様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １１ 日現在

機関番号：３２６８９
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１５７８４

研究課題名（和文）Identification and characterization of lncRNAs involved in genetic compensation

研究課題名（英文）Identification and characterization of lncRNAs involved in genetic compensation

研究代表者
曽　超（ZENG, CHAO）

早稲田大学・理工学術院総合研究所（理工学研究所）・次席研究員（研究院講師）

研究者番号：８０８２２４６０

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：遺伝子発現降下（deGene）が遺伝子補償応答（GCR）を誘発し、類似配列適応遺伝子（adGene）の転写発現が増加する可能性があります。本研究では、RNA-seqデータを用いてヒト細胞のGCRに関与する遺伝子ペアを予測しました。複数のRNA分解因子のノックダウンデータから、数百のdeGene-adGeneペアを同定し、データセット間での重複が少ないことが示唆されました。この結果は、ヒト細胞における遺伝子発現調節の複雑なダイナミクスに対する新たな理解を提供します。

研究成果の学術的意義や社会的意義
この研究は、特に遺伝的補償応答（GCR）の役割に焦点を当て、ヒト細胞における遺伝子発現調節の複雑なダイナミクスについての理解を深める可能性があります。GCRに関与する数百の遺伝子ペアを特定することで、遺伝子調節を理解し、可能ならば治療目的で操作するための新たな途が開かれます。これは、遺伝性疾患や病気の治療に広範な影響を及ぼす可能性があり、社会の健康と福祉に寄与することになります。

研究成果の概要（英文）：Degraded transcript (referred to as deGene) may trigger the genetic compensation response (GCR), potentially leading to an increase in the transcriptional expression of sequence-similar adapting genes (adGene). In this study, we aimed to predict gene pairs involved in GCR in human cells using RNA-seq data. From the knockdown data of multiple RNA degradation factors, we identified hundreds of deGene-adGene pairs, with little overlap suggested between different datasets. These results provide new insights into the complex dynamics of gene expression regulation in human cells.

研究分野：オミックス解析

キーワード：GCR　genetic compensation　遺伝子補償応答

１．研究開始当初の背景

Genetic compensation response (GCR), a mutant mRNA degradation (deGene) activates the increasing expression of its adapting gene (adGene), was first experimentally validated in zebrafish and mouse [1][2]. It was reported that two indispensable conditions for triggering the GCR: RNA degradation (mainly NMD) and sequence similarity (Fig. 1). RNA degradation promotes the nuclear translocation of the mRNA sequence or fragment, which then further tracks to the adapting gene by sequence similarity. The entire process may finally lead to a chromatin remodeling (e.g. histone modification), which up-regulates the transcriptional level of the adapting gene. However, the occurrence of GCR in human cells has not been studied.

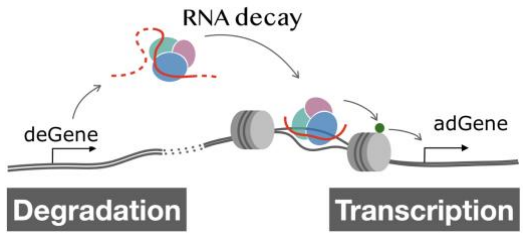

Figure 1 Mechanism of genetic compensation response

２．研究の目的

The purpose of this study is to predict GCR-related genes and investigate their biological functions in human cells. RNA-seq data of

３．研究の方法

3.1 Data

Human genomic sequences (hg38) and gene annotation (v35, comprehensive) were downloaded from GENCODE project [3]. RNA-seq of knockdown of RNA decay factors (UPF1, SMG6, SMG7, SMG6/7, XRN) were downloaded from the GEO database under access numbers GSE86148 and GSE115471.

3.2 Alignment, quantification, and differential expression

We applied STAR (2.6.0c) [4] to align RNA-seq reads to the reference genome and set *outFilterMultimapNmax* to 1 to filter for uniquely mapped reads. In addition, we utilized the parameters quantMode and *twopassMode* for TranscriptomeSAM and Basic, respectively. RSEM (v1.3.1) [5] estimated gene expression from the alignment results. Then, EBSeq (v1.2.0) [6] calculated the fold change and the significance (False Discovery Rate, FDR) for the differential expression between knockdown of NMD factor and control. We used default parameters for both RSEM and EBSeq.

To visualize aligned reads in the UCSC Genome Browser [7], we first indexed the coordinate-sorted alignment files (BAM format) with Samtools (v1.7) [8], normalized them as aligned reads in bedGraph format using bamCoverage (v3.5.0, *--binSize 1 --normalizeUsing RPKM*) [9], then converted into bigWig format with bedGraphToBigWig (v4) [10]. The chromosome size information required for the above processing can be obtained from the genome index generated by STAR. Next, biological replicates (bigWig format) were merged into a single bedGraph file with bigWigMerge (v2) [11] and finally converted to bigWig format with bedGraphToBigWig.

3.3 Homologous genes

Paralogues and orthologues between huamans, zebrafish and mice were obtained from Ensembl BioMart (v101) [12]. In mapping zebrafish and mouse genes to human homologs (or homologous genes), we considered both human orthologues and their corresponding paralogues in the human genome.

3.4 Mutations and gene expression

Somatic mutations and gene expression were obtained from COSMIC (v92) [11]. A custom Python script extracted samples with genes containing nonsense mutations and corresponding gene expression.

3.5 GO enrichment analysis

Gene ontology enrichment analysis was performed with g:Profiler (version: e101_eg48_p14_baf17f0) [11] using Bonferroni correction (significance threshold: 0.01). MF for molecular function, BP for biological process, and CC for cellular component.

４．研究成果

4.1 Identifying gene pairs involved in transcriptional adaptation

We first asked whether the mechanism of transcriptional adaptation is conserved among homologous genes between species. We tried to detect gene pairs involved in the transcriptional adaptation from RNA-seq

data with knockdown of RNA decay factors. The rational is that, first, when the RNA decay factor is knocked down, we consider that deGenes will have increased expression due to its blocked RNA decay process, and correspondingly adGenes will have decreased expression due to the repressed transcriptional adaptation. This is because RNA decay has been shown to be a key factor in triggering transcriptional adaptation. Second, we limited the search for reGene:adGene to between homologous genes because previous studies found that transcriptional adaptation exists between homologous genes that enjoy sequence similarity and are used to balance phenotype. For example, for the experimentally validated hbegfa:hbegfb gene pair in zebrafish [1], we restricted the search to the homologous genes of this pair in humans based on the above criteria, and manually screened the AREG:HBEGF gene pair (Figure 2A) for possible association with transcriptional adaptation. Among

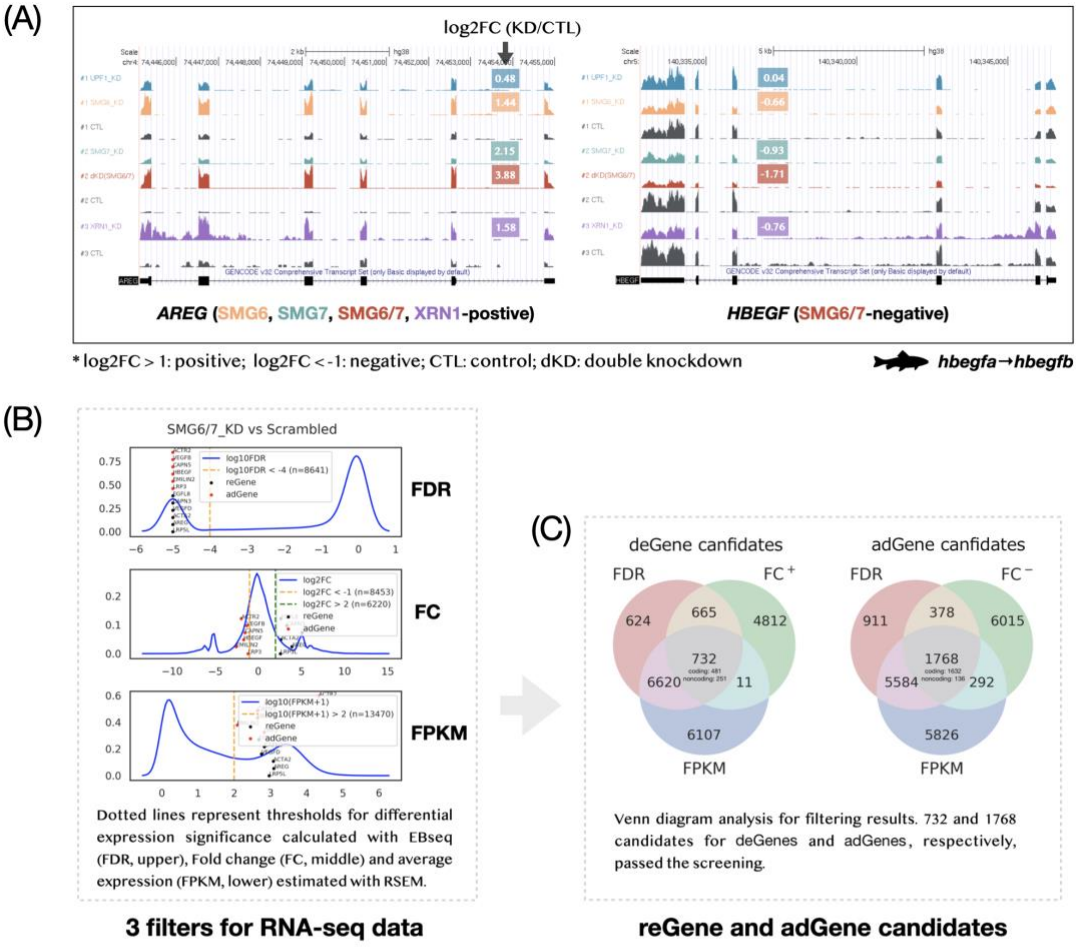

Figure 2. Identification of GCR gene pairs in HeLa cells. (A) AREG:HBEGF gene pair was predicted as GCR in human cells. **(B-C)** Criteria for screening candidates of deGenes and adGenes. **(D)** Venn diagram analysis of deGene:adGene pairs associated with different RNA decay factors (UPF1, SMG6, SMG7, SMG6/7).

them, most of the tested RNA decay factors (excluding UPF1) were knocked down followed by a substantial increase in AREG expression (log2FC > 1) indicating that it may be a target for RNA degradation. Interestingly, HBEGF expression was significantly reduced (log2FC < -1) when SMG6 and SMG7 were simultaneously knocked down, implying that HBBGF may be intrinsically regulated by transcriptional adaptation.

We next sought to determine the number of gene pairs involved in transcriptional adaptation present in human cells. To correct for thresholds, we employed the positive reference outlined earlier and established the following criteria to define candidate reGenes and adGenes from RNA-seq data associated with knockdown of RNA decay factors (Figure 2B). Firstly, candidate genes should exhibit significant expression differences (log10FDR < -4). Secondly, the knockdown of RNA decay should induce considerable expression changes (log2FC > 2) of reGene, while the expression of the corresponding adGene should vary within a relatively moderate range (log2FC < -1). Lastly, to exclude genes that showed significant difference due to low expression, we focused only on those genes with relatively high average expression (log10(FPKM+1) > 2). As shown in Figure 2C, within the SMG6/7 knockdown comparison data, after three levels of filtering, we identified 732 reGene candidates and 1768 adGene candidates.

In this study, we utilized a relaxed fold change (FC) threshold to identify potential adapting gene (adGene) candidates. This approach was chosen due to the indirect regulation of adGene expression by RNA degradation, and consequently, we identified more regulated gene (reGene) candidates in comparison to adGene candidates. By examining the homologous relationships among these candidates, we ultimately identified 6, 34, 156, and 345 reGene-adGene pairs within the knockdown datasets for SMG6, SMG7, UPF1, and SMG6/7, respectively (Figure 2D). We observed a low rate of duplication between the detections in different datasets, with 33.3%-91.3% of the reGene-adGene pairs being dataset specific. Notably, gene pairs identified in the SMG6 and SMG7 datasets were also found in the SMG6/7 dataset at rates of 66.7% (4/6) and 47.1% (16/34), respectively. Interestingly, there was no overlap of reGene-adGene pairs detected in the SMG6 dataset with those from the SMG7 dataset. Moreover, the results from concurrent knockdown of SMG6 and SMG7 were notably different from those obtained when these two factors were knocked down separately. For instance, 94.2% (325/345) of the gene pairs were only observed following simultaneous SMG6/7 knockdown. In addition, although a small number of reGene and adGene candidates were extracted from XRN1 knockdown experiments, we did not identify any reGene-adGene pairs among these candidates via homologous linkage. This suggests that gene pairs involved in transcriptional adaptation may be broadly present in human cells.

4.2 Validation with nonsense mutation and gene expression

Accounting for the possibility that Nonsense-Mediated Decay (NMD) prompted by nonsense mutations could also induce transcriptional adaptation, we utilized COSMIC data [12] to authenticate the presumptive reGene-adGene pairs identified in this study. COSMIC data encompasses information concerning the somatic mutation and gene expression of individual samples. Our working hypothesis is that the occurrence of a nonsense mutation in a reGene might lead to an increase in the expression of the corresponding adGene. For instance, the gene pair RBMS3-HNRNPR, involved in transcriptional adaptation, was identified in both UPF1 and SMG6/7 knockdown data, where RBMS3 is the reGene and HNRNPR

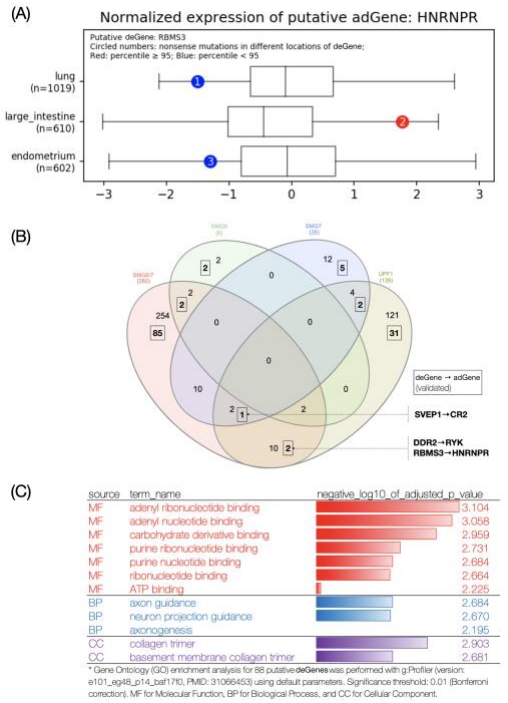

Figure 3. Validation of predicted GCR gene pairs with COSMIC. (A) NMD in RBMS3 induces an increase in expression of HNRNPR in large intestine. **(B)** Overview of GCR pairs validated by COSMIC. **(C)** GO analysis of deGenes.

is the adGene (Figure 2D). We discovered one sample in each of three tissues exhibiting a nonsense mutation in the RBMS3 gene, with the mutations occurring at distinct locations within the gene. As anticipated, we noted a significant increase in HNRNPR gene expression (≥ 95th percentile) in samples with nonsense mutations in the RBMS3 gene compared to other samples in the large intestine (Figure 3A). This observation bolsters the potential for transcriptional adaptation between RBMS3 and HNRNPR due to nonsense mutation. We refer to such gene pairs present in arbitrary tissues as "validated", primarily derived from nonsense mutation data. While a more robust statistical test could potentially be achieved with a larger sample size of nonsense mutation data, the available nonsense mutation data can provide a viable testing ground. However, the number of samples with nonsense mutations in the COSMIC database is limited, and the objective of this study is to identify as many gene pairs exhibiting transcriptional adaptation in humans as possible, while tolerating some degree of false positives. Ultimately, we identified a total of 130 reGene-adGene pairs validated with COSMIC data (Figure 3B), suggesting the potential for intrinsic, previously unobserved, transcriptional adaptation between these gene pairs. We then sought to understand the biological functions of these 130 gene pairs. We segregated these gene pairs into two groups: reGenes (n=88) and adGenes (n=102), and conducted separate gene ontology enrichment analyses. Interestingly, adGenes are primarily involved in axon guidance and genesis (padj ≤ 10e-2.195, Fig 3C), while reGenes were predominantly associated with spindle organization and assembly (padj ≤ 10e-2.145, not shown).

〈引用文献〉

[1] EI-Brolosy, M. A. et al. Genetic compensation triggered by mutant mRNA degradation. Nature 568, 193-197 (2019)
[2] Ma, Z. et al. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature 568, 259-263 (2019)
[3] Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research 47, D766-D773 (2019)
[4] Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013)
[5] Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics 12, 323 (2011)
[6] Leng, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29, 1035-1043 (2013)
[7] Kent, W. J. et al. The human genome browser at UCSC. Genome research 12, 996-1006 (2002)
[8] Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079 (2009)
[9] Ramírez, F. et al. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research 42, W187-W191 (2014)
[10] Kent, W. J. et al. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26, 2204-2207 (2010)
[11] Yates, A. D. et al. Ensembl 2020. Nucleic acids research 48, D682-D688 (2020)
[12] Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic acids research 47, D941-D94710 (2019)
[13] Raudvere, U. et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research 47, W191-W198 (2019)

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 Onoguchi Masahiro、Zeng Chao、Matsumaru Ayako、Hamada Michiaki	４．巻 3
２．論文標題 Binding patterns of RNA-binding proteins to repeat-derived RNA sequences reveal putative functional RNA elements	５．発行年 2021年
３．雑誌名 NAR Genomics and Bioinformatics	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nargab/lqab055	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Hamaguchi Yu、Zeng Chao、Hamada Michiaki	４．巻 22
２．論文標題 Impact of human gene annotations on RNA-seq differential expression analysis	５．発行年 2021年
３．雑誌名 BMC Genomics	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12864-021-08038-7	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Zeng Chao、Hamada Michiaki	４．巻 11
２．論文標題 RNA-Seq Analysis Reveals Localization-Associated Alternative Splicing across 13 Cell Lines	５．発行年 2020年
３．雑誌名 Genes	６．最初と最後の頁 820～820
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/genes11070820	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Zeng Chao、Hamada Michiaki	４．巻 2254
２．論文標題 Detection and Characterization of Ribosome-Associated Long Noncoding RNAs	５．発行年 2020年
３．雑誌名 Methods in Molecular Biology	６．最初と最後の頁 179～194
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-0716-1158-6_11	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Zeng Chao、Onoguchi Masahiro、Hamada Michiaki	4．巻 12
2．論文標題 Association analysis of repetitive elements and R-loop formation across species	5．発行年 2021年
3．雑誌名 Mobile DNA	6．最初と最後の頁 3
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s13100-021-00231-5	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Takahashi Satoshi、Noro Rintaro、Seike Masahiro、Zeng Chao、Matsumoto Masaru、Yoshikawa Akiko、Nakamichi Shinji、Sugano Teppei、Hirao Mariko、Matsuda Kuniko、Hamada Michiaki、Gemma Akihiko	4．巻 22
2．論文標題 Long Non-Coding RNA CRNDE Is Involved in Resistance to EGFR Tyrosine Kinase Inhibitor in EGFR-Mutant Lung Cancer via eIF4A3/MUC1/EGFR Signaling	5．発行年 2021年
3．雑誌名 International Journal of Molecular Sciences	6．最初と最後の頁 4005～4005
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/ijms22084005	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Chao Zeng, Atsushi Takeda, Kotaro Sekine, Naoki Osato, Tsukasa Fukunaga, Michiaki Hamada	4．巻 2509
2．論文標題 Bioinformatics Approaches for Determining the Functional Impact of Repetitive Elements on Non-coding RNAs	5．発行年 2022年
3．雑誌名 Methods in Molecular Biology	6．最初と最後の頁 315-340
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-0716-2380-0_19	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　2件）

1．発表者名 Masahiro Onoguchi, Chao Zeng, Yukiteru Ono, Michiaki Hamada
2．発表標題 Binding patterns of RNA binding proteins to repeat-derived RNA sequences reveal putative functional RNA elements
3．学会等名 Noncoding RNAs: Biology and Applications (Keystone Symposia)（国際学会）
4．発表年 2021年

１．発表者名
Masahiro Onoguchi, Chao Zeng, Yukiteru Ono, Michiaki Hamada

２．発表標題
Binding Patterns of RNA Binding Proteins To Repeat-Derived RNA Sequences Reveal Putative Functional RNA Elements

３．学会等名
第22回日本RNA学会年会

４．発表年
2021年

１．発表者名
Chao Zeng, Michiaki Hamada

２．発表標題
Searching for Repeat-derived Functional Elements in RNAs

３．学会等名
第22回日本RNA学会年会

４．発表年
2021年

１．発表者名
Yu Hamaguchi, Chao Zeng and Michiaki Hamada

２．発表標題
ヒト遺伝子アノテーションがRNA-seq解析に与える影響の評価

３．学会等名
第10回生命医薬情報学連合大会(IIBMP2021)

４．発表年
2021年

１．発表者名
曽 超、浜田 道昭

２．発表標題
リピート配列由来の機能エレメントの探索

３．学会等名
第10回生命医薬情報学連合大会(IIBMP2021)

４．発表年
2021年

１．発表者名
Ryo Niwa, Chao Zeng and Michiaki Hamada

２．発表標題
転写因子結合部位として働く反復配列の発見とその機能の解明

３．学会等名
第10回生命医薬情報学連合大会(IIBMP2021)

４．発表年
2021年

１．発表者名
Chao Zeng, Masahiro Onoguchi, Michiaki Hamada

２．発表標題
Bioinformatic approaches for understanding RNA reincarnation

３．学会等名
第43回日本分子生物学会

４．発表年
2020年

１．発表者名
Chao Zeng, Masahiro Onoguchi, Risa Maemura, Michiaki Hamada

２．発表標題
Sequence characterization of R-loop-forming RNAs in mammals

３．学会等名
Transposable Elements（国際学会）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関