnol.* **2001**, *19*, 242.
(22) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; World, B. *Nature Methods* **2008**, *5*, 621.
(23) Ota, T.; Suzuki, Y.; Nishikawa, T.; Otsuki, T.; Sugiyama, T.; Irie, R.; Wakamatsu, A.; Hayashi, K.; Sato, H.; Nagai, K.; al., e. *Nat. Genet.* **2004**, *36*, 40.
(24) Ponjavic, J.; Ponting, C.; Lunter, G. *Genome Res.* **2007**, *17*, 556.
(25) Hess, D. T.; Matsumoto, A.; Kim, S.; Marshall, H. E.; Stamler, J. S. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 150.