慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	Time-course low input RNA-Seq analysis of Ramazzottius varieornatus for the comprehensive identification of genes related to gamma ray irradiation response
Sub Title	
Author	吉田, 祐貴(Yoshida, Yuki)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2015
Jtitle	生命と情報 No.22 (2015.) ,p.95- 106
JaLC DOI	
Abstract	Several terrestrial tardigrades enter a state called anhydrobiosis when the surrounding environment becomes desiccated. In this state, the tardigrade Ramazzottius varieornatus is known to tolerate various extreme environments (high and low temperatures/pressures, high dosage of ultraviolet and gamma rays). Recent studies have reported several novel protection proteins that may play important roles in the mechanism of R. varieornatus anhydrobiosis. Nevertheless, a single cell gel electrophoresis conducted on R. varieornatus showed double strand breaks in the DNA when irradiated with 4000 Gy Co60 gamma rays, a dosage that does not strongly affect the life span, implying that there is also a prominent repair system. In order to determine the main factors of this repair system, we conducted a transcriptomic analysis for the identification of genes responding to severe stress. We irradiated R. varieornatus with 500 Gy of Co60 gamma ray, a dosage below the LD50 of eggs, and conducted a time-course low-input RNA-Seq analysis for every three hours. We have found a quick response group, which contains the gene Ku 70/80. These genes are known as the first responder in the non-homologous end joining pathway, which works as the main pathway for double strand repair in eukaryotes. In addition, a novel gene family with no similar proteins in public databases was found to have the same expression profiles with Ku80. Homologs of this gene family in Hypsibius dujardini showed increase in expression during the rehydration phase, implying that this putative gene family is related to cellular repair. This is one of the first reports of a multi-replicate time-course RNA-Seq analysis of tardigrades, and these results would shed new light into the response to excessive stress of extremophiles.
Notes	慶應義塾大学湘南藤沢キャンパス先端生命科学研究会 2015年度学生論文集 卒業論文ダイジェスト
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001004-00000022-0095

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

Graduation thesis digest

Time-course low input RNA-Seq analysis of *Ramazzottius varieornatus* for the comprehensive identification of genes related to gamma ray irradiation response

Keio University
Faculty of Environment and Informational studies

Yuki Yoshida
t12968yy@sfc.keio.ac.jp

Abstract

Several terrestrial tardigrades enter a state called anhydrobiosis when the surrounding environment becomes desiccated. In this state, the tardigrade *Ramazzottius varieornatus* is known to tolerate various extreme environments (high and low temperatures/pressures, high dosage of ultraviolet and gamma rays). Recent studies have reported several novel protection proteins that may play important roles in the mechanism of *R. varieornatus* anhydrobiosis. Nevertheless, a single cell gel electrophoresis conducted on *R. varieornatus* showed double strand breaks in the DNA when irradiated with 4000 Gy Co^{60} gamma rays, a dosage that does not strongly affect the life span, implying that there is also a prominent repair system. In order to determine the main factors of this repair system, we conducted a transcriptomic analysis for the identification of genes responding to severe stress. We irradiated *R. varieornatus* with 500 Gy of Co^{60} gamma ray, a dosage below the LD_{50} of eggs, and conducted a time-course low-input RNA-Seq analysis for every three hours. We have found a quick response group, which contains the gene Ku 70/80. These genes are known as the first responder in the non-homologous end joining pathway, which works as the main pathway for double strand repair in eukaryotes. In addition, a novel gene family with no similar proteins in public databases was found to have the same expression profiles with Ku80. Homologs of this gene family in *Hypsibius dujardini* showed increase in expression during the rehydration phase, implying that this putative gene family is related to cellular repair. This is one of the first reports of a multi-replicate time-course RNA-Seq analysis of tardigrades, and these results would shed new light into the response to excessive stress of extremophiles.

Keywords : *Ramazzottius varieornatus,* extremophile, gamma ray, low-input RNA-Seq

1. Background

It is commonly known that water comprises nearly 70% of the cell [1], followed by carbohydrates, nucleic acids, lipids, and proteins. The water molecule acts as a solvent for most of the chemical reactions in the cell, therefore loss of water will cause cell death. 300 years ago, Antony van Leeuwenhoek observed the first anhydrobiotic rotifer [2], which can tolerate complete desiccation. Tardigrades, an organism that forms the phylum tardigrada, is also capable of anhydrobiosis, and is known to tolerate extreme environments [3-5]. In particular, *Ramazzottius varieornatus* can tolerate 4000 Gy of He ion beam even in the active stage, implying that it has a prominent protection and repair mechanisms. An protein localized in the nucleus called Dsup has been reported to bind to the DNA molecule, and when transfected into a human cell, enhanced gamma ray tolerance [6]. However, when irradiated with 4000Gy of ^{60}Co gamma ray, the DNA showed signs of fragmentation, implying that even with Dsup functioning, damage to the DNA molecule can be induced [7]. 4000Gy of ^{60}Co does not vary the life of *R. varieornatus* greatly, thus implying that the repair system is too prominent, with the possibility of novel genes aiding.

Therefore, we have planned and conducted a time-course RNA-Seq analysis of gamma ray irradiated tardigrades, to extract genes that are regulated by irradiation. We would be able to characterize the transcriptomic response of known pathways, i.e. DNA repair pathways, cell cycle regulation, etc. Furthermore, analysis of tardigrade specific genes that have been significantly up regulated in a short time frame may be related to repair of cellular molecules.

One major problem is that RNA-Seq requires few µgrams of total RNA input. RNA extracted from a single tardigrade averages at about 2 ng / animal, which will come to about 1000 animals. If one was to conduct a triple replicate, 5 time point RNA-seq, nearly 15,000 animals would be needed. Species like *H. dujardini* that duplicates at a rapid pace can be used for this kind of analysis, but *R. varieornatus* cannot. Therefore, previous studies have only been used single replicate for analysis. By using low-input RNA-seq methods, we have been able to conduct RNA-Seq by only 30 animals, enabling a triple replicate, multiple time point RNA-Seq. Using this method, we have irradiated *R. varieornatus* with 500 Gy, which is below the LD_{50} of eggs, and conducted a time-course RNA-Seq analysis.

2. Methods

2.1 Low-input time-course RNA-Seq

The workflow of the experimental protocol is shown in Fig. 1A. Tardigrades were reared based on the method established by Horikawa [8]. Animals were placed on 90 mm plastic plates layered with 2% volvic agar at 22°C, and fed with Chlorella vulgaris at Keio University, Institute of Advanced Biosciences (IAB). These plates were sealed with parafilm and transported to Japan Atomic Energy Agency (JAEA), Takasaki Advanced Radiation Research Institute. On arrival, the plates were incubated at 22°C for more than 12 hours to eliminate the effects of the transportation.

2,000 tardigrades were suspended in a 1.5 mL tube, and filled with 100 µL of distilled water. These tubes were then irradiated with 500 Gy of ^{60}Co gamma ray, at a dose rate of 2,008 Gy/hour at a distance of 63 cm from the ^{60}Co source. The contents of the tubes were then spread over a 90 mm agar layered plate, and incubated at 22°C until further sampling. At every three hours until 24 hours, 30 animals were picked with 10 µL Pipettes with the least water as possible, and placed into 200 µL tubes containing 150 µL of TRIzol (Life Technologies, N=3 per condition), and was perserved at -20°C. These tubes were transported back to IAB in a foamed styro box, with dry ice to maintain low temperature, and was placed into -20°C freezer upon arrival. RNA extraction was conducted using Direct-Zol (Zymo Research), and cDNA synthesis, sequencing adaptor and index ligation, PCR amplification of library was conducted by NEBNext Ultra RNA Kit (New England BioLabs Japan) according to manufacture protocol. RNA was extracted within 1 week of irradiation. 20 ng of Total RNA was used as the input

of cDNA synthesis, and cDNA was amplified by at least 15 cycles of PCR (h1_1~h6_2 : 15 cycles, h6_3~h21_3 : 21 cycles, h24_1~h24_3 : 23 cycles). Quality check using Nanodrop (Thermo Scientific) and nucleic acid quantification by Qubit RNA High Sensitivity or dsDNA Broad Range (Life Technologies) for both total RNA and sequence library, and electrophoresis of the sequence library was done by Tapestation D1000 for fragment size estimation. Since sample h09_1 showed signs of adapter dimers, further electrophoresis was done by E-Gel 2.0% (Life Technologies) after sample pooling, and extraction of the library was conducted using NucleoSpin Gel and PCR Clean-up (MACHEREY-NAGEL). Sequencing was preformed with NextSeq 500 (Illumina) following manufacture's instructions, using NextSeq 500 High Output Kit (75bp single end).

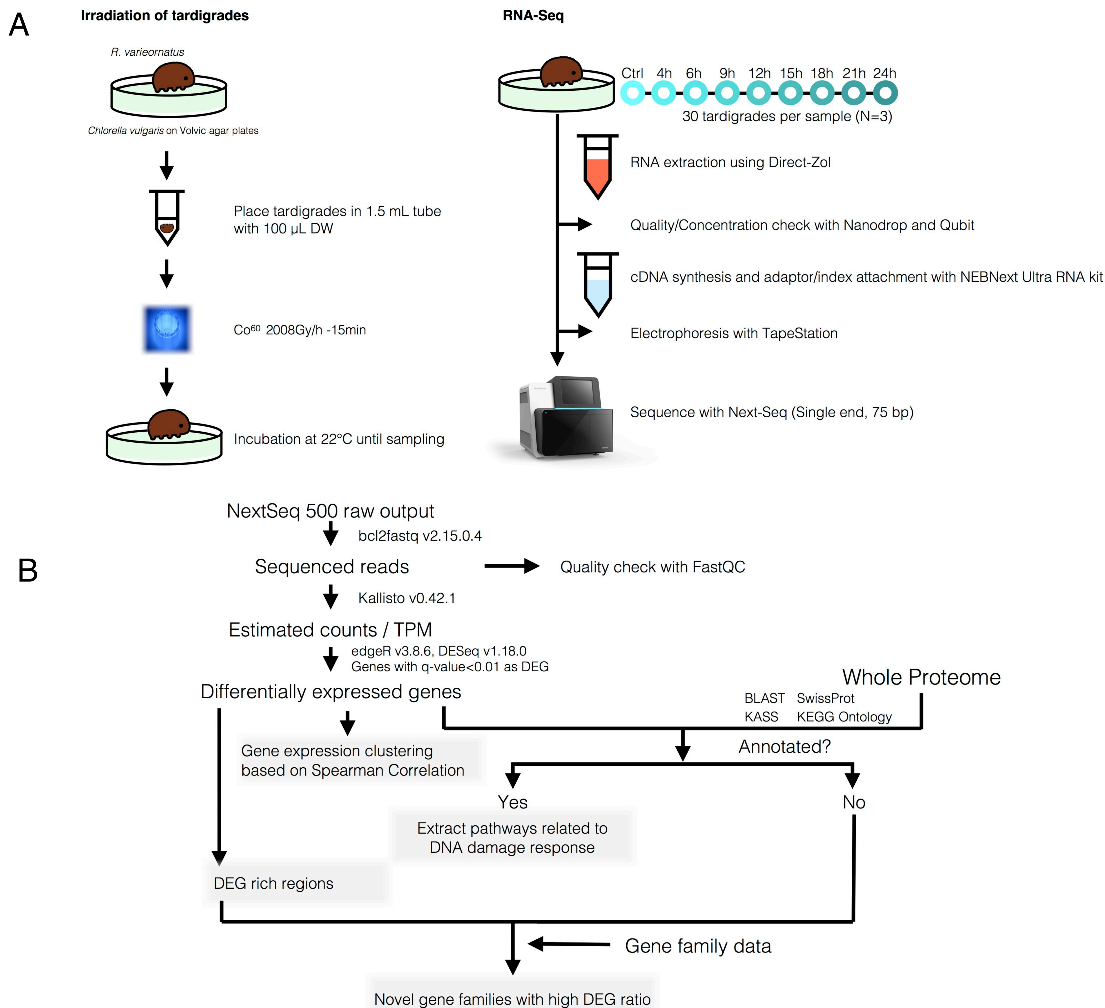

Figure 1. Experimental workflow and informatics analysis workflow.
[A] Experimental protocol of this research. Gamma ray irradiation to tardigrade was conducted at JAEA, and the RNA-Seq at IAB. ©TARUDI, ©Illumina Inc. **[B]** Informatics analysis workflow of this study. Gene expression quantification and differentially expressed testing were done, and using SwissProt/KEGG Ortholog annotation, analysis of cellular damage response related genes were conducted.

2.2. Expression analysis

Custom perl scripts and G-language Genome Analysis Environment v1.9.0 [9, 10] were used for sequence manipulation and data parsing. The workflow of the informatics protocol is shown in Fig. 1B. Next-Seq raw output includes several file formats, but to gain sequence files for each samples, de-multiplexing by bcl2fastq v2.15.0.4 (Illumina) and each sequencing lane were merged. Quality check of each sequenced sample were done by FastQC v0.11.2 [11].

Gene expression quantification (Transcripts Per Million reads : TPM) was done against *R. varieornatus* transcriptome [12] using Kallisto v0.42.1 [13], and statistical tests were done with edgeR v3.10.5 [14] and DESeq

v1.20.0 [15]. Genes that had FDR corrected q-value below 0.01 in both methods were called as statistically differentially expressed genes (DEGs). Gene expression profiles were clustered based on spearman correlations and ward method between both samples and transcripts using R v3.2.2 [16]. Assuming that the early responsive groups were related to cellular damage repair, an enrichment analysis of functional domains were done using Fisher's exact test against Pfam-A domain [17] annotation made by HMMER v3.0 [18] of the whole proteome and DEGs. As for annotation of *R. varieornatus* genes, BLAST search [19, 20] of the whole proteome sequence against SwissProt and Uniref90 [21] and KAAS search [22] against KEGG ortholog database [23] were conducted. We then extracted DEGs that were homologous to genes in known DNA repair pathways (E-value < 1e-15).

2.3. Screening of novel gene families

To investigate if novel paralogs were found in the DEGs, we first screened *R. varieornatus* genome for multi-paralogous gene families. From the results of a self-robin similarity search of the whole proteome using SSEARCH v36.3.6 [24], we clustered genes with identity of over 40 with mclblastline (--mcl-I=2.0, [25]).

Furthermore, it has been implied that *R. varieornatus* genome has several tandem paralogs. Therefore, we screened the genome of *R. varieornatus* for regions with high DEG ratio, and searched for the previous gene families. From these gene families, we filtered clusters with counts below 10 and DEG ratio below 0.2. Localization signals of these gene families predicted using TargetP v1.1b, March 2006 [26] and NLStradamus v1.8 [27]. Of the three clusters found, further sequence analysis was done against Clust_14. Multiple alignment using clustalo v1.1.0 [28, 29] and phylogenetic tree construction with PhyML v20120412 [30] were done using Seaview 4.5.2 [31], with 1,000 bootstraps. For the three groups, MEME v4.10.1 [32] was used for motif search in the amino acid sequences.

3 Results

3.1. Clustering analysis showed two responding groups

Clustering of gene expression profiles were conducted to characterize the response to gamma ray irradiation. From the clustering between each sample using the whole transcriptome (Fig. 2A), the correlation between each sample were over 0.9, whereas h09_1 and h18_2 showed low similarity. Meanwhile, when clustering only DEGs (Fig. 2B), we observed two response groups, before 6 hours (h06) and after 9 hours (h09). Since the latter response continues until 24 hours, we determined this group as the late response, and the former group as the early response group. This grouping does not show in the DEG count (Tab. 4-1), but clustering expression profiles of DEGs between transcripts (Fig. 3) showed that nearly a third of the DEGs were early response groups (Group 6, 4, 7). Housekeeping genes were not found to be DE.

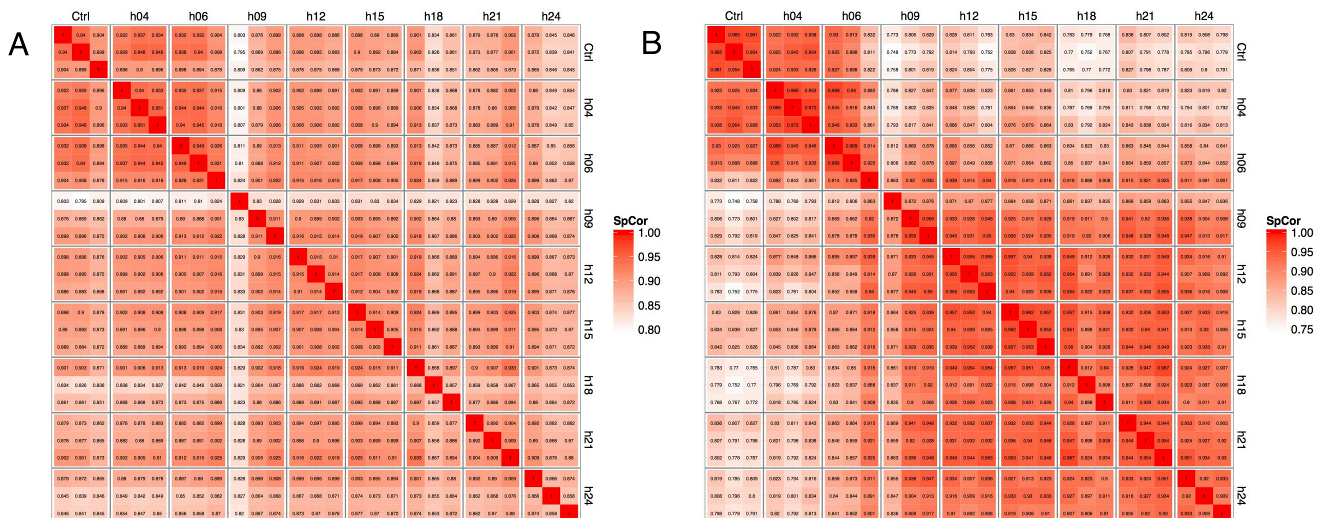

Figure 2. Two gamma ray irradiation response groups were observed.
Heat map of spearman correlation between samples, using the whole transcriptome **[A]**, and only DEGs **[B]**.

To determine the functions of the early response genes, we first analyzed the group 6. 219 genes (230 isoforms) were differentially expressed, and from those genes 84 Pfam-A domain were found. By conducting a Fisher's exact test between the whole genome and these 230 genes, we found 31 domain to be significantly enriched in this group 6 (FDR corrected q-value < 0.05, Table 1A). These domains were associated with 23 genes, with 7 genes with fold change over 2 (Table 1B). In this 7 genes, XRCC5 and XRCC6 are also known as Ku80 and Ku70, respectively, which are the first responders to DSB in the non-homologous end joining pathway. The gene ZN12171.1 has a Kazal domain, has functions similar to the other genes with a NAS annotation. ZN03630.1 was annotate as BCS1, which is a mitochondrial protein with chaperone like functions [33].

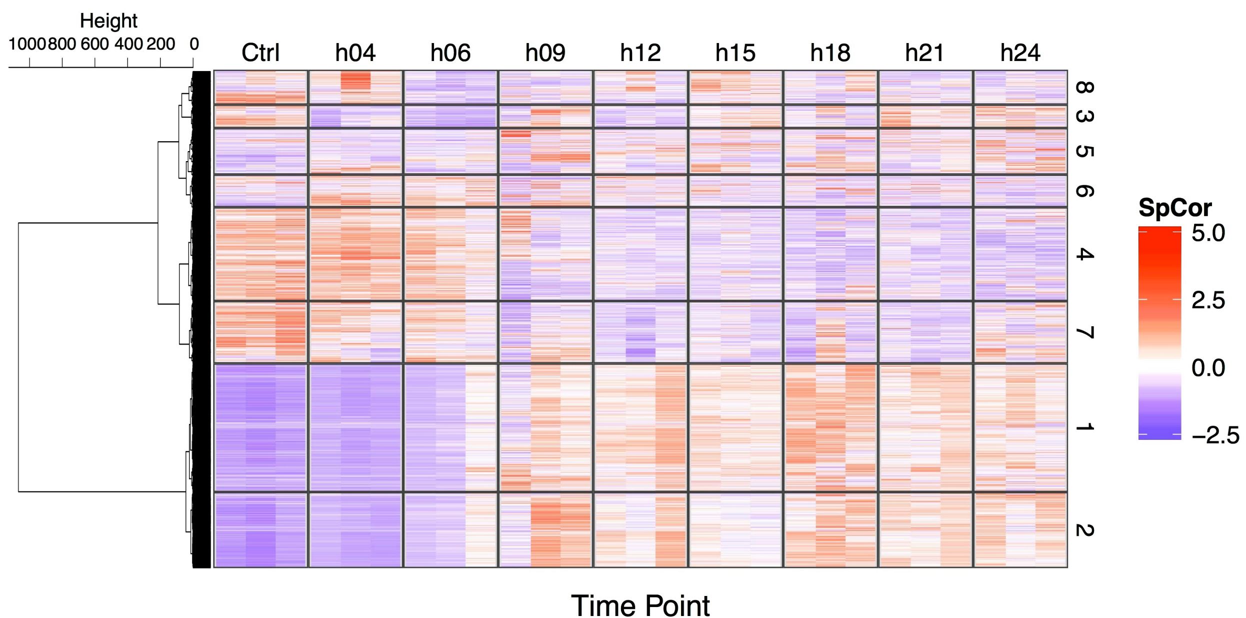

Figure 3. Clustering of TPM expression revealed two response groups.
Clustering based on spearman correlation between transcripts. Groups 1 and 2 were late responders, and Groups 4, 6, 7 were early responders.

A

Domain	DEG	Genome	*p*-value	q-value
EGF_2	3	3	2.21E-05	1.00E-03
Kazal_1	4	13	2.47E-05	1.00E-03
Peptidase_C1	4	13	2.47E-05	1.00E-03
ZZ	3	6	9.08E-05	2.77E-03
Annexin	4	22	1.44E-04	3.52E-03
Astacin	5	61	6.24E-04	6.49E-03
Kazal_2	3	14	6.91E-04	6.49E-03
Ku	2	2	6.50E-04	6.49E-03
Ku_C	2	2	6.50E-04	6.49E-03
Ku_N	2	2	6.50E-04	6.49E-03
Propeptide_C1	2	2	6.50E-04	6.49E-03
SPRY	3	11	3.79E-04	6.49E-03
VIT	2	2	6.50E-04	6.49E-03
AAA_16	2	4	1.60E-03	1.15E-02
ATP-grasp	2	4	1.60E-03	1.15E-02
RuvB_N	2	4	1.60E-03	1.15E-02
VWA_3	2	4	1.60E-03	1.15E-02
Calpain_III	2	5	2.23E-03	1.36E-02
Notch	2	5	2.23E-03	1.36E-02
VWA_2	2	5	2.23E-03	1.36E-02
ATPgrasp_Ter	2	6	2.95E-03	1.50E-02
HSP20	2	6	2.95E-03	1.50E-02
PAN_4	2	6	2.95E-03	1.50E-02
RyR	2	6	2.95E-03	1.50E-02
MIT	2	7	3.77E-03	1.84E-02
AAA_17	2	8	4.68E-03	2.04E-02
CPSase_L_D2	2	8	4.68E-03	2.04E-02
CPSase_L_chain	2	8	4.68E-03	2.04E-02
ATP-grasp_4	2	10	6.77E-03	2.85E-02
C2-set_2	2	11	7.95E-03	3.23E-02
AAA	3	40	1.03E-02	4.06E-02

B

Gene	Swissprot Annot	Ctrl	h04	Fold Change
ZN04375.1	HSPB1_CHICK	206.57	197.34	0.955
ZN13480.1	HSPBB_DANRE	268.63	276.98	1.031
ZN10121.1	VWA5A_RAT	10.47	11.33	1.082
ZN18381.1	CAN7_MOUSE	8.02	8.69	1.084
ZN10472.1	SQSTM_RAT	159.15	184.29	1.158
ZN10508.1	CATB_MOUSE	216.05	257.48	1.192
ZN00629.1	TERA_DANRE	75.49	94.76	1.255
ZN12213.1	HCE1_ORYLA	25.30	32.69	1.292
ZN06741.1	ANXB9_DROME	22.53	29.85	1.325
ZN06010.1	CATZ_MOUSE	369.20	512.79	1.389
ZN16394.1	SQSTM_RAT	105.60	152.43	1.443
ZN15099.1	CFAD_DICDI	47.26	69.48	1.470
ZN01002.1	SQSTM_PONAB	223.26	341.22	1.528
ZN16458.1	HCE1_ORYLA	22.57	37.31	1.653
ZN02602.1	VMA5A_MOUSE	37.40	66.30	1.773
ZN16509.1	CATB_BOVIN	16.70	32.29	1.933
ZN01971.1	XRCC6_CHICK	12.40	24.87	2.006
ZN12171.1	-	128.06	357.76	2.794
ZN13210.1	NAS6_CAEEL	7.63	21.97	2.882
ZN16459.1	NAS4_CAEEL	22.39	76.27	3.406
ZN11805.1	NAS14_CAEEL	52.60	234.16	4.452
ZN03630.1	BCS1_DANRE	26.43	119.73	4.531
ZN19171.1	XRCC5_MOUSE	8.56	78.96	9.229

Table 1. 31 domains were significantly enriched in the genes in group 6.

[A] 87 domain were detected from the 230 isoforms in group 6, and an enrichment analysis showed that 31 of the domains were significantly enriched. DEG : # of domains in the group. Genome : # of domain in the whole genome. pval : p-value determined from Fisher's exact test. qval : FDR corrected p-value using the p.adjust() function in R. **[B]** Genes associated with enriched domains. The expression of each genes of the control and h04 are shown with the fold change. ZN12171.1 was not annotated by SwissProt search.

3.2. Gene expression profiles of known pathways

Genes in the non-homologous end joining pathway (NHEJ:ko03450) and homologous recombination repair pathway (HR:ko03440) were obtained from the KAAS annotation. The genes in NHEJ were up regulated shortly after irradiation, where HR genes were done so after 9~12 hours (Fig. 4). Most of the genes in each pathway are conserved in *R. varieornatus*, although no homologs of XRCC4 and XLF in the NHEJ could be found (Fig. S4-2). Furthermore, several genes that play important roles in the DNA damage response (DDR), such as p53 and ATM, Nbs1 in the MRN complex were found to be deficit. On the other hand, we found several genes to be multi-paralogous, i.e. 4 copies of MRE11, 3 copies of DNA ligase 4, 2 copies of DNA polymerase Lamda, and 2 copies of Ku80 (which one was not annotated with Ku80 with SwissProt search). Out of the many copies of previously reported protection related proteins CAHS and SAHS, only one CASH2 (ZN10625.1) was differentially expressed between Ctrl and h09.

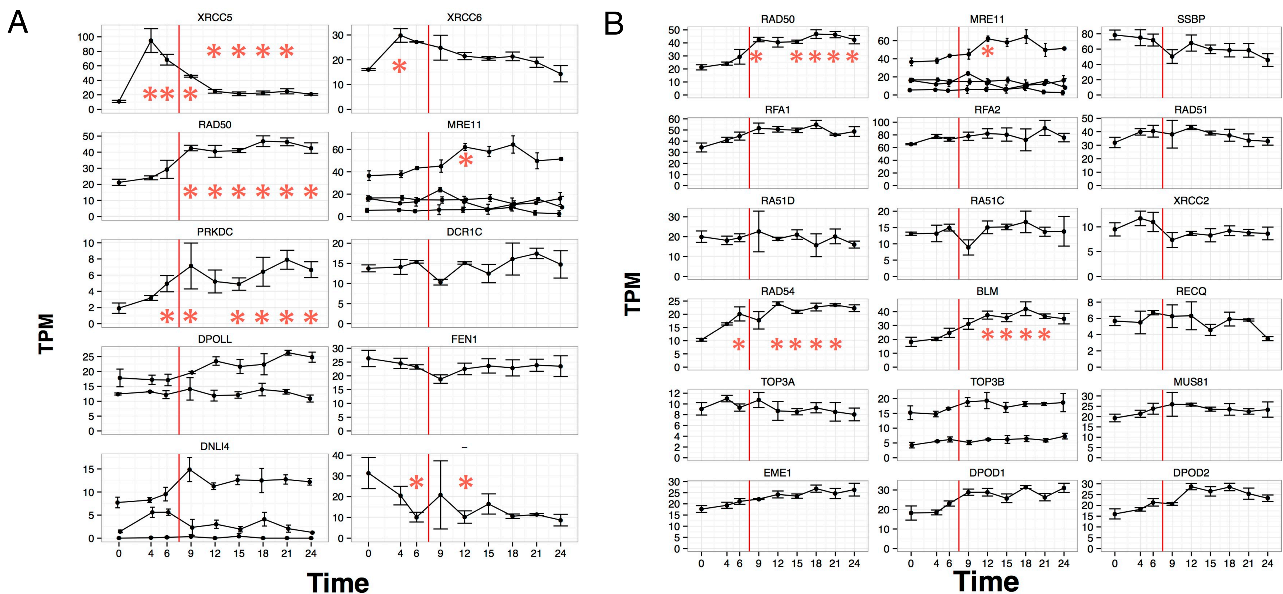

Figure 4. Expression of DNA repair related pathways.
Expression profiles of NHEJ **[A]** and HR **[B]**. DEGs between Control and the particular time point are shown as a red asterisk.

3.3. Novel gene family Clust_14

Screening of DEG rich regions revealed 9 regions with over DEG ratio of 30%. In these regions, we have found 33 novel gene families constructed in chapter 3. Out of the 33 gene families were found, and 19 families had over 5 copies, and only 5 families had DEG ratio of 20%. Further analysis revealed that only 3 families had homologs in the tardigrade *H. dujardini* (Table 2). Similarity search against Uniref90 determined Clust_49 as the CAHS gene family, and functional domain search against Pfam-A found astacin domain in the genes in the Clust_49. No annotation could be made for Clust_14 genes, therefore we have determined Clust_14 as a novel tardigrade specific gene family.

Multiple alignment and phylogenetic tree computation suggested that this group has three subgroups (Fig 5A). Each subgroup was mostly located in the same Supercontig, and several of the genes were neighboring genes. We observed highly conserved motifs in each group (Fig. 5B).

Clust ID	# Paralogs in DRR	# Paralogs in genome	Ratio	Orthologs in *H. dujardini*	Annotation
Clust_49	6	17	0.35	9	CAHS
Clust_14	8	25	0.32	36	-
Clust_9	6	29	0.21	67	Astacin domain

Table 2. Screening of DEG rich regions within *R. varieornatus* genome.
Clusters found in DEG rich regions. Gene families with DEG ratio of over 0.2 and 5 paralogs in the genome.

Gene expression clustering of the genes in Clust_14 revealed that half of the genes were slightly up regulated within 3 hours of irradiation (Fig. 6A), although only 2 genes were significantly up regulated (Fig. 6B). Additionally, nearly half of the genes had an expression of over 100.

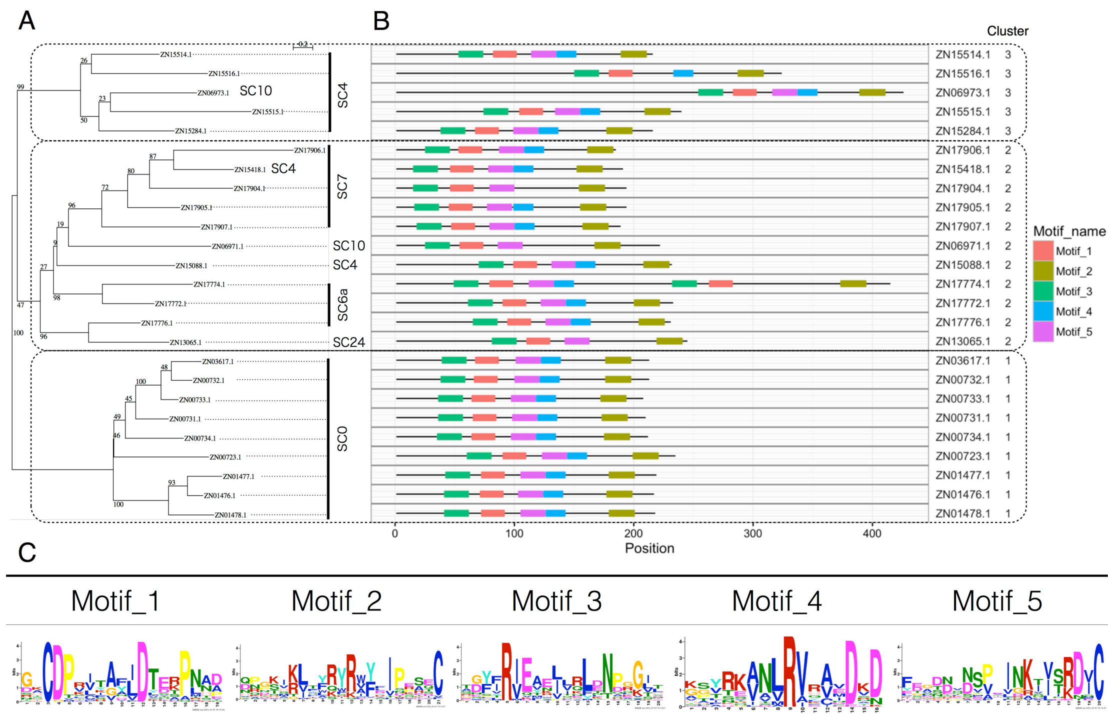

Figure 5. Sequence analysis of Clust_14 genes.
Amino acid sequences were highly conserved in each of the three subgroups of Clust_14. **[A] Dendrogram made with phyML with 1000 bootstraps, based on the** multiple alignment of transcripts in clust_14,The name of the contig the gene is located is written beside the name of the gene. Numbers shown on the dendrogram are bootstrap values. **[B]** Motif structure of each gene. **[C] Sequence logo of motifs found by MEME.** 5 motifs were found with MEME are shown for each group.

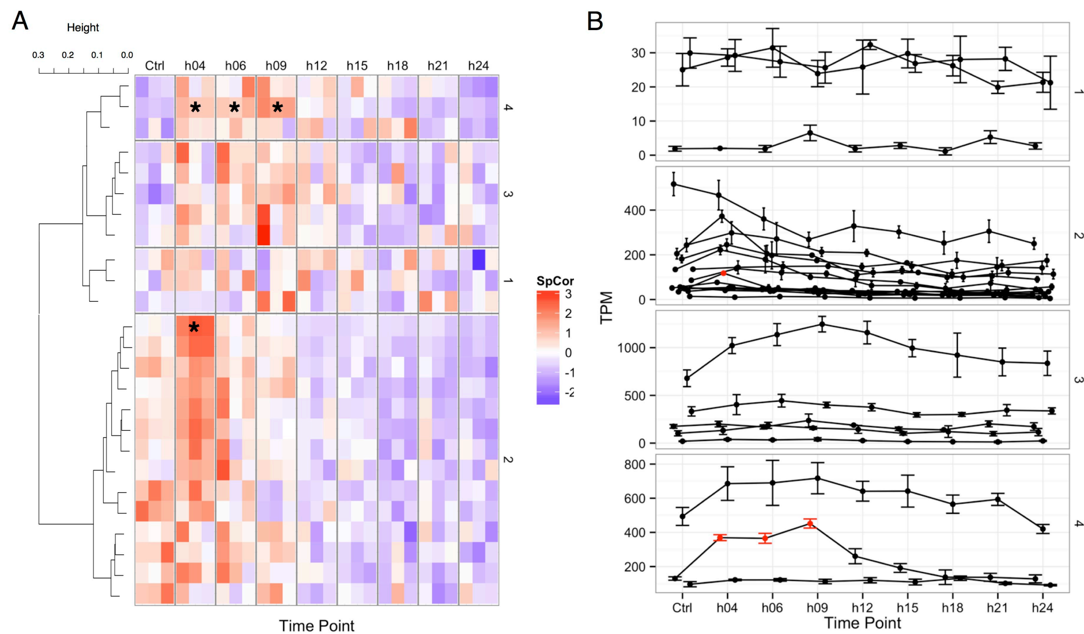

Figure 6. Expression profile of Clust_14.
[A] Clustering of gene expression profiles of Clust_14 genes based on Spearman correlation. Asterisk shows DEGs. **[B]** Expression of each genes. S.D. is shown as error bars, and isoforms that were differentially expressed compared with Control are shown in red.

4. Discussion

This research is one of the first reports of a time-course RNA-Seq with multiple replicates. The RNA-Seq methods used in previous tardigrade RNA-Seq studies required few thousands of animals per sample. In our study, we have a total of 27 samples in this study, which would sum up to nearly 60 to 80 thousand animals. Using this number of *R. varieornatus* for a single experiment is nearly impossible, therefore we have chosen a low input RNA-Seq method using NEBNext Ultra RNA prep, which uses 10 ng of total RNA to create a sequencing library. Using improved methods for RNA-Seq, we were able to gain high quality sequence data. Using Kallisto for TPM quantification, we confirmed a high correlation (0.9 spearman correlation) in between replicates. However, we have observed bias between the samples with different PCR cycle counts in the same condition. This may be due to PCR bias, therefore in further RNA-Seq experiments, it would be necessary to unify the cycle count for all samples. None of the housekeeping genes (β-tublin, EF1-α, GAPDH) were differentially expressed, although all three of the genes were slightly up regulated until 12 hours, and simultaneously down regulated between 12 and 15 hours, although housekeeping genes are believed to have a constant expression. The DSB repair pathway homologous recombination, which is coupled with DNA replication, was up regulated in a similar time frame. These observations may imply cell division occurring at 12 to 15 hours. Furthermore, the transcriptome profile of h24 is not similar to the control profile, suggesting that the transcriptome requires more time to return to its original state.

Functional enrichment analysis of early response genes showed up regulation of both Ku70 and Ku80 immediately after irradiation. Homologous recombination repair is also known as a pathway of double strand breaks (DSB) repair, but this pathway is up regulated after 9 hours, which would be classified as late responding genes. Repair of DSB in DNA by NHEJ, which is initiated by the localization of Ku70/80 homo-dimer to DSB ends is widely known throughout eukaryotes [34], and with the fact that homologous recombination occurs simultaneously with DNA replication, these findings suggest that NHEJ may be the main repair mechanism against double strand breaks in *R. varieornatus*. On the other hand, XRCC4 and XLF, proteins that assists the localization of DNA ligase 4 to DSB ends at the end of the NHEJ pathway, were not found in *R. varieornatus* genome. XRCC4 deficiency would lead to failure of loading of DNA ligase 4 to DSB ends [35], therefore would leave nicks in the double strand DNA. Furthermore, ZN00647.1 was determined as a XRCC6 like protein, but not

so in BLAST searches. This gene was similar to an uncharacterized protein conserved in arthropods, and nuclear localization signal could not be predicted by NLStradamus [36], whereas a secretory signal predicted by TargetP [26].

We also determined that several genes that play important roles in DSB signaling, mainly p53 and ATM, are deficit in *R. varieornatus*. These genes relay DSB signals sent from MRN complexes and DNA-PKcs to downstream pathways, such as cell cycle regulation, apoptosis. Arakawa has shown that the apoptosis pathway in *R. varieornatus* has several deficiencies, as well as several pathways that synthesize ROS (Arakawa, Personal communication). These findings suggests the survival adaption against severe stress, such as desiccation and in this case, high dosage of gamma ray irradiation.

By our DEG rich region screening, we have observed 9 regions with over 30% DEG ratio. However, we have not conducted a statistical analysis of this DEG rich region screening, thus this 30% is an arbitrary threshold, but has been able to reveal several gene families. We found that a novel gene family Clust_14 as an early response gene family with 25 paralogs. Functional domains registered in public databases were not found in these proteins, thus these functions can not be predicted, but we have observed highly conserved regions throughout the cluster.

Genomic analysis together with RNA-Seq analysis revealed the gene family Clust_14. The expression of genes in Clust_14 was up regulated immediately after irradiation, which implies that these genes are related to cellular repair. RNA-Seq analysis of before and after anhydrobiosis in *H. dujardini* showed the same expression profile of Clust_14 homologs (Fig. S3-5), which supports the hypothesis. Further analysis by TargetP predicted a secretory signal in Clust_14 genes, which suggests that these genes may repair extracellular molecules.

This is one of the first reports of a multi-replicate time-course RNA-Seq analysis of tardigrades using gamma ray as a stress inducer, and these results would shed new light into the response to excessive stress of extremophiles.

Acknowledgements

This research could not have been done without the help of Dr. Kazuharu Arakawa and Dr. Daiki D. Horikawa. Dr. Horikawa has given me the chance of using gamma ray irradiation, which I had been interested in for a long time. Dr. Arakawa has supported me in not only the ongoing research, but has taught me the very fundamentals of conducting my own research. This tardigrade project if funded by the Grant-in-Aid for Scientific Research on Priority Areas ‘Genome’ of the Japanese Ministry of Education, Culture, Sports, Science and Technology. This work was funded by the ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan) and in part by research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan. Finally, I personally thank Professor Masaru Tomita, Department of Environment and Information Studies, Keio University for giving me this splendid chance to have my own research.

Reference

[1] Sheng HP, Huggins RA. A review of body composition studies with emphasis on total body water and fat. Am J Clin Nutr. 1979; 32(3): 630-647.

[2] Tunnacliffe A, Lapinski J. Resurrecting van leeuwenhoek's rotifers: A reappraisal of the role of disaccharides in anhydrobiosis. Philos Trans R Soc Lond B Biol Sci. 2003; 358(1438): 1755-1771.

[3] Mobjerg N, Halberg KA, Jorgensen A, Persson D, Bjorn M, Ramlov H, *et al.* Survival in extreme environments - on the current knowledge of adaptations in tardigrades. Acta Physiol (Oxf). 2011; 202(3): 409-420.

[4] Horikawa DD, Cumbers J, Sakakibara I, Rogoff D, Leuko S, Harnoto R, *et al.* Analysis of DNA repair and protection in the tardigrade ramazzottius varieornatus and hypsibius dujardini after exposure to uvc radiation. PLoS One. 2013; 8(6): e64793.

[5] Beltran-Pardo E, Jonsson KI, Harms-Ringdahl M, Haghdoost S, Wojcik A. Tolerance to gamma radiation in the tardigrade hypsibius dujardini from embryo to adult correlate inversely with cellular proliferation. PLoS One. 2015; 10(7): e0133658.

[6] Hashimoto T, Saito Y, Oyama M, Kozuka HH, Enomoto A, Miyagawa K, *et al.* Suppression of DNA damage by a novel protein dsup from a radiation resistant tardigrade. 13th International Symopsium on Tardigrada; Jun 23-26; Modena, Italy2015.

[7] Horikawa D, D. Study on the repair mechanism of ionizing radiation-induced DNA damage in tardigrades. JAEA-Review. 2007; 049: 4.

[8] Horikawa DD, Kunieda T, Abe W, Watanabe M, Nakahara Y, Yukuhiro F, *et al.* Establishment of a rearing system of the extremotolerant tardigrade *ramazzottius varieornatus*: A new model animal for astrobiology. Astrobiology. 2008; 8(3): 549-556.

[9] Arakawa K, Mori K, Ikeda K, Matsuzaki T, Kobayashi Y, Tomita M. G-language genome analysis environment: A workbench for nucleotide sequence data mining. Bioinformatics. 2003; 19(2): 305-306.

[10] Arakawa K, Tomita M. G-language system as a platform for large-scale analysis of high-throughput omics data. J Pestic Sci. 2006; 31(3): 7.

[11] Andrews S. Fastqc a quality-control tool for high-throughput sequence data. 2015 [cited 2015 May 21]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[12] Katayama T, Kunieda T. Genome sequencing and assembly of anhydrobiotic extremotolerant tardigrade. J Jpn Soc Extremophiles. 2010; 9(2): 7.

[13] Bray N, Pimentel. H, P. M, Pachter L. Near-optimal rna-seq quantification. ArXiv e-prints. 2015; 2015(Jun): 21.

[14] Robinson MD, McCarthy DJ, Smyth GK. Edger: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1): 139-140.

[15] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10): R106.

[16] R CT. R: A language and environment for statistical computing. R Foundation for Statistical Computing R Core Team 2014. Available from: http://www.R-project.org. .

[17] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, *et al.* Pfam: The protein families database. Nucleic Acids Res. 2014; 42(Database issue): D222-230.

[18] Eddy SR. Accelerated profile hmm searches. PLoS Comput Biol. 2011; 7(10): e1002195.

[19] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3): 403-410.

[20] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al.* Gapped blast and psi-blast: A new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17): 3389-3402.

[21] UniProt C. Uniprot: A hub for protein information. Nucleic Acids Res. 2015; 43(Database issue): D204-212.

[22] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. Kaas: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007; 35(Web Server issue): W182-185.

[23] Nakaya A, Katayama T, Itoh M, Hiranuka K, Kawashima S, Moriya Y, *et al.* Kegg oc: A large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res. 2013; 41(Database issue): D353-357.

[24] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988; 85(8): 2444-2448.

[25] Dongen SV. Graph clustering by flow simulation. In: Utrecht Uo, editor. PhD thesis 2000.

[26] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. J Mol Biol. 2000; 300(4): 1005-1016.

[27] Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. Nlstradamus: A simple hidden markov model for nuclear localization signal prediction. BMC Bioinformatics. 2009; 10: 202.

[28] Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, *et al.* A new bioinformatics analysis tools framework at embl-ebi. Nucleic Acids Res. 2010; 38(Web Server issue): W695-699.

[29] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011; 7: 539.

[30] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. Syst Biol. 2010; 59(3): 307-321.

[31] Gouy M, Guindon S, Gascuel O. Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010; 27(2): 221-224.

[32] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, *et al.* Meme suite: Tools for motif discovery and searching. Nucleic Acids Res. 2009; 37(Web Server issue): W202-208.

[33] Ho LH, Giraud E, Uggalla V, Lister R, Clifton R, Glen A, *et al.* Identification of regulatory pathways controlling gene expression of stress-responsive mitochondrial proteins in arabidopsis. Plant Physiol. 2008; 147(4): 1858-1873.

[34] Koike M. Dimerization, translocation and localization of ku70 and ku80 proteins. J Radiat Res. 2002; 43(3): 223-236.

[35] Andres SN, Vergnes A, Ristic D, Wyman C, Modesti M, Junop M. A human xrcc4-xlf complex bridges DNA. Nucleic Acids Res. 2012; 40(4): 1868-1878.

[36] Boulikas T. Putative nuclear localization signals (nls) in protein transcription factors. J Cell Biochem. 1994; 55(1): 32-58.