UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA

Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

DINAR YUNUSOV

Caracterização do *HIPSTR* destaca o padrão de expressão heterogênea de lncRNAs em embriões humanos e linhagens estáveis de células

Versão corrigida da Tese defendida

São Paulo

20/05/2016

DINAR YUNUSOV

Characterization of *HIPSTR* highlights the heterogeneous expression pattern of lncRNAs in human embryos and stable cell lines

Tese apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioquímica)

Orientador: Prof. Dr. Sergio Verjovski-Almeida

São Paulo

2016

Ficha Catalográfica
Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP.

Yunusov, Dinar
Y95e Characterization of HIPSTR highlights the heterogeneous expression pattern of IncRNAs in human embryos and stable cell lines. / Dinar Yunusov. -- São Paulo, 2016.
89p.

Tese (doutorado) Instituto de Química da Universidade de São Paulo. Departamento de Bioquímica.
Orientador: Verjovski-Almeida, Sergio

1. Expressão gênica I. T. II. Verjovski-Almeida, Sergio, orientador.

574.88 CDD

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

"Caracterização do *HIPSTR* destaca o padrão de expressão heterogênea de lncRNAs em embriões humanos e linhagens estáveis de células"

DINAR YUNUSOV

Tese de Doutorado submetida ao Instituto de Química da Universidade de São Paulo como parte dos requisitos necessários à obtenção do grau de Doutor em Ciências no Programa de Ciências Biológicas (Bioquímica) - Área de Concentração: Bioquímica.

Aprovado (a) por:

Prof. Dr. Sergio Verjovski de Almeida
(Orientador e Presidente)

Profa. Dra. Carla Columbano de Oliveira
IQ - USP

Profa. Dra. Bettina Malnic
IQ - USP

Profa. Dra. Lygia da Veiga Pereira Carramaschi
IB - USP

Profa. Dra. Chao Yun Irene Yan
IB - USP

SÃO PAULO
10 de junho de 2016

DEDICATION

This modest research work is dedicated to my family

ACKNOWLEDGEMENTS

I would like to acknowledge the immense contribution of Prof. Sergio Verjovski-Almeida into the present work. Without his help, patience and dedication, my stay in Brazil would be much harder, if not absolutely impossible. I will always be grateful to him for being one of the few best friends on whom I could always rely, for the incredible support and excellent guidance. His impact on my life I am not able to describe within this tiny paragraph.

I would like to say special thanks to my colleagues and simply friends Felipe and Ana Ayupe aka casal Beckedorff, Letícia, Carlos, Lauren Camargo, Anya, Ludmila, Lucas, Bianca, Alexey and Nastya, and Ana Paula for everything they have done for me, all the help through the years of my PhD studentship, and again – patience with my stubbornness and sometimes – plain stupidity and arrogance. I can only wonder what I did to deserve your kindness, loyalty, and friendship.

I would like to address my best wishes to my close collaborators Prof. R. Michael Roberts, Prof. Toshihiko Ezashi, Prof. Andrei Alexenko and Kate Shipova, and Prof. Eduardo Reis. Additional thanks to Prof. Joanna Wysocka for giving me an eye-opening experience of working in one of the best labs in the world. I also would like to thank Prof. Carla Columbano, Prof. Solange Serrano and Eduardo Kitano, and my friend Maíra Nagai for their help with the experiments.

Thanks to Ângela and Ana Tahira for their incredible company in the lab on many late nights. Thanks to Adriana for being an excellent billiards partner. Thanks to all my colleagues, especially to Mariana, Santiago, Renan, Yuri, Katia, Murilo, and Jefferson for welcoming me to the lab and for creating a great environment to work in. Thanks to Marina and Lauren Ragel for much more than just inspiration and great memories.

Thanks to Fundação de Amparo e Pesquisa do Estado de São Paulo – FAPESP for making this work financially possible by awarding grants #2010/51152-7 and # 2015/01397-7.

Finally, I would like to thank my family. I do not know, and I doubt I will ever get to know more loving and caring people in this world. Like real friends, you were always there, so modest and patient. I would like to thank my father Shamil Yunusov for being the best father one can ever wish for, for me you will always remain the most successful person I have ever known. I would like to thank my mother Venera Yunusova for never giving up – stay strong, you know how to do this better than anybody! I would like to thank Rifhat Shakurov for all this naïve carelessness I inherited, for all the paintings and sincere love. I would like to thank Illuza Shakurova for all the years of endless energy, and irrational support in all crazy ideas I ever had, *davani*, you rock! I would like to thank Lilya Shakurova for all that seemingly invisible, but incredible and irreplaceable help and protection.

RESUMO

Yunusov, D. **Caracterização do *HIPSTR* destaca o padrão de expressão heterogênea de lncRNAs em embriões humanos e linhagens estáveis de células.** 2016. 87 p. Tese (Doutorado) – Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

Tem sido cada vez mais reconhecido que a transcrição dos genomas eucarióticos produz múltiplos transcritos novos, anteriormente não detectados e ainda não caracterizados, sendo que a maioria é constituida de RNAs não-codificantes longos (lncRNAs) regulatórios. Estudos recentes estão focados principalmente nos lncRNAs transcritos de regiões intergênicas e *enhancers*; assim, o grupo dos lncRNAs antisenso permanece o menos estudado de todos. Ao mesmo tempo, a transcrição antisenso ocorre em até 74% dos *loci* de genes humanos, frequentemente – a partir da fita oposta de genes que codificam proteínas envolvidas na regulação da transcrição. No presente trabalho, nós identificamos *HIPSTR* (*Heterogeneously expressed from the Intronic Plus Strand of the* TFAP2A-*locus RNA*), um lncRNA novo conservado que é transcrito a partir da fita antisenso do gene TFAP2A. Ao contrário do anteriormente relatado para os lncRNAs antisenso, a expressão de *HIPSTR* não está correlacionada com a expressão do gene da fita oposta. *HIPSTR* e *TFAP2A* são co-expressos em células da crista neural e em trofoblastos derivadas *in vitro*, mas somente *HIPSTR* e não *TFAP2A* está especificamente expresso num subconjunto de células de embriões humanos nos estágios de 8-células e mórula. Mostramos que, semelhante a *HIPSTR*, a expressão de lncRNAs é mais altamente heterogênea que a expressão de mRNAs em células individuais de embriões humanos em desenvolvimento ou em linhagens estáveis de células. Finalmente, nós demonstramos que a depleção de *HIPSTR* em células HEK293 e $H1_{BP}$, uma linhagem de células tronco embrionárias humanas, afeta predominantemente os níveis de genes envolvidos no início do desenvolvimento do organismo e na diferenciação de células. No conjunto, nós mostramos que a expressão de *HIPSTR* e de centenas de outros lncRNAs é altamente heterogênea em embriões humanos e linhagens celulares. Usamos *HIPSTR* para exemplificar a relevância funcional de lncRNAs com padrões de expressão heterogêneos e estágio-de-desenvolvimento específicos.

Palavras chave: RNAs longos não-codificadores, RNAs antisenso, *TFAP2A*, desenvolvimento embrionário, variabilidade da expressão em células individuais.

ABSTRACT

Yunusov, D. **Characterization of *HIPSTR* highlights the heterogeneous expression pattern of lncRNAs in human embryos and stable cell lines.** 2016. 87 p. PhD Thesis – Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

There is a growing appreciation that eukaryotic genomes are transcribed into numerous, previously undetected – and thus uncharacterized regulatory long non-coding RNAs (lncRNAs). Recent studies are primarily focused on lncRNAs transcribed from intergenic regions and enhancers, leaving antisense lncRNAs the least studied group of lncRNAs. At the same time, antisense transcription occurs in up to 74 % of human gene loci, frequently – from the opposite strand of genes encoding proteins involved in regulation of transcription. Here, we identified *HIPSTR* (Heterogeneously expressed from the Intronic Plus Strand of the TFAP2A-locus RNA), a novel conserved lncRNA that is transcribed antisense to the *TFAP2A* gene. Unlike previously reported antisense lncRNAs, *HIPSTR* expression does not correlate with the expression of its antisense counterpart. Although *HIPSTR* and *TFAP2A* are co-expressed in *in vitro* derived neural crest and trophoblast cells, only *HIPSTR* and not *TFAP2A* is specifically expressed in a subset of cells within 8-cell- and morula-stage human embryos. We show that, similar to *HIPSTR*, in the individual cells of developing human embryos or of stable cell lines the expression of lncRNAs is more highly heterogeneous than the expression of mRNAs. Finally, we demonstrate that *HIPSTR* depletion in HEK293 and $H1_{BP}$, a human embryonic stem cell line, predominantly affects the expression levels of genes involved in early organismal development and cell differentiation. Together, we show that expression of *HIPSTR* and hundreds other lncRNAs is highly heterogeneous in human embryos and cell lines. We use *HIPSTR* to exemplify the functional relevance of lncRNAs with heterogeneous and developmental stage-specific expression patterns.

Keywords: long non-coding RNAs, antisense RNAs, *TFAP2A*, early embryonic development, single-cell expression variability

LIST OF ABBREVIATIONS

ASO:	Antisense oligonucleotide
ATRA:	All-*trans* retinoic acid
ChIP:	Chromatin Immunoprecipitation
EGF:	Epidermal growth factor
ESCs:	Embryonic stem cells
FBS:	Fetal bovine serum
FGF2:	Fibroblast growth factor 2
FPKM:	Fragments per kilobase of transcript per million mapped reads
GO:	Gene ontology
iPSCs:	Induced pluripotent stem cells
LncRNA:	Long non-coding RNA
MEF:	Mouse embryonic fibroblast
NCCs:	*in vitro* derived neural crest cells
ORF:	Open reading frame
PRC2:	Polycomb Repressive Complex 2
RNA:	Pol II RNA Polymerase II
TBCs:	*in vitro* derived trophoblast cells
TF:	Transcription factor
TSS:	Transcription start site

Table of contents

1. Introduction

1.1 Pervasive eukaryotic transcription

It is now widely accepted that eukaryotic genomes are pervasively transcribed (Berretta and Morillon, 2009; Clark *et al.*, 2011; Djebali *et al.*, 2012), producing thousands of uncharacterized transcripts, the majority of which are classified as long non-coding RNAs (lncRNAs) (for human cells, the most complete catalog is published in (Iyer *et al.*, 2015)). A story of universal obsession with regulatory lncRNAs begins in the year 1991 with the discovery of *XIST* lncRNA (Borsani *et al.*, 1991; Brown *et al.*, 1991; Brockdorff *et al.*, 1992; Brown *et al.*, 1992), and this story is far from its happy ending. Nowadays, more than two decades after the discovery of *XIST*, we are still struggling to identify a complete set of proteins that interact with this lncRNA (Chu *et al.*, 2015; McHugh *et al.*, 2015).

LncRNAs are simply defined as long (> 200 nt) non-protein coding transcripts, and as such they represent a very broad, widely uncharacterized group that includes non-functional transcripts resulting from transcriptional noise (random transcription initiation by RNA Pol II throughout the genome), and lncRNAs exerting their function either passively through the act of their transcription, or actively in *cis* and in *trans* (Quinn and Chang, 2015). Similar to mRNAs in many aspects, such as transcription by RNA Pol II, presence of 5'-cap and poly(A)-tail (Guttman *et al.*, 2009; Ayupe *et al.*, 2015), lncRNAs are usually shorter than mRNAs and have fewer exons (Derrien *et al.*, 2012). Consistent with transcription of lncRNA genes by RNA Pol II, they are often marked by trimethylation of lysine 4 of histone 3 (H3K4me3) in their promoter regions, and by trimethylation of lysine 36 of histone 3 (H3K36me3) in their gene bodies (Guttman *et al.*, 2009). Notably, for a group of human lncRNA genes, such H3K4me3-H3K36me3 demarcation was evolutionarily preserved in orthologous regions of the mouse genome, and this was proposed to serve as one of the possible indicators of conserved functionality (Guttman *et al.*, 2009).

INTRODUCTION

While important for the functionality of protein-coding genes, sequence conservation is only modestly pronounced in lncRNAs (Pang *et al.*, 2006; Cabili *et al.*, 2011), with the conservation of genomic position apparently being predominant instead (Carninci *et al.*, 2005). For example, well characterized lncRNAs, such as *XIST*, *Cyrano/OIP5-AS1*, *MIAT*, *TUNAR*, and *HOTAIR* have conserved function even in the absence of broad sequence conservation (Li, L. *et al.*, 2013a; Ulitsky and Bartel, 2013; Kapusta and Feschotte, 2014). Nonetheless, sequence conservation of lncRNA exons was reported in several studies, and is stronger than conservation of intergenic regions or lncRNA introns, with the greatest evolutionary pressure near splice sites (Guttman *et al.*, 2009; Chodroff *et al.*, 2010). Additionally, studies of single nucleotide polymorphisms in primate-specific lncRNA exons showed lower derived allele frequencies than those from intergenic regions (Necsulea *et al.*, 2014). For the oldest lncRNAs, conservation of exonic sequences approaches that of coding exons (Necsulea *et al.*, 2014), while lncRNA promoter sequences are equally (Necsulea *et al.*, 2014; Chen, J. *et al.*, 2016), or even more conserved (Carninci *et al.*, 2005) than promoter sequences of protein-coding genes, depending on the set of lncRNAs used for analysis. The latter observation stands true even for younger lncRNAs (Necsulea *et al.*, 2014).

When compared to mRNAs, lncRNAs are expressed at lower levels with considerably higher organ-, tissue- and developmental stage specificity (Ravasi *et al.*, 2006; Cabili *et al.*, 2011; Derrien *et al.*, 2012; Yan *et al.*, 2013; Necsulea *et al.*, 2014), although tissue-specificity of expression is weakly preserved for orthologous lncRNAs across eutherians (Necsulea *et al.*, 2014). To explain evolutionarily preserved functionality of lncRNAs, four dimensions of the lncRNA conservation were proposed, and include conservation of the sequence, structure, function and transcription from a syntenic region (Diederichs, 2014). In other words, conservation of genomic position, structure and sequence could be considered as good indicators of conserved functionality of a given lncRNA, but are not necessarily required.

1.2 LncRNAs in development and cell differentiation

Among other processes the importance of lncRNAs is clearly shown for organismal development and cell differentiation. This section briefly outlines several major features and functions of well-characterized lncRNAs in these processes.

In 2007, Rinn *et al.* used primary adult fibroblasts from 11 anatomic sites to study expression patterns of *HOX* genes, and discovered 231 lncRNAs transcribed in *HOX* loci, of which 64% were differentially expressed along the developmental axis of the body (Rinn *et al.*, 2007). A specific example, that of *HOTAIR* lncRNA, was used to show that such lncRNAs are capable of establishing mutually exclusive domains of silent and active chromatin in *HOX* loci by recruiting Polycomb Repressive Complex 2 (PRC2) to its target genes (Rinn *et al.*, 2007). Later, in 2010, the same group showed that lncRNAs in *HOX* loci become dysregulated in breast cancer, leading to genome-wide alteration in PRC2 binding profile with consequent increase in PRC2-dependent invasiveness of cancer cells (Gupta *et al.*, 2010).

Following the discovery of *HOTAIR* lncRNA and its interaction with PRC2, the latter was shown to interact with *XIST* (Zhao *et al.*, 2008). In 2004, yeast telomerase was used as an example to propose a model where lncRNAs function as modular scaffolds for protein complexes (Zappulla and Cech, 2004). This was later demonstrated for *Kcnq1ot1* that interacts with H3K9-specific histone methyltransferase G9a and PRC2 (Pandey *et al.*, 2008), and for *HOTAIR* that binds to PRC2 via its 5'-domain, and to LSD1/CoREST/REST complex – through its 3'-domain (Tsai *et al.*, 2010). This concept is further developed and explained in a landmark review by Guttman and Rinn (Guttman and Rinn, 2012). Interactions of lncRNAs with chromatin modifiers appear to be another common feature of lncRNAs as thousands of lncRNAs were shown to interact with PRC2 in mouse embryonic stem cells (ESCs) (Zhao *et al.*, 2010). Although the fidelity of PRC2-RNA interactions is still a subject of active debate

(Davidovich *et al.*, 2013; Cifuentes-Rojas *et al.*, 2014), it is clear that at least some of these interactions are not promiscuous (Davidovich *et al.*, 2015).

In agreement with the ability of lncRNAs to act as scaffolds for proteins, it is not surprising that interactions of lncRNAs with various transcription factors (TFs) were demonstrated. For example, *PAUPAR* lncRNA co-regulates a subset of its target genes in association with PAX6 (Vance *et al.*, 2014), *RMST* lncRNA co-operates with SOX2 to recruit the latter to promoters of neurogenic TFs (Ng *et al.*, 2013), and *PANDAR* lncRNA interacts with NF-YA and regulates senescence (Puvvula *et al.*, 2014). LncRNAs are intimately linked to important TFs not only by physical interactions with them, but also by genomic co-localization, and often – by regulating their expression. For instance, not only *PAUPAR* lncRNA associates with PAX6, but also regulates *PAX6* gene that is located in the vicinity (Vance *et al.*, 2014). Another example is *utNgn1* lncRNA that is required for the expression of the downstream *Neurog1* gene during neuronal differentiation (Onoguchi *et al.*, 2012). General tendency is that gene deserts surrounding genes of developmentally regulated TFs harbor multiple lncRNA genes in human, mouse and zebrafish, but the functional relevance of these lncRNAs remains mostly unstudied (Ulitsky and Bartel, 2013).

LncRNAs are tightly incorporated into networks regulating pluripotency and differentiation. By comparing transcriptomes of human ESCs and neuronal cells derived from them *in vitro*, lncRNAs required for pluripotency and neurogenesis were identified (Ng *et al.*, 2012). Similar approach in mouse ESCs led to identification of *TUNA*, whose sequence and CNS-restricted expression pattern is conserved in vertebrates (Lin *et al.*, 2014). Similarly, a pioneering study of lncRNAs expression in the process of reprogramming of human fibroblasts into induced pluripotent stem cells (iPSCs) showed regulated activation of hundreds of pluripotency-associated lncRNAs (Loewer *et al.*, 2010). During reprogramming, lncRNAs are activated during reprogramming in a dynamic fashion, downregulating lineage-

specific genes and modulating expression of metabolic genes (Kim, D. H. *et al.*, 2015). Concordantly, in mouse ESCs, lncRNAs were shown to regulate gene expression in *trans*, maintain pluripotency by repressing differentiation programs, and by acting downstream of ESC-specific TFs, likely through interaction with numerous chromatin-modifying enzymes (Guttman *et al.*, 2011).

1.3 Expression of lncRNAs and mRNAs at the level of single cells

In Synthetic Biology, only dynamic and not static data provides necessary parameter and network connection constraints for modeling of signaling and gene networks (Bennett, M. R. and Hasty, 2009). On the other hand, a cell is the fundamental unit of life, and single-cell resolution is the resolution of choice for synthetic biologists (Bennett, M. R. and Hasty, 2009). Therefore, further advances in the area of Synthetic Biology required development and improvement of single-cell isolation techniques, including microfluidic devices (reviewed in (Bennett, M. R. and Hasty, 2009)), which in turn made possible automated unbiased high-throughput isolation of single cells, and are predicted to serve as a basis for the sequencing-based single-cell analyses (Shapiro *et al.*, 2013). A recent review (Shapiro *et al.*, 2013) summarizes the main findings in the areas of single-cell genomics, transcriptomics, and epigenomics, which are out of scope of this work. Instead, below we provide a brief overview of the key single-cell transcriptomic studies of human and mouse early embryos and cell lines.

In 2009, based on several observations for short and long ncRNAs, including 849 heterogeneously expressed brain lncRNAs detectable with *in situ* hybridization, it was proposed that seemingly lowly expressed lncRNAs have high expression levels in a particular subset of cells (Dinger *et al.*, 2009). Later, it was shown that in mouse lipopolysaccharide (LPS)-stimulated bone-marrow-derived dendritic cells (BMDCs), lncRNAs with both – high (GAS5) and low (Gm8773, 2810025M15Rik) population-level expression are indeed present only in a subset of cells (Shalek *et al.*, 2013). This was true not only for lncRNAs, but even

for several highly expressed mRNAs (TPM > 250), and this was validated in the independent RNA-FISH experiments for a set of representative genes (Shalek *et al.*, 2013). Such expression heterogeneity unlikely resulted from the lack of cell cycle synchronicity of LPS-stimulated BMDCs, as the latter are post-mitotic and their response to LPS is synchronous in time (Shalek *et al.*, 2013). It is interesting to mention that Pearson correlation of gene expression for different individual cells was only 0.48, while reaching 0.98 for populations of 10^5 cells (Shalek *et al.*, 2013). Such difference is believed to originate from a random assembly of RNA polymerase factors and that results in differences in efficiency of a given gene expression (reviewed in (Levine *et al.*, 2013; Sanchez and Golding, 2013).

Since the processes of differentiation during embryonic development are essentially a consequence of division of a single cell (zygote), understanding gene expression patterns in early embryos is impossible without techniques that allow precise and reliable high-throughput quantification of transcripts in single cells (Saliba *et al.*, 2014). One of such methods, developed by Tang *et al.* (Tang *et al.*, 2009) was used for sequencing of germ cells (Guo, F. *et al.*, 2015), and hESCs and human preimplantation embryos (Yan *et al.*, 2013). This single-cell RNA-seq method was able to capture significant gene expression differences between 4- and 8-cell stages, which is associated with the major wave of embryonic genome activation (EGA) (Yan *et al.*, 2013). Authors used 4- to 8-cell stage transition to show that lncRNAs with heterogeneous and developmental stage-specific expression (> 0.1 FPKM, fragments per kilobase per million) show consistent expression in all sampled embryos, and thus unlikely represent leaky transcription (Yan *et al.*, 2013). Another study that used strand-specific single-cell-tagged reverse-transcription STRT-seq approach showed that during oocyte to 4-cell stage transition proportionally more of maternal coding than non-coding transcripts are degraded (Tohonen *et al.*, 2015). Moreover, while coding transcripts expression increased during the major wave of EGA, noncoding transcripts increased during

the transition from oocyte to 4-cell stage embryos (Tohonen *et al.*, 2015). Aside from dissecting the differences in protein-coding and non-coding transcripts in human embryos, single-cell analyses facilitated the discovery of bimodal expression of dozens of protein-coding genes in sister mouse blastomeres (Biase *et al.*, 2014), identification of the earliest marker genes of the inner and outer mouse blastocyst cells (Guo, G. *et al.*, 2010), and determination of modules of co-expressed genes that define specific mouse and human developmental stages (Xue *et al.*, 2013). Despite significant progress in the area of single-cell transcriptomics, we are just beginning to understand the complexity and dynamics of transcription in single cells, and further improvements, including development of strand-specific full-length RNA-seq technologies, is required to study, for example, antisense transcription in rare and transient cell states, such as totipotent blastomeres.

1.4 Antisense lncRNAs

As recent research focuses on long intergenic non-coding RNAs and enhancer RNAs (reviewed in (Ulitsky and Bartel, 2013) and (Lam *et al.*, 2014)), antisense lncRNAs remain the least studied group of all lncRNAs. Antisense transcription was proposed to occur in 74 % of human gene loci (Nakaya *et al.*, 2007). These antisense transcription events were shown to coincide with the presence of promoter-associated chromatin marks, CpG islands, and RNA Pol II binding (Tahira *et al.*, 2011; Fachel *et al.*, 2013), and therefore can be considered to be independent transcription units. Interestingly, 28 % of antisense transcripts were detected in the absence of their overlapping genes (Ayupe *et al.*, 2015), further supporting the independence of these transcription units. Our previous work demonstrates that such antisense transcription units frequently produce monoexonic lncRNAs (Louro *et al.*, 2007). Antisense lncRNAs are tissue specific, and the most highly expressed of them are transcribed antisense to genes coding for regulators of transcription (Nakaya *et al.*, 2007). The importance of antisense lncRNAs is illustrated by their differential expression in pancreatic cancer (Tahira *et*

al., 2011), and renal cell carcinoma (Fachel *et al.*, 2013). The expression of antisense lncRNAs was shown to correlate with expression (Louro *et al.*, 2007; Nakaya *et al.*, 2007; Beckedorff *et al.*, 2013; Fachel *et al.*, 2013), or alternative splicing of their sense counterparts (Louro *et al.*, 2007).. Additionally, if a given antisense lncRNA is expressed in another species, its expression would be, by definition, syntenic to its sense counterpart. Syntenic transcription would in turn increase the likelihood of a *cis*-regulatory effect of such antisense lncRNA (Diederichs, 2014). Nonetheless, the widely accepted assumption that a large portion of antisense lncRNAs regulate their overlapping genes (Magistri *et al.*, 2012) might be a poor predictor of function for any yet uncharacterized antisense lncRNA.

1.5 *HIPSTR* and its overlapping *TFAP2A* gene

In this study, we report identification of a novel lncRNA, which we named *HIPSTR*, that is expressed from the opposite strand of *TFAP2A* gene, the gene encoding a transcription factor (TF) AP-2alpha that is essential for vertebrate neural crest development (Schorle *et al.*, 1996; Zhang, J. *et al.*, 1996; Rada-Iglesias *et al.*, 2012; Prescott *et al.*, 2015), and that is also induced in mouse (Guo, G. *et al.*, 2010) and human (Cheng *et al.*, 2004; Aghajanova *et al.*, 2012) trophectoderm. AP-2alpha (TFAP2A) belongs to a family of five related TFs that are encoded by five retinoic acid-inducible, developmental genes. This family is composed of: AP-2alpha, AP-2beta, AP-2gamma, AP-2delta, and AP-2epsilon (reviewed in (Eckert *et al.*, 2005)). *TFAP2A* gene has three alternative first exons conserved in vertebrates, which in turn give rise to *TFAP2A* isoforms 1a, 1b, and 1c, and only isoform 1a encodes a TF that is capable of acting as both – repressor and activator (others only function as activators) (Berlato *et al.*, 2011). Significant TFAP2A expression is observed in the developing epidermis, kidney, cerebellum, spinal cord, and eye (Zhang, J. and Williams, 2003).

Despite its role as an important developmental regulator and in the light of our previous findings that connect antisense lncRNAs and cancer, the most intriguing was the

association of aberrant regulation of *TFAP2A* with tumorigenesis (Yu *et al.*, 2002). Expression of TFAP2A is detectable in several adult tissues, for example, in the ductal epithelium of the mammary gland, where altered TFAP2A expression is linked to the progression of breast cancer (Zhang, J. and Williams, 2003). TFAP2A expression is progressively lost in primary breast tumors with tumor progression from non-malignant epithelium to invasive breast cancer. Similarly, TFAP2A protein was lost in advanced stage colon tumors (McPherson *et al.*, 2002), and CREB-dependent loss of *TFAP2A* expression is considered as a hallmark of malignant progression of cutaneous melanoma (Melnikova *et al.*, 2010). While TFAP2A loss in melanoma, breast and colon cancers is a rather late event, in prostate cancer TFAP2A expression is lost early, and its re-expression in TFAP2A-negative LNCaP-LN3 prostate cancer cell line eliminated tumorigenicity of these cells in nude mice (Ruiz *et al.*, 2004). *TFAP2A* is a transcriptional target of p53 (Li, H. *et al.*, 2006), and tumor suppressor activity of TFAP2A protein is achieved through tight cooperation with p53 (McPherson *et al.*, 2002), and consequent co-regulation of target gene promoters of both p53 and TFAP2A (Li, H. *et al.*, 2006), including upregulation of *CDKN1A* (Scibetta *et al.*, 2010). At the same time, TFAP2A overexpression results in inhibition of growth and stable colony formation *in vitro*, apoptosis induction, cell cycle arrest in G1 and G2 phase in various cancer cells (McPherson *et al.*, 2002; Wajapeyee and Somasundaram, 2003). On the contrary, in head and neck squamous cell carcinoma (HNSCC) TFAP2A epigenetically silences tumor suppressive genes and induces microsatellite instability, while downregulation of TFAP2A in these HNSCC results in decreased cell proliferation (Bennett, K. L. *et al.*, 2009).

We therefore hypothesized that *HIPSTR* lncRNA transcribed antisense to *TFAP2A* might be involved into cancer-related deregulation of *TFAP2A* expression. In the present work we found that *HIPSTR* has conserved expression patterns between human and mouse, and its promoter demarcation is conserved in the amniotes. Unlike previously characterized

antisense lncRNAs, *HIPSTR* levels do not correlate with the expression of its overlapping *TFAP2A* gene in cell lines and tissues, and *HIPSTR* expression could not be associated with tumor or normal phenotypes in cell lines. Silencing of *HIPSTR* led to differential expression of a group of genes involved in development and differentiation. *HIPSTR* and *TFAP2A* were weakly co-induced in *in vitro* developmental models, such as *in vitro* derived neural crest cells and trophoblasts, but such co-induction was absent in retinoic acid-treated NT2/D1 cells. Moreover, we show that *HIPSTR* is activated independently from *TFAP2A* during early development in a group of cells within totipotent 8-cell- and morula-stage human embryos. Analyses of expression patterns of *HIPSTR* and hundreds of other lncRNAs in totipotent human embryos, human embryonic stem cells (hESCs), and myelogenous leukemia (K562) cells provide additional evidence that cell-to-cell variability is an inherent feature of lncRNAs.

2. Aims and objectives

2.1 Aims

The aim of the current study was to characterize *HIPSTR* antisense lncRNA, expressed from the antisense strand in the *TFAP2A* gene locus, and to understand the relationship of these two genes.

2.2 Objectives

1. Characterization of *HIPSTR* as an antisense lncRNA, including evaluation of coding potential, expression patterns in cell lines and tissues, and conservation.
2. Identification of *HIPSTR* promoter sequences and their conservation in various species.
3. Analysis of *HIPSTR* knockdown and overexpression effect on *TFAP2A* locus genes expression, and on global gene expression patterns.
4. Assessment of *HIPSTR* expression in developmental models and its relation to the expression of the overlapping developmentally regulated *TFAP2A* gene.
5. Comparison of *HIPSTR* expression patterns with other lncRNAs with similar expression levels.

3. Materials and methods

3.1 Cell culture

DU 145, 769-P, 786-O, MCF7, HepG2, NT2/D1, HEK293, HeLa (all – ATCC), RC-124 (CLS Cell Lines Service, GmbH) cell lines and HES, human endometrial cells, were cultured in DMEM medium (Vitrocell Embriolife) supplemented with 10 U/ml Penicillin, 0.01 mg/ml Streptomycin (1x Pen-Strep; Vitrocell Embriolife) and 10 % FBS (Vitrocell Embriolife). HES human endometrial cell line was a kind gift from Dr. Douglas Kniss (Ohio State University, Columbus, USA). H9 human embryonic stem cells (hESCs, WiCell) were cultured as described in (Rada-Iglesias *et al.*, 2011). $H1_{BP}$ cells were derived from H1 hESCs (WiCell) as described previously (Yang *et al.*, 2015).

LNCaP and K562 cell lines (both – ATCC) were cultured in RPMI-1640 medium (Gibco) supplemented with 1x Pen-Strep (Vitrocell Embriolife) and 10 % FBS (Vitrocell Embriolife), and for LNCaP an additional 10 mM HEPES (Gibco). For RNA-seq experiments, LNCaP cells were grown in RPMI-1640 medium supplemented with 10 mM HEPES (Gibco) and 10 % charcoal stripped FBS (Sigma) for 48 h prior to RNA extraction.

RWPE-1 cells (ATCC) were cultured in K-SFM medium (Gibco) containing 0.05 mg/ml bovine pituitary extract (Gibco), 5 ng/ml EGF (Gibco), and 1x Pen-Strep (Vitrocell Embriolife).

MCF10A cells (ATCC) were cultured in DMEM/F12 medium (Vitrocell Embriolife) supplemented with 20 ng/ml EGF (Invitrogen), 10 μg/ml insulin (Invitrogen), 0.5 μg/ml hydrocortisone (Sigma), 0.1 μg/ml cholera toxin (Sigma), 1x Pen-Strep (Vitrocell Embriolife), and 5 % horse serum (Gibco).

RL95-2 cells (ATCC) were cultured in DMEM/F12 medium (Vitrocell Embriolife) supplemented with 1x Pen-Strep (Vitrocell Embriolife) and 10 % FBS (Vitrocell Embriolife).

THLE-3 cells (ATCC) were cultured on flasks precoated with FNC coating mix (AthenaES), and in BEGM medium (Clonetics) supplemented with 1x Pen-Strep (Vitrocell Embriolife), 10 % FBS (Vitrocell Embriolife), 5 ng/ml EGF (Invitrogen), 70 ng/ml Phosphoethanolamine (Sigma), supplemented with all additives from BEGM bullet kit (Clonetics), except for Epinephrine and Gentamycin/Amphotericin.

All cell lines were grown at 37 °C in 5 % CO_2-humidified atmosphere.

3.2 LNCaP RNA-seq

LNCaP RNA-seq libraries were prepared as described in (Beckedorff *et al.*, 2013). Briefly, LNCaP poly(A)$^+$ RNA was extracted with FastTrack MAG Maxi mRNA Isolation Kit (Invitrogen), as per manufacturer's protocol, treated with 25 U of DNase I, Amplification Grade (Invitrogen) for 1 h at room temperature, quantified with Quant-iT RiboGreen RNA Reagent (Invitrogen) and assessed for integrity on 2100 Bioanalyzer (Agilent). Obtained RNA samples were used for strand-specific paired-end RNA-seq library preparation, in accordance with the standard illumina protocol and two biological replicates were sequenced on a HiSeq 2000. Data were processed as described below.

3.3 5' and 3' rapid amplification of cDNA ends (RACE)

Human Prostate Marathon-Ready cDNA (Clontech) was used to validate strand-specific RNA-seq identification of *HIPSTR* in LNCaP prostate carcinoma cell line. The first round of the 5′ and 3′ RACE PCRs was done in complete agreement with Marathon-Ready cDNA library user manual (Clontech). The second round of RACE PCR was performed with nested strand-specific primers to increase the specificity of target product detection (Additional file 1: Table 1). Obtained PCR products were gel-purified (Wizard SV Gel and PCR Clean-Up System; Promega), cloned into pGEM T-Easy vector (Promega), and sequenced.

3.4 *HIPSTR* coding potential analysis and polyadenylation signal prediction

To assess *HIPSTR* coding potential, we first searched for potential open reading frames (ORFs) within *HIPSTR* gene sequence by using the ORF Finder on-line tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). To screen for similarities with any known proteins, all found ORFs were then subjected to blastp search against Non-redundant (nr) protein sequences database (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins).

ORF shuffling was done essentially as described in (Klattenhoff *et al.*, 2013). Briefly, *HIPSTR* sequence was split into groups of 3 nucleotides, which were subsequently shuffled 1000 times. Considering only ORFs that begin with a canonical ATG start codon, maximum ORF sizes were retrieved after each shuffling, and their distribution was plotted. ORF sizes are expressed as fractions of *HIPSTR* length.

HCpolya, Hamming Clustering poly-A prediction in Eukaryotic Genes on-line tool (http://bioinfo4.itb.cnr.it/~webgene/wwwHC_polya.html) with pattern length parameter set at *12* was used to predict *HIPSTR* polyadenylation signal position (Milanesi *et al.*, 1996).

3.5 Biogenesis by RNA-Polymerase II, HIPSTR 5'-capping status, half-life estimation, and cell fractionation

Confirmation of *HIPSTR* transcription by RNA-Polymerase II, test for the presence of 5′-methylguanosine cap, as well as determination of *HIPSTR* sub-cellular localization were performed in parallel with analogous experiments for *INXS* antisense lncRNA characterization, and by using essentially the same samples and controls as described in detail (DeOcesano-Pereira *et al.*, 2014), except for primers required for specific detection of *TFAP2A* locus genes (Additional file 1: Table 1). Stability of transcripts of *TFAP2A* locus genes was assessed in HEK293 cells after 1, 3, 6, 9, and 12 h of treatment with 10 μg/ml actinomycin D (Sigma) or vehicle alone (0.05 % DMSO). Half-lives of transcripts were calculated as described in (Beckedorff *et al.*, 2013).

3.6 RNA extraction, cDNA synthesis, and quantitative PCR (qPCR)

Total RNA was extracted with TRIzol (Invitrogen) and purified with RNeasy Micro Kit (QIAGEN) according to manufacturer's protocol, with on-column DNAse I treatment time extended to 1 h. Total RNA was quantified on ND-1000 (NanoDrop), and its integrity was checked with 2100 Bioanalyzer (Agilent). Total RNA was reverse transcribed with SuperScript III First-Strand Synthesis System (Invitrogen) and oligo$(dT)_{20}$ primer for detection of any transcript mentioned in this study, except for *HIPSTR*. To detect human *HIPSTR*, 100 to 500 ng total RNA and 20 pmol of strand-specific Primer #1 (Additional file 1: Table 1) were annealed at 60 °C for 5 min, and cDNA was then synthesized at 55 °C for 1 h with ImProm-II Reverse Transcription System (Promega) and Mg^{2+} concentration of 6 mM. To detect mouse *Hipstr*, 1 μg total RNA and 20 pmol of strand-specific Primer #2 (Additional file 1: Table 1) were annealed at 62.5 °C for 5 min, and cDNA was then synthesized at 50 °C for 1 h with ImProm-II Reverse Transcription System (Promega) and Mg^{2+} concentration of 6 mM. Strand-specific primers #1 and #2 contained a tag sequence (ATGGCGAGAATCAATGCG) at the 5′-end that has no complementarity to the human or mouse genome. This tag sequence served as a target for annealing of the reverse qPCR primer, ensuring the strand specificity and eliminating non-specific background amplification (Lanford *et al.*, 1994) in the human or mouse *HIPSTR* detection assays.

Transcripts expression levels were measured by using Power SYBR Green (Applied Biosystems) on the 7500 Real Time PCR System (Applied Biosystems), with the default reaction setup for 20 μl reactions. Absolute expression levels of human and mouse *HIPSTR* were determined by comparison with an amplification of dilution curve points of a corresponding PCR product of known concentration. To measure human *HIPSTR* expression levels, qPCR extension step was performed for 30 s at 65 °C; to measure mouse *Hipstr* expression, qPCR extension step was done for 1 min at 60 °C. For all other qPCR reactions

GAPDH was used for normalizing the data, unless stated otherwise. Normalized data are represented as relative abundances determined by using delta Ct method (Pfaffl, 2001). Threshold cycle measurements were done by the 7500 System software with the default setup.

3.7 Total RNA libraries

Human Total RNA Master Panel II (20 tissues) and Mouse Total RNA Master Panel (15 tissues) (both – Clontech) were used to screen for tissue-specific expression of *HIPSTR* in human and mouse tissue samples, correspondingly.

3.8 Derivation of human neural crest-like cells (hNCCs) *in vitro*

H9 human embryonic stem cells (hESCs, WiCell) cultured as described in (Rada-Iglesias *et al.*, 2011) were subsequently differentiated into H9 hNCCs as described in (Bajpai *et al.*, 2010; Rada-Iglesias *et al.*, 2012). Briefly, H9 hESCs were grown in mTeSR-1 (STEMCELL Technologies) feeder- and serum-free medium. Cells were passaged 1:7 every 5-6 days by accutase detachment (Invitrogen) with subsequent replating of the resultant clusters of 50-200 cells on tissue culture dishes coated overnight with growth-factor-reduced Matrigel (BD Biosciences). To derive H9 hNCCs, H9 hESCs were incubated with 2 mg/ml collagenase (Gibco). Once detached, clusters of 100-200 cells were plated in hNCC differentiation medium: 1:1 Neurobasal medium/D-MEM F-12 medium (Invitrogen), 0.5x B-27 supplement with Vitamin A (50x stock, Invitrogen), 0.5x N-2 supplement (100x stock, Invitrogen), 20 ng/ml FGF2 (Peprotech), 20 ng/ml EGF (Sigma), 5 μg/ml bovine insulin (Sigma) and 1x Glutamax-I supplement (Invitrogen). Medium was changed every other day. After six-seven days of differentiation, resultant neuroepithelial spheres attached and gave rise to migratory hNCCs, as previously described (Rada-Iglesias *et al.*, 2012). Four-five days after the appearance of the first hNCCs, cells were collected for subsequent analyses.

3.9 $H1_{BP}$ cells culture and derivation of human trophoblast-like cells (hTBCs) *in vitro*

$H1_{BP}$ cells were derived from H1 hESCs (WiCell), cultured and differentiated into hTBCs as described previously (Yang *et al.*, 2015). Briefly, $H1_{BP}$ cells were maintained in the hESC basal medium (Amit *et al.*, 2000; Ezashi *et al.*, 2005), which had been conditioned by a monolayer of γ-irradiated mouse embryonic fibroblast (MEF) feeder cells for 24 h, and then supplemented with 10 ng/ml FGF2. Medium was changed every day. For passaging, $H1_{BP}$ cells were detached with Gentle Cell Dissociation Reagent (STEMCELL Technologies) for 6-7 min at 37 °C, dispersed into clusters of 5-10 cells, and plated on 0.1 % gelatin-coated culture dishes of desired size.

For hTBCs derivation, $4x10^4$ $H1_{BP}$ cells were passaged onto 5 cm^2 culture dishes and cultured for the next 48 h as described above, after which the medium was changed to one lacking FGF2 but containing 0.1 μM PD173074 (Sigma-Aldrich) in hESC basal medium not conditioned with MEF feeder cells. Media of both – untreated and PD173074-treated cells – was changed every day. Cells were collected for subsequent analyses after 1, 2, 4, 6, and 8 d of PD173074 treatment.

3.10 All-*trans* retinoic acid (ATRA) treatment of NT2/D1 cells

For ATRA treatment, $1x10^6$ NT2/D1 cells were plated per 75 cm^2 tissue culture flask. Four hours after plating, ATRA in DMSO was added to complete growth medium to the final concentration of 10 μM, essentially as described in (Andrews, 2006). Medium containing ATRA was replaced every 7 days of treatment. Increase in *HOXB5* mRNA expression levels was used to control for successful ATRA treatment, as in (Luscher *et al.*, 1989).

3.11 Antisense oligonucleotide (ASO)-mediated silencing

For ASO-mediated silencing of *HIPSTR* $4.5x10^5$ HEK293 cells or $2.4x10^5$ LNCaP cells were plated on 6-well plates 24 h before transfection. Transfections were performed by

using 0.025 µl of Lipofectamine RNAiMAX (Invitrogen) per 1 pmol of transfected ASO. Transfection mixes were prepared in OptiMEM I Reduced Serum Medium (Gibco).

To silence *HIPSTR* expression in $H1_{BP}$ cells, $4x10^4$ cells were plated on 6-well plates 48 h before transfection, and cultured as described above; 0.013 µl of GenMute siRNA Transfection Reagent (SignaGen) per 1 pmol of ASO were used for transfection. Transfection mixes were prepared in 1x GenMute Transfection Buffer (SignaGen).

A total of 300 pmol of ASO or mix of ASOs per well on 6-well plates was used for transfection. In all silencing experiments cells were collected for subsequent RNA or protein extraction 24 h after transfection with ASOs. For time-course *HIPSTR* knockdown assay in HEK293 cell line, cells were collected 6, 12, 24, 48, and 72 h after transfection with ASOs.

3.12 Oligonucleotide sequences

All oligonucleotide sequences (primers and ASOs) are listed in Additional file 1: Table 1.

3.13 TFAP2A protein and *HIPSTR* lncRNA transient ectopic overexpression

Full-length *HIPSTR* sequence was amplified from HEK293 genomic DNA with KpnI-FL-HIPSTR-F and HindIII-FL-HIPSTR-R primers (Additional file 1: Table 1) and cloned into pCEP4 vector (Invitrogen) between *Kpn*I and *Hind*III sites.

Approximately $5x10^5$ HEK293 cells were transfected with 3 µg of pCEP4-HIPSTR or pCEP4 empty vector for *HIPSTR* overexpression assays, or pcDNA3-TFAP2A-1a, pcDNA3-TFAP2A-1b, pcDNA3-TFAP2A-1c, or pcDNA3 empty vector for TFAP2A overexpression assays; pcDNA3-TFAP2A-1a, pcDNA3-TFAP2A-1b, and pcDNA3-TFAP2A-1c expression vectors used for TFAP2A isoforms overexpression were kindly provided by Dr. Chiara Berlato (Queen Mary University of London, London, UK). Cells were collected for RNA and protein extraction 72 h after transfection.

Transfections were carried out by using FuGENE HD Reagent (Promega) at 3:1 transfection reagent:DNA ratio in the corresponding complete growth media.

3.14 Promoter predictions and dual-luciferase assays

Genomic sequence containing *TFAP2A* isoform 1c (NM_001042425), and sequences 10 kb up- and downstream of it (chr6:10386916 – 10427411 in human genome assembly hg19) were used as an input for transcription start site (TSS) prediction with the TSSG program (Solovyev and Salamov, 1997). By using H3K4me3 ChIP-seq peaks from ENCODE Project around the predicted TSS as a guideline, *HIPSTR* candidate promoter sequences were amplified from HEK293 genomic DNA and cloned into pGL3-Basic vector (Promega) between *Kpn*I and *Nhe*I sites. Inserts were generated as follows: for pGL3-P1 the insert was generated with KpnI-promoter-primer-A and NheI-promoter-primer-G, for pGL3-P2 – with KpnI-promoter-primer-A and NheI-promoter-primer-E, for pGL3-P3 – with KpnI-promoter-primer-F and NheI-promoter-primer-G, for pGL3-P4 – with KpnI-promoter-primer-H and NheI-promoter-primer-B, for pGL3-P5 – with KpnI-promoter-primer-F and NheI-promoter-primer-B, for pGL3-P6 – with KpnI-promoter-primer-A and NheI-promoter-primer-B, and for pGL3-P7 – with NheI-promoter-primer-C and KpnI-promoter-primer-D (Additional file 1: Table 1). For the assay, $1x10^5$ cells per well were seeded on 24-well plates 24 h before transfection. Cells were co-transfected with 650 ng of empty pGL3-Basic vector, pGL3-SV40 plasmid, or one of the above-described constructs, and 150 ng of pRL-SV40 plasmid (Promega). Transfections were carried out by using FuGENE HD Reagent (Promega) at 3:1 transfection reagent:DNA ratio in corresponding complete growth media. Cells were lysed and assayed in accordance with Dual-Luciferase Reporter Assay System (Promega) protocol 48 h after transfection. Firefly luciferase signal was normalized to *Renilla* luciferase activity from the same lysate. Lysates of the cells transfected with pGL3-Basic and pGL3-Promoter

plasmids served as a negative and a positive control of the Firefly luciferase activity, respectively.

In the overexpression assays of TFAP2A isoforms, 800 ng of TFAP2A-overexpressing plasmid were co-transfected with luciferase genes-carrying constructs at transfection reagent:DNA ratio 1.5:1. Firefly luciferase activity in the lysates of the cells transfected with 3xAP2-Bluc plasmid served as a positive control for TFAP2A transactivation activity. 3xAP2-Bluc plasmid was a kind gift from Dr. Trevor Williams (University of Colorado Denver, Aurora, USA).

3.15 Western blotting analysis

For western blot analysis, collected cells were washed twice with ice-cold PBS, resuspended in RIPA buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1 % SDS, 0.5 % sodium deoxycholate, 1 % Triton X-100, 1mM EDTA), and sonicated. Protein content of the lysates was quantified with Micro BCA Protein Assay Kit (Thermo Fisher Scientific). Equal protein amounts (40 µg) were resolved on 12 % SDS-polyacrylamide gel and transferred onto nitrocellulose membranes (Amersham Biosciences). Membranes were blocked with Tris-buffered saline containing 0.1 % Tween 20 and 2 % Bovine Serum Albumin (BSA) (TBST/2 % BSA), and incubated overnight in TBST/2 % BSA with primary antibody. Membranes were then washed five times with TBST and incubated for 1 hour in TBST/2 % BSA with goat anti-Mouse IgG secondary antibody, Alexa Fluor® 680 conjugate (Thermo Fisher Scientific) (1:10000). Membranes were next washed again with TBST, and the signal intensities were captured with Odyssey Infrared Imaging System (LI-COR Biosciences). Primary antibodies were anti-TFAP2A (Santa Cruz, sc12726) (1:100), and anti-Actin (Millipore, MAB1501) (1:5000).

3.16 Microarray experiments

200 ng of total RNA from HEK293 cells or 100 ng of total RNA from $H1_{BP}$ cells transfected with ASOs targeting *HIPSTR* were converted into Cy3- and Cy5-labeled cRNA with the Agilent Low Input Quick Amp Labeling Two Color Kit. Dye-swap technical replicates were created for each biological replicate. Three biological replicates of HEK293 cells transfected with each ASO were used for microarray experiments. In experiments with $H1_{BP}$ cells, three biological replicates for control ASO, and two – for each of the targeting ASOs were assayed. Obtained cRNA samples were then hybridized to Agilent SurePrint G3 Gene Expression Microarrays (G4851B) 8x60K as per manufacturer's instructions. Data intensities were extracted from the slide images with Feature Extraction Software (Agilent Technologies) and normalized by using the Lowess method (Agilent Technologies).

All probes whose mean signal was lower than background on at least one array were filtered out. Signal intensities were normalized by 40 % trimmed mean. Significance Analysis of Microarrays (SAM) with two-class comparison was then used to identify differentially expressed genes (Tusher *et al.*, 2001). SAM q-value $\leq$ 0.01 and fold change $\geq$ 2 were considered as a threshold for identification of differentially expressed genes. Hierarchical clustering of differentially expressed genes was done with TIBCO Spotfire software by applying Z-score transformation of the normalized data intensities for each gene across all samples.

3.17 Gene Ontology (GO) analysis

GO and tissue-specific expression analyses of annotated differentially expressed genes were performed with DAVID (https://david.ncifcrf.gov/) (Huang da *et al.*, 2009) with GOTERM_BP_ALL and UP_TISSUE tables, respectively. Benjamini-Hochberg adjusted p-value $\leq$ 0.01 was used as a significance threshold. Genes are referred to as "annotated" if they have a HGNC symbol in Agilent annotation.

3.18 Motif search

To search for known TF recognition motifs around TSSs of genes differentially expressed upon *HIPSTR* knockdown, the *findMotifs.pl* module from Homer package v.4.7.2 (Heinz *et al.*, 2010) was used with the following parameters: *human -len 8,10,12 -p 12*. We searched for enrichment of TF motifs at different positions relative to TSSs, which are indicated for each TF on the corresponding figures.

3.19 Public RNA-seq and ChIP-seq analysis

ENCODE Project (Consortium, 2012) human long polyadenylated RNA-seq data for the indicated cell lines were obtained from GEO entry GSE30567, and mouse long RNA-seq – from GEO entry GSE36025. Ribosome profiling data from (Stumpf *et al.*, 2013) were downloaded from SRA entry SRA099816. K562 single-cell RNA-seq data were downloaded from SRA entry SRX495504 (Luo *et al.*, 2014). Early human and mouse embryo single-cell RNA-seq data were retrieved from ENA entry PRJEB8994 (Tohonen *et al.*, 2015), and from GEO entries GSE44183 (Xue *et al.*, 2013), GSE36552 (Yan *et al.*, 2013), and GSE57249 (Biase *et al.*, 2014). RNA-seq of DRB- (RNA Pol II elongation inhibitor) or vehicle-treated HEK293 cells from (Werner and Ruthenburg, 2015) were obtained from GEO entry GSE66478. H3K4me3 ChIP-seq data for liver samples of 10 mammalian species were downloaded from Array Express website entry E-MTAB-2633 (Villar *et al.*, 2015), for testis samples of mouse and rooster – from GEO entry GSE44588 (Li, X. Z. *et al.*, 2013b), for frog blastula, gastrula, neurula and tailbud stage embryos – from GEO entry GSE41161 (van Heeringen *et al.*, 2014), and for zebrafish 256 cell, oblong and dome stage embryos – from GEO entry GSE44269 (Zhang, Y. *et al.*, 2014). TFAP2A and H3K4me3 ChIP-seq data for chimpanzee NCCs and hNCCs were obtained from GEO entry GSE70751 (Prescott *et al.*, 2015).

Sequencing data were preprocessed with Trimmomatic v.0.30 (Bolger *et al.*, 2014) with parameters *-phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15*. Trimmomatic parameter *MINLEN:* was set at *16* for ChIP-seq reads, at *20* – for RNA-seq reads, except for RNA-seq data from (Stumpf *et al.*, 2013), for which it was set at *30*. Additional clipping of adapter sequence CTGTAGGCACCATCAAT was done for preprocessed RNA-seq reads from (Stumpf *et al.*, 2013) with fastx_clipper from FASTX Toolkit v.0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/). Human RNA-seq reads were mapped with TopHat v.2.0.12 (Kim, D. *et al.*, 2013) and Bowtie v.2.2.3 (Langmead and Salzberg, 2012) and a custom GTF file to guide transcriptome assembly. This custom GTF file was built by using the human transcriptome annotation GTF file downloaded from Ensembl Project web-site (http://www.ensembl.org/) and modified to include all lncRNAs reported in (Cabili *et al.*, 2011). Mouse RNA-seq reads were mapped as described above by using a GTF file for mouse genome assembly mm9. This GTF file was fetched from the illumina support site (https://support.illumina.com/). The following parameters for TopHat were used: *--no-coverage-search --b2-sensitive*; for paired-end strand-specific RNA-seq data (except LNCaP RNA-seq), *--library-type fr-firststrand* parameter was used in addition to the mentioned above; for LNCaP RNA-seq data *--library-type fr-secondstrand* parameter was added. ChIP-seq reads were mapped by Bowtie v.2.2.3 with parameter: *--sensitive*. Read densities were retrieved with *genomecov* command from bedtools package v.2.20.1 (Quinlan and Hall, 2010), and UCSC Genome Browser tracks were built with bedGraphToBigWig v.4 (Kent *et al.*, 2010). To count RNA-seq reads, TopHat paired-end RNA-seq data alignment output files were first sorted by read names with *sort* command from SAMtools package v. 0.1.19-44428cd (Li, H. *et al.*, 2009). RNA-seq reads were counted with htseq-count v.0.6.1p1 (Anders *et al.*, 2015), with parameter *-s yes* for single-end strand-specific data sets, *-s reverse* – for paired-end strand-specific data sets, and *-s no* for non-stranded data sets. Gene

expression levels were calculated in FPKM, considering gene length as a sum of all exonic non-overlapping sequences of all isoforms of a given gene. Unless stated otherwise, ChIP-seq and RNA-seq data are presented as aggregates of biological replicates for each indicated condition to increase resulting genome and transcriptome coverage, respectively.

To map RNA-seq and ChIP-seq data, the following reference genome assemblies were downloaded from UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/downloads.html): galGal4 (chicken), panTro4 (chimpanzee), bosTau7 (cow), canFam3 (dog), xenTro3 (frog), hg19/GRCh37 (human), calJac3 (marmoset), mm9 (mouse), monDom5 (opossum), susScr3 (pig), oryCun2 (rabbit), rn5 (rat), rheMac3 (rhesus), danRer7 (zebrafish).

For single-cell RNA-seq data analyses, genes were considered as protein-coding if they were assigned RefSeq accession prefix NM_ (mRNA) or XM_ (mRNA predicted). Genes were considered as non-coding if they were assigned RefSeq accession prefix NR_ (ncRNA) or XR_ (ncRNA predicted), or were annotated as novel lncRNAs (prefix XLOC_) in (Cabili *et al.*, 2011). For comparisons of expression profiles of non-coding and protein-coding genes in single cells, we considered only genes generating transcripts with total length of non-overlapping exonic sequences longer than 200 nt.

3.20 Expression heterogeneity comparisons

To evaluate heterogeneity of gene expression in single cells, we used single-cell RNA-seq data sets for totipotent blastomeres from 8-cell and morula-stage human embryos, hESCs (both – from (Yan *et al.*, 2013)), or K562 cells (from (Luo *et al.*, 2014)). For each gene in each data set, we calculated the number of cells N, in which a given gene was expressed. LncRNA genes are usually expressed at lower levels than protein-coding genes, and to make them comparable we considered only lncRNAs and protein-coding genes with expression levels in the same range (Cabili *et al.*, 2011); therefore, we did not consider genes whose expression was > 30 FPKM in at least one cell of a data set under analysis. Of the remaining

genes, we only considered those with expression > 3 FPKM in at least one cell of a data set. We counted a cell as *positive* for expression of a given gene if the expression level of that gene was > 3 FPKM in that cell.

We observed that, when assessed for all genes, the distribution of their corresponding *N* values is a mixture distribution. We used the *normalmixEM* function from mixtools v.1.0.4 R package (Benaglia *et al.*, 2009) to fit a model mixture distribution with two populations of genes – those with high or low heterogeneity of expression. Parameters used were: *number_of_components=2, lambda=0.5, sigma=0.5*. We next applied the resultant model to calculate the posterior probability of each gene under analysis to belong to either the high or the low heterogeneity of expression population. If a given gene could be associated with one of the abovementioned populations with a posterior probability > 0.99, it was assigned the "H" or "L" flag (for high or low heterogeneity of expression, respectively; Additional file 1: Tables 3 – 5), otherwise the "U" (uncertain) flag was assigned.

3.21 RNA-pulldown of the chromatin-associated portion of *HIPSTR*

To identify potential protein partners of *HIPSTR*, we used essentially the same approach as in (Klattenhoff *et al.*, 2013). First, we *in vitro* generated sense (target) and antisense (control) biotinylated RNA probes of the chromatin-associated candidate fragment (the first 1000 nt) of the *HIPSTR* sequence. For this, we used T7 MEGAScript Kit (Ambion) as per manufacturer's protocol, with the following modifications: we used 7.5 mM ATP, GTP and UTP, 6.75 mM CTP, and 0.1 mM biotin-14-CTP (Invitrogen), and *in vitro* transcription time was 2 h. Templates for *in vitro* transcription were generated by TA-cloning of the probe sequences into pGEM T-Easy vector (Promega). Obtained clones were sequenced, and those containing the inserts in the desired orientation were linearized with 10 U *Spe*I (NEB) overnight, and subsequently transcribed *in vitro* as described above. The length and integrity of the generated biotinylated probes was assessed on 2100 Bioanalyzer (Agilent).

To prepare nuclear fractions, 6 x 10^7 pluripotent NT2/D1 cells per pulldown were collected, washed once with ice-cold PBS, pelleted, resuspended in 2 ml nuclear isolation buffer (1.28 M sucrose; 40 mM Tris-HCl pH 7.5; 20 mM $MgCl_2$; 4 % Triton X-100), 2 ml PBS, 6 ml of DEPC-treated water, and incubated on ice for 20 min. Nuclei were pelleted by centrifugation at 2500 g for 15 min at 4 °C. Nuclear pellets were resuspended in 1 ml RIP buffer (150 mM KCl, 25 mM Tris pH 7.5, 0.5 % NP-40, 0.5 mM DTT, 1 mM PMSF, 1 X protease inhibitor cocktail (Roche) and 40 U/ml RNaseOUT (Invitrogen)), and homogenized by 30 strokes in a dounce homogenizer. Nuclear extracts were separated from nuclear membrane debris by centrifugation at 13000 g for 10 min at 4 °C. The supernatants (containing nuclear proteins) were pre-cleared by incubation with equilibrated Streptavidin Magnetic Beads (60 µl per pulldown; NEB) for 30 min at 4 °C with end-to-end mixing. Unbound proteins were next mixed with pre-blocked and equilibrated Streptavidin Magnetic Beads (60 µl per pulldown; NEB), and biotinylated RNA probes (40 pmol per pulldown), and pulldowns were performed for 1 h at room temperature with end-to-end mixing. Pre-blocking of magnetic beads was done with 7.5 µg of yeast tRNA per pulldown, and 10 µg of salmon sperm DNA per pulldown for 30 min at room temperature. To allow for the proper secondary structure formation, prior to pulldown step, biotinylated RNA probes were incubated at 90 °C for 2 min, transferred on ice for 2 min, mixed with pre-chilled RNA structure buffer (10 mM Tris pH 7, 0.1 M KCl, 10 mM $MgCl_2$), and incubated at room temperature for 20 min. After pulldown, beads were washed 5 times with wash buffer (5 mM Tris pH 7.5, 500 µM EDTA, 1 M NaCl), and proteins were eluted into 30 µl of water by incubation at 65 °C for 5 min. Pulldown experiments were performed in duplicates.

Eluted proteins were subjected to digestion with trypsin (Sigma Aldrich), in accordance with the protocol adapted from (Wisniewski *et al.*, 2009), also known as Filter-Aided Sample Preparation (FASP). Briefly, protein samples were mixed with 200 µl of 8 M

urea in 0.1 M Tris-HCl, pH 8.5 (solution UA), added to equilibrated Microcon YM-10 columns (Millipore), gently mixed, and centrifuged at 14000 g for 15 min. Subsequently, another 200 µl of solution UA were added to the columns, and centrifugation repeated. Next, 100 µl of 0.02 M DTT in solution UA were added to the columns, followed by centrifugation at 14000 g for 10 min, after which the columns were incubated for 30 min in the dark at room temperature with 100 µl of 0.05 M iodoacetamide (IAA) in solution UA. After incubation with IAA, columns were centrifuged at 14000 g for 10 min, washed three times with 100 µl of 0.05 M NH_4HCO_3 (solution ABC), and centrifugation repeated. The digestion of the column-bound proteins by trypsin was done for 18 h at 37 °C in 60 µl of solution ABC. After the incubation, columns were centrifuged at 14000 g for 10 min with subsequent addition of 50 µl of solution ABC and centrifugation at 14000 g for another 10 min. Obtained peptides were acidified with trifluoroacetic acid to pH ≤ 3, desalinized with StageTip C18 in accordance with the protocol from (Rappsilber *et al.*, 2007), and submitted to liquid chromatography-tandem mass spectrometry (LC-MS/MS) on a LTQ-Orbitrap Velos (Thermo Scientific) mass spectrometer coupled with a nanoflow liquid chromatography system Easy-nLCII (Thermo Scientific). LC-MS/MS experimental runs and analyses were done by Eduardo Shigueo Kitano at the laboratory of Dr. Solange Serrano at Instituto Butantan. Each pulldown sample was analyzed twice.

3.22 Accession numbers

The microarray data reported in this work were deposited in Gene Expression Omnibus (GEO) under accession GSE77937. RNA-seq data from LNCaP prostate cancer cell line were deposited in GEO under accession GSE79301. *HIPSTR* sequence is deposited in GenBank with accession number KU904338.

4. Results

4.1 *HIPSTR* is a *bona fide* antisense lncRNA

We have previously shown that expression of antisense lncRNAs correlates with the degree of tumor differentiation in prostate cancer (Reis *et al.*, 2004). Moreover, such antisense lncRNAs are frequently expressed from the opposite strand of genes encoding proteins involved in the regulation of transcription (Nakaya *et al.*, 2007). Aiming at the identification of novel antisense lncRNAs possibly associated with prostate cancer, we obtained strand-specific RNA-seq data from LNCaP prostate cancer cell line and searched for antisense transcription events in loci encoding TFs. *TFAP2A* encodes a TF known to be involved in various cancers (reviewed in (Pellikainen and Kosma, 2007)), including prostate cancer (Ruiz *et al.*, 2004; Makhov *et al.*, 2011). We focused on a putative monoexonic antisense lncRNA gene located between exons 2 and 5 of *TFAP2A* on the opposite genomic strand (Figure 1), a locus where no lncRNAs had been annotated so far. We named this lncRNA gene *HIPSTR* (Heterogeneously expressed from the Intronic Plus Strand of the TFAP2A-locus RNA).

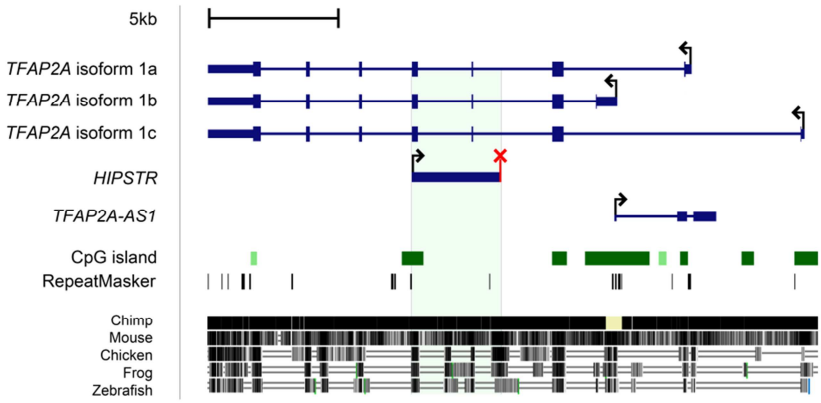


Figure 1. *HIPSTR* is a novel antisense lncRNA. UCSC Genome Browser snapshot showing genomic position of human *HIPSTR* relative to *TFAP2A* isoforms and *TFAP2A-AS1* gene, positions of CpG islands, repetitive sequences defined by RepeatMasker, and regions of vertebrate conservation within *TFAP2A* locus. The predicted *HIPSTR* polyadenylation signal is marked with a red "X" sign; genomic coordinates of the region shown are hg19 chr6:10396400 – 10420700.

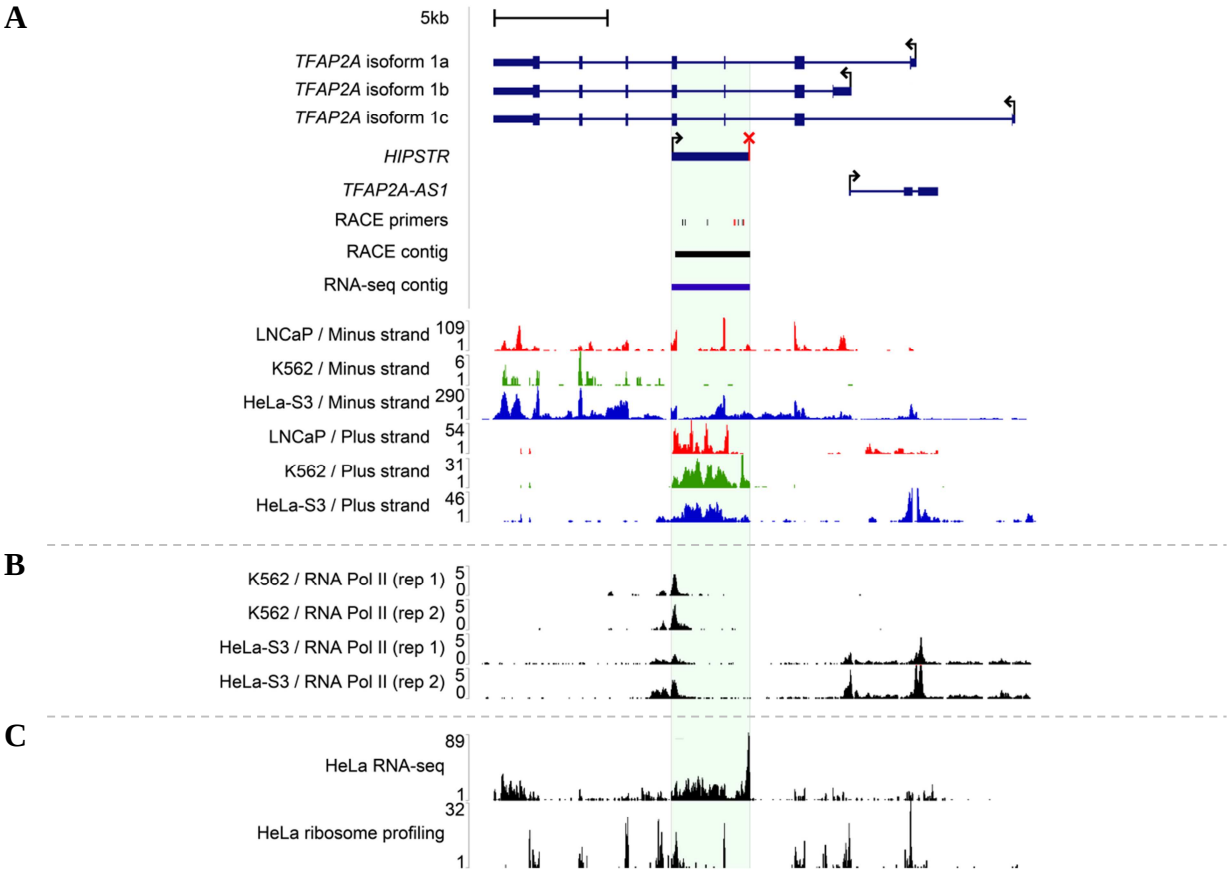


Figure 2. Analysis of both – our and public data evidences that *HIPSTR* is transcribed by RNA Pol II and is not associated with ribosomes. (A) UCSC Genome Browser snapshot showing RACE and RNA-seq contigs, genomic positions of primers used for 5′- (black) and 3′-(red) RACE, as well as positions of mapped strand-specific RNA-seq reads from K562 and HeLa-S3 (both – from (Consortium, 2012)), and from LNCaP cells (this work). (B) Genomic positions in the *TFAP2A* locus of RNA Pol II ChIP-seq peaks (data from (Consortium, 2012)). (C) Analysis of ribosome profiling data from (Stumpf *et al.*, 2013) shows no significant continuous association of ribosomes with *HIPSTR* sequence in HeLa cells.

To validate the *HIPSTR* sequence contig that was obtained after mapping the LNCaP RNA-seq reads (Figure 2 A), we performed 5′- and 3′-RACE PCR using a normal prostate poly(A) RACE cDNA library (Figure 2 A). We extended *HIPSTR* contig by 1 nt before we reached the poly(A) tail in this cDNA library. In agreement with RNA-seq and 3′-RACE PCR data, two potential polyadenylation signals were predicted to be located 13 nt and 9 nt upstream of the 3′-end of the *HIPSTR* gene (Figure 1). The most 5′ RACE clone obtained did not extend the LNCaP RNA-seq contig (Figure 2 A); however, analyses of the STRT-seq data, which preferentially sequences the 5′ end of transcripts (Islam *et al.*, 2011), for early

human embryos (Tohonen *et al.*, 2015) allowed us to further extend *HIPSTR* gene sequence by only 4 nt in the 5′-direction (Figure 2 A and Figure 16). In agreement with these data, the nucleotide at position chr6:10404790, had the highest TSS prediction score by TSSG tool (Solovyev and Salamov, 1997) across the entire *TFAP2A* genomic locus. This predicted TSS is located only 50 bp downstream from the 5′-end of the *HIPSTR* contig from LNCaP RNA-seq data. Notably, analysis of data from another publication (Consortium, 2012) showed that *HIPSTR* has an alternative TSS in HeLa-S3 cells located more than 600 bp upstream of the *HIPSTR* TSS in K562 or LNCaP cells (Figure 2 A). It remains to be investigated whether this alternative *HIPSTR* isoform is functionally different from the *HIPSTR* isoform described in this study (chr6:10404735 – 10408161 in human genome assembly hg19, 3427 nt). It is also evident from RNA-seq data that *HIPSTR* transcripts are unspliced (Figure 1, and Figure 2 A).

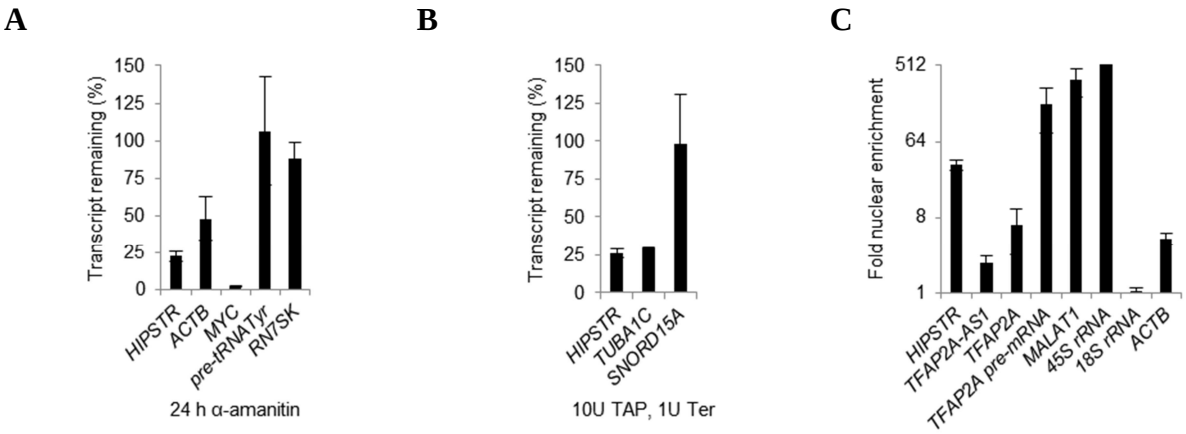


Figure 3. *HIPSTR* is transcribed by RNA Pol II, is capped and enriched in the nucleus. (A) RNA Pol II inhibition by α-amanitin in HeLa cells results in dramatic decrease in *HIPSTR* levels, as measured by RT-qPCR; known RNA Pol II-transcribed RNAs (*ACTB*, *MYC*) and RNA Pol III-transcribed RNAs (pre-tRNA^{Tyr}, *7SK*) served as controls. (B) 5′-cap structure removal by co-treatment of HeLa cells total RNA with Terminator 5′-phosphate-dependent exonuclease (Ter) and tobacco acid pyrophosphatase (TAP) reduces levels of *HIPSTR*, as measured by RT-qPCR; capped *TUBA1C* and uncapped *SNORD15A* transcripts served as controls. (C) HeLa cells fractionation into nuclear and cytoplasmic extracts shows nuclear enrichment of *HIPSTR*, as measured by RT-qPCR; nuclear enrichment of *TFAP2A* and *TFAP2A-AS1* is comparable with that of *ACTB*; we used *TFAP2A* pre-mRNA, *MALAT1* lncRNA, and 45S rRNA as nuclear fraction controls, and 18S rRNA – as cytoplasmic fraction control. The same RNA samples were used as in (DeOcesano-Pereira *et al.*, 2014), and data shown on (A – C) for control transcripts, except for *TFAP2A* locus genes, are the same as presented on Fig. 3A, 3B, and 3D in (DeOcesano-Pereira *et al.*, 2014). Experiments were performed in triplicate, error bars represent SD.

RESULTS

HIPSTR TSS is located within an 818-bp-long CpG island and overlaps RNA Pol II ChIP-seq peaks from ENCODE Project data (Consortium, 2012) (Figure 2 B). We confirmed that *HIPSTR* is transcribed by RNA Pol II (Figure 3 A), and has a 5′-cap structure (Figure 3 B), a typical feature of RNA Pol II transcripts.

We next examined *HIPSTR* coding potential. Both CPC (Kong *et al.*, 2007) and CPAT (Wang, L. *et al.*, 2013) coding potential evaluation tools classified *HIPSTR* as non-coding. None of the potential ORFs within *HIPSTR* sequence showed any similarity to known proteins in a blastx search (not shown). Recently, Stumpf *et al.* used synchronized HeLa cells to perform ribosome profiling of G1, S and M phases of cell cycle (Stumpf *et al.*, 2013). In these RNA-seq data, we did not find any evidence of significant ribosome association with either the entire *HIPSTR* sequence, or with the longest potential ORF (345 nt) in the *HIPSTR* sequence (Figure 2 C). Moreover, *in silico* analysis (see Methods) demonstrated that such 345 nt-long ORF can be expected to occur by chance in a 3427 nt-long transcript (Figure 4). Finally, we observed a strong nuclear enrichment of *HIPSTR* transcript (~33.5-fold, Figure 3 C), similar to some previously described regulatory lncRNAs (see Table 1 in (Fatica and Bozzoni, 2014)). Altogether, these data argue that *HIPSTR* is a *bona fide* lncRNA.

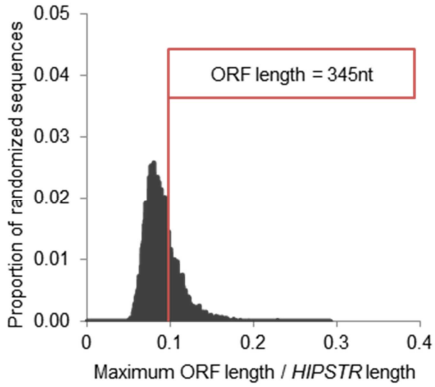


Figure 4. The appearance of the longest ORF within *HIPSTR* sequence can be expected to occur by chance. Plotted is the distribution of the longest ORFs generated by random shuffling of *HIPSTR* sequence.

RESULTS

Considering the proposed roles for antisense RNAs in cancer (Tahira *et al.*, 2011; Fachel *et al.*, 2013), we hypothesized that *HIPSTR* may be differentially expressed in tumor and non-tumor cell lines. We found that *HIPSTR* expression was not associated with tumor or non-tumor phenotype in prostate, kidney, breast, liver, or endometrial cell lines (Figure 5 A). Moreover, *HIPSTR* expression did not correlate with its overlapping gene (*TFAP2A*) across the cell lines tested (Figure 5 B). The latter observation was further supported by analysis of *HIPSTR* and *TFAP2A* expression in ENCODE Project RNA-seq data sets (Consortium, 2012) (Figure 5 C) and in a panel of human tissue RNA samples (Figure 5 D). Consistent with previous reports for lncRNAs (Ravasi *et al.*, 2006; Cabili *et al.*, 2011), *HIPSTR* population-level expression was low and exceeded the value of 1 FPKM only in two (HeLa-S3 and K562) out of eleven ENCODE cell lines (Consortium, 2012) (Figure 5 C).

Finally, we note here that *HIPSTR* has a tissue-specific expression pattern, also reported previously for some lncRNAs (Ravasi *et al.*, 2006; Cabili *et al.*, 2011), and that it is predominantly expressed in human testis and placenta (Figure 6 A). Strikingly, this expression pattern is evolutionarily conserved, as evident from Mouse ENCODE Project RNA-seq data (Consortium, 2012) (Figure 6 B). Additionally, we successfully detected *HIPSTR* transcription with RT-qPCR in a panel of mouse tissue RNA samples (Figure 6 C).

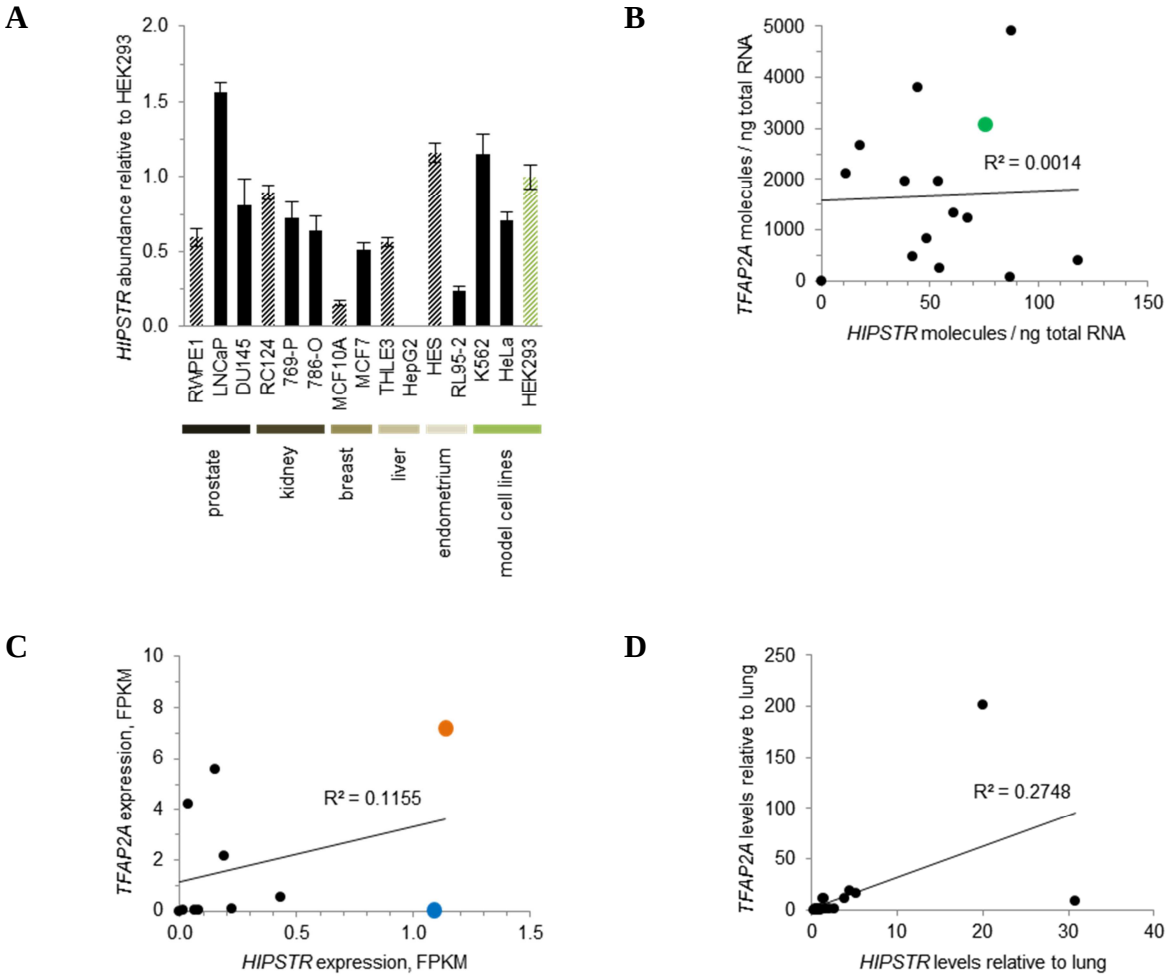


Figure 5. *HIPSTR* expression is not different in tumor cell lines and is not correlated with the expression of its antisense counterpart gene *TFAP2A*. (A) *HIPSTR* expression cannot be associated with tumor or non-tumor phenotype, as measured by RT-qPCR in human tumor (solid bars) and non-tumor (hatched bars) cell lines. *HIPSTR* expression in non-tumor human embryonic kidney HEK293 cell line (hatched green bar) is shown for comparison. Experiments were done in triplicate, error bars represent SD. (B) *HIPSTR* expression does not correlate with *TFAP2A* levels in the human cell lines shown in (A), as measured with RT-qPCR. HEK293 cells (green dot) express high levels of *HIPSTR* and *TFAP2A*, and were used for subsequent *HIPSTR* silencing experiments. (C) *HIPSTR* expression does not correlate with *TFAP2A* levels in human cell lines from the ENCODE Project (A549, GM12878, H1 hESCs, HeLa-S3, HepG2, HMEC, HSMM, HUVEC, K562, MCF7, NHEK) (Consortium, 2012). *HIPSTR* expression in HeLa-S3 (orange dot) and K562 (blue dot) cell lines is > 1 FPKM. (D) *HIPSTR* expression does not correlate with *TFAP2A* levels in the human tissues shown on Figure 6 A, as measured by RT-qPCR.

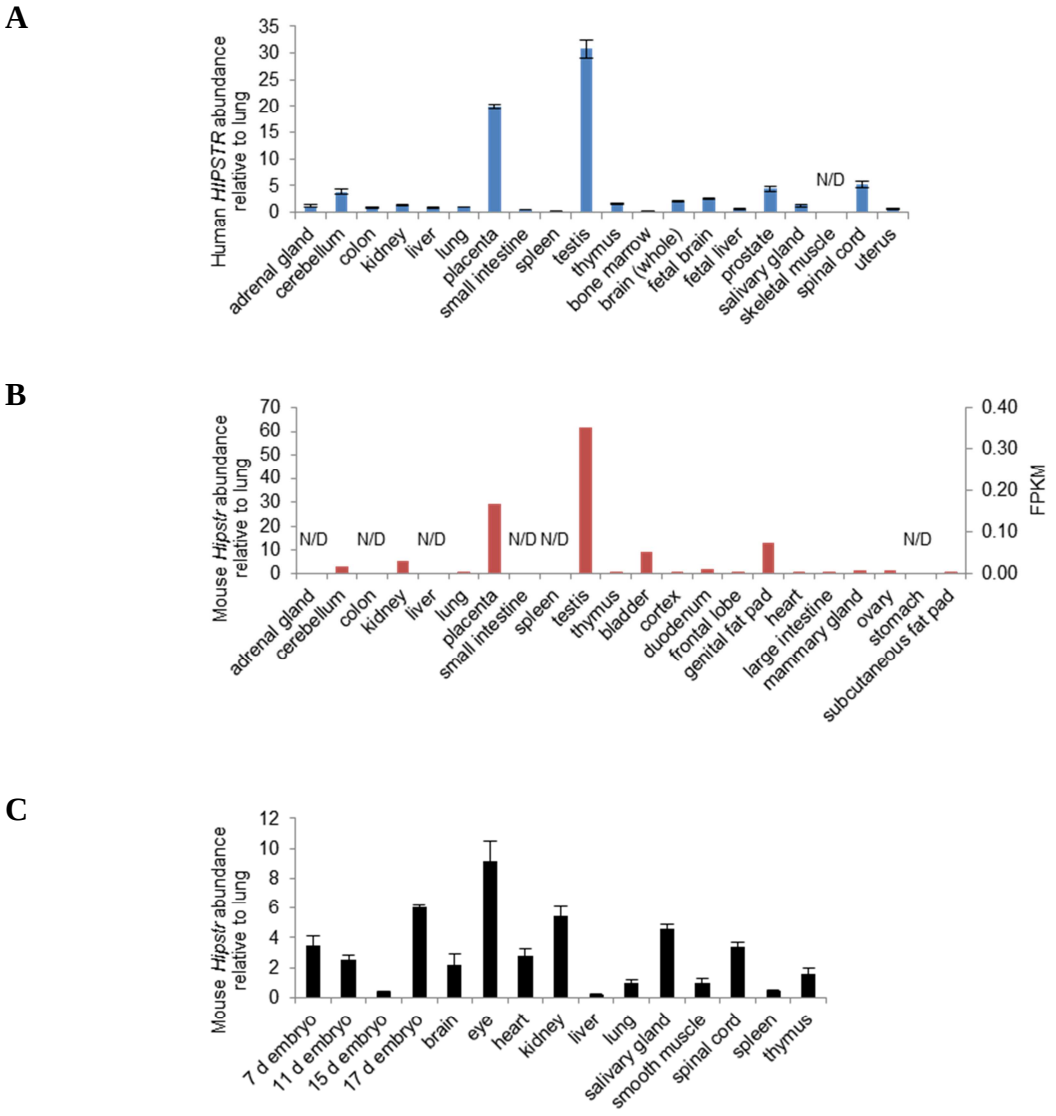


Figure 6. *HIPSTR* expression patterns in human and mouse tissue samples. (A) *HIPSTR* expression across a panel of human tissue RNA samples, as assessed with RT-qPCR; N/D – not detected. (B) Mouse *Hipstr* (chr13:40818458 – 40821725) ortholog expression across a panel of mouse tissue RNA samples from ENCODE Project RNA-seq data (Consortium, 2012). (C) Mouse *Hipstr* ortholog expression across a panel of mouse tissue RNA samples, as measured with RT-qPCR; error bars represent SD in three independent measurements.

4.2 *HIPSTR* promoter demarcation is conserved between human and chicken

The highest level of turnover among all classes of functional elements identified by the ENCODE Project (Rands *et al.*, 2014) and the lack of known orthologs in other species are common features of lncRNAs (reviewed in (Kapusta and Feschotte, 2014) and (Ulitsky and Bartel, 2013)). For example, only 19 % of lncRNA families expressed in at least three out

of eleven tetrapod species studied by Necsulea *et al.* (Necsulea *et al.*, 2014) have originated more than 90 million years ago (Ma), and only 21 % of lncRNA loci that are present in human, chimpanzee and macaque have an orthologous lncRNA outside of primates (Necsulea *et al.*, 2014). Interestingly, human lncRNAs transcribed from canonical RNA Pol II promoters emit strong and consistent signal of purifying selection, as opposed to lncRNAs transcribed from enhancers (Marques *et al.*, 2013). Of all ENCODE cell lines (Consortium, 2012), HeLa-S3 and K562 cells have the highest *HIPSTR* expression (Figure 5 C). In agreement with *HIPSTR* transcription by RNA Pol II, K562 cells exhibit a characteristic promoter-associated H3K4me3 mark (Schneider *et al.*, 2004; Barski *et al.*, 2007) surrounding *HIPSTR* TSS (Figure 7 A). Notably, H3K4me3 does not mark just active, but also silent promoters (Schneider *et al.*, 2004; Barski *et al.*, 2007).

Since *HIPSTR* expression patterns are conserved between human and mouse (Figure 6 A and B), we asked whether other mammalian species also have the *HIPSTR* gene. Due to the absence of publicly available deep strand-specific RNA-seq data sets for placenta and testis for organisms other than human and mouse, we hypothesized that the presence of a H3K4me3 mark may help to indirectly estimate the degree of *HIPSTR* promoter conservation and hence – of *HIPSTR* transcription unit itself.

To this end, we first questioned the ability of DNA sequences surrounding *HIPSTR* TSS and occupied by H3K4me3 mark in HeLa-S3 and K562 (Consortium, 2012) to drive reporter gene transcription in four human cell lines (HeLa, HEK293, HepG2 and NT2/D1). We cloned sequences surrounding *HIPSTR* TSS upstream of the firefly luciferase gene, and compared the luminescence signal produced by cells transfected with different constructs (Figure 7 A). We tested seven sequences, and all of them produced a stronger luminescence signal than negative control plasmid (pGL3-Basic) in all four cell lines (Figure 7 B – E).

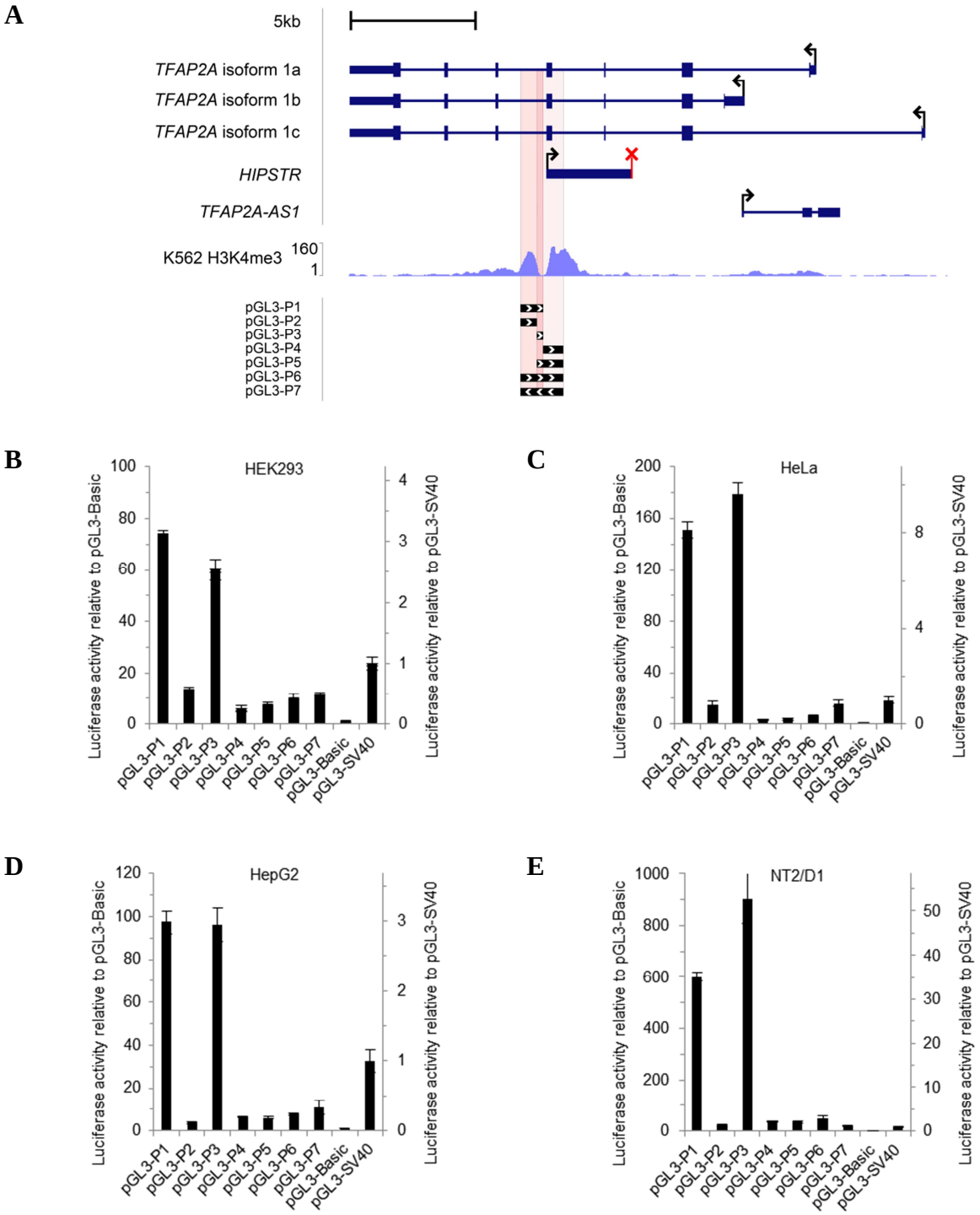


Figure 7. *HIPSTR* promoter-reporter assays. (A) Genomic positions of H3K4me3 ChIP-seq peaks around *HIPSTR* TSS (data from (Consortium, 2012)) and of the DNA sequences used for *HIPSTR* promoter-reporter assays (pGL3-P1 to -P7). (B – E) *HIPSTR* promoter-reporter assays in HEK293 (B), HeLa (C), HepG2 (D), and NT2/D1 (E) cells. Experiments were performed in triplicate, error bars represent SD.

We next analyzed public ChIP-seq data for H3K4me3 mark distribution in: (i) liver samples of 10 mammals (Villar *et al.*, 2015), (ii) mouse and rooster testis samples (Li, X. Z. *et al.*, 2013b), as well as in (iii) frog (van Heeringen *et al.*, 2014) and (iv) zebrafish (Zhang, Y. *et al.*, 2014) embryos. We found H3K4me3 ChIP-seq peaks around *HIPSTR* TSS orthologous region in the liver samples of all 10 mammals tested, in testis samples of mouse and, surprisingly, rooster, but not in any of the frog or zebrafish embryos (Figure 8). These results suggest that functional *HIPSTR* promoter demarcation existed approximately 325 Ma in a common ancestor of human and chicken (Kapusta and Feschotte, 2014), and that therefore other amniotes likely have the *HIPSTR* gene.

4.3 *HIPSTR* silencing in HEK293 cells upregulates developmental genes

HIPSTR levels show no distinctive pattern between tumor and non-tumor cell lines (Figure 5 A) and do not correlate with *TFAP2A* gene expression (Figure 5 B – D). We reasoned that *HIPSTR* might regulate other genes in the *TFAP2A* locus, or elsewhere in the genome in *trans*. We used HEK293 human embryonic kidney cells that express high levels of *TFAP2A* mRNA and *HIPSTR* (Figure 5 B). HEK293 cells were also successfully used to characterize *trans*-acting lncRNAs (Orom *et al.*, 2010; Lai *et al.*, 2013). Thus, we used microarrays to simultaneously monitor changes in gene expression in the *TFAP2A* locus and genome-wide after *HIPSTR* depletion with two ASOs (ASO #1 and ASO #2) in HEK293 cells. Consistent with a relatively short half-life of *HIPSTR* in HEK293 cells (38 min) (Figure 9 A), efficient *HIPSTR* knockdown with a pool of targeting ASOs was achieved as early as 6 h after ASOs transfection (~71 %, Figure 9 B). The highest knockdown efficiency was reached 24 h after transfection (89 %, Figure 9 B), with a decrease in efficiency over time to 49 % at 72 h (Figure 9 B).

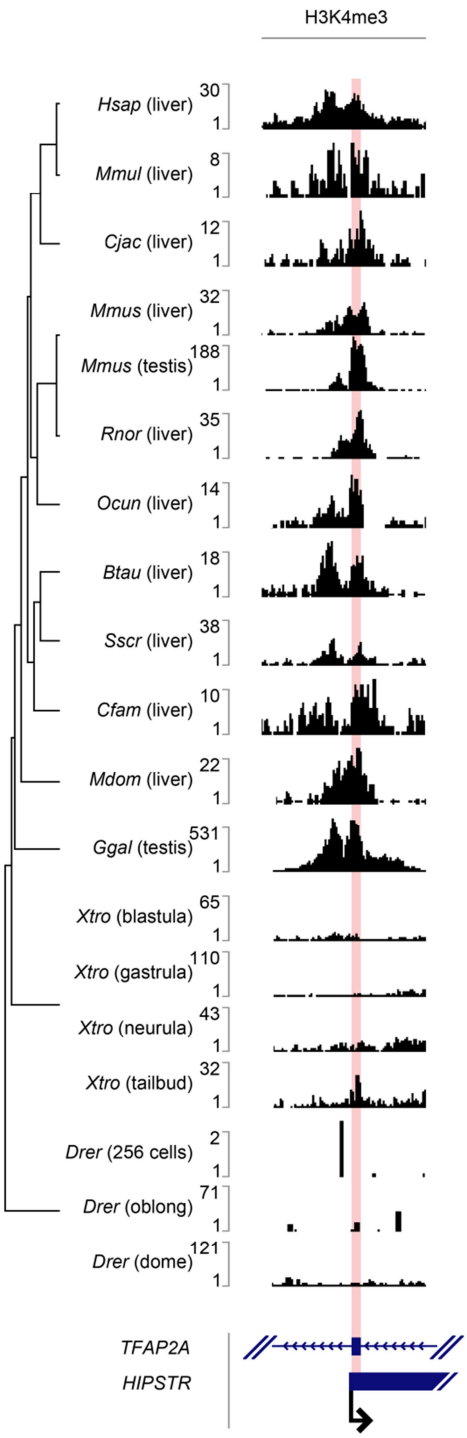
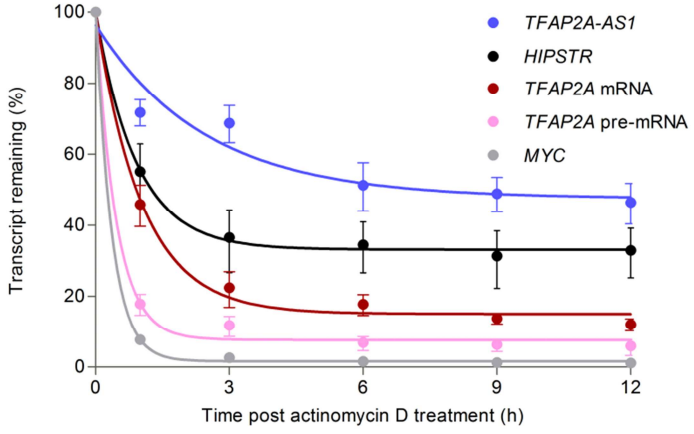


Figure 8. *HIPSTR* promoter demarcation is conserved in Amniota. Analysis of H3K4me3 ChIP-seq data from (Li, X. Z. *et al.*, 2013b; van Heeringen *et al.*, 2014; Zhang, Y. *et al.*, 2014; Villar *et al.*, 2015) reveals conserved *HIPSTR* promoter demarcation across the genomes of 10 mammalian species and chicken, and absence of H3K4me3 mark around *HIPSTR* TSS orthologous region in frog and zebrafish. For each species, the maximal value on the y-axis scale corresponds to the highest H3K4me3 peak across the entire *TFAP2A* locus for that species.

A



B

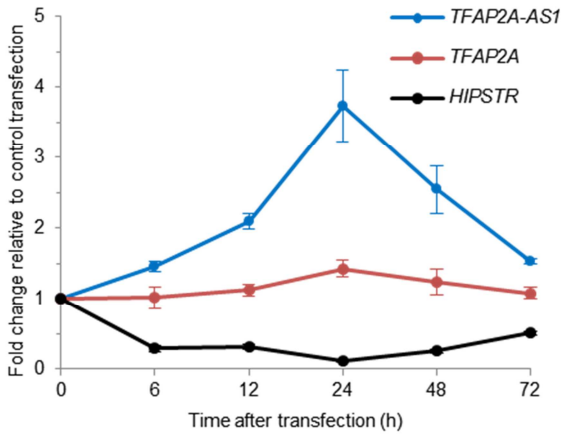


Figure 9. Titration of *HIPSTR* knockdown experiment in HEK293 cells. (A) HEK293 cells were treated with actinomycin D, and decay rates of different transcripts were measured with RT-qPCR. Half-life of *HIPSTR* (38 min) is shorter than of *TFAP2A-AS1* lncRNA (102 min), comparable with half-life of *TFAP2A* mRNA (43 min), and longer than that of *MYC* (15 min) or *TFAP2A* pre-mRNA (19 min). (B) Changes in the *TFAP2A* locus genes over time after *HIPSTR* knockdown. For *HIPSTR* knockdown experiments HEK293 cells were transfected with a combination of *HIPSTR*-targeting ASO #1 and ASO #2 or with non-targeting ASO CTL. Expression of the *TFAP2A* locus genes on (A, B) was determined with RT-qPCR. Experiments were performed in triplicate, error bars represent SD.

HIPSTR knockdown with either of the two ASOs (Figure 10 A) resulted in a moderate (~4.0 to 5.5-fold) upregulation of *TFAP2A-AS1*, another lncRNA in the *TFAP2A* locus (Figure 10 A). Overall *TFAP2A* mRNA (Figure 10 A) and pre-mRNA (Figure 10 B) remained unchanged, and TFAP2A protein levels remained undetectable (Figure 10 C). Although we observed a weak (~1.5 to 1.9-fold) upregulation of the *TFAP2A* isoform 1b (Figure 10 B), expression of the predominant *TFAP2A* isoform 1a was not affected by *HIPSTR* knockdown

(Figure 10 B). More importantly, silencing of *HIPSTR* resulted in a significant differential expression of 381 (437 probes) annotated genes located outside of the *TFAP2A* locus (Figure 11 A; Additional file 1: Table 2). Of these, 378 (~99.2 %) were upregulated, suggesting a repressive function for *HIPSTR* in HEK293 cells (Figure 11 A). Gene ontology analysis of the protein-coding genes differentially expressed upon *HIPSTR* knockdown revealed their enrichment in "Developmental Process" and "Cell Differentiation" categories (Figure 11 B). Similar to HEK293 cells, *HIPSTR* knockdown in LNCaP prostate carcinoma cells resulted in upregulation of developmentally regulated TFs, such as *SNAI1*, *ZSCAN10* and several others (Figure 11 C). Moreover, genomic regions surrounding TSSs of the differentially expressed genes were enriched in recognition motifs of NF-Y, POU5F1 and KLF4 (Figure 11 D), TFs important for pluripotency maintenance (Takahashi *et al.*, 2007; Oldfield *et al.*, 2014). Although the endogenous *HIPSTR* gene is not expressed in NT2/D1 embryonal carcinoma cells, in this pluripotent cell line, two *HIPSTR* promoter-luciferase constructs (pGL3-P1 and pGL3-P3) produced ~35 – 50-times stronger luminescence signal than did positive control construct (pGL3-SV40) (Figure 7 E). This signal was also ~600 – 900-times stronger than the signal from NT2/D1 cells transfected with negative control plasmid (Figure 7 E). We observed a similar trend for these two constructs, yet to a much lower extent, in non-pluripotent HeLa, HEK293 and HepG2 cells (Figure 7 B – D).

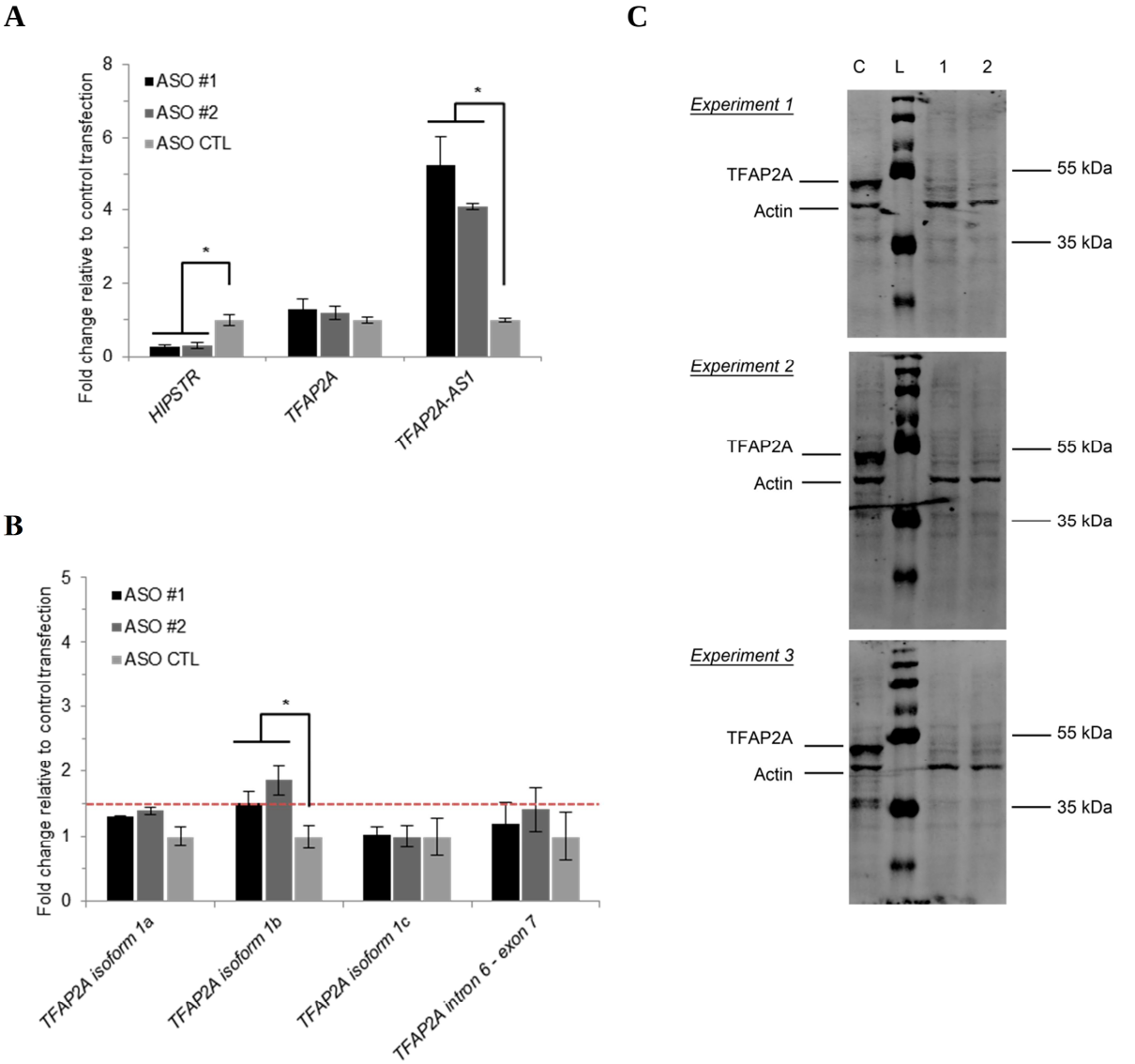
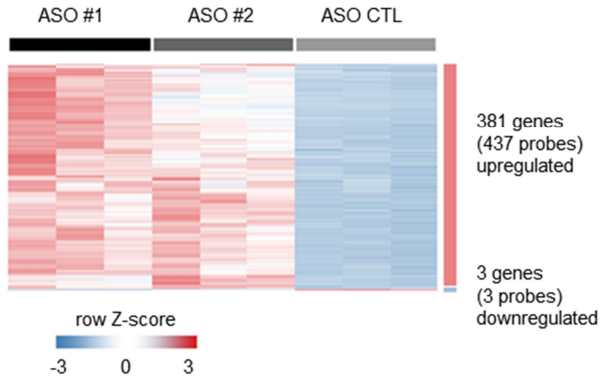
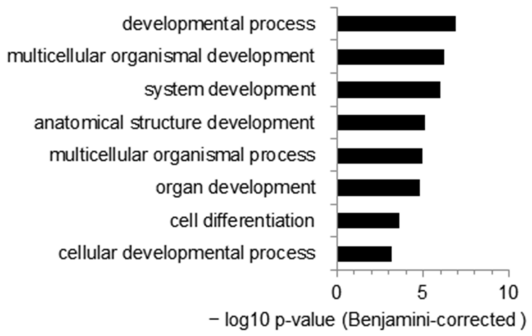


Figure 10. Knockdown of *HIPSTR* significantly upregulates *TFAP2A-AS1* levels, but not overall *TFAP2A* mRNA, pre-mRNA or TFAP2A protein levels. (A) Effect of *HIPSTR* knockdown on the expression of *TFAP2A* locus genes in HEK293 cells, as measured by RT-qPCR. (B) *HIPSTR* knockdown upregulates *TFAP2A* isoform 1b, but not predominant *TFAP2A* isoform 1a, or *TFAP2A* pre-mRNA (primers to intron 6 – exon 7 junction), as measured with RT-qPCR. Experiments shown on (A, B) were performed in triplicate, error bars represent SD; the asterisks indicate statistical significance of the expression differences calculated with two-tailed t-test, equal variance (p-value < 0.01 on A, p-value < 0.05 on B). (C) *HIPSTR* knockdown does not affect TFAP2A protein levels. We used total protein extracts from HEK293 cells transfected with a combination of ASO #1 and ASO #2 (*lane 1*) or with ASO CTL (*lane 2*) to perform western blot with anti-TFAP2A and anti-Actin antibodies; total protein extract from HEK293 cells overexpressing TFAP2A isoform 1a served as positive control for TFAP2A antibody (*lane C*); detection of Actin served as loading control; PageRuler Plus Prestained Protein Ladder was used to estimate approximate MW of the proteins (*lane L*).

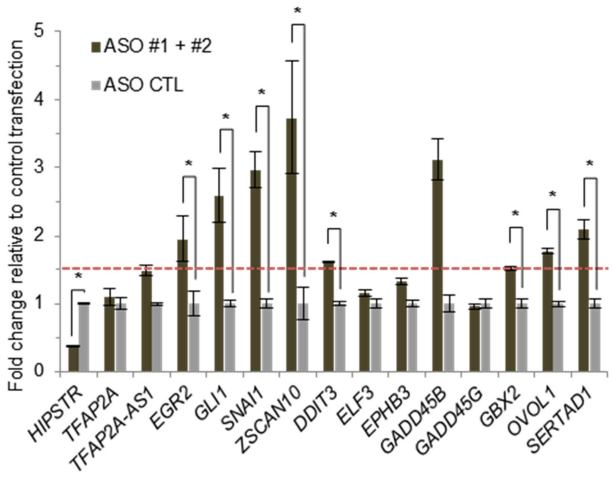
A



B



C



D

Known motif	Transcription factor	Enrichment q-value (Benjamini)	Position relative to TSS
	NF-Y	0.0223	-100..+400
	POU5F1	0.0223	-100..+400
	KLF4	0.0370	-400..+100

Figure 11. Developmental genes are affected by *HIPSTR* knockdown in HEK293 cells. (A) *HIPSTR* knockdown in HEK293 cells leads to a significant upregulation of 381 annotated genes (380 of them outside of *TFAP2A* locus) (1 % FDR, fold-change > 2; Additional file 1: Table 2). (B) GO categories significantly enriched with genes upregulated upon *HIPSTR* knockdown in HEK293 cells. (C) *HIPSTR* knockdown in LNCaP results in upregulation of developmental genes, as measured with RT-qPCR. Experiments were performed in triplicate, error bars represent SD; the asterisks indicate statistical significance of the observed changes calculated with two-tailed t-test, equal variance (p-value < 0.05). (D) Motif analysis reveals significant enrichment of NF-Y, POU5F1, and KLF4 recognition motifs around TSSs of genes differentially expressed upon *HIPSTR* knockdown in HEK293 cells.

4.4 *HIPSTR* is not consistently co-activated with *TFAP2A* in developmental models *in vitro*

HIPSTR overlaps *TFAP2A*, the gene encoding a TF that is essential for vertebrate neural crest development, as exemplified by mouse knockout studies (Schorle *et al.*, 1996; Zhang, J. *et al.*, 1996), and by epigenetic profiling of chimp and human neural crest-like cells (NCCs) derived from pluripotent cells *in vitro* (Rada-Iglesias *et al.*, 2012; Prescott *et al.*, 2015). *TFAP2A* gene is also induced in mouse (Guo, G. *et al.*, 2010) and human (Cheng *et al.*, 2004; Aghajanova *et al.*, 2012) trophectoderm. This induction can be reproduced *in vitro* by human trophoblast-like cells (hTBCs) derivation from pluripotent cells (Xu *et al.*, 2002; Marchand *et al.*, 2011). Finally, *TFAP2A* expression can be transiently induced in human embryonal carcinoma NT2/D1 cells grown in the presence of ATRA (Luscher *et al.*, 1989).

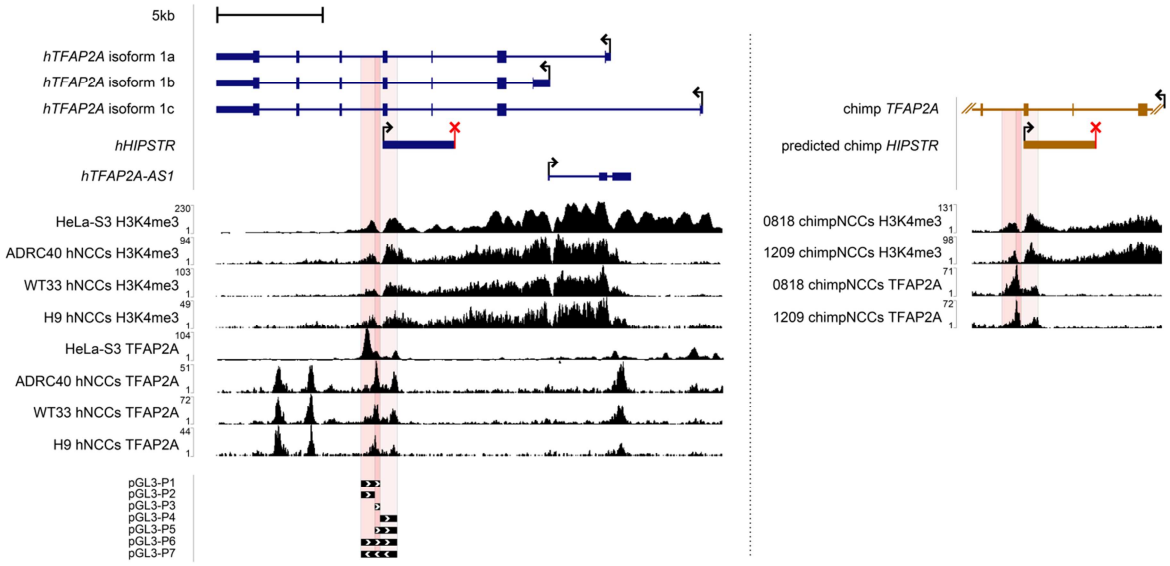
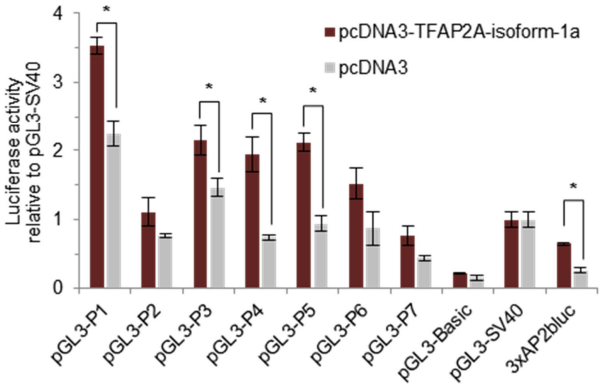


Figure 12. Conserved TFAP2A binding to *HIPSTR* promoter region. UCSC Genome Browser snapshot showing positions of *HIPSTR*, three *TFAP2A* isoforms and *TFAP2A-AS1*, as well as H3K4me3 and TFAP2A ChIP-seq reads mappings from (Consortium, 2012) (HeLa-S3 cells) and (Prescott *et al.*, 2015) (three hNCCs and two chimp NCCs lines), and positions of the DNA sequences used for *HIPSTR* promoter-reporter assays (pGL3-P1 to -P7).

A



B

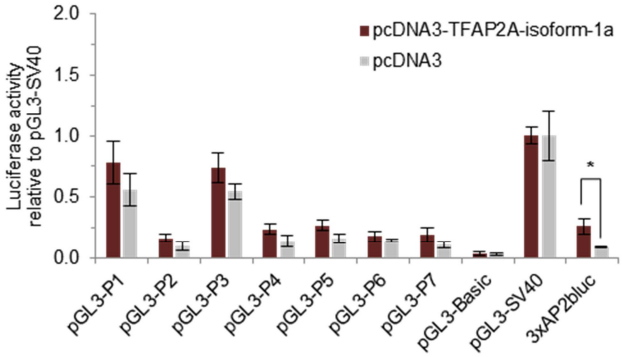


Figure 13. TFAP2A isoform 1a overexpression increases *HIPSTR* promoter activity in luciferase reporters. Luciferase reporter assays in HEK293 (A) or HepG2 (B) cells. DNA sequences surrounding *HIPSTR* TSS cloned upstream of the firefly luciferase gene were co-transfected with the plasmid expressing *Renilla* luciferase and TFAP2A isoform 1a overexpressing plasmid or empty vector; pGL3-Basic served as negative control (no promoter upstream of the firefly luciferase); pGL3-SV40 served as positive control (SV40 promoter upstream of the firefly luciferase); 3xAP2bluc served as positive control for transactivation by TFAP2A isoform 1a.

Interestingly, we found that TFAP2A ChIP-seq peaks were mapped to sequences upstream and downstream of *HIPSTR* TSS in HeLa-S3 cells (Consortium, 2012), hNCCs and an orthologous region in chimp NCCs (Rada-Iglesias *et al.*, 2012; Prescott *et al.*, 2015) (Figure 12). We therefore asked whether TFAP2A could regulate *HIPSTR* expression. For this, we overexpressed TFAP2A isoform 1a (predominant in HEK293), and observed a significant increase in the luminescence signal from *HIPSTR* promoter-luciferase constructs in HEK293 (Figure 13 A), but not in HepG2 hepatocellular carcinoma cells (Figure 13 B).

Overexpression of the three TFAP2A isoforms described to date (Berlato *et al.*, 2011) also upregulated endogenous *HIPSTR* levels in HEK293 (Figure 14 A and C), but was insufficient to start *HIPSTR* expression in HepG2 cells that lack endogenous *HIPSTR* expression (Figure 14 B and C).

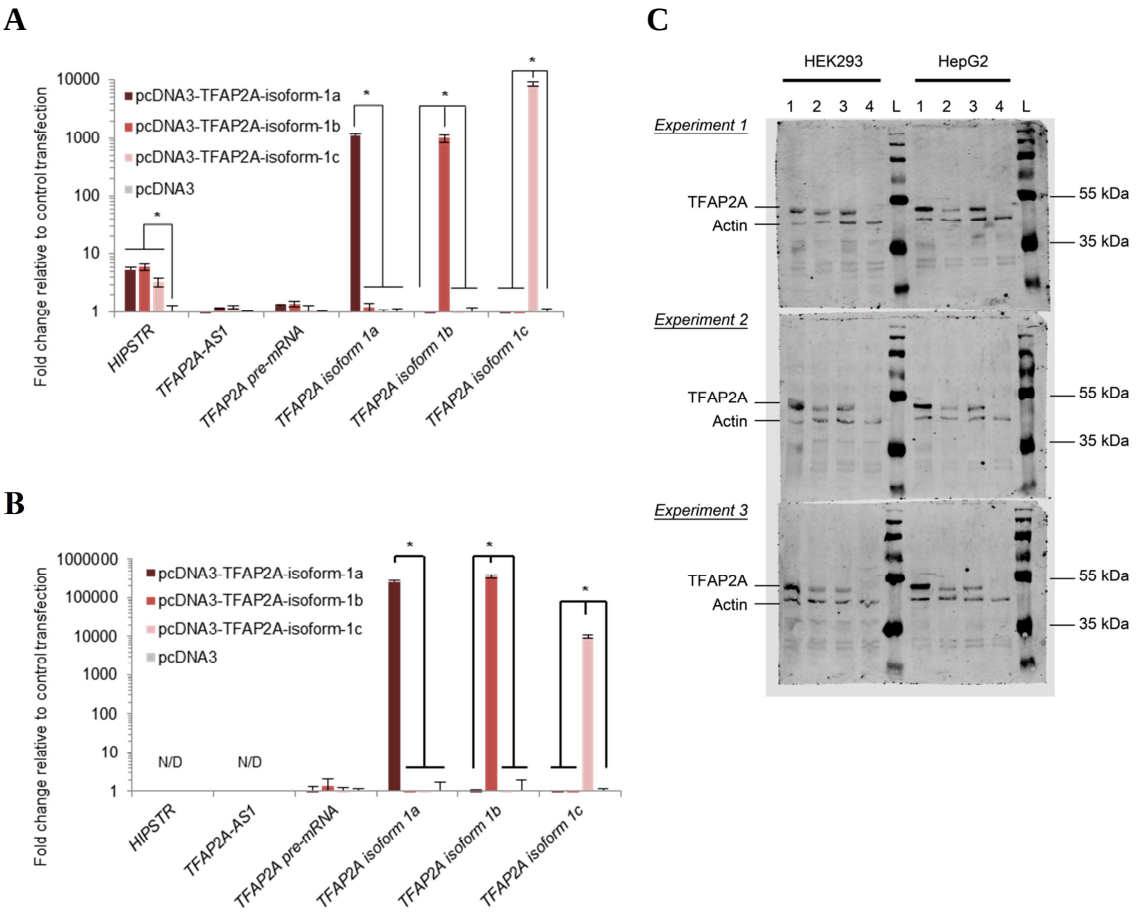


Figure 14. TFAP2A is capable of upregulating endogenous *HIPSTR* expression, but is not sufficient for starting *HIPSTR* transcription. HEK293 (A) and HepG2 (B) cells were transfected with plasmids overexpressing TFAP2A isoforms 1a (dark red), 1b (red), or 1c (pink). Shown are expression levels of *HIPSTR*, *TFAP2A-AS1*, *TFAP2A* isoforms and pre-mRNA, relative to cells transfected with empty plasmid, as measured by RT-qPCR. Experiments were performed in triplicate, and error bars represent SD; the asterisks indicate statistical significance of the observed changes calculated with two-tailed t-test, equal variance (p-value < 0.01); N/D – not detected. (C) Western blot showing efficient overexpression of TFAP2A isoforms 1a (*lane 1*), 1b (*lane 2*), and 1c (*lane 3*) in the indicated cell lines in three independent experiments, as compared to cells transfected with empty vector (*lane 4*); detection of Actin served as loading control; PageRuler Plus Prestained Protein Ladder was used to estimate approximate MW of the proteins (*lane L*).

RESULTS

We showed above that *HIPSTR* knockdown in HEK293 cells unexpectedly led to upregulation of development-related genes in these cells. Since *HIPSTR* gene is completely overlapped by the *TFAP2A* gene and can be regulated by the protein product of the latter, we next reasoned that both genes could be simultaneously induced during development. Thus, we induced *TFAP2A* expression *in vitro* by differentiating hESCs into hNCCs and hTBCs, as well as by treating NT2/D1 cells with ATRA. Strong induction of *TFAP2A* transcription in *in vitro* derived hNCCs (> 200-fold, Figure 15 A) and hTBCs (> 40-fold, Figure 15 B) was accompanied by only a moderate (~9.4-fold, Figure 15 A) and a weak (~1.8-fold after 1 day of differentiation, Figure 15 B) upregulation of *HIPSTR*, respectively. In turn, ATRA treatment of NT2/D1 cells led to upregulation of *TFAP2A* gene, but not of *HIPSTR* (Figure 15 C). Notably, the upregulation of *TFAP2A-AS1* lncRNA divergently transcribed from *TFAP2A* isoform 1b promoter was comparable to that of *TFAP2A* in hNCCs (Figure 15 A), hTBCs (Figure 15 B), and ATRA-treated NT2/D1 cells (Figure 15 C).

Together, upregulation of developmental genes in HEK293 cells upon *HIPSTR* knockdown and lack of consistent co-activation of *HIPSTR* and *TFAP2A* in three developmental models (differentiation of hESCs into hNCCs or hTBCs and treatment of human embryonal carcinoma with ATRA), raised the question whether *HIPSTR* is expressed and functions in early embryonic development independently of *TFAP2A* gene.

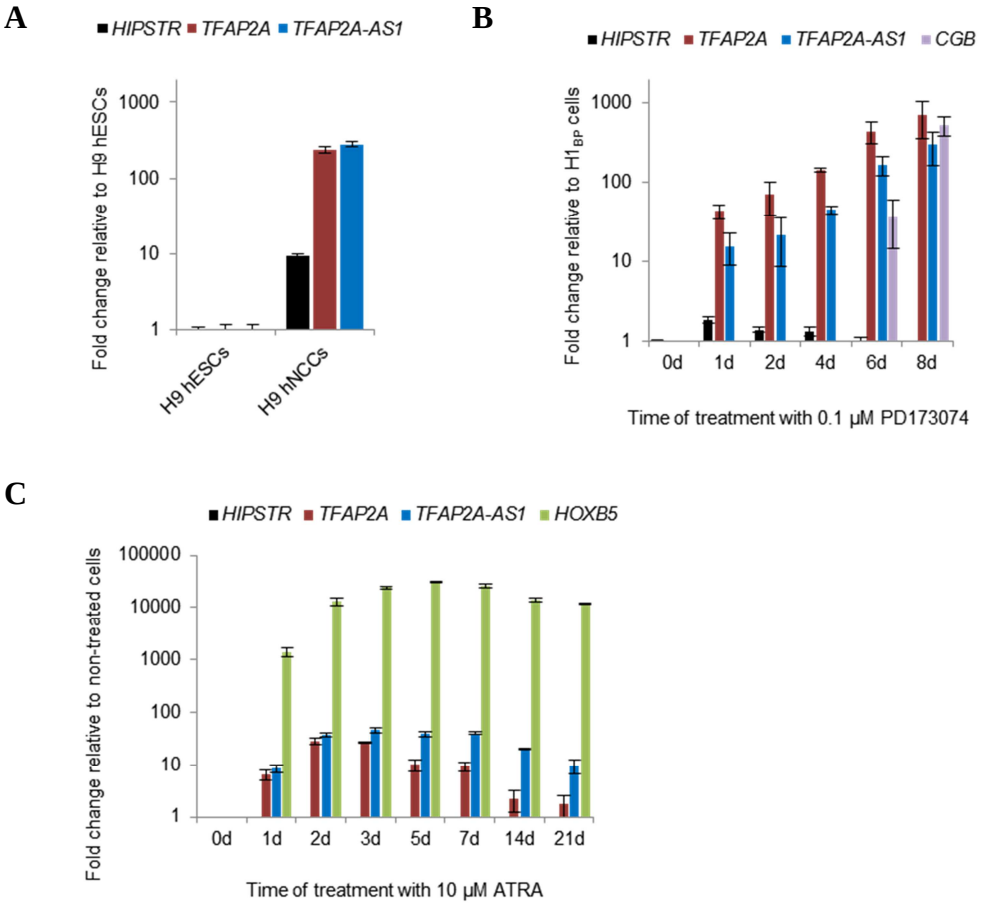


Figure 15. *HIPSTR* and *TFAP2A* are not consistently co-induced in *in vitro* developmental models. *HIPSTR* is moderately co-upregulated with *TFAP2A* in *in vitro* derived hNCC (A), weakly co-upregulated with *TFAP2A* in *in vitro* derived hTBCs (B), and not co-upregulated with *TFAP2A* in NT2/D1 cells treated with ATRA (C), as measured by RT-qPCR. Upregulation of *TFAP2A* gene itself (hNCCs marker), of *CGB* (hTBCs marker), or *HOXB5* gene (induced by ATRA treatment in NT2/D1 cells (Luscher *et al.*, 1989)) served as positive controls. Experiments were performed in triplicate, error bars represent SD.

4.5 *HIPSTR* expression in the early human embryo is restricted to a subset of cells

If *HIPSTR* acts independently from *TFAP2A* gene, activation of the former may occur prior to activation of the entire *TFAP2A* locus during human development. To address this possibility, we sought evidence of *HIPSTR* transcription during early stages of human embryonic development in public data. In the past few years, several studies reported successful transcriptome sequencing of individual blastomeres of early human and mouse embryos (Xue *et al.*, 2013; Yan *et al.*, 2013; Biase *et al.*, 2014; Tohonen *et al.*, 2015). We first screened for *HIPSTR* expressing cells in the strand-specific single-cell-tagged reverse-transcription (STRT) RNA-seq libraries from (Tohonen *et al.*, 2015). Surprisingly, we found that *HIPSTR* and not *TFAP2A* or *TFAP2A-AS1* was present in 2-3 days old human embryos (in one cell from a 4-cell stage embryo, and in eight cells from five separate 8-cell stage embryos) (Figure 16). To visualize and estimate *HIPSTR* and *TFAP2A* expression during early human embryonic development, we mapped (Figure 17) and quantified (Figure 18 A and B) RNA-seq reads from two other data sets (Xue *et al.*, 2013; Yan *et al.*, 2013). As these RNA-seq data sets are not strand-specific, we present *HIPSTR* expression as “underestimated” and “overestimated” FPKM values, by accordingly excluding or including the reads mapping to exons of *TFAP2A* that overlap *HIPSTR*. We found that *HIPSTR* is specifically upregulated in 8-cell and morula stage human embryos (Figure 17, Figure 18 A and B). Moreover, *HIPSTR* expression is restricted to only a subset of cells within 8-cell and morula stage embryos (Figure 19 A and B). We also noticed that *HIPSTR* was expressed by only few K562 cells within a population (23 out of 96 cells), when analyzed at the single-cell level (Figure 19 C). A similar pattern of expression was reported earlier for several mouse lncRNAs with low population-level expression in bone-marrow-derived dendritic cells (Shalek *et al.*, 2013). Importantly, *TFAP2A-AS1*, *TFAP2A* mRNA and pre-mRNA levels do

not change significantly throughout the human embryonic development time course analyzed (Figure 18 A – D). Hence, *HIPSTR* gene is activated independently from and prior to *TFAP2A* during the course of development shortly after a major wave of human EGA (Yan *et al.*, 2013).

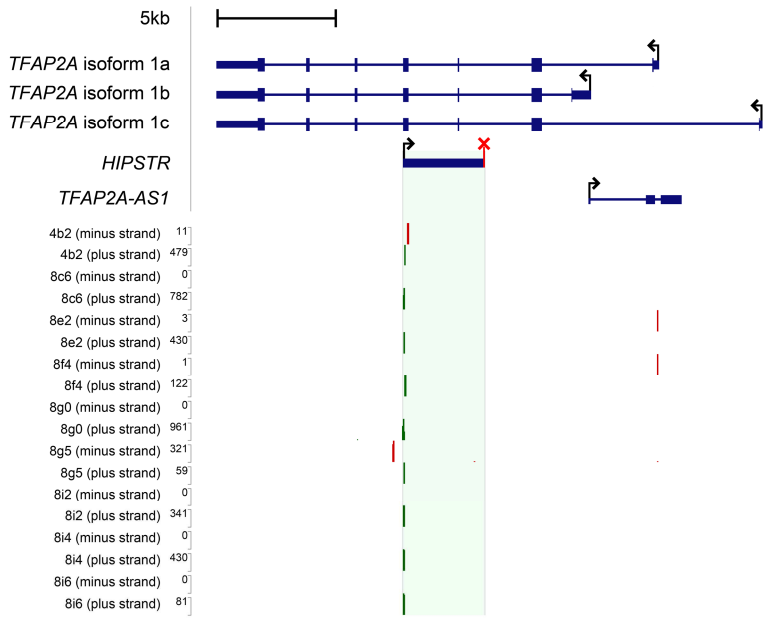


Figure 16. *HIPSTR* is expressed in 2-3 days old human embryos. Mapping of the 5′-ends of transcripts with strand-specific STRT-seq data from (Tohonen *et al.*, 2015) shows specific expression of *HIPSTR* in one cell (4b2) from a 4-cell human embryo, and in eight cells (8c6 through 8i6) originating from five different 8-cell human embryos; cell names are as in (Tohonen *et al.*, 2015).

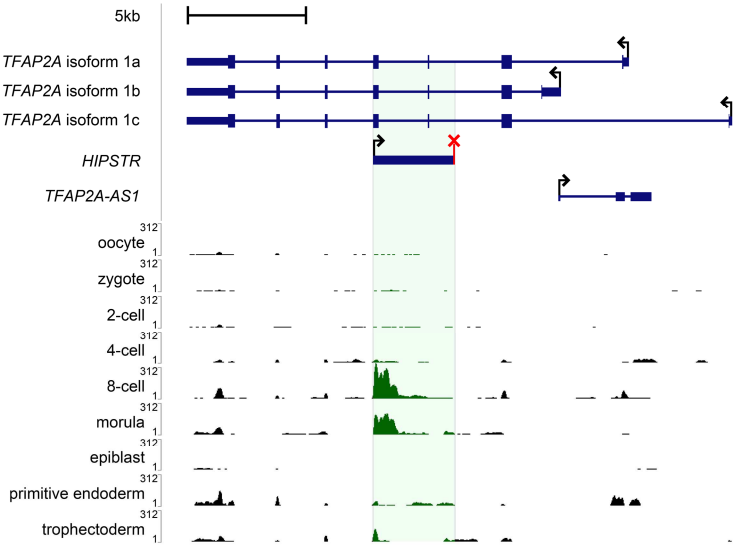


Figure 17. *HIPSTR* expression in 8-cell and morula stage embryos. Mapping of RNA-seq reads from (Yan *et al.*, 2013) illustrates specific expression of *HIPSTR*, and not *TFAP2A* or *TFAP2A-AS1*, in 8-cell and morula-stage human embryos.

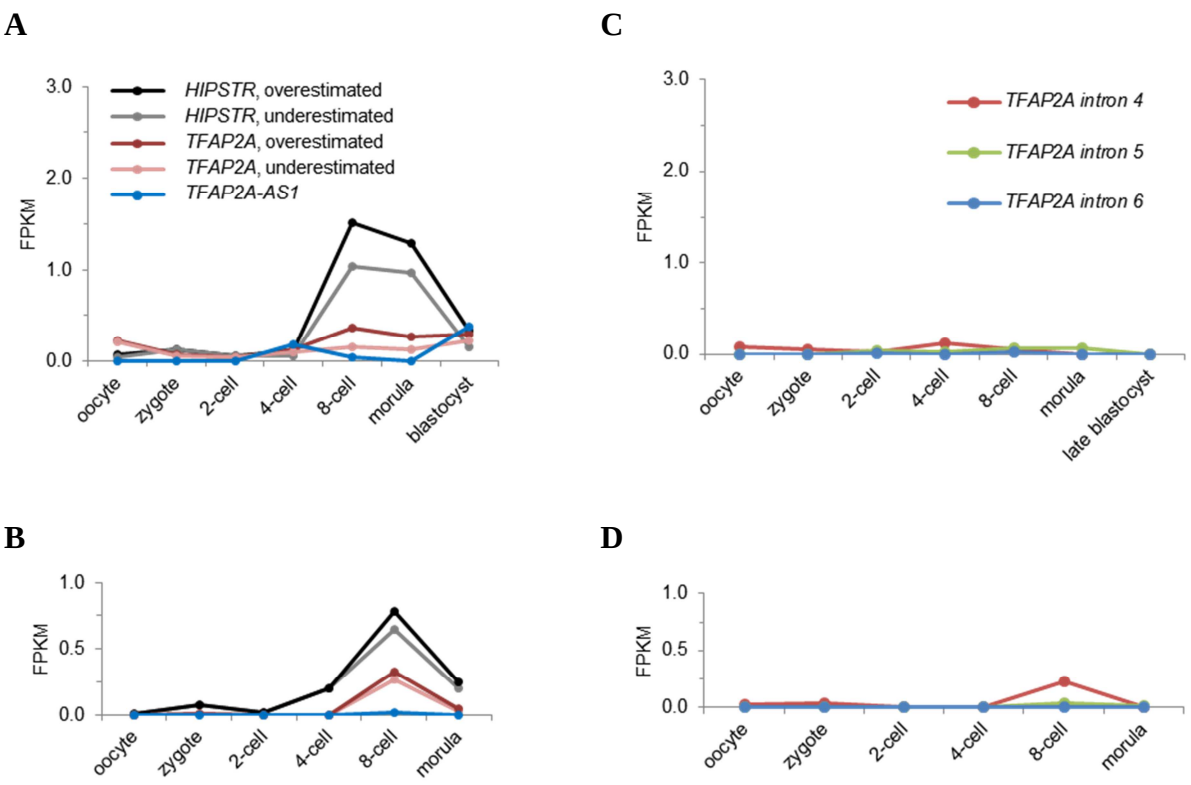


Figure 18. Quantification of *HIPSTR* expression in human oocytes and early embryos. Average *HIPSTR* expression through early human embryonic development, as estimated by analyzing RNA-seq data from (Yan *et al.*, 2013) in (A) or from (Xue *et al.*, 2013) in (B). Plotted are under- and overestimated FPKM values for *HIPSTR* and *TFAP2A* expression (see text). *TFAP2A* pre-mRNA is not detectable at significant levels in the corresponding data sets; data in (C) are from (Yan *et al.*, 2013), and in (D) – from (Xue *et al.*, 2013).

Intriguingly, multiple single-cell RNA-seq reads from a public data set from (Biase *et al.*, 2014) mapped within *HIPSTR* orthologous region in the mouse genome specifically in the 2-cell embryos, the stage at which mouse EGA is initiated (Biase *et al.*, 2014) (Figure 20 A – C). These observations are in conflict with mouse single-cell RNA-seq data from (Xue *et al.*, 2013), where no evidence of expression in the *HIPSTR* orthologous region was detected at all stages, including 2-cell stage (not shown). Therefore, these results suggest that *HIPSTR* likely functions after a major wave of EGA in human embryos, but whether it is the case for mouse embryonic development remains an open question.

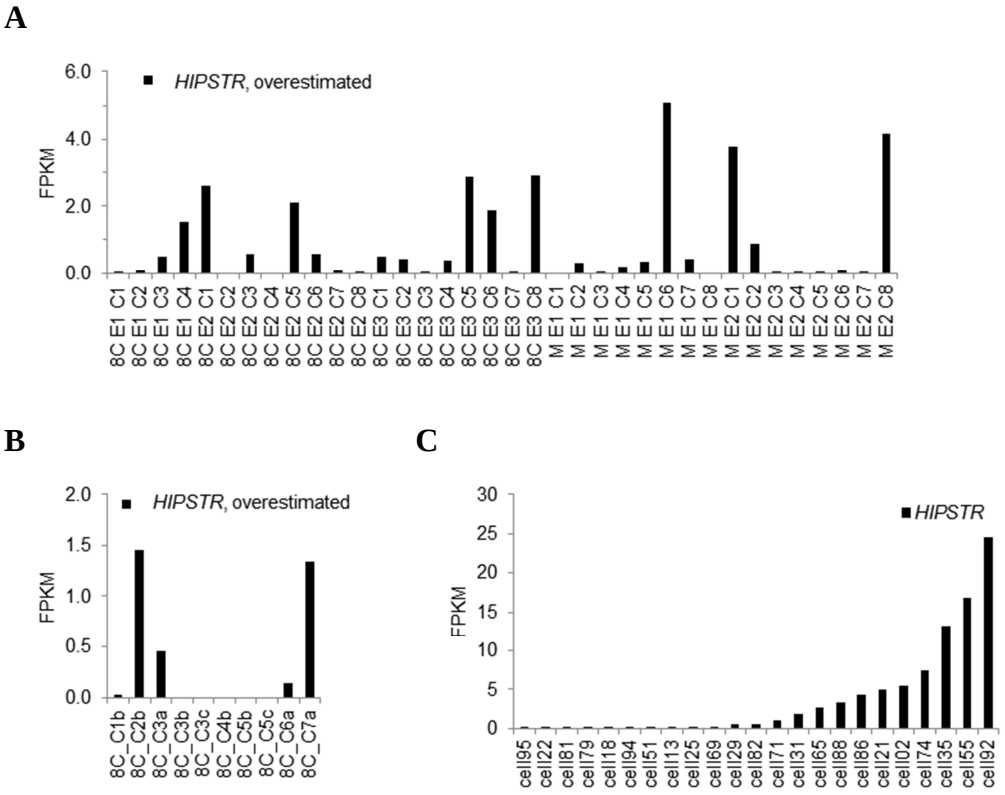
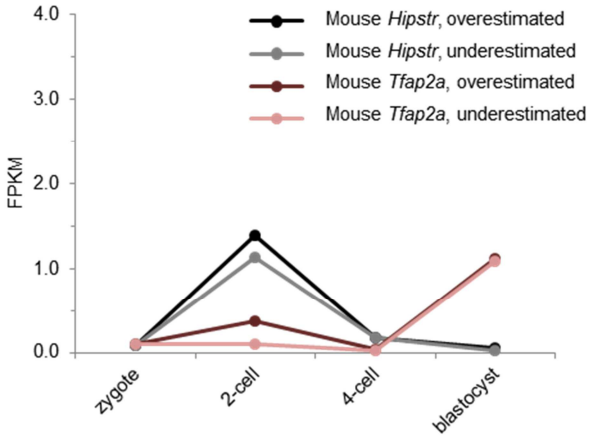
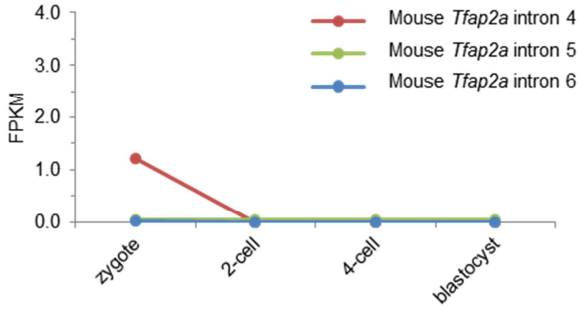


Figure 19. *HIPSTR* expression is restricted to a subset of cells within early human embryos and within a population of K562 cells. Quantification of RNA-seq data for 8-cell- and morula-stage embryos from (Yan *et al.*, 2013) in (A) or RNA-seq data for 8-cell-stage embryos from (Xue *et al.*, 2013) in (B); plotted are overestimated FPKM values for *HIPSTR* expression (see main text for Figure 18). (C) *HIPSTR* is expressed by a subset of cells within a population of K562 cells. Only cells where *HIPSTR* expression is detected are shown (23 cells, FPKM > 0); 73 cells out of 96 do not express *HIPSTR* and are not shown. Data are from (Luo *et al.*, 2014).

A



B



C

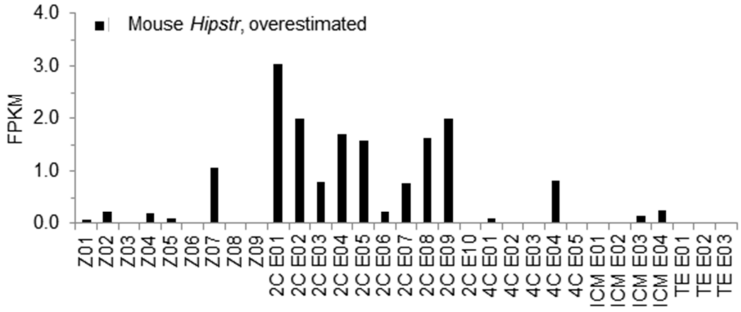


Figure 20. Expression of mouse *Hipstr* ortholog in early mouse embryo. (A) Mouse *Hipstr* ortholog is induced during the major wave of mouse EGA (2-cell stage); analyses of aggregate data for each stage from (Biase *et al.*, 2014) is shown; these data are in conflict with the data from (Xue *et al.*, 2013) where we did not detect mouse *Hipstr* at any stage; (B) *TFAP2A* pre-mRNA is not detectable in 2-cell stage mouse embryos; data from (Biase *et al.*, 2014). (C) Mouse *Hipstr* ortholog expression is induced in nine out of ten 2-cell embryos from (Biase *et al.*, 2014). Plotted on (A, C) are under- and overestimated FPKM values for *Hipstr* and *Tfap2a* expression (see main text for Figure 18).

4.6 Cell-to-cell variability in expression of lncRNAs is higher than that of mRNAs

In a recent work, Cabili *et al.* used single-molecule RNA-FISH approach and concluded that no difference exists in cell-to-cell variability in expression of mRNAs and lncRNAs (Cabili *et al.*, 2015). This argues against a hypothesis that lncRNAs with low population-level abundance are instead expressed at high levels by a subset of cells within that population (Dinger *et al.*, 2009). In agreement with the latter hypothesis, expression pattern of *HIPSTR* in the early human embryos (Figure 19 A and B), and in K562 cell line (Figure 19 C) is restricted to a subpopulation of cells. Similarly, Yan *et al* have demonstrated that expression of lncRNAs is heterogeneous among individual human cells (Yan *et al.*, 2013).

To resolve this discrepancy between single-molecule RNA-FISH results and observations from single-cell RNA-seq data, we next systematically explored patterns of cell-to-cell expression variability of lncRNAs and mRNAs in human cells. For this, we used three single-cell RNA-seq data sets – from human totipotent blastomeres (8-cell and morula-stage embryos) and hESCs (both from ref. (Yan *et al.*, 2013)), and from K562 cell line (from ref. (Luo *et al.*, 2014)). As lncRNAs are generally less abundant than mRNAs (Cabili *et al.*, 2011), we considered only genes expressed in the range 3 – 30 FPKM. We noted that the distribution of the numbers of cells in which the genes were expressed (> 3 FPKM, see Methods) was a mixture distribution. We fitted this mixture distribution with a finite mixture model with two populations, having high or low heterogeneity of expression and used this model to compare the expression heterogeneity of lncRNAs and mRNAs.

Of the lncRNAs expressed in the range 3 – 30 FPKM only a tiny fraction showed low heterogeneity of expression – 1.5 %, 2.1 %, and 0.6 % in human totipotent blastomeres (Figure 21 A), hESCs (Figure 21 B), and K562 cells (Figure 21 C), respectively. For example, known pluripotency regulators *LINC-ROR* (Loewer *et al.*, 2010) and *TUNAR* (Lin *et al.*,

2014) were associated with high heterogeneity of expression in hESCs in our model and with the transcriptome annotation used in the present work (Additional file 1: Table 3), and *HIPSTR* showed high heterogeneity of expression in 8-cell and morula-stage human embryos, and in K562 cells (Additional file 1: Tables 4 and 5). Additionally, 129 lncRNAs were highly heterogeneous, being strongly expressed with FPKM value > 5 in a single totipotent blastomere of a single embryo, while expressed with FPKM value < 1 in all other cells of all other sampled totipotent (8-cell- or morula-stage) embryos (Additional file 1: Table 4). Interestingly, blastomeres from the same 8-cell embryo are more similar to each other than to blastomeres from a separate 8-cell embryo (Xue *et al.*, 2013). In this context, strong expression of several lncRNAs detected at high levels in only one totipotent blastomere across several embryos likely illustrates an extremely specific spatiotemporal expression pattern of lncRNAs. The remarkably high heterogeneity of expression of lncRNAs was in a stark contrast to the much lower heterogeneity of expression of mRNAs with comparable expression levels (3 – 30 FPKM), of which 23.9 %, 27.3 %, and 8.8 % were associated with low heterogeneity in human totipotent blastomeres, hESCs, and K562 cells, respectively (Figure 21 A – C).

Based on these data, we conclude that in addition to previously reported tissue and developmental stage expression specificity (Ravasi *et al.*, 2006; Cabili *et al.*, 2011; Yan *et al.*, 2013), heterogeneous expression in a population of seemingly identical cells is another common feature of human lncRNAs. This observation is of special importance for human totipotent embryos (e.g. 8-cell or morula stage), where the number of cells is finite, and where heterogeneity of expression of lncRNAs is strongly pronounced (Figure 21 A) and therefore might have important physiological implications.

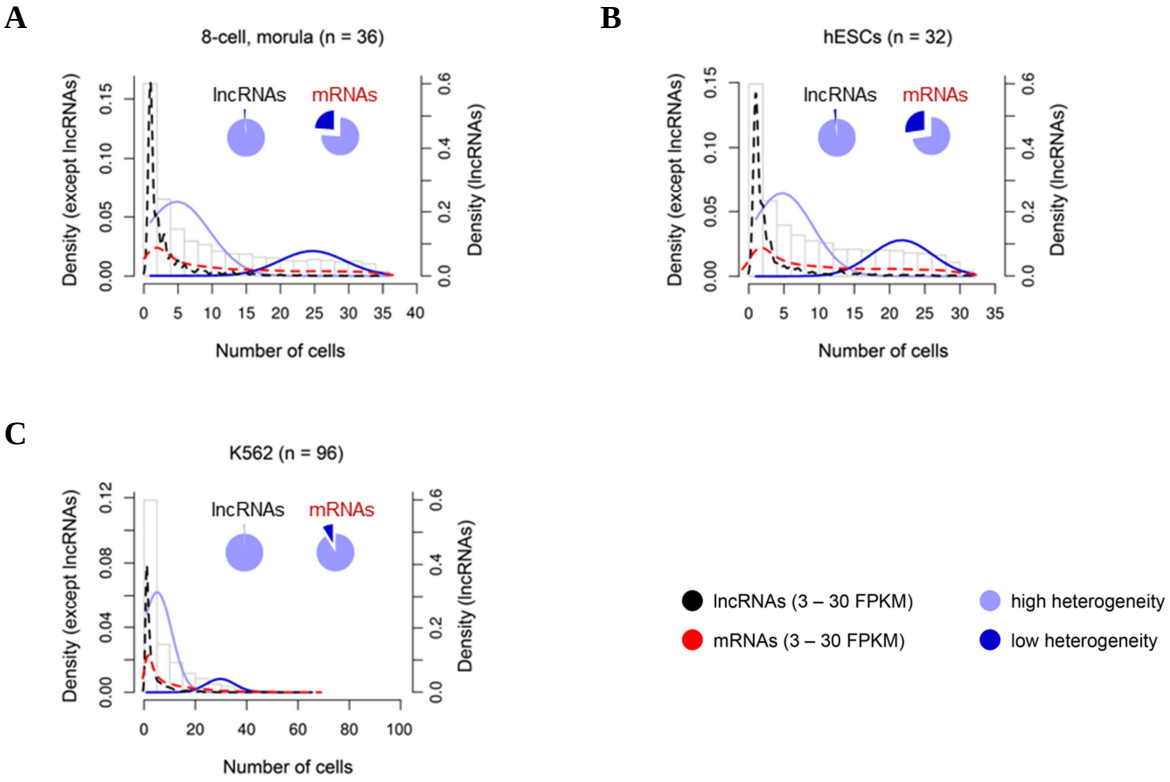


Figure 21. LncRNAs show higher heterogeneity of expression than mRNAs. (A – C) LncRNAs are more heterogeneously expressed than mRNAs, as evidenced by single-cell RNA-seq analyses. Plotted are density distributions of numbers of expressing cells calculated for lncRNAs (black dashed line), mRNAs (red dashed line), lncRNAs and mRNAs together (grey bars), and for modeled populations of genes with high (solid light blue line) or low (solid dark blue line) heterogeneity of expression. Pie charts demonstrate that a much lower fraction of lncRNAs was associated with the population of genes with low heterogeneity of expression, as compared to mRNAs. Genes used for this analysis had expression > 3 FPKM in at least one cell, and < 30 FPKM in all cells of the corresponding data set: totipotent human embryos (8-cell and morula stage) (A), hESCs (B) (both – data from (Yan *et al.*, 2013)), and K562 cells (C) (data from (Luo *et al.*, 2014)). Genes that contributed to the plots and pie charts on (A – C) were associated with one of the above-mentioned modeled populations of genes with a posterior probability > 0.99. Number of individual cells used for each analysis is given in parentheses.

4.7 *HIPSTR* is capable of activating and repressing genes in the pluripotent cells

Single-cell transcriptome analyses revealed that *HIPSTR* is expressed only by a subset of cells within human embryos and within a population of K562 cells. It is also evident that such expression pattern is typical for lncRNAs in general (Figure 21 A – C). Thus, we wanted to explore the functional importance of a lncRNA with such restricted pattern in a biologically relevant system. Functional studies of *HIPSTR* in early human embryos would be complicated by the relatively large amounts of material required for such experiments. Conveniently, $H1_{BP}$ cells have a normal karyotype, they express higher levels of *HIPSTR* than the H1 hESCs (Figure 22) from which they were derived by transient (24–36 h) exposure to bone morphogenetic protein 4 (BMP4) plus inhibitors of ACTIVIN signaling (A83-01) and of FGF2 (PD173074). Most importantly, $H1_{BP}$ cells have been proposed to have a totipotent potential, analogous to the outer cells of the 16-cell morula (Yang *et al.*, 2015).

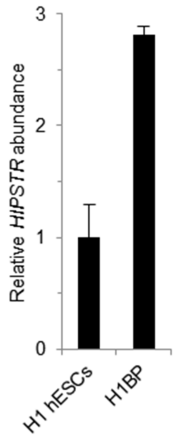
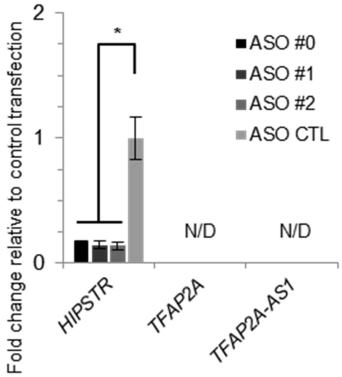


Figure 22. *HIPSTR* is expressed at higher levels in $H1_{BP}$ cells, compared to H1 hESCs. Quantification of *HIPSTR* levels with RT-qPCR. Experiments were performed in triplicates, error bars represent SD.

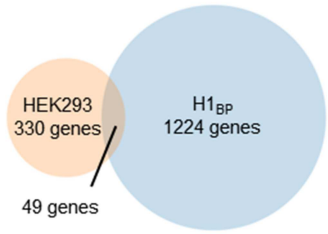
We silenced *HIPSTR* expression in $H1_{BP}$ cells with three ASOs – ASO #1 and ASO #2 used for knockdown of *HIPSTR* in HEK293 cells along with an additional ASO #0 (Figure 23 A), and analyzed global expression changes with microarrays. Surprisingly, 53 probes for 49 annotated genes differentially expressed upon *HIPSTR* knockdown in both – $H1_{BP}$ and HEK293 cells, were downregulated in $H1_{BP}$ and upregulated in HEK293 cells (Figure 23 B

and C; Additional file 1: Table 6). We validated such opposite differential expression for a group of these genes after *HIPSTR* knockdown in $H1_{BP}$ and HEK293 cells with RT-qPCR (Figure 23 C).

A



B



C

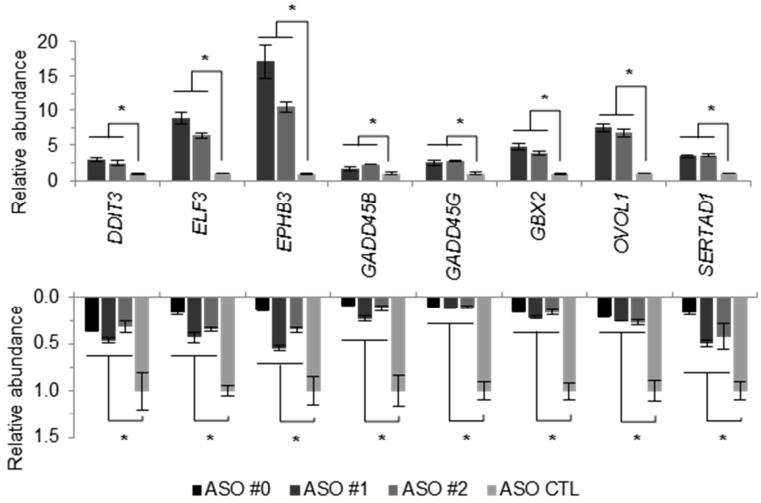


Figure 23. Silencing of *HIPSTR* in $H1_{BP}$ and in HEK293 cells demonstrates different modes of *HIPSTR* action. (A) Efficiency of *HIPSTR* knockdown in $H1_{BP}$ cells, as measured by RT-qPCR; N/D – not detected. (B) Overlap between genes differentially expressed upon *HIPSTR* silencing in HEK293 and $H1_{BP}$ cells (also see Additional file 1: Table 6). (C) RT-qPCR validation of a group of genes, whose expression is significantly up- and downregulated by *HIPSTR* knockdown in HEK293 and $H1_{BP}$ cells, correspondingly. Experiments on (A, C) were performed in triplicate, error bars represent SD, and the asterisks indicate statistical significance of the expression differences calculated with two-tailed t-test, equal variance (p-value < 0.05).

In addition, transient overexpression of *HIPSTR* in HEK293 cells (Figure 24 A) led to downregulation by at least 25% of eight out of twelve genes otherwise upregulated by *HIPSTR* knockdown in these cells (Figure 24 B).

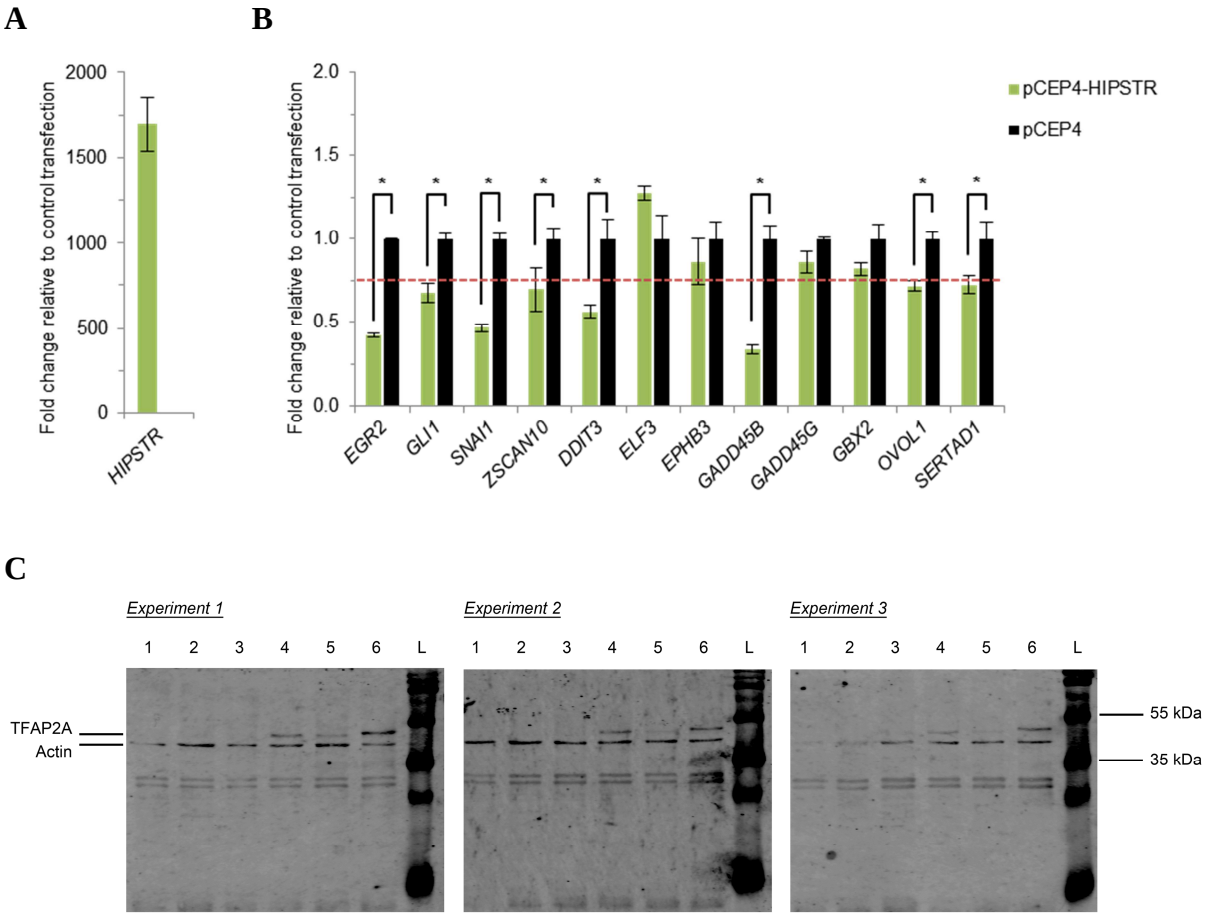


Figure 24. *HIPSTR* overexpression in HEK293 cells downregulates genes that are upregulated by *HIPSTR* knockdown in these cells. (A). *HIPSTR* overexpression efficiency in HEK293 cells, as measured with RT-qPCR. (B) *HIPSTR* ectopic overexpression downregulates developmental genes that are upregulated by *HIPSTR* knockdown in HEK293 cells, as measured with RT-qPCR. Experiments shown on (A, B) were performed in triplicate, error bars represent SD. For experiments on (B) the asterisks indicate statistical significance of the observed changes (reduction by at least 25 %) calculated with two-tailed t-test, equal variance (p-value < 0.05). (C) *HIPSTR* overexpression does not affect TFAP2A protein levels. We used total protein extracts from HEK293 cells transfected with pCEP4-HIPSTR (*lane 1*) or with empty pCEP4 vector as a negative control (*lane 2*) to perform western blot with anti-TFAP2A and anti-Actin antibodies; total protein extract from HEK293 cells transfected with pcDNA3 served as an additional negative control (*lane 3*); total protein extracts from HEK293 cells overexpressing TFAP2A isoforms 1c (*lane 4*), 1b (*lane 5*), 1a (*lane 6*) served as positive controls for TFAP2A antibody; detection of Actin served as loading control; PageRuler Plus Prestained Protein Ladder was used to estimate approximate MW of the proteins (*lane L*).

Similar reversal of knockdown effect by transient overexpression was observed for developmentally regulated *trans*-acting *PAUPAR* lncRNA (Vance *et al.*, 2014). Notably, *HIPSTR* overexpression did not affect TFAP2A protein levels (Figure 24 C).

We next considered the overall effect of *HIPSTR* silencing in $H1_{BP}$ cells, and detected 1349 significantly differentially expressed annotated genes (Figure 25 A; Additional file 1: Table 7). The majority of the transcripts (985 probes; ~62.2 %) was downregulated, corresponding to 777 annotated genes (Figure 25 A). The remaining differentially expressed transcripts (598 probes; ~37.8 %) corresponding to 572 annotated genes were upregulated (Figure 25 A). Importantly, genes downregulated by *HIPSTR* knockdown in $H1_{BP}$ cells are enriched in "Regulation of macromolecule biosynthetic process" and "Developmental process" GO categories (Figure 25 B). At the same time, genes downregulated by *HIPSTR* in pluripotent $H1_{BP}$ cells, have skin-, placenta-, lung-, and brain-specific expression (Figure 25 C). As in HEK293 cells, we found that TSS-surrounding regions of genes differentially expressed after *HIPSTR* knockdown were significantly enriched in NF-Y recognition motifs (Figure 25 D).

These results suggest that in the context of a pluripotent cell ($H1_{BP}$ cells), and likely in the early, totipotent human embryo, *HIPSTR* is capable of both activating and repressing its target genes, whereas in a cell lacking pluripotency network associated factors (HEK293 cells) *HIPSTR* acts solely as a repressor. Analysis of genes differentially expressed upon depletion of *HIPSTR* in a biologically relevant system, such as $H1_{BP}$ cells, further highlights the likely functional importance of lncRNAs with low population-level expression.

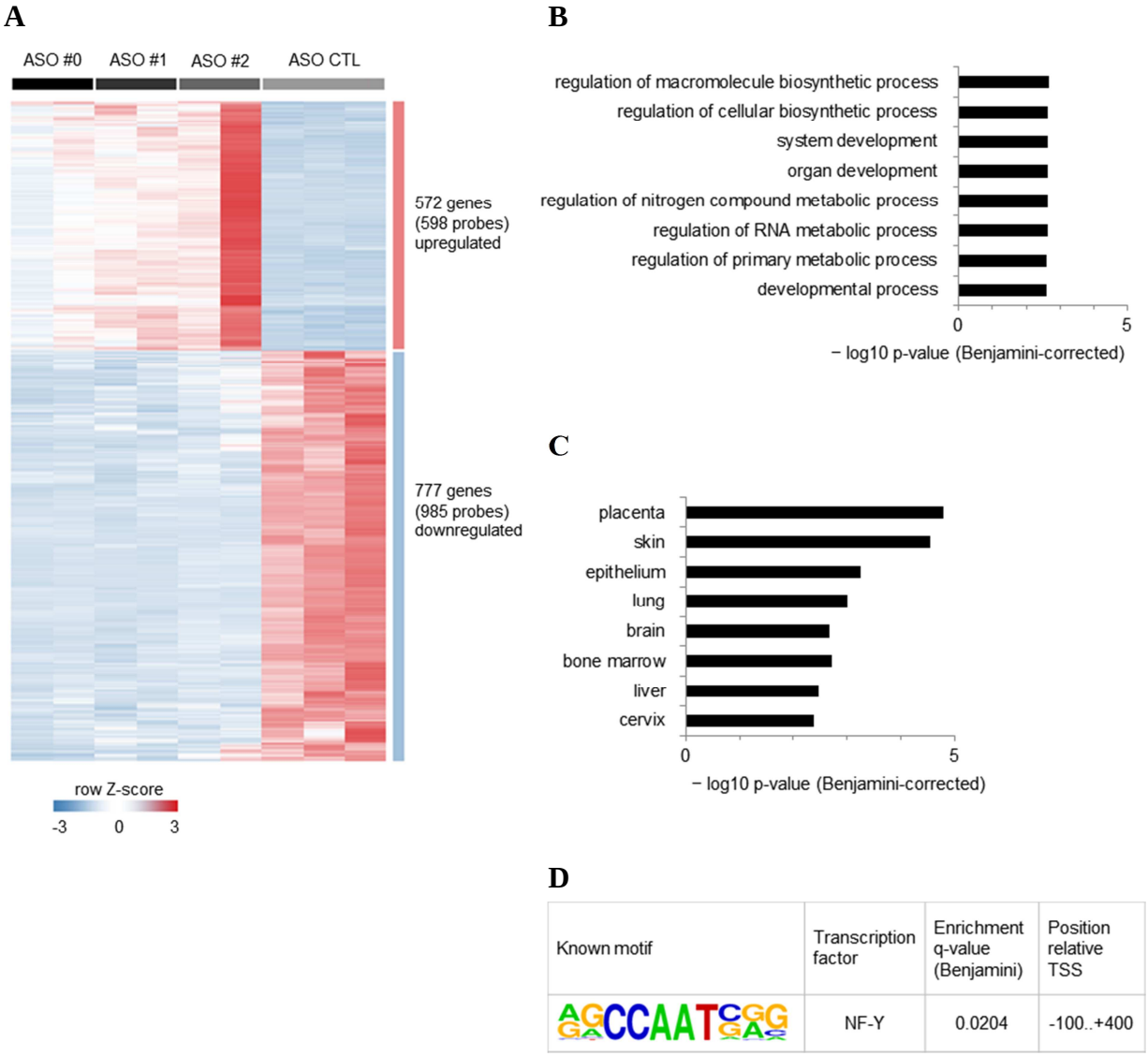


Figure 25. *HIPSTR* is capable of repressing and activating genes in $H1_{BP}$ cells. (A) *HIPSTR* knockdown in $H1_{BP}$ cells leads to significant upregulation of 572 and downregulation of 777 genes (1 % FDR, fold-change > 2, Additional file 1: Table 7). (B) GO categories significantly enriched with genes downregulated upon *HIPSTR* knockdown in $H1_{BP}$ cells. (C) Significantly enriched "Uniprot tissue" (UP_TISSUE) database entries for genes upregulated after *HIPSTR* silencing in $H1_{BP}$ cells. (D) NF-Y recognition motif is significantly enriched in regions surrounding TSSs of genes differentially expressed upon *HIPSTR* knockdown in $H1_{BP}$ cells.

4.8 Work on identification of potential protein partners of *HIPSTR*

Analyses of public data from (Werner and Ruthenburg, 2015) showed that the first 1000 nt of the *HIPSTR* sequence are stably associated with chromatin in HEK293 cells, and that treatment with RNA Pol II elongation inhibitor DRB does not affect the association of this portion of *HIPSTR* with chromatin (Figure 26 A, red). We next questioned whether such association is mediated by a protein, other than RNA Pol II, which in conjunction with *HIPSTR* could possibly regulate target genes of the latter. For this, we prepared nuclear protein extracts from pluripotent (and therefore, likely containing pluripotent cell-specific *HIPSTR*-interacting proteins) NT2/D1 cells, mixed them with biotinylated RNA probes (for the first 1000 nt of the *HIPSTR* sequence), and subjected proteins captured by the probes to LC-MS/MS analysis. The sense (target) and antisense (control) probes were used for these pulldown experiments, and a pulldown without RNA probes was used as an additional control. Together, we were able to identify 8 peptides corresponding to potential protein partners of *HIPSTR* (Additional file 1: Table 8) that were not present in the control samples from pulldown experiments without RNA probes. Of those 8 peptides, 1 was excluded as a potential contaminant. Notably, none of the identified peptides was specific to the sense probe. Among peptides that appeared in both pulldowns (with sense and antisense probes), and not in the control pulldowns without probes, a peptide corresponding to TARDBP captured our attention for two reasons: (i) TARDBP is a known RNA-binding (Sephton *et al.*, 2011) and DNA-binding protein (Fiesel *et al.*, 2010), and (ii) the list of genes differentially expressed upon TARDBP knockdown in (Fiesel *et al.*, 2010) significantly overlaps with the list of genes that were upregulated upon *HIPSTR* knockdown in $H1_{BP}$ cells (Figure 26 B), as identified with Enrichr on-line tool (Chen, E. Y. *et al.*, 2013).

Although it is tempting to speculate that *HIPSTR* might act through an interaction with TARDBP, which, in turn, is known to interact with Polycomb Repressive Complex 1 or 2

(Cao *et al.*, 2014) and therefore may recruit the Polycomb Repressive Complex to *HIPSTR* targets, further work, such as RNA-IP with anti-TARDBP antibody is required to validate such *HIPSTR*-TARDBP interaction. Additionally, in our pulldown assays we did not detect any known nuclear RNA-binding protein that could act as an activating partner of *HIPSTR* in pluripotent cells.

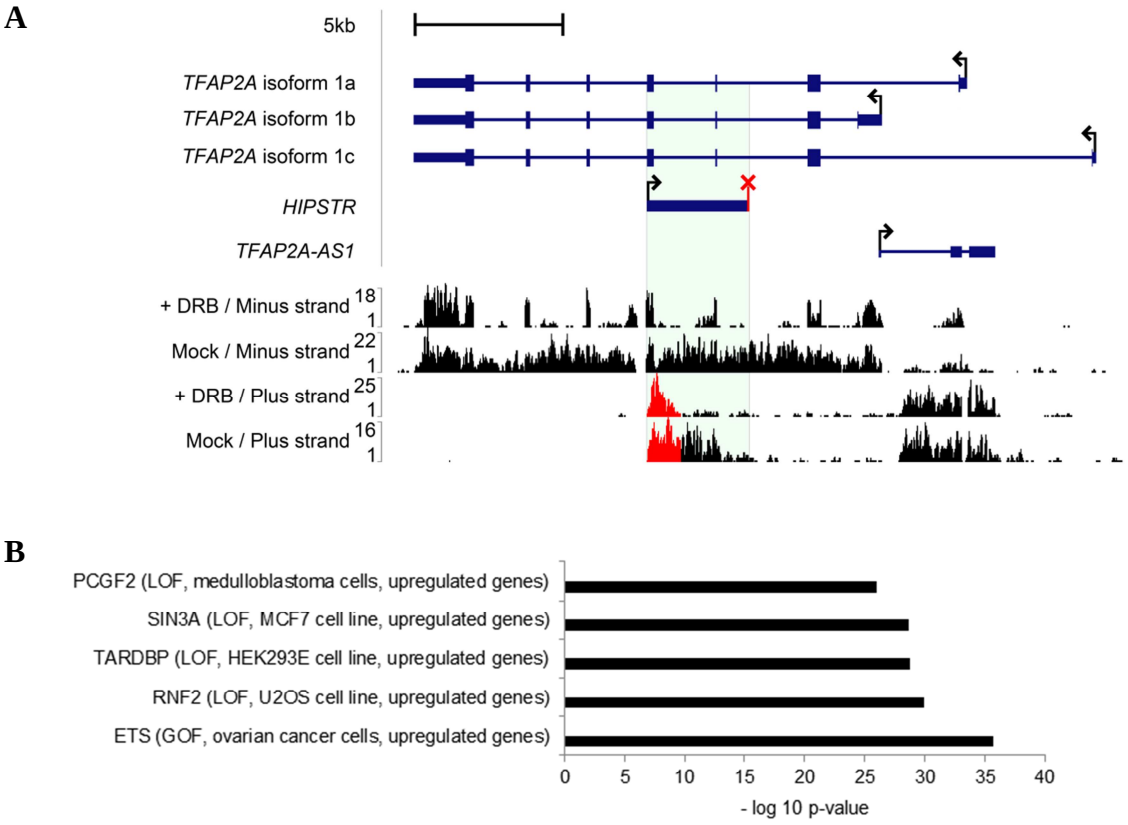


Figure 26. First 1000 nt of *HIPSTR* are associated with chromatin. (A) Mapping of strand-specific RNA-seq reads from chromatin-associated fraction of total RNA from HEK293 cells; data from (Werner and Ruthenburg, 2015). Treatment with RNA Pol II elongation inhibitor DRB does not affect the association of the first 1000 nt (red) of *HIPSTR* lncRNA with chromatin. (B) The list of genes, upregulated upon *HIPSTR* knockdown in $H1_{BP}$ cells significantly overlaps with genes upregulated upon TARDBP knockdown in HEK293E cells (Fiesel *et al.*, 2010). LOF – loss-of-function study, GOF – gain-of-function study.

5. Discussion

In the present work, we searched for novel antisense lncRNAs in the loci encoding TFs and identified *HIPSTR* (Heterogeneously expressed from the Intronic Plus Strand of the TFAP2A-locus RNA) gene that is located on the opposite strand of *TFAP2A* gene. *HIPSTR* is transcribed by RNA Pol II into a capped, monoexonic, nuclear-enriched antisense lncRNA (Figures 1 and 3). *HIPSTR* does not possess ORFs that could potentially encode any known polypeptide, moreover the longest potential ORF within *HIPSTR* sequence can be expected to occur by chance in a 3427-nt-long transcript (Figure 4). Publicly available ribosome profiling analysis did not show binding of ribosomes along the sequence of *HIPSTR* lncRNA (Figure 2 C).

Unexpectedly, *HIPSTR* expression did not correlate with the expression of its overlapping *TFAP2A* gene in cell lines and tissues (Figure 5 B – D). In agreement with these data, *HIPSTR* expression perturbations in HEK293 and $H1_{BP}$ cells did not affect overall levels of *TFAP2A* mRNA (Figure 10 A, and Figure 23 A), pre-mRNA (Figure 10 B) or TFAP2A protein levels (Figure 10 C). On the contrary, and unlike other antisense transcripts shown to regulate their overlapping or divergently transcribed genes (reviewed in (Pelechano and Steinmetz, 2013)), *HIPSTR* promoter and endogenous *HIPSTR* expression can be positively regulated by the protein product of its overlapping gene (Figure 13 A, and Figure 14 A). Such regulation was only observed in HEK293 cells, and not in HepG2 cells, suggesting that *TFAP2A* alone is not sufficient to regulate the *HIPSTR* promoter (Figure 13 B, and Figure 14 B). Finally, we did not find any evidence of *HIPSTR* differential expression in human tumor and non-tumor cell lines (Figure 5 A).

TFAP2A was first isolated from HeLa cells as a DNA-binding protein activating transcription from SV40 and metallothionein IIA promoters (Mitchell *et al.*, 1987). During embryonic development, *TFAP2A* gene is expressed in extraembryonic tissues

(trophectoderm) and in the embryo proper (premigratory neural crest and its derivatives) (reviewed in (Hilger-Eversheim *et al.*, 2000)). We evaluated *HIPSTR* gene activation in *in vitro* derived hNCCs and hTBCs, and ATRA-treated NT2/D1 cells, and did not observe consistent strong co-induction of *TFAP2A* and *HIPSTR* (Figure 15 A, B).

Tfap2a-null mice have been generated by two independent groups. In each case, the affected animals died perinatally with severe congenital defects (Schorle *et al.*, 1996; Zhang, J. *et al.*, 1996). Most interestingly, *Tfap2a*-null mice were generated by targeting exons 5 and 6 of the *Tfap2a* gene, which are located upstream of the *Hipstr* gene and its promoter region. Human *HIPSTR* expression is induced independently of *TFAP2A* during the major wave of human EGA (8-cell stage) (Figures 16 – 18, and Figure 19 A, B). At the same time, *HIPSTR* expression pattern with predominant expression in testis and placenta is conserved between human and mouse (Figure 6 A, B), and promoter demarcation of the *HIPSTR* transcription unit is conserved between human and chicken (Figure 8). Whether conservation of *HIPSTR* expression pattern extends to the major wave of mouse EGA (2-cell stage) remains to be established, since existing RNA-seq data for early mouse embryos is inconsistent with respect to *Hipstr* expression (Figure 20). The variability of gene expression patterns among different studies may be related to the known stochastically based lack of synchrony in cell cycle progression between the two cells in twin blastomeres from 2-cell stage mouse embryos (Roberts *et al.*, 2011). Should mouse *Hipstr* be induced in 2-cell embryos (and thus – prior to *Tfap2a* induction in trophectoderm or neural crest), genetic knockout studies would provide the ultimate evidence for the functional importance of *HIPSTR* during early embryonic development. Knockout studies would also be useful for phenotypic comparisons of *Tfap2a*$^{-/-}$ and *Hipstr*$^{-/-}$ mice.

We used microarray and qPCR analyses to show that *HIPSTR* knockdown in HEK293 and LNCaP cells that do not express TFs associated with pluripotency leads to upregulation of

development- and differentiation-related genes (Figure 11 B, C, and Figure 23 C). In turn, *HIPSTR* silencing in $H1_{BP}$ cells that express pluripotency TFs results in downregulation of development- and metabolism-related genes (Figure 23 C, and Figure 25 B), and upregulation of genes whose expression is associated with differentiated tissues (Figure 25 C). Nonetheless, a mechanism of *HIPSTR* action in pluripotent and totipotent cells, and in non-pluripotent cells (e.g., HEK293) remains to be investigated. It is tempting to speculate that in the context of a pluripotent cell (and probably in the totipotent cells of an early human embryo) *HIPSTR* is capable of both activating and repressing its target genes, whereas in a cell lacking pluripotency TFs *HIPSTR* acts solely as a repressor. This would be possible if in undifferentiated cells nuclear, chromatin-associated *HIPSTR* lncRNA (Figure 26 A) is directly or indirectly connected to one of the components of pluripotency network (absent from differentiated cells) to positively regulate its target genes. Previously, several lncRNAs with activating (e.g., *HOTTIP* (Rinn *et al.*, 2007)), repressing (e.g., *HOTAIR* (Wang, K. C. *et al.*, 2011)), and both – activating and repressing (e.g., *FENDRR* (Grote *et al.*, 2013)) functions have been described. These and other lncRNAs were proposed to function as modular scaffolds for chromatin modifying enzymes and TFs (Tsai *et al.*, 2010). Ubiquitously expressed, pioneer TF NF-Y is an essential component of the core pluripotency maintenance network (Oldfield *et al.*, 2014) and was also shown to act as activator and repressor (Ceribelli *et al.*, 2008). Significant enrichment of NF-Y recognition motif around TSSs of genes differentially expressed after *HIPSTR* silencing in HEK293 (Figure 2 D) and $H1_{BP}$ cells (Figure 4 G) suggests that this TF is a promising candidate partner of *HIPSTR*.

We attempted to identify *HIPSTR*-associated proteins with RNA-pulldown technique. Considering that the first 1000 nt of the *HIPSTR* sequence are associated with chromatin, even after RNA Pol II inhibition (Werner and Ruthenburg, 2015) (Figure 26 A, red), we hypothesized that *HIPSTR* interaction with chromatin, and probably with target genes, could

be mediated through this 1000 nt-long sequence. We used nuclear extracts of NT2/D1 pluripotent cells to identify proteins that could act as mediators of the activating and repressing activities of *HIPSTR* in pluripotent (and possibly – totipotent) cells. Surprisingly, RNA-pulldown with sense (target) and antisense (control) biotinylated probes for the first 1000 nt of the *HIPSTR* sequence followed by LC-MS/MS identified a set of only 8 peptides corresponding to several proteins (Additional file 1: Table 8), of which to our knowledge only TARDBP is a known and well-studied RNA-binding protein (Sephton *et al.*, 2011). However, none of the identified peptides was specific to the sense probe, although all 8 peptides were not present in the control pulldown performed in the absence of biotin-labeled RNA. These results suggest that *HIPSTR* likely interacts with its partner proteins through its 3'-regions, or the full-length *HIPSTR* probe is required for the proper secondary structure formation and protein pulldown. Whether this is the case, has to be determined by future studies. We note here that in our experience the *in vitro* transcription of the last 2000 nt of the *HIPSTR* sequence is challenging, probably due to stable RNA structures that are formed during the *in vitro* transcription reaction. This latter phenomenon remained independent of whether T7 or SP6 RNA Polymerase was used. We believe that an introduction of new methods, allowing for enrichment of *HIPSTR*-expressing cells (see below), in combination with ChIRP-MS (Chu *et al.*, 2011) or similar techniques would be an optimal solution for the future high-throughput search of *HIPSTR*-interacting proteins.

Our work shows that lncRNAs with low population-level expression frequently have high expression in individual cells in totipotent human embryos and stable human cell lines (Figure 21 A – C). For example, *HIPSTR* expression was absent from 73 out of 96 individual K562 cells, but was as high as 24.5 FPKM in one out of twenty-three *HIPSTR*-expressing cells (Figure 19 C). In agreement with ENCODE data for K562 cells (Figure 5 C), the population-average expression of *HIPSTR* in these 96 individual K562 cells was 0.91 FPKM.

DISCUSSION

Single-cell analysis has revealed that transcription is dynamic and stochastic, with transcription occurring as individual bursts of transcriptional activity inside a cell (Larson *et al.*, 2013), and this might be the dominant source of heterogeneity in RNA abundance from cell-to-cell. In fact, the complete absence of a given lncRNA in multiple cells in a population complicates statistical analyses, and the high cell-to-cell variability in lncRNAs levels suggests that analyses of hundreds or even thousands of individual cells might be required to reveal meaningful expression correlations between heterogeneously expressed lncRNAs and other genes. Low population-level and tissue specificity of lncRNAs expression (Ravasi *et al.*, 2006; Cabili *et al.*, 2011) might also be a serious obstacle for identification of partner proteins in RNA-Immunoprecipitation and endogenous RNA-pulldown assays (such as ChIRP (Chu *et al.*, 2011)), possibly resulting in false-negative results. For this, development of reliable and easy-to-use techniques facilitating enrichment for subpopulations of live cells expressing a lncRNA of interest will be required to uncover the exact mechanism of action of heterogeneously expressed lncRNAs, such as *HIPSTR*.

6. Conclusion

In the present study, we identified a novel antisense lncRNA gene that we named *HIPSTR*, we characterized conservation and expression patterns of its transcript, and we showed that *HIPSTR* lncRNA exemplifies the functional relevance of lncRNAs with heterogeneous and developmental stage-specific expression patterns.

HIPSTR is a monoexonic lncRNA, it is transcribed by RNA Pol II, and possesses 5'-cap structure and poly(A) tail. Based on the conservation of *HIPSTR* promoter demarcation, we estimated that *HIPSTR* gene appeared approximately 325 Ma, and its expression patterns are conserved at least between human and mouse. In agreement with recent studies demonstrating the involvement of other nuclear lncRNAs in gene expression regulation, we demonstrated that the silencing of *HIPSTR* in HEK293 and $H1_{BP}$ cells leads to up- and downregulation of important developmental genes, respectively. These observations were supported by overexpression experiments in HEK293 cells. We demonstrated that *HIPSTR* expression can be stimulated by TFAP2A protein, but such stimulation is not essential for *HIPSTR* expression, which is expressed independently from *TFAP2A* gene in human embryos, specifically at the 8-cell and morula stages. Similar to *HIPSTR*, in the individual cells of totipotent human embryos, the expression of lncRNAs is more highly heterogeneous than the expression of mRNAs. We further explored public data and presented evidence that high cell-to-cell expression variability is one of the characteristic features of lncRNAs.

Overall, heterogeneity in gene expression may be essential during early stages of embryonic development and may create distinct expression "footprints" for individual yet undifferentiated blastomeres. We conclude that the development of new techniques that will allow for enrichment of cells expressing a specific gene or a set of genes is required to facilitate mechanistic studies of lncRNA with low population level expression.

7. References

Aghajanova, L., S. Shen, A. M. Rojas, S. J. Fisher, J. C. Irwin and L. C. Giudice (2012). "Comparative transcriptome analysis of human trophectoderm and embryonic stem cell-derived trophoblasts reveal key participants in early implantation." Biol Reprod **86**(1): 1-21.

Amit, M., M. K. Carpenter, M. S. Inokuma, C. P. Chiu, C. P. Harris, M. A. Waknitz, J. Itskovitz-Eldor and J. A. Thomson (2000). "Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture." Dev Biol **227**(2): 271-278.

Anders, S., P. T. Pyl and W. Huber (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics **31**(2): 166-169.

Andrews, P. W. (2006). Chapter 23 - TERA2 and Its NTERA2 Subline: Pluripotent Human Embryonal Carcinoma Cells. Cell Biology (Third Edition). J. E. Celis. Burlington, Academic Press**:** 183-190.

Ayupe, A. C., A. C. Tahira, L. Camargo, F. C. Beckedorff, S. Verjovski-Almeida and E. M. Reis (2015). "Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome." RNA Biol **12**(8): 877-892.

Bajpai, R., D. A. Chen, A. Rada-Iglesias, J. Zhang, Y. Xiong, J. Helms, C. P. Chang, Y. Zhao, T. Swigut and J. Wysocka (2010). "CHD7 cooperates with PBAF to control multipotent neural crest formation." Nature **463**(7283): 958-962.

Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Beckedorff, F. C., A. C. Ayupe, R. Crocci-Souza, M. S. Amaral, H. I. Nakaya, D. T. Soltys, C. F. M. Menck, E. M. Reis and S. Verjovski-Almeida (2013). "The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation." PLoS Genet **9**: e1003705.

Benaglia, T., D. Chauveau, D. R. Hunter and D. S. Young (2009). "mixtools: An R package for analyzing finite mixture models." J Stat Softw **32**: 1-29.

Bennett, K. L., T. Romigh and C. Eng (2009). "AP-2alpha induces epigenetic silencing of tumor suppressive genes and microsatellite instability in head and neck squamous cell carcinoma." PLoS One **4**(9): e6931.

Bennett, M. R. and J. Hasty (2009). "Microfluidic devices for measuring gene network dynamics in single cells." Nat Rev Genet **10**(9): 628-638.

Berlato, C., K. V. Chan, A. M. Price, M. Canosa, A. G. Scibetta and H. C. Hurst (2011). "Alternative TFAP2A isoforms have distinct activities in breast cancer." Breast Cancer Res **13**(2): R23.

Berretta, J. and A. Morillon (2009). "Pervasive transcription constitutes a new level of eukaryotic genome regulation." EMBO Rep **10**(9): 973-982.

Biase, F. H., X. Cao and S. Zhong (2014). "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing." Genome Res **24**(11): 1787-1796.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Borsani, G., R. Tonlorenzi, M. C. Simmler, L. Dandolo, D. Arnaud, V. Capra, M. Grompe, A. Pizzuti, D. Muzny, C. Lawrence, H. F. Willard, P. Avner and A. Ballabio (1991).

"Characterization of a murine gene expressed from the inactive X chromosome." Nature **351**(6324): 325-329.

Brockdorff, N., A. Ashworth, G. F. Kay, V. M. McCabe, D. P. Norris, P. J. Cooper, S. Swift and S. Rastan (1992). "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**(3): 515-526.

Brown, C. J., A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi and H. F. Willard (1991). "A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome." Nature **349**(6304): 38-44.

Brown, C. J., B. D. Hendrich, J. L. Rupert, R. G. Lafreniere, Y. Xing, J. Lawrence and H. F. Willard (1992). "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**(3): 527-542.

Cabili, M. N., M. C. Dunagin, P. D. McClanahan, A. Biaesch, O. Padovan-Merhar, A. Regev, J. L. Rinn and A. Raj (2015). "Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution." Genome Biol **16**: 20.

Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev and J. L. Rinn (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." Genes Dev **25**(18): 1915-1927.

Cao, Q., X. Wang, M. Zhao, R. Yang, R. Malik, Y. Qiao, A. Poliakov, A. K. Yocum, Y. Li, W. Chen, X. Cao, X. Jiang, A. Dahiya, C. Harris, F. Y. Feng, S. Kalantry, Z. S. Qin, S. M. Dhanasekaran and A. M. Chinnaiyan (2014). "The central role of EED in the orchestration of polycomb group complexes." Nat Commun **5**: 3127.

Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N.

Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai and Y. Hayashizaki (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563.

Ceribelli, M., D. Dolfini, D. Merico, R. Gatta, A. M. Vigano, G. Pavesi and R. Mantovani (2008). "The histone-like NF-Y is a bifunctional transcription factor." Mol Cell Biol **28**(6): 2047-2058.

Chen, E. Y., C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark and A. Ma'ayan (2013). "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool." BMC Bioinformatics **14**: 128.

Chen, J., A. A. Shishkin, X. Zhu, S. Kadri, I. Maza, M. Guttman, J. H. Hanna, A. Regev and M. Garber (2016). "Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs." Genome Biol **17**(1): 19.

Cheng, Y. H., B. J. Aronow, S. Hossain, B. Trapnell, S. Kong and S. Handwerger (2004). "Critical role for transcription factor AP-2alpha in human trophoblast differentiation." Physiol Genomics **18**(1): 99-107.

Chodroff, R. A., L. Goodstadt, T. M. Sirey, P. L. Oliver, K. E. Davies, E. D. Green, Z. Molnar and C. P. Ponting (2010). "Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes." Genome Biol **11**(7): R72.

Chu, C., K. Qu, F. L. Zhong, S. E. Artandi and H. Y. Chang (2011). "Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions." Mol Cell **44**(4): 667-678.

Chu, C., Q. C. Zhang, S. T. da Rocha, R. A. Flynn, M. Bharadwaj, J. M. Calabrese, T. Magnuson, E. Heard and H. Y. Chang (2015). "Systematic discovery of Xist RNA binding proteins." Cell **161**(2): 404-416.

Cifuentes-Rojas, C., A. J. Hernandez, K. Sarma and J. T. Lee (2014). "Regulatory interactions between RNA and polycomb repressive complex 2." Mol Cell **55**(2): 171-185.

Clark, M. B., P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn, C. P. Ponting, P. F. Stadler, K. V. Morris, A. Morillon, J. S. Rozowsky, M. B. Gerstein, C. Wahlestedt, Y. Hayashizaki, P. Carninci, T. R. Gingeras and J. S. Mattick (2011). "The reality of pervasive transcription." PLoS Biol **9**(7): e1000625; discussion e1001102.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Davidovich, C., X. Wang, C. Cifuentes-Rojas, K. J. Goodrich, A. R. Gooding, J. T. Lee and T. R. Cech (2015). "Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA." Mol Cell **57**(3): 552-558.

Davidovich, C., L. Zheng, K. J. Goodrich and T. R. Cech (2013). "Promiscuous RNA binding by Polycomb repressive complex 2." Nat Struct Mol Biol **20**(11): 1250-1257.

DeOcesano-Pereira, C., M. S. Amaral, K. S. Parreira, A. C. Ayupe, J. F. Jacysyn, G. P. Amarante-Mendes, E. M. Reis and S. Verjovski-Almeida (2014). "Long non-coding RNA INXS is a critical mediator of BCL-XS induced apoptosis." Nucleic Acids Res **42**(13): 8343-8355.

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow and R. Guigo (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res **22**(9): 1775-1789.

Diederichs, S. (2014). "The four dimensions of noncoding RNA conservation." Trends Genet **30**(4): 121-123.

Dinger, M. E., P. P. Amaral, T. R. Mercer and J. S. Mattick (2009). "Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications." Brief Funct Genomic Proteomic **8**(6): 407-423.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo and T. R. Gingeras (2012). "Landscape of transcription in human cells." Nature **489**(7414): 101-108.

Eckert, D., S. Buhl, S. Weber, R. Jager and H. Schorle (2005). "The AP-2 family of transcription factors." Genome Biol **6**(13): 246.

Ezashi, T., P. Das and R. M. Roberts (2005). "Low O2 tensions and the prevention of differentiation of hES cells." Proc Natl Acad Sci U S A **102**(13): 4783-4788.

Fachel, A. A., A. C. Tahira, S. A. Vilella-Arias, V. Maracaja-Coutinho, E. R. Gimba, G. M. Vignal, F. S. Campos, E. M. Reis and S. Verjovski-Almeida (2013). "Expression analysis and in silico characterization of intronic long noncoding RNAs in renal cell carcinoma: emerging functional associations." Mol Cancer **12**(1): 140.

Fatica, A. and I. Bozzoni (2014). "Long non-coding RNAs: new players in cell differentiation and development." Nat Rev Genet **15**(1): 7-21.

Fiesel, F. C., A. Voigt, S. S. Weber, C. Van den Haute, A. Waldenmaier, K. Gorner, M. Walter, M. L. Anderson, J. V. Kern, T. M. Rasse, T. Schmidt, W. Springer, R. Kirchner, M. Bonin, M. Neumann, V. Baekelandt, M. Alunni-Fabbroni, J. B. Schulz and P. J. Kahle (2010). "Knockdown of transactive response DNA-binding protein (TDP-43) downregulates histone deacetylase 6." EMBO J **29**(1): 209-221.

Grote, P., L. Wittler, D. Hendrix, F. Koch, S. Wahrisch, A. Beisaw, K. Macura, G. Blass, M. Kellis, M. Werber and B. G. Herrmann (2013). "The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse." Dev Cell **24**(2): 206-214.

Guo, F., L. Yan, H. Guo, L. Li, B. Hu, Y. Zhao, J. Yong, Y. Hu, X. Wang, Y. Wei, W. Wang, R. Li, J. Yan, X. Zhi, Y. Zhang, H. Jin, W. Zhang, Y. Hou, P. Zhu, J. Li, L. Zhang, S. Liu, Y. Ren, X. Zhu, L. Wen, Y. Q. Gao, F. Tang and J. Qiao (2015). "The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells." Cell **161**(6): 1437-1452.

Guo, G., M. Huss, G. Q. Tong, C. Wang, L. Li Sun, N. D. Clarke and P. Robson (2010). "Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst." Dev Cell **18**(4): 675-685.

Gupta, R. A., N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar and H. Y. Chang (2010). "Long non-coding RNA HOTAIR

reprograms chromatin state to promote cancer metastasis." Nature **464**(7291): 1071-1076.

Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn and E. S. Lander (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.

Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root and E. S. Lander (2011). "lincRNAs act in the circuitry controlling pluripotency and differentiation." Nature **477**(7364): 295-300.

Guttman, M. and J. L. Rinn (2012). "Modular regulatory principles of large non-coding RNAs." Nature **482**(7385): 339-346.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Mol Cell **38**(4): 576-589.

Hilger-Eversheim, K., M. Moser, H. Schorle and R. Buettner (2000). "Regulatory roles of AP-2 transcription factors in vertebrate development, apoptosis and cell-cycle control." Gene **260**(1-2): 1-12.

Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Islam, S., U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg and S. Linnarsson (2011). "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq." Genome Res **21**(7): 1160-1167.

Iyer, M. K., Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, A. Poliakov, X. Cao, S. M. Dhanasekaran, Y. M. Wu, D. R. Robinson, D. G. Beer, F. Y. Feng, H. K. Iyer and A. M. Chinnaiyan (2015). "The landscape of long noncoding RNAs in the human transcriptome." Nat Genet **47**(3): 199-208.

Kapusta, A. and C. Feschotte (2014). "Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications." Trends Genet **30**(10): 439-452.

Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs and D. Karolchik (2010). "BigWig and BigBed: enabling browsing of large distributed datasets." Bioinformatics **26**(17): 2204-2207.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol **14**(4): R36.

Kim, D. H., G. K. Marinov, S. Pepke, Z. S. Singer, P. He, B. Williams, G. P. Schroth, M. B. Elowitz and B. J. Wold (2015). "Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming." Cell Stem Cell **16**(1): 88-101.

Klattenhoff, C. A., J. C. Scheuermann, L. E. Surface, R. K. Bradley, P. A. Fields, M. L. Steinhauser, H. Ding, V. L. Butty, L. Torrey, S. Haas, R. Abo, M. Tabebordbar, R. T. Lee, C. B. Burge and L. A. Boyer (2013). "Braveheart, a long noncoding RNA required for cardiovascular lineage commitment." Cell **152**(3): 570-583.

Kong, L., Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei and G. Gao (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucleic Acids Res **35**(Web Server issue): W345-349.

Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel and R. Shiekhattar (2013). "Activating RNAs associate with Mediator to enhance chromatin architecture and transcription." Nature **494**(7438): 497-501.

Lam, M. T., W. Li, M. G. Rosenfeld and C. K. Glass (2014). "Enhancer RNAs and regulated transcriptional programs." Trends Biochem Sci **39**(4): 170-182.

Lanford, R. E., C. Sureau, J. R. Jacob, R. White and T. R. Fuerst (1994). "Demonstration of in vitro infection of chimpanzee hepatocytes with hepatitis C virus using strand-specific RT/PCR." Virology **202**(2): 606-614.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Larson, D. R., C. Fritzsch, L. Sun, X. Meng, D. S. Lawrence and R. H. Singer (2013). "Direct observation of frequency modulated transcription in single cells using light activation." Elife **2**: e00750.

Levine, J. H., Y. Lin and M. B. Elowitz (2013). "Functional roles of pulsing in genetic circuits." Science **342**(6163): 1193-1200.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, H., G. S. Watts, M. M. Oshiro, B. W. Futscher and F. E. Domann (2006). "AP-2alpha and AP-2gamma are transcriptional targets of p53 in human breast carcinoma cells." Oncogene **25**(39): 5405-5415.

Li, L., B. Liu, O. L. Wapinski, M. C. Tsai, K. Qu, J. Zhang, J. C. Carlson, M. Lin, F. Fang, R. A. Gupta, J. A. Helms and H. Y. Chang (2013a). "Targeted disruption of Hotair leads to homeotic transformation and gene derepression." Cell Rep **5**(1): 3-12.

Li, X. Z., C. K. Roy, X. Dong, E. Bolcun-Filas, J. Wang, B. W. Han, J. Xu, M. J. Moore, J. C. Schimenti, Z. Weng and P. D. Zamore (2013b). "An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes." Mol Cell **50**(1): 67-81.

Lin, N., K. Y. Chang, Z. Li, K. Gates, Z. A. Rana, J. Dang, D. Zhang, T. Han, C. S. Yang, T. J. Cunningham, S. R. Head, G. Duester, P. D. Dong and T. M. Rana (2014). "An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment." Mol Cell **53**(6): 1005-1019.

Loewer, S., M. N. Cabili, M. Guttman, Y. H. Loh, K. Thomas, I. H. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, P. D. Manos, S. Datta, E. S. Lander, T. M. Schlaeger, G. Q. Daley and J. L. Rinn (2010). "Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells." Nat Genet **42**(12): 1113-1117.

Louro, R., H. I. Nakaya, P. P. Amaral, F. Festa, M. C. Sogayar, A. M. da Silva, S. Verjovski-Almeida and E. M. Reis (2007). "Androgen responsive intronic non-coding RNAs." BMC Biol **5**: 4.

Luo, H., J. Li, B. K. Chia, P. Robson and N. Nagarajan (2014). "The importance of study design for detecting differentially abundant features in high-throughput experiments." Genome Biol **15**(12): 527.

Luscher, B., P. J. Mitchell, T. Williams and R. Tjian (1989). "Regulation of transcription factor AP-2 by the morphogen retinoic acid and by second messengers." Genes Dev **3**(10): 1507-1517.

Magistri, M., M. A. Faghihi, G. St Laurent, 3rd and C. Wahlestedt (2012). "Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts." Trends Genet **28**(8): 389-396.

Makhov, P. B., K. V. Golovine, A. Kutikov, D. J. Canter, V. A. Rybko, D. A. Roshchin, V. B. Matveev, R. G. Uzzo and V. M. Kolenko (2011). "Reversal of epigenetic silencing of AP-2alpha results in increased zinc uptake in DU-145 and LNCaP prostate cancer cells." Carcinogenesis **32**(12): 1773-1781.

Marchand, M., J. A. Horcajadas, F. J. Esteban, S. L. McElroy, S. J. Fisher and L. C. Giudice (2011). "Transcriptomic signature of trophoblast differentiation in a human embryonic stem cell model." Biol Reprod **84**(6): 1258-1271.

Marques, A. C., J. Hughes, B. Graham, M. S. Kowalczyk, D. R. Higgs and C. P. Ponting (2013). "Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs." Genome Biol **14**(11): R131.

McHugh, C. A., C. K. Chen, A. Chow, C. F. Surka, C. Tran, P. McDonel, A. Pandya-Jones, M. Blanco, C. Burghard, A. Moradian, M. J. Sweredoski, A. A. Shishkin, J. Su, E. S. Lander, S. Hess, K. Plath and M. Guttman (2015). "The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3." Nature **521**(7551): 232-236.

McPherson, L. A., A. V. Loktev and R. J. Weigel (2002). "Tumor suppressor activity of AP2alpha mediated through a direct interaction with p53." J Biol Chem **277**(47): 45028-45033.

Melnikova, V. O., A. S. Dobroff, M. Zigler, G. J. Villares, R. R. Braeuer, H. Wang, L. Huang and M. Bar-Eli (2010). "CREB inhibits AP-2alpha expression to regulate the malignant phenotype of melanoma." PLoS One **5**(8): e12452.

Milanesi, L., M. Muselli and P. Arrigo (1996). "Hamming-Clustering method for signals prediction in 5' and 3' regions of eukaryotic genes." Comput Appl Biosci **12**(5): 399-404.

Mitchell, P. J., C. Wang and R. Tjian (1987). "Positive and negative regulation of transcription in vitro: enhancer-binding protein AP-2 is inhibited by SV40 T antigen." Cell **50**(6): 847-861.

Nakaya, H. I., P. P. Amaral, R. Louro, A. Lopes, A. A. Fachel, Y. B. Moreira, T. A. El-Jundi, A. M. da Silva, E. M. Reis and S. Verjovski-Almeida (2007). "Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription." Genome Biol **8**(3): R43.

Necsulea, A., M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker, F. Grutzner and H. Kaessmann (2014). "The evolution of lncRNA repertoires and expression patterns in tetrapods." Nature **505**(7485): 635-640.

Ng, S. Y., G. K. Bogu, B. S. Soh and L. W. Stanton (2013). "The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis." Mol Cell **51**(3): 349-359.

Ng, S. Y., R. Johnson and L. W. Stanton (2012). "Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors." EMBO J **31**(3): 522-533.

Oldfield, A. J., P. Yang, A. E. Conway, S. Cinghu, J. M. Freudenberg, S. Yellaboina and R. Jothi (2014). "Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors." Mol Cell **55**(5): 708-722.

Onoguchi, M., Y. Hirabayashi, H. Koseki and Y. Gotoh (2012). "A noncoding RNA regulates the neurogenin1 gene locus during mouse neocortical development." Proc Natl Acad Sci U S A **109**(42): 16939-16944.

Orom, U. A., T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytnicki, C. Notredame, Q. Huang, R. Guigo and R. Shiekhattar (2010). "Long noncoding RNAs with enhancer-like function in human cells." Cell **143**(1): 46-58.

Pandey, R. R., T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-Dinardo and C. Kanduri (2008). "Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation." Mol Cell **32**(2): 232-246.

Pang, K. C., M. C. Frith and J. S. Mattick (2006). "Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function." Trends Genet **22**(1): 1-5.

Pelechano, V. and L. M. Steinmetz (2013). "Gene regulation by antisense transcription." Nat Rev Genet **14**(12): 880-893.

Pellikainen, J. M. and V. M. Kosma (2007). "Activator protein-2 in carcinogenesis with a special reference to breast cancer--a mini review." Int J Cancer **120**(10): 2061-2067.

Pfaffl, M. W. (2001). "A new mathematical model for relative quantification in real-time RT-PCR." Nucleic Acids Res **29**(9): e45.

Prescott, S. L., R. Srinivasan, M. C. Marchetto, I. Grishina, I. Narvaiza, L. Selleri, F. H. Gage, T. Swigut and J. Wysocka (2015). "Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest." Cell **163**(1): 68-83.

Puvvula, P. K., R. D. Desetty, P. Pineau, A. Marchio, A. Moon, A. Dejean and O. Bischof (2014). "Long noncoding RNA PANDA and scaffold-attachment-factor SAFA control senescence entry and exit." Nat Commun **5**: 5323.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Quinn, J. J. and H. Y. Chang (2015). "Unique features of long non-coding RNA biogenesis and function." Nat Rev Genet **17**(1): 47-62.

Rada-Iglesias, A., R. Bajpai, S. Prescott, S. A. Brugmann, T. Swigut and J. Wysocka (2012). "Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest." Cell Stem Cell **11**(5): 633-648.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn and J. Wysocka (2011). "A unique chromatin signature uncovers early developmental enhancers in humans." Nature **470**(7333): 279-283.

Rands, C. M., S. Meader, C. P. Ponting and G. Lunter (2014). "8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage." PLoS Genet **10**(7): e1004525.

Rappsilber, J., M. Mann and Y. Ishihama (2007). "Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips." Nat Protoc **2**(8): 1896-1906.

Ravasi, T., H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, S. M. Grimmond, D. A. Hume, Y. Hayashizaki and J. S. Mattick (2006). "Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome." Genome Res **16**(1): 11-19.

Reis, E. M., H. I. Nakaya, R. Louro, F. C. Canavez, A. V. Flatschart, G. T. Almeida, C. M. Egidio, A. C. Paquola, A. A. Machado, F. Festa, D. Yamamoto, R. Alvarenga, C. C. da Silva, G. C. Brito, S. D. Simon, C. A. Moreira-Filho, K. R. Leite, L. H. Camara-Lopes, F. S. Campos, E. Gimba, G. M. Vignal, H. El-Dorry, M. C. Sogayar, M. A. Barcinski, A. M. da Silva and S. Verjovski-Almeida (2004). "Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer." Oncogene **23**(39): 6684-6692.

Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal and H. Y. Chang (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell **129**(7): 1311-1323.

Roberts, R. M., M. Katayama, S. R. Magnuson, M. T. Falduto and K. E. Torres (2011). "Transcript profiling of individual twin blastomeres derived by splitting two-cell stage murine embryos." Biol Reprod **84**(3): 487-494.

Ruiz, M., C. Pettaway, R. Song, O. Stoeltzing, L. Ellis and M. Bar-Eli (2004). "Activator protein 2alpha inhibits tumorigenicity and represses vascular endothelial growth factor transcription in prostate cancer cells." Cancer Res **64**(2): 631-638.

Saliba, A. E., A. J. Westermann, S. A. Gorski and J. Vogel (2014). "Single-cell RNA-seq: advances and future challenges." Nucleic Acids Res **42**(14): 8845-8860.

Sanchez, A. and I. Golding (2013). "Genetic determinants and cellular constraints in noisy gene expression." Science **342**(6163): 1188-1193.

Schneider, R., A. J. Bannister, F. A. Myers, A. W. Thorne, C. Crane-Robinson and T. Kouzarides (2004). "Histone H3 lysine 4 methylation patterns in higher eukaryotic genes." Nat Cell Biol **6**(1): 73-77.

Schorle, H., P. Meier, M. Buchert, R. Jaenisch and P. J. Mitchell (1996). "Transcription factor AP-2 essential for cranial closure and craniofacial development." Nature **381**(6579): 235-238.

Scibetta, A. G., P. P. Wong, K. V. Chan, M. Canosa and H. C. Hurst (2010). "Dual association by TFAP2A during activation of the p21cip/CDKN1A promoter." Cell Cycle **9**(22): 4525-4532.

Sephton, C. F., C. Cenik, A. Kucukural, E. B. Dammer, B. Cenik, Y. Han, C. M. Dewey, F. P. Roth, J. Herz, J. Peng, M. J. Moore and G. Yu (2011). "Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes." J Biol Chem **286**(2): 1204-1215.

Shalek, A. K., R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park and A. Regev (2013). "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells." Nature **498**(7453): 236-240.

Shapiro, E., T. Biezuner and S. Linnarsson (2013). "Single-cell sequencing-based technologies will revolutionize whole-organism science." Nat Rev Genet **14**(9): 618-630.

Solovyev, V. and A. Salamov (1997). "The Gene-Finder computer tools for analysis of human and model organisms genome sequences." Proc Int Conf Intell Syst Mol Biol **5**: 294-302.

Stumpf, C. R., M. V. Moreno, A. B. Olshen, B. S. Taylor and D. Ruggero (2013). "The translational landscape of the mammalian cell cycle." Mol Cell **52**(4): 574-582.

Tahira, A. C., M. S. Kubrusly, M. F. Faria, B. Dazzani, R. S. Fonseca, V. Maracaja-Coutinho, S. Verjovski-Almeida, M. C. Machado and E. M. Reis (2011). "Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer." Mol Cancer **10**: 141.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda and S. Yamanaka (2007). "Induction of pluripotent stem cells from adult human fibroblasts by defined factors." Cell **131**(5): 861-872.

Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao and M. A. Surani (2009). "mRNA-Seq whole-transcriptome analysis of a single cell." Nat Methods **6**(5): 377-382.

Tohonen, V., S. Katayama, L. Vesterlund, E. M. Jouhilahti, M. Sheikhi, E. Madissoon, G. Filippini-Cattaneo, M. Jaconi, A. Johnsson, T. R. Burglin, S. Linnarsson, O. Hovatta and J. Kere (2015). "Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development." Nat Commun **6**: 8207.

Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal and H. Y. Chang (2010). "Long noncoding RNA as modular scaffold of histone modification complexes." Science **329**(5992): 689-693.

Tusher, V. G., R. Tibshirani and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-5121.

Ulitsky, I. and D. P. Bartel (2013). "lincRNAs: genomics, evolution, and mechanisms." Cell **154**(1): 26-46.

van Heeringen, S. J., R. C. Akkers, I. van Kruijsbergen, M. A. Arif, L. L. Hanssen, N. Sharifi and G. J. Veenstra (2014). "Principles of nucleation of H3K27 methylation during embryonic development." Genome Res **24**(3): 401-410.

Vance, K. W., S. N. Sansom, S. Lee, V. Chalei, L. Kong, S. E. Cooper, P. L. Oliver and C. P. Ponting (2014). "The long non-coding RNA Paupar regulates the expression of both local and distal genes." EMBO J **33**(4): 296-311.

Villar, D., C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek and D. T. Odom (2015). "Enhancer evolution across 20 mammalian species." Cell **160**(3): 554-566.

Wajapeyee, N. and K. Somasundaram (2003). "Cell cycle arrest and apoptosis induction by activator protein 2alpha (AP-2alpha) and the role of p53 and p21WAF1/CIP1 in AP-2alpha-mediated growth inhibition." J Biol Chem **278**(52): 52093-52101.

Wang, K. C., Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms and H. Y. Chang (2011). "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression." Nature **472**(7341): 120-124.

Wang, L., H. J. Park, S. Dasari, S. Wang, J. P. Kocher and W. Li (2013). "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model." Nucleic Acids Res **41**(6): e74.

Werner, M. S. and A. J. Ruthenburg (2015). "Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes." Cell Rep **12**(7): 1089-1098.

Wisniewski, J. R., A. Zougman, N. Nagaraj and M. Mann (2009). "Universal sample preparation method for proteome analysis." Nat Methods **6**(5): 359-362.

Xu, R. H., X. Chen, D. S. Li, R. Li, G. C. Addicks, C. Glennon, T. P. Zwaka and J. A. Thomson (2002). "BMP4 initiates human embryonic stem cell differentiation to trophoblast." Nat Biotechnol **20**(12): 1261-1264.

Xue, Z., K. Huang, C. Cai, L. Cai, C. Y. Jiang, Y. Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, J. Y. Liu, S. Horvath and G. Fan (2013). "Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing." Nature **500**(7464): 593-597.

Yan, L., M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, J. Qiao and F. Tang (2013). "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells." Nat Struct Mol Biol **20**(9): 1131-1139.

Yang, Y., K. Adachi, M. A. Sheridan, A. P. Alexenko, D. J. Schust, L. C. Schulz, T. Ezashi and R. M. Roberts (2015). "Heightened potency of human pluripotent stem cell lines created by transient BMP4 exposure." Proc Natl Acad Sci U S A **112**(18): E2337-2346.

Yu, Y., Y. Wang, M. Li and P. Kannan (2002). "Tumorigenic effect of transcription factor hAP-2alpha and the intricate link between hAP-2alpha activation and squelching." Mol Carcinog **34**(4): 172-179.

Zappulla, D. C. and T. R. Cech (2004). "Yeast telomerase RNA: a flexible scaffold for protein subunits." Proc Natl Acad Sci U S A **101**(27): 10024-10029.

Zhang, J., S. Hagopian-Donaldson, G. Serbedzija, J. Elsemore, D. Plehn-Dujowich, A. P. McMahon, R. A. Flavell and T. Williams (1996). "Neural tube, skeletal and body wall defects in mice lacking transcription factor AP-2." Nature **381**(6579): 238-241.

Zhang, J. and T. Williams (2003). "Identification and regulation of tissue-specific cis-acting elements associated with the human AP-2alpha gene." Dev Dyn **228**(2): 194-207.

Zhang, Y., N. L. Vastenhouw, J. Feng, K. Fu, C. Wang, Y. Ge, A. Pauli, P. van Hummelen, A. F. Schier and X. S. Liu (2014). "Canonical nucleosome organization at promoters forms during genome activation." Genome Res **24**(2): 260-266.

Zhao, J., T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky and J. T. Lee (2010). "Genome-wide identification of polycomb-associated RNAs by RIP-seq." Mol Cell **40**(6): 939-953.

Zhao, J., B. K. Sun, J. A. Erwin, J. J. Song and J. T. Lee (2008). "Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome." Science **322**(5902): 750-756.

8. Supplementary information

Additional file 1: Supplementary Tables are contained in the CD that is present at the inside back cover of the Thesis.

Table 1, Sequences of the oligonucleotides used in this study; Table 2, List of genes differentially expressed in HEK293 cells after *HIPSTR* knockdown (q-value < 0.01, fold-change > 2); Tables 3, 4 and 5, Lists of genes from hESCs passages (P) 0 and 10, from 8-cell (8C) and morula (M) stage human embryos (E), and from K562 cells, respectively, with expression levels 3 – 30 FPKM used for heterogeneity of expression analysis, and their corresponding heterogeneity flags; Table 6, List of genes differentially expressed – both in $H1_{BP}$ cells and HEK293 after *HIPSTR* knockdown (q-value < 0.01, fold-change > 2); Table 7, List of genes differentially expressed in $H1_{BP}$ cells after *HIPSTR* knockdown (q-value < 0.01, fold-change > 2); Table 8 (on CD and see below), List of proteins identified by *HIPSTR* RNA-pulldown followed by mass spectroscopy.

pulldown				
no probes	sense probe	antisense probe	Potential contaminant	Protein IDs, corresponding to detected peptide
absent	detected	detected		tr|B4DDC8|B4DDC8_HUMAN;tr|B3KXL8|B3KXL8_HUMAN;tr|Q6IAU5|Q6IAU5_HUMAN;tr|B2R665|B2R665_HUMAN;sp|O15355|PP1G_HUMAN;tr|Q96IN7|Q96IN7_HUMAN;tr|Q59GB2|Q59GB2_HUMAN
absent	detected	absent		sp|O60613|SEP15_HUMAN
absent	absent	detected	+	tr|J3KN47|J3KN47_HUMAN;tr|B4DI57|B4DI57_HUMAN;tr|B4E1B2|B4E1B2_HUMAN;tr|Q53H26|Q53H26_HUMAN;tr|Q06AH7|Q06AH7_HUMAN;sp|P02787|TRFE_HUMAN;tr|C9JVG0|C9JVG0_HUMAN;tr|H7C5E8|H7C5E8_HUMAN;tr|A0PJA6|A0PJA6_HUMAN;tr|B4DHZ6|B4DHZ6_HUMAN;CON__Q2HJF0;CON__Q29443;CON__Q0IIK2
absent	absent	detected		tr|Q59FC6|Q59FC6_HUMAN;tr|Q5CAQ5|Q5CAQ5_HUMAN;tr|V9HWP2|V9HWP2_HUMAN;sp|P14625|ENPL_HUMAN;tr|B4DHT9|B4DHT9_HUMAN;tr|B4DU71|B4DU71_HUMAN
absent	detected	detected		tr|Q2F838|Q2F838_HUMAN;tr|B4DSR4|B4DSR4_HUMAN;tr|B4DUK7|B4DUK7_HUMAN;tr|Q53YD7|Q53YD7_HUMAN;sp|P26641|EF1G_HUMAN
absent	detected	detected		tr|B4DJ45|B4DJ45_HUMAN;tr|B4DRW3|B4DRW3_HUMAN;sp|Q13148|TADBP_HUMAN;tr|K7EJM5|K7EJM5_HUMAN;tr|K7EN94|K7EN94_HUMAN;tr|B1AKP7|B1AKP7_HUMAN;tr|G3V162|G3V162_HUMAN
absent	detected	detected		sp|Q9H009|NACA2_HUMAN
absent	detected	detected		tr|E7FL39|E7FL39_RUBV;tr|C7G0C9|C7G0C9_RUBV;tr|B5BNX3|B5BNX3_RUBV