様式19 別紙1

課題番号	LS127

先端研究助成基金助成金（最先端・次世代研究開発支援プログラム）
実施状況報告書（平成 24 年度）

本様式の内容は一般に公表されます

研究課題名	細胞分化に関するノンコーディングRNAの全ゲノム解析
研究機関・ 部局・職名	独立行政法人理化学研究所 オミックス基盤研究領域　ゲノム機能研究チーム　チームリーダー
氏名	カルニンチ　ピエロ

1. 当該年度の研究目的

After the completion of the transcriptome collection, we planned to complete the bioinformatics analysis to maximize the chance to find the best candidates for experimental perturbation, and then to proceed with the experimental perturbation. In particular, we aimed at:

- Doing a deep, comprehensive bioinformatics analysis to establish the type, nature of the RNAs, and understand which genomic region produce them in order to fully characterize the extent of the non-coding transcriptome. We aimed also at bioinformatically characterize the retrotransposon class (mostly LTR transposons) to understand their function in a possible control of the epigenome.
- We next planned positive and negative perturbation (overexpression of cloned full-length cDNA newly isolated; and siRNAs). We planned to follow these experiments by CAGE-sequencing and by cell biology experiments (cell morphology, stem-cell markers), for the largest number possible that was feasible with the proposed budget. We planned to publish a transcriptome paper with the description of ncRNAs.

2. 研究の実施状況

Introduction

We are investigating the role of non-coding RNAs (ncRNAs) for the pluripotent state, with particular emphasis on nuclear and retrotransposon-derived transcripts. Retrotransposons are transposable elements of the genome (such as LINE, SINE, LTR) that require reverse transcription for propagation. Upon integration in the genome these elements often acquire mutations that render them “transposition incompetent”, consequently the genome is littered with degraded fragments of these elements collected over evolution and thought to be inactive. We have previously reported that a very large number of these repeated elements are used as promoters and surprisingly show very specific tissue and developmental stage restricted expression patterns (Faulkner *et al.*, Nat Genet, 2009). More recently, the ENCODE consortium, observed in a large-scale transcriptomics research, that an important proportion of human transcripts initiate from repeat elements (Djebali *et al.*, Nature, 2012). It has also been shown that specific retrotransposon elements are highly expressed in embryonic stem cells and switched off upon differentiation (Cloonan *et al*, Nat. Meth. 2008). Others have demonstrated, by blocking endogenous reverse transcriptase, that embryonic development is blocked at the 2-cell stage (Pittogi *et al.*, Mol. Rep. Dev. 2003) and also promotes differentiation of tumor cells

(Oricchio *et al.*, Oncogene 2007). Taken together, these findings suggest a key role for retrotransposons derived transcripts in cellular differentiation and maintenance of pluripotency state; however little is known as to the molecules and mechanisms involved.

Research methodology

The first phase of our project consists in the generation of an exhaustive and representative profiling of mammalian stem cell transcriptome. For this purpose, we selected 11 different pluripotent cell lines from mouse and human origins (Table 1).

Aiming to detect abundant as well as rare compartment specific transcripts, we analyzed nuclear enriched and cytoplasmic RNA fractions from all pluripotent and 6 differentiated cell lines. Deep transcriptome profiling of the 34 samples were produced combining four complementary highthroughput transcriptomics methods. First, genome wide transcription start sites (TSSs) activity was defined using Cap Analysis of Gene Expression (CAGE, Takahashi *et al.*, Nat. Protoc.). Second, CAGEscan (CAGE combined with paired-end sequencing, Plessy *et al.*, Nat. Methods, 2010) and RNA-seq (Cloonan *et al*, Nat. Methods 2008) were used to generate *de novo* transcript assemblies. Finally, short-RNA□seq data were produced to assess post-transcriptional RNA processing events. All libraries were sequenced using multiplex sequencing on HiSeq2000 (Illumina) platform. Using state of the art bioinformatics tools, we performed integrative analyzes of these large datasets describing deeply the non-coding transcriptome of stem cell and identifying potential novel stem cell specific transcripts.

The second phase of the project comprises the functional screen of a large number of stem specific non-coding transcripts, identified in phase 1, for their putative implication in the genetic regulation of pluripotency.

Table 1 | Cell lines used for deep transcriptome profiling and sequencing depth

Cell line (clone name)	Cell type	Strain/sex	Aligned tags (x10^6) CAGE (Nu/Cy)	CAGEscan (Nu/Cy)	sRNA-seq (Nu/Cy)	RNA-seq (Nu/Cy)
mouse						
mESR08 (Nanog^(βgeo/+)ES)	ESC	129 SV Jae	19.7/16.6	20.5/27.3	26.9/12.3	50/46.3
mESB6G-2	ESC	C57Bl/6	16.2/16.2	23.1/7.3	38.1/20.4	77.4/60.9
mESFVB-1	ESC	FVB	19.9/16.2	19.8/11	31.7/18.1	
miPS.F (iPS_MEF-Ng-20D-17)	iPSC	C57Bl/6	17.9/14.8	20.2/28.1	25.5/22.5	
miPS.B (iPS_LymB_44.1B4e)	iPSC	C57Bl/6	14.7/16.7	21/23.1	26.9/14	
miPS.T (iPS_LymT_i103 H12)	iPSC	C57Bl/6	15.1/17.4	8.9/25.5	28.1/22.1	
MEF (MEF-Ng-20D-17)	fibroblasts	C57Bl/6	23.9/15.8	29.5/20.3	22.2/25.3	
Primary Lymphocytes B	B cells	C57Bl/6	18/16.2	15.8/27.1	28/21.1	
Primary Lymphocytes T	T cells	C57Bl/6	18.4/15.5	27.9/25.2	26.4/44.1	
human						
KhES-1	ESC	female	22.9/16.2	24.3/27.2	21.6/20	
KhES-2	ESC	female	19.4/15.4	29.4/33.1	23.3/19.5	105.1/46.4
KhES-3	ESC	male	20/16.3	41.6/17	14.6/15.6	49.3/33.3
hiPS.F (iPS_HDF-f_hi6)	iPSC	male	19.6/15.1	26.6/8.6	28.1/19.3	
hiPS.B (iPS_LymB_hi-68)	iPSC	male	19.2/18.6	28.3/19.1	26.4/17.7	
HDF-f	fibroblasts	male	10.7/26.6	26.5/3.5	18.1/28.6	
Primary Lymphocytes B	B cells	male	16.3/8.4	2.0/2.8	19.1/22.0	
Primary Lymphocytes T	T cells	male	19.5/27.5	26.2/25.9	61.3/54	

Current research status

We have currently completed the data collection and bioinformatics analyzes. A manuscript presenting a deep transcriptomics analyzes of a representative selection of human and mouse stem cells analyzed by various high-throughput state of the art methods has been submitted.

At first, identification of CAGE-tag-clusters up-regulated in stem cells (Figure 1a,b) revealed a high complexity of the stem cell nuclear transcriptomes, composed largely of non-annotated transcripts (unknown from RefSeq, GENCODE, Ensembl databases). We thus focus our data analyzes on the identification and description of putative novel stem specific transcripts enriched in the nuclear fraction. It appears that 8,873 mouse and 3,043 human nuclear stem cell specific CAGE-clusters are found either in antisense relative to annotated genes or reside in intronic and intergenic regions. We named these putative novel stem specific transcripts NASTs for Non-Annotated-Stem-Transcripts (Figure 1c). We looked at NASTs expression within the FANTOM5 atlas and realized these transcript show very much stem specific expression patterns, being express in no other somatic cell types/tissues than testis (Figure 1d).

Using publically available histone marks ChIP-seq data (The ENCODE Project Consortium, Nature, 2012), we have shown that over 80% of human and mouse NASTs carry histone marks for enhancer, promoter or other combination of active transcription marks. As additional characterization, we analyzed NASTs relative expression and length of associated *de novo* assembled transcripts. We found that there are significantly shorter and expressed at lower levels than annotated mRNAs.

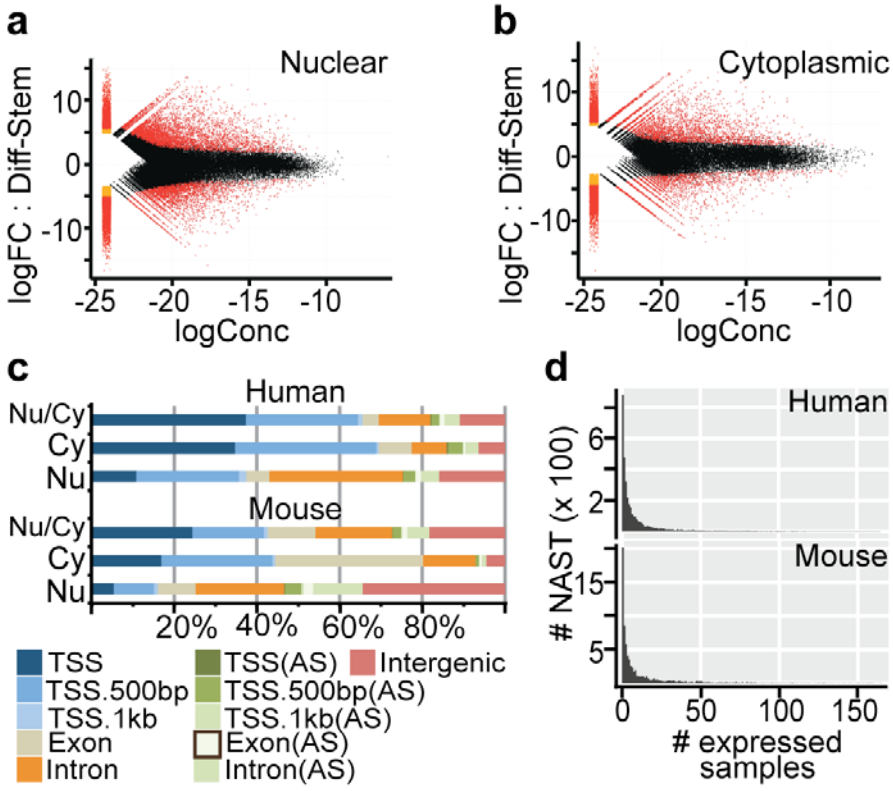

Figure1: a-b. MAplots of differentially expressed CAGE-clusters (FDR<0.01 marked in red) for mouse nuclear (**a.**) and cytoplasmic (**b.**) datasets. **c.** Annotation of CAGE-tag-clusters significantly up-regulated in stem cells and detected in nucleus (Nu), cytoplasm (Cy) or in both compartments (Nu/Cy). AS: antisense. **d.** Number of tissues and differentiated cell type samples from the FANTOM5 atlas in which non-annotated stem transcripts (NAST) are expressed. Bin-width=1.

We analyzed the NASTs repeat composition and found that their promoters localized more often than expected by chance in specific LTR retrotransposons families. Of interest, such enrichments are generally not observed among un-annotated transcripts up-regulated in control-differentiated cells. In detail, novel transcripts associated with LTR-ERVK and LTR-MaLR elements appear clearly enriched in mouse stem cells, while they are more often associated with LTR-ERV1 in human (Figure 2a). In mouse, such enrichments for ERVK are observed for nuclear clusters presenting histone marks for enhancer, promoters and other combinations of active transcription marks as well as cluster lacking such epigenetic marks (Exact Fisher test Bonferroni corrected p<0.05). Comparable enrichments patterns are observed for CAGE-clusters associated with MaLR elements in mouse. In human, nuclear NASTs carrying histone marks for active transcription are significantly enriched for ERV1 elements (Exact Fisher test Bonferroni corrected p<0.05) but not for MaLRs.

It has been recently reported that *Setb1* mediates repression of numerous noncoding and/or repetitive elements in mouse ESC regulating tri-methylation of H3K9 (Karimi et al., Cell Stem Cell, 2011). In this light, we found that NASTs associated with mouse ERVK, mouse MaLR and human ERV1 are deprived of H3K9me3 marks, while the non-expressed elements are indeed carrying this repressive marks (Figure 2b,c).

To investigate further the implication of retrotransposon derived transcripts in stem cells, we performed differential CAGE-cluster expression analyzes focusing exclusively on repeated elements, including sequences mapping to multiple genomic loci. For this purpose, we assessed CAGE-based expression values for each repeat family and sub-family mapping CAGE-tags to all repeated elements of the human and mouse genome as defined by RepeatMasker (Jurka *et al.*, Cytogenet. Genome Res., 2005). When considering expression values calculated for nuclear samples, ERVK and MaLR families appear significantly up-regulated (t-test, Bonferroni corrected p<0.05) in mouse stem cells (Figure 2d). DNA repeats, MuDR, appears also significantly over-expressed while LINE-L1, LTR-ERVL and satellite repeats show similar tendencies but do not pass the strict statistical significance threshold. In human, ERV1 and ERVK depict analogous trends, being expressed at higher levels in stem compare to differentiated cells. More specifically, when focusing at the sub-family level, BGLII, RLTR9E and RLTR17 elements were identified as members of the murine ERVK family being the most significantly up-regulated in stem cells and showing the highest expression levels relative to other repeat elements (Figure 2e). In human, ERV1-LTR7, LTR7B, LTR7Y and HERVH-int elements appear clearly expressed at highest levels and present lowest

FDR values. Importantly, these mouse ERVK and human ERV1 elements are not among the most abundant in the genome (Figure 2f), suggesting that we are in presence of precisely regulated transcription event and not observing products of random pervasive transcription.

Together, these observations suggest that the nuclear transcriptome of stem cell is more complex than previously thought and that an important fraction of the newly identified transcriptomics complexity is composed of genes with promoters associated to a few specific types of mouse ERVK and human ERV1 elements.

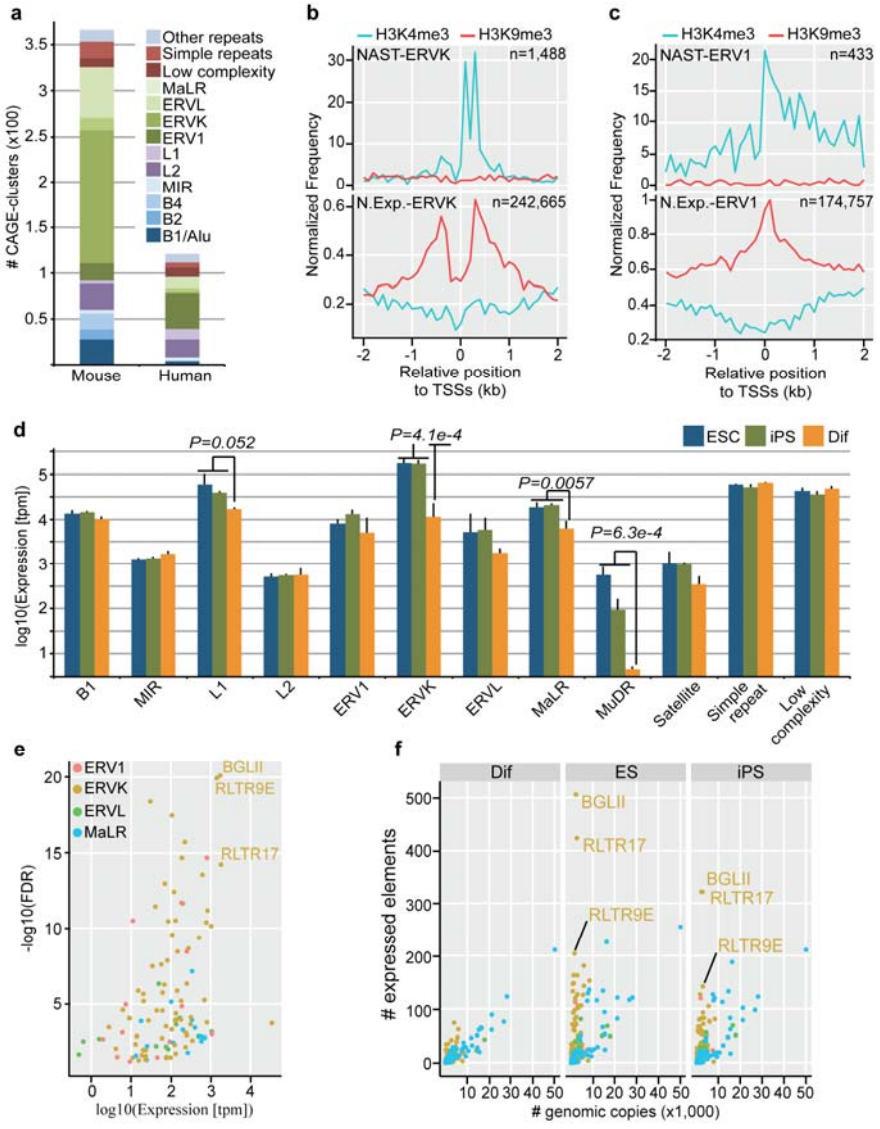

Figure 2: a. Repeat composition of non-annotated stem transcripts (NAST). **b-c.** Frequency plot of normalized tag counts for H3K4me3 (promoter) and H3K9me3 (repressive) ChIP-Seq (ENCODE date on mouse ES-Bruce4) at NAST loci associated with mouse ERVK (**b.**) and human ERV1 (**c.**). **d.** Repeat family expression values in tag per million (tpm) for mouse ESC, iPS and differentiated cells (Dif). Error-bars show S-D, indicated *P*-values are from t-test, two-sided, Bonferroni corrected, n=3. **e.** Relative expressions for selected mouse sub-family repeats expressions, in tpm, are plotted against associated false discovery rate (FDR). **f.** Amounts of repeat elements counting at least 5 CAGE-tags are plotted against copy number found in the genome for mouse LTRs.

Of great interest, when plotting the CAGE-tags distributions over the length of specific ERVK elements, we observed specific divergent transcription pattern for mouse BGLII and RLTR17 elements (Figure 3a) that has been previously reported as landmark of enhancer regions (Kim *et al.*, Nature, 2010). We thus looked for cluster-pairs showing divergent transcription, separated by less than 400bp and overlapping LTR repeats (Figure 3b). In detail, 1498 and 217 of such divergent transcription loci were identified in the mouse and human datasets respectively. In mouse the top three most represented ERVK elements are RLTR17 (97 loci), BGLII (85 loci) and RLTR9E (53 loci), while in human LTR7 (49 loci), HERVH-int (37 loci) and LTR9 (15 loci) are the most abundant. To support regulatory potential functions of these mouse ERVK and human ERV1 associated loci, we assessed whether they present open chromatin configuration and if stem cell specific transcription factors and enhancer binding protein p300 are actually bound to these potential regulatory elements. Publically available DNase-HS data (The ENCODE Project Consortium, Nature, 2012) confirms that these loci presenting divergent transcription and overlapping LTR elements in mouse are on an open chromatin state specifically in ESC but not in differentiated cells (Figure 3c). In addition, ChIPseq data for the main stem specific transcription factors and enhancer associated protein, p300 (The ENCODE Project Consortium, Nature, 2012; Marson *et al*., Cell, 2008) show enriched signal at these potential LTR associated regulatory regions (Figure 3d). Finally, histone marks associated with enhancers (K3K27ac) were clearly enriched at these loci unlike repressive marks (H3K36me3, H3K9me3).

Taken together these results suggest that in stem cells some retrotransposons have been recruited as regulatory elements and can be detected as enhancer-RNA among novel stem cell specific transcripts identified in this study.

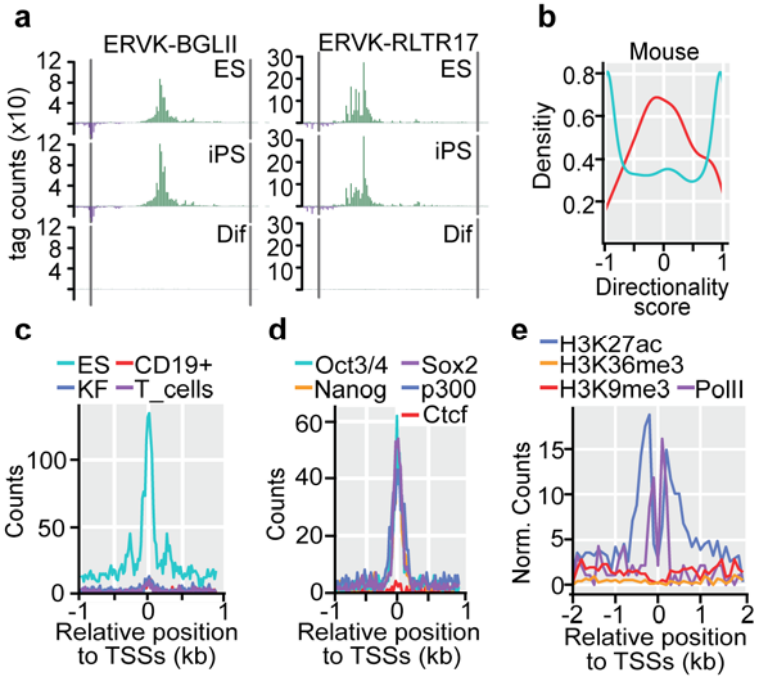

Figure 3: a. Relative CAGE-tags distribution from 5' to 3' extremities (grey bars, +/-10%) of mouse intergenic and intronic BGLII and RLTR17 elements. Green and purple bars indicate CAGE-tags mapping to the (+) and (-) strand respectively. **b.** Density plot for directionality score at loci showing divergent transcription overlapping intergenic LTR (red) and from annotated TSSs (blue). Perfectly balanced transcription is reflected as a directionality score of 0; $[Exp_f - Exp_r]/[Exp_f + Exp_r]$ (Exp_f, Exp_r : expressions from forward and reverse strands). **c-e.** Tag counts density plots for mouse DNaseI-HS (**c.**) and ChIP-Seq (**d, e.**) at loci presenting divergent transcription and overlapping LTR repeats.

Based on the results presented above, we have started the second phase of our project, aiming to functionally screen a large panel of the newly identified stem specific transcript for putative role in the regulation of pluripotency maintenance. For this purpose, we designed RNAi knock-down assays for over 150 candidate transcripts. We are currently testing these candidates by knock-down and gain of function experiments in mouse iPS carrying a GFP reporter gene under the control of a Nanog promoter (Okita *et al.*, Nature, 2007).

We have tested 131 candidates by knock-down experiments and identified 26 transcripts for which Nanog expression is decreased upon perturbation (Figure 4a-c). Their mechanisms of action are currently under investigation, profiling gene-network modification after knock-down.

In addition, we have cloned 66 full-length ncRNA candidates, for which we previously performed 3'RACE aiming to define precisely their 3'ends. We are currently overexpressing them in mouse fibroblasts carrying a Nanog-GFP cassette to assess their potential in cell de-differentiation.

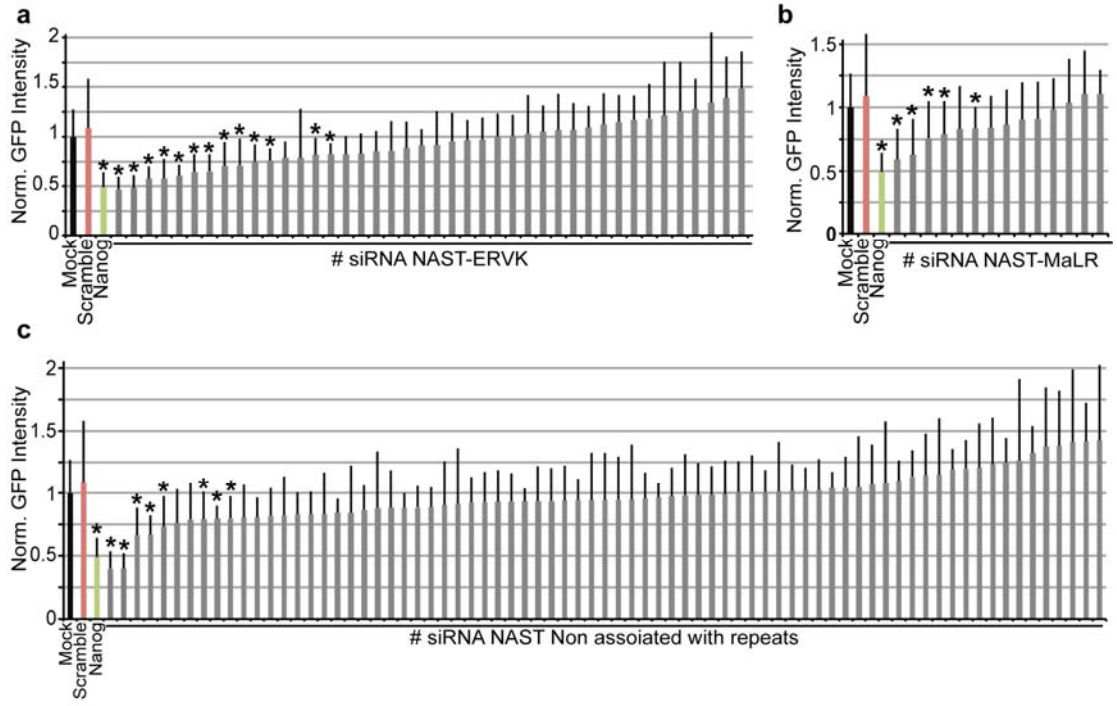

Figure 4: a-c. Relative GFP intensities of transiently transfected miPS-20D17 with 20nM of siRNA targeting NAST associated with ERVK (a.), MaLR (b.) and NAST non-associated with repeated elements (c.). Error-bars show S-D, * $p<0.05$, t-test, two-sided, n=9.

様式19 別紙1

3. 研究発表等

雑誌論文 計2件	（掲載済み－査読有り） 計2件 A Fadloun, S Le Gras, B Jost, C Ziegler-Birling, H Takahashi, E Gorab, P Carninci, M Torres-Padilla. (2013) “Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA” Nature Structural & Molecular Biology, 20, 332-338 B Lenhard, A Sandelin, P Carninci. (2012) “Metazoan promoters: emerging characteristics and insights into transcriptional regulation” Nature Reviews Genetics, 13, 233-245 （掲載済み－査読無し） 計0件 （未掲載） 計0件
会議発表 計12件	専門家向け 計12件 Carninci P. “COMPLEXITY OF MAMMALIAN TRANSCRIPTION ANALYZED BY DEEPCAGE” Keystone Symposium on Non-Coding RNAs. April 1, 2012. Snowbird Resort , Snowbird, United States. Carninci P. “Complexity of the mammalian transcriptome” SRP Diabetes Mini Symposium at Karolinska Institute. May 29, 2012. Karolinska Institute, Stockholm, Sweden. Carninci P. “The complexity of the mammalian transcriptome” Inserm workshop 215 ″Diversity of non coding transcriptomes revealed by RNA-seq″ May 31, 2012. Hotel Mercure Bordeaux Centre, Bordeaux, France. Carninci P. “DISCOVERY OF THE RNA WORLD AND NEW BIOTECHNOLOGY OPPORTUNITIES” Discovery of the RNA World and New Biotechnology Opportunities at ACCJ. June 7, 2012, ACCJ Tokyo Office, Tokyo, Japan. Iwasaki Y, Sato K, Shibuya A, Komai M, Carninci P, Siomi H, Siomi M. “An essential role of a Tudor domain-containing protein, Krimper, in piRNA mediated transposable element silencing in Drosophila germline” ISSCR Annual Meeting. June 13, 2012. Pacifico Yokohama, Yokohama, Japan. Carninci P. “A brief introduction to proteome complexity using high-throughput transcriptome data” The Japan Human Proteome Organization 2012 Annual Meeting / 日本プロテオーム学会 2012 年大会 (10thJHUPO). July 26, 2012. Miraikan, Tokyo, Japan. Iwasaki Y, Sato K, Shibuya A, Kamatani M, Tsuchizawa Y, Carninci P, Siomi H, Siomi M. “An essential role of a Tudor domain-containing protein, Krimper, in Drosophila piRNA biogenesis” Regulatory & Non-Coding RNAs. August 28, 2012. Cold Spring Harbor Laboratory, New York, United States. Ghosheh Y, Ryu T, Clinton M, Carninci P, Faulkner G, Ravasi T. “Genome-wide discovery of piRNA clusters dynamically regulated during brain development” ECCB’12 – the European Conference on Computational Biology 2012. September 9, 2012. Congress Center Basel, Basel, Switzerland. Carninci P. “Miniaturization of CAGE technologies towards single cell profiling” The 2nd Annual BioTechniques ″Virtual″ Symposium (Webinar). October 24, 2012. (Webinar) Online, Yokohama, Japan. Carninci P. “THE COMPLEXITY OF MAMMALIAN TRANSCRIPTION” 第 35 回日本分子生物学会年会; The 35nd Annual Meeting of the Molecular Biology Society of Japan (MBSJ). December 11, 2012. Fukuoka International Congress Center/Marinemesse Fukuoka, Fukuoka, Japan. Carninci P. “Complexity of mammalian transcription” ISCB-Asia/SCCG 2012. December 17, 2012. Kingkey Palace Hotel, Shenzhen, China.

様式19 別紙1

	Carninci P. “Complexity of mammalian transcription” The 34th Annual Lorne Genome Conference 2013. February 17, 2013. Lorne, Victoria, Australia. 一般向け 計 0 件
図書 計0件	
産業財産権 出願・取得状況 計0件	（取得済み） 計0件 （出願中） 計0件
Webページ （URL）	http://www.riken.jp/research/labs/clst/genom_tech/life_sci_accel/transcript_tech/l
国民との科学・技術対話の実施状況	内 容: 横浜サイエンスフロンティア高等学校第4回文化祭への出展 （セントラル・ドグマについての3D 映画の放映、遺伝子やセントラル・ドグマについてのパネル展示） 実施日: 2012 年 9 月 15 日＆16 日 場 所: 横浜サイエンスフロンティア高等学校
新聞・一般雑誌等掲載 計1件	Newton2013 年 5 月号、2013 年 3 月 26 日、ページ 44－63、「あなたは究極の個人情報を手に入れたいか？新・ゲノム革命」
その他	「ヒトゲノムの 80 %に機能 解析プロジェクト「ENCODE」が解明」、理化学研究所、http://www.riken.jp/pr/press/2012/20120906/

4. その他特記事項

様式19 別紙2

課題番号	LS127

実施状況報告書(平成24年度) 助成金の執行状況

本様式の内容は一般に公表されます

1. 助成金の受領状況(累計) (単位:円)

	①交付決定額	②既受領額(前年度迄の累計)	③当該年度受領額	④(=①−②−③)未受領額	既返還額(前年度迄の累計)
直接経費	138,000,000	47,338,000	42,577,000	48,085,000	0
間接経費	41,400,000	14,201,400	12,773,100	14,425,500	0
合計	179,400,000	61,539,400	55,350,100	62,510,500	0

2. 当該年度の収支状況 (単位:円)

	①前年度未執行額	②当該年度受領額	③当該年度受取利息等額(未収利息を除く)	④(=①+②+③)当該年度合計収入	⑤当該年度執行額	⑥(=④−⑤)当該年度未執行額	当該年度返還額
直接経費	857,590	42,577,000	0	43,434,590	42,511,253	923,337	0
間接経費	0	12,773,100	0	12,773,100	12,773,100	0	0
合計	857,590	55,350,100	0	56,207,690	55,284,353	923,337	0

3. 当該年度の執行額内訳 (単位:円)

	金額	備考
物品費	20,873,771	シーケンス用キット、実験試薬、実験用消耗品等
旅費	2,460,184	海外出張費、国内旅費
謝金・人件費等	18,753,283	職員人件費
その他	424,015	学会等参加費、発送費、外勤交通費
直接経費計	42,511,253	
間接経費計	12,773,100	
合計	55,284,353	

4. 当該年度の主な購入物品(1品又は1組若しくは1式の価格が50万円以上のもの)

物品名	仕様・型・性能等	数量	単価(単位:円)	金額(単位:円)	納入年月日	設置研究機関名