博士論文

Investigation of genome functions and their evolution using high-throughput sequencing

（超並列シーケンシングを用いたゲノム機能とその進化に関する研究）

尾崎　遼

Investigation of genome functions and their evolution using high-throughput sequencing

（超並列シーケンシングを用いたゲノム機能とその進化に関する研究）

by

Haruka Ozaki

尾崎　遼

A Ph. D. Thesis

博士論文

Submitted to
Graduate School of Frontier Sciences, The University of Tokyo
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

March, 2015

Acknowledgments

My research would not have been possible without help of people around me. I am using this opportunity to express my gratitude to everyone who supported me throughout the Ph.D. course.

Firstly, I would like to show my respect and gratitude to my supervisor, Dr. Toshihisa Takagi. When I planned to change my major from biochemistry and molecular biology to bioinformatics, he kindly welcomed me to his lab. Moreover, he gave me an opportunity to select research themes independently. I appreciate his beneficial advice on career design from a broader perspective.

I am sincerely grateful to Dr. Wataru Iwasaki as my advisor for his help both officially and personally. He gave me constructive comments on my research and many opportunities to present my research in academic meetings. He also offered advice on career design and personal matters. My research would not have done without his continuous encouragement.

I thank all co-authors of our paper published in *Molecular and Cellular Biology*. In particular, Dr. Yoshitaka Fukada, Dr. Hikari Yoshitane, and Mr. Hideki Terajima kindly provided the data and made continuous discussion on the analyses. Their comments helped me a lot to conduct research in depth.

I would like to pay a compliment to Dr. Shotaro Hirase, a co-author of our paper published in *BMC Genomics*. He discussed interpretation of results with me. His broad and profound knowledge on ecology and evolutionary biology helped me to deepen insights into biology of sticklebacks and parallel selection.

I would also like to thank Dr. Yutaka Suzuki, Ms. Kiyomi Imamura, Ms. Terumi Horiuchi, and Ms. Makiko Tosaka for helpful information on experimental design and preprocessing of sequencing data.

I would like to show my respect and gratitude to members in the Iwasaki laboratory, especially Mr. Tsukasa Fukunaga and Dr. Motomu Matsui. They encouraged me in conducting research and gave me constructive comments on writing my thesis.

I also express my warm thanks to the former and current members in the Takagi laboratory. Ms. Naoko Tomioka helped me a lot with processing of documents and accounts, and she taught me common sense and social skills. I have to express my gratitude to Dr. Thanet Praneenararat for his advice on programming and English writing.

My special thanks go to members of the Department of Computational Biology. In particular, Mr. Hirotaka Matsumoto and Ms. Risa Kawaguchi provided grateful suggestions on my research.

I am full of praise of the super computer resources provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo and National Institute of Genetics, Research Organization of Information and Systems.

Last but not least, I appreciate my parent for finical support.

Abstract

To date, more than 30,000 genomes have been sequenced since the publication of the first free-living organism genome. However, understanding of how genomes exert their functions is lagging behind. This is due to laboriousness of measuring dynamics of entities that link genomes and phenotypes (e.g., genes and proteins). Such information is required to investigate mechanisms of genome functions, and also should provide clues for understanding how genome functions have evolved. Recent advent of high-throughput sequencing enables us to obtain genome-wide information on not only DNA sequences but also gene expressions and DNA-binding protein occupancies more easily than ever before. Thus, by taking advantage of the high-throughput sequencing data, now we can investigate genome functions and their evolution.

In this thesis, I describe research on genome functions and their evolution using of high-throughput sequencing data in two topics.

First, I investigated contribution of CLOCK, a fundamental transcription factor in the mammalian circadian oscillator, to circadian rhythms in gene expression levels in a genome-wide manner. Circadian rhythms are oscillation with a period of approximately 24 hours in biochemistry, physiology, and behavior of organisms. The circadian rhythms are generated by circadian clock. Mammalian circadian clock have been well characterized: many key genes are identified and systems-biological approaches have been applied. In that, transcriptional and translational feedback loops (TTFLs) with dozens of genes are well modeled. Meanwhile, early microarray experiments found that many genes out of the TTLFs showed circadian expressions in cells and tissues, which raises a question: how these many genes are regulated on a genome-wide manner? To address this question, I analyzed high-throughput sequencing data including CLOCK ChIP-Seq, mRNA-Seq, and small RNA-Seq data from mouse liver. The new method to enumerate DNA-binding motifs from ChIP-Seq data was developed. Application of the method to the ChIP-Seq data revealed comprehensive set of CLOCK-binding motifs. In addition, I found that contribution of CLOCK to the transcriptome-wide circadian gene expressions as a direct transactivator was smaller than expected. Several plausible mechanisms of CLOCK indirectly regulating rhythmically expressed genes expressions were discussed.

Second, I show the possibility of positive selection on gene copy number variations by taking advantage of the system of parallel evolution of three-spined sticklebacks. Positive selection is an evolutionary process by which an allele increases its frequency in a population. Copy number variations (CNVs) constitute a significant proportion of genomic diversity. In particular, gene copy

number variations (GCNVs), which change the numbers of gene loci in genomes, can significantly alter gene functions and dosages, and thus can undergo positive selection. However, positive selection of CNVs has not been proved. One way of assessing the possibility is to search for increase or decrease of copy numbers in parallel evolution of freshwater groups of three-spined stickleback. Parallel evolution is the adaptive evolution in which the same genotypes or phenotypes are selected in different but related lineages, and provides strong evidence of positive selections. To address the possibility of positive selection on GCNVs using the system of parallel evolution of three-spined stickleback, I analyzed resequencing data of multiple individuals from freshwater and marine populations. A novel approach was devised to detect GCNVs under parallel selection from whole-genome sequencing data with low coverage by comparing two resequencing datasets. Application of the method to the resequencing data of sticklebacks revealed GCNVs that were likely under parallel selection. Many of the identified GCNVs showed increase in gene copy numbers in freshwater individuals, which is consistent with the notion that increase in gene copy number would facilitate adaptive evolution. These results suggest that contribution of GCNVs should be considered in studies on adaptive evolution.

Contents

List of figures

List of tables

Abbreviations

AUC: area under curve

ChIP-Seq: chromatin immunoprecipitation sequencing

CNV: copy number variation

DD: constant dark condition

FPKM: fragments per kilobase of exon per million fragments

GCNV: gene copy number variation

GO: gene ontology

LD: light-12-hour:dark-12-hour condition

MOCCS: Motif centrality analysis of ChIP-Seq

RNA-Seq: RNA sequencing

SNV: single nucleotide variation

Transcriptional and translational feedback loop: TTFLs

Transcription factor: TF

Transcription factor binding site: TFBS

Chapter 1: General introduction

Studying functions and mechanisms in biology

Biology is studies of living organisms and biological entities including the biological molecules, organelles, cells, tissues, organs. One of major objectives of biology is to reveal functions of biological entities, which are observed from various aspects such as chemistry, physiology and behavior. Studies of functions are usually accompanied by studies of mechanisms by which biological entities exert their functions.

Another major goal of biology is understanding of diversity of living organisms. Organisms show intra- and inter-species differences in their phenotypes in terms of chemistry, morphology, physiology, and behavior. The fundamental step for interpretation of these diversities is understanding evolution of the functions and mechanisms, as the title of Dobzhansky's famous essay, "Nothing in Biology Makes Sense Except in the Light of Evolution" (Dobzhansky, 1973).

In short, investigation of functions of biological entities, the underlying mechanisms and evolution of the functions is curtail in biology.

Understanding of genomes lagging behind

Molecular bases of phenotypes are of general interest, because it should help understanding the functions of biological entities and the underlying mechanisms, and also may provide clues to understand evolution of the functions. Advances in genetics, molecular biology, and genome sciences enabled biologists to investigate relationships between genes and phenotypes.

Genetics has proven inheritance of genetic substance. In addition, geneticists have identified genes underlying or associated with various phenotypes. On the other hand, they generally ignored the molecular mechanisms in which the genes cause the phenotypes. Such attitude is exemplified by what François Jacob wrote about Thomas Morgan's discovery of the role of chromosome in heredity: "Rather than asking questions about physiology and chemistry of genes or speculating about possible theories of heredity, he stuck to the facts, thereby founding a genetics that interpreted Mendelian inheritance in terms of chromosomal theory" (Jacob, 1998). Thus, using genetics, one often knows that genes *somehow* cause phenotypes and changes in genes *somehow* cause changes in phenotypes.

Molecular biology is the branch of biology that searches for molecular basis underlying biological processes. One of the major discoveries in molecular biology is the central dogma, in which genetic information is transferred from DNA to RNA through transcription and from RNA to proteins through translation (Crick, 1970). The central dogma as well as the subsequent advances in biotechnology provided the way to uncover causative links between genes and phenotypes. Using molecular biology, biologists have identified genes involved in a particular biological process and revealed how the genes or gene products exert their functions.

Since the genome of *Haemophilus influenzae* was completely sequenced in 1995 (Fleischmann *et al*., 1995), more than 30,000 organisms have had its genomes sequenced (the Genomes OnLine Database; https://gold.jgi-psf.org). As genome contains all genetic information including coding and non-coding sequences, understanding of genome is the fundamental step toward understanding of life especially in the post-genomic era. Genomics has raised two problems. One is the difficulty in predicting controls, functions, and interactions of genes and their products from genome sequences themselves. Such information has to be obtained by molecular biological experiments one by one. The other is the increased number of genes and their combination that should be considered in studies of genome functions and their evolution.

Molecular biology has revealed regulatory relationships of genes and interactions among gene products including proteins separately. However, measuring dynamics of entities that link genomes and phenotypes (*e.g*., genes and proteins) is laborious. Consequently, understanding of how genomes exert their functions is lagging behind for diverse research topics. Similarly, how evolutionary changes in genome caused evolution of genome functions remain often unexplained from the point of view of molecular level. Therefore, what we know is often "genome *somehow* causes phenotypes". Similarity, we know that evolution of genome *somehow* resulted in evolution of phenotypes.

How to investigate genome functions and their evolution

Recent advent of high-throughput sequencing enables us to obtain genome-wide information on not only DNA sequences but also dynamics of biological molecules more easily than ever before. For example, RNA Sequencing (RNA-Seq) can reveal not only expression levels of all genes but also alternative splicing patterns. Another example is chromatin immunoprecipitation sequencing (ChIP-Seq), which can provide genome-wide information on occupancies of DNA-binding proteins. Comprehensive measurement of molecular phenotypes under biological processes of interest should

facilitate understanding how genome exert its functions and how genome functions have evolved. Thus, by taking advantage of the high-throughput sequencing data, now we can investigate genome functions and their evolution.

Outline of this thesis

This thesis describes researches on genome functions and their evolution using of high-throughput sequencing data in two topics. In Chapter 2, I investigated contribution of CLOCK, a fundamental transcription factor in the mammalian circadian oscillator, to circadian rhythms in gene expression levels in a genome-wide manner. In Chapter 3, I show the possibility of positive selection on gene copy number variations by taking advantage of the system of parallel evolution of three-spined sticklebacks.

Chapter 2: CLOCK-controlled regulations of circadian rhythms through canonical and non-canonical E-boxes

Circadian rhythms are oscillation with a period of approximately 24 hours in biochemistry, physiology, and behavior of organisms. The circadian rhythms are generated by circadian clock. Mammalian circadian clock have been well characterized: many key genes are identified and systems-biological approaches have been applied (Ukai and Ueda, 2010). In that, transcriptional and translational feedback loops (TTFLs) with dozens of genes are well modeled (Brown *et al*., 2012). Meanwhile, early microarray experiments found that many genes out of the TTLFs showed circadian expressions in cells and tissues (Miller *et al*., 2007; Hughes *et al*., 2009), which raises a question: how these many genes are regulated on a genome-wide manner? In this chapter, I investigated contribution of CLOCK, a fundamental transcription factor in TTFL, to circadian rhythms in gene expression levels in a genome-wide manner. To this end, I analyzed high-throughput sequencing data including CLOCK ChIP-Seq, mRNA-Seq, and small RNA-Seq data from mouse liver. The new method to enumerate DNA-binding motifs from ChIP-Seq data was developed. Application of the method to the ChIP-Seq data revealed comprehensive set of CLOCK-binding motifs. In addition, I found that contribution of CLOCK to the transcriptome-wide circadian gene expressions as a direct transactivator was smaller than expected. Several plausible mechanisms of CLOCK indirectly regulating rhythmically expressed genes expressions were discussed.

Background

Diverse species from bacteria to human show nearly 24-hour rhythms in their physiology and behavior even in the condition where no external cues or "zeitgebers" (*e.g*., light) are available. These rhythms are called circadian rhythms. The mechanisms underlying circadian rhythms are designated as circadian clocks. A circadian clock is composed of three parts: input pathways, the circadian oscillator, and output pathways (Lowrey and Takahashi, 2004). The input pathways convey signals to synchronize the circadian oscillator depending on the environments. The

---

This chapter is in part published in the following publication:

Hikari Yoshitane*, Haruka Ozaki*, Hideki Terajima* *et al*., CLOCK-controlled polyphonic regulation of circadian rhythms through canonical and noncanonical E-boxes. *Molecular and cellular biology* **34**, 1776–87 (2014).
(* Equal contributions)

circadian oscillator generates circadian rhythms. The output pathways were used to regulate circadian rhythms in biochemistry, physiology, and behavior.

In mammal, the circadian oscillator is the interconnected regulatory network of transcription factors (Ukai and Ueda, 2010). Among these transcription factors, CLOCK and BMAL1 are fundamental for the circadian oscillator to generate circadian rhythms (Lowrey and Takahashi, 2004). CLOCK and BMAL1 form heterodimers and activate expression of their target genes, including *Per* and *Cry*. mRNAs of *Per* and *Cry* are then translated to produce PER and CRY proteins, which in turn suppress the transactivation by the CLOCK-BMAL1 complex. This negative feedback loop makes the CLOCK-BMAL1 complex bind to the target gene loci in a circadian manner. Other transcription factors are also shown or thought to function in the regulation of the circadian oscillator and the output pathway of circadian gene expression (Ukai and Ueda, 2010).

CLOCK binding is thought to be one of the output pathways of circadian gene expression. To date, dozens of rhythmically expressed genes such as *Dbp* have been shown to be directly targeted by the CLOCK-BMAL1 complex via the DNA-binding motif E-box (Ripperger and Schibler, 2006). The canonical E-box (CACGTG) was the sequence most strongly bound by CLOCK and BMAL1 (Gekakis *et al*., 1998). However, several non-canonical E-boxes have been proposed (Yoo *et al*., 2005; Kiyohara *et al*., 2008; Kumaki *et al*., 2008; Ueda *et al*., 2005). These findings raise the question of what is the entire set of CLOCK-binding motifs. The understanding of diversity of E-boxes would provide insight into genome-wide regulation of circadian rhythm.

Previous studies reported that 10-30% of expressed genes show circadian oscillation in the expression levels using microarray (Miller *et al*., 2007; Hughes *et al*., 2009). However, it was unclear to what extent these rhythmically expressed genes are targeted by CLOCK in a genome wide manner. On the other hand, the consequences of the CLOCK binding to gene loci also remain largely unknown. As circadian gene expressions are involved in proper functioning of metabolism, and the disruption of them results in metabolic disorders and diseases (Takahashi *et al*., 2008; Bass and Takahashi, 2010), detail knowledge on the mechanisms and consequences of CLOCK binding is of great importance for understanding regulation of metabolism.

In this study, I investigated the relationship of CLOCK-binding and rhythmic gene expression using CLOCK ChIP-Seq, mRNA-Seq, and small RNA-Seq data from mouse liver. I first identified nearly 8,000 CLOCK-binding sites on the mouse genome. Then, I developed MOCCS, a bioinformatic method to enumerate all binding motifs of DNA-binding proteins from ChIP-Seq data. By applying the method to the sequences around the CLOCK-binding sites, I found novel

CLOCK-binding motifs and revealed nucleotide selectivity within known and novel CLOCK-binding motifs. In addition, I detected rhythmically expressed genes using mRNA-Seq and small RNA-Seq data, and found that more than 70% of these rhythmically expressed genes were not directly targeted by CLOCK. Last, I discussed indirect regulations of the rhythmic genes which are not directly targeted by CLOCK.

Materials and Methods

High-throughput sequencing data

All Illumina sequencing data used in this study were obtained at the Fukada lab (Graduate School of Science, The University of Tokyo) and the Suzuki lab (Graduate School of Frontier Sciences, The University of Tokyo). The sequencing data were generated from the mouse liver sample under the light-12-hour:dark-12-hour (LD) condition or the constant dark (DD) condition. Note that time is represented by zeitgeber time (ZT) and circadian time (CT) under LD and DD conditions, respectively: ZT0 corresponding to the lights-on time and CT0 corresponding to the lights-on time in the previous LD cycle.

The CLOCK ChIP-Seq data (DRP001092) were derived from mouse liver sampled at ZT8, ZT22, CT2, CT5, CT8, CT11, CT14, CT17, CT20 and CT22. The mRNA-Seq (DRP001093) and small RNA-Seq (DRP001094) data were derived from mouse liver sampled at CT2, CT5, CT8, CT11, CT14, CT17, CT20 and CT22. The RNA-Seq data of *Bmal1*-KO mice and WT littermates (DRP001349) were derived from mouse liver sampled at CT2, CT8, CT14, and CT20.

The CLOCK ChIP-Seq (including ChIP-Seq using 1D4 control IgG and input DNA), mRNA-Seq and small RNA-Seq were generated by Illumina GA IIx sequencer (36 bp, single end). The RNA-Seq data of *Bmal1*-KO mice and WT littermates were generated by Illumina HiSeq 2000 (125 bp for *Bmal1*-KO mice and 101 bp for WT littermates, paired end).

Genome sequence and gene annotation

The mouse genome sequence was obtained from UCSC Genome Browser (mm9) (http://genome.ucsc.edu/). The annotated gene models (NCBIM37) and the annotations of snRNAs, snoRNAs, and rRNAs were taken from Ensembl (release 64) (http://www.ensembl.org/). The annotations of tRNAs were retrieved from GTRNADB (http://gtrnadb.ucsc.edu/) (Chan and Lowe, 2009). The mouse precursor and mature miRNA sequences were downloaded from miRBase (release 18) (http://www.mirbase.org/) (Kozomara and Griffiths-Jones, 2011).

Motif centrality analysis of ChIP-Seq

Motif centrality analysis of ChIP-Seq (MOCCS) aims at enumerating all significant DNA binding motifs of DNA-binding proteins using ChIP-Seq data. The inputs of MOCCS are (1) the locations of binding sites of DNA-binding proteins (TFBSs) inferred from peak calling of ChIP-Seq data and (2) length of motifs in search (k). First, DNA sequences within the region of $\pm d$ bp centered at all

TFBSs were searched for occurrences of each *k*-mer to compute frequency distribution $f(x)$ of the *k*-mer around TFBSs, where x ($-d \le x \le d$) is the position around TFBSs and is zero if the position is on the TFBSs. Then, cumulative relative frequency distribution $F(x)$ of each *k*-mer is calculated as follows:

$$F(x) = \sum_{-x \le i \le x} f(i) / \sum_{-d \le j \le d} f(j) \ (-d \le x \le d). \quad (1)$$

When a *k*-mer overrepresents around TFBSs, the frequency distribution of the *k*-mer takes peak-like shapes. To quantify such shape, area under curve (AUC) is calculated for the cumulative relative frequency distribution $F(x)$ of each *k*-mer as follows:

$$AUC = \sum_{0 \le x \le d} \{F(x) - x/d\}. \quad (2)$$

Finally, *k*-mers with AUC higher than a threshold are selected as significant motifs. Note that AUC for *k*-mer pair in that the suffix of one *k*-mer of length *l* and the prefix of the other *k*-mer of length *l* are identical would take similar values, and such pairs should be merged for interpretation. Therefore, for such pair, only the *k*-mer with higher AUC value than that of the other is reported.

For the analysis of CLOCK ChIP-Seq data, I set $k = 6$ because primary interest of this study is how mach mismatches to E-box are allowed for CLOCK to binding to the motifs. In addition, I focused on *k*-mers that did not overlap more overrepresented motifs and had at most two mismatches to the canonical E-box motif CACGTG. The sequences within the region of ±250 bp centered at all CLOCK-binding sites at CT8 or ZT8 were used. The AUC for each *k*-mer were normalized by dividing the value by the standard deviation of the AUC of all 6-mers within sequences 501 bp upstream of the transcription start sites of all protein-coding genes. I set the threshold for a normalized AUC of >5.

Motif discovery

DNA sequences in the window of ±80 bp and ±5 kbp around each CLOCK-binding site were used for motif search by MEME (Bailey and Elkan, 1994) and POSMO (Ma *et al*., 2012), respectively.

ChIP-Seq data analysis

The sequenced reads were mapped to the mouse genome by using Burrows-Wheeler Aligner version 0.5.9 (Li and Durbin, 2009) with default parameters. Genome Positioning System (GPS) version 1.0 (Guo *et al*., 2010) was used for peak calling with the options "–s 2100000000 –nrf –q

2" (q-value < 0.01). GPS accurately predicts protein-binding sites from ChIP-Seq data at single-base resolution by using the expectation-maximization algorithm. For each predicted binding site, GPS uses control data to calculate the p-value, which is adjusted for multiple testing by a Benjamini-Hochberg correction. In this study, ChIP-Seq data with 1D4 control IgG were used as the control data. The mapped reads were visualized by using Integrative Genomics Viewer (Thorvaldsdóttir *et al*., 2013). The CLOCK-binding sites were defined by using zeitgeber time 8 (ZT8) or circadian time (CT8) results, because CLOCK binds to DNA most strongly at ZT8 or CT8.

mRNA-Seq data analysis

MapSplice version 1.15.2 (Wang *et al*., 2010) was used for mapping of RNA-Seq data (WT at eight time points), and Cufflinks version 2.0.0 (Trapnell *et al*., 2010) was used for quantifying the expression level of each gene as fragments per kilobase of exon per million fragments (FPKM). A gene was defined as an expressed one if the sum of its FPKM values across the eight time points was more than 5.

A gene was defined as a rhythmically expressed one if its maximal and minimal expression values were significantly different (q-value < 0.1), and its expression profile was fitted with cosine curves (p-value < 0.01). The q-values were calculated by Cuffdiff (Trapnell *et al*., 2010). An in-house R script was applied to estimate periods and phases of gene expression profiles by fitting curves using the equation $A * \cos[2\pi(t + t_0)/T] + B$, where A and B are means and standard deviations of the gene expression profiles, respectively, T is the period and ranges from 23 to 25 h with increments of 0.1 h, t is time, and t_0 is the phase and ranges from 0 to 23.9 h with increments of 0.1 h.

The sequenced reads of RNA-Seq data of *Bmal1*-KO mice (4 time points) were trimmed from the 3′ ends so that their lengths became the same as those of their WT littermates (i.e., 101 nucleotides) by using an in-house Perl script. TopHat2 version 2.0.6 (Kim *et al*., 2013) was used to map RNA-Seq data of *Bmal1*-KO mice and WT littermates with default parameters. Cuffdiff 2 version 2.0.2 (Trapnell *et al*., 2013) was used to detect differentially expressed genes between *Bmal1*-KO mice and WT littermates at each time point (q-value < 0.1).

The differences between two phase distributions of rhythmically expressed genes were test using Mardia-Watson-Wheeler test (p-value < 0.05 after Bonferroni correction) with R package ‘circular’.

Detection of rhythmic alternative splicing events

I focused on sequenced reads that were mapped to splice junctions by MapSplice. The transcripts with and without cassette-type exons were referred to as long forms and short forms, respectively (see **Figure 2.4A**). The sequenced reads mapped to 5-prime and 3-prime ends of the cassette-type exons are referred to as long form reads (5p) and (3p) represent, respectively. Inclusion ratio was calculated by counting the number of sequenced reads that were mapped to splice junctions, 5p or 3p, relative to the total numbers of reads of short form plus 5p (left ratio), or 3p (right ratio), respectively. I used two different criteria for circadian variations of the splicing patterns. The first criterion was to find out the large change in ratios of constitutively active splicing variants: (i) Inclusion ratios were not 0 or 1 at any time point, and (ii) maximal and minimal values differ by more than 0.4 in both left ratio and right ratio. The second criterion was aimed at finding a significant change (even if it is small) in ratio of the temporarily active splicing variants: (i) inclusion ratios were 0 or 1 at a time of at least one time point, (ii) inclusion ratios were in the range of 0.2 to 0.8 at any time point, and (iii) the sum of the total numbers of reads of short forms plus 5p across all the time points exceeded 80. Among the rhythmic alternative exons, those included in the Ensembl gene models were counted.

small RNA-Seq data analysis

Because the lengths of the sequenced reads could be longer than the lengths of mature miRNAs, adaptor sequences were included on the 3′ side of the sequenced reads, and hence, they were trimmed by cutadapt (Martin, 2011). Bowtie version 0.12.7 (Langmead, 2009) was used for mapping, and unmapped reads were discarded. From different pre-miRNA sequences, the same sequences of mature miRNAs can be generated, and 1,141 sequences were unique among 1,157 mature miRNAs stored in miRBase (release 18). The read numbers at various time points were normalized, as the total read number mapped to the mouse genome at each time point was the same as the number at CT2. A miRNA was determined to be “expressed” if the sum of its normalized read numbers across the 8 time points exceeded 100. The cosine fitting analysis was performed to calculate p-values of rhythmicity ($p < 0.05$) and phases.

Candidate target genes of miRNAs were estimated by TargetScan (release 18) (Garcia et al., 2011), and those whose ±3 kb from the transcription start sites contained neither E/E’-box (CACGT[GT]), E-like box (CAC[ATGC]TG), D-box (TTATG[TC]AA), nor RRE ([AT]A[AT][ATGC]T[AG]GGTCA) were analyzed.

Gene ontology analysis

Gene ontology (GO) annotations of the mouse genes were retrieved from Ensembl BioMart. To perform a hypergeometric test to identify overrepresented gene ontology terms, an R package, "GOstats" (Falcon and Gentleman, 2007), was used with a *p*-value cutoff of 0.01 . Genes that were assigned to each of the following GO terms were designated "transcription factors": GO:0000122, GO:0000978, GO:0000981, GO:0000982, GO:0000983, GO:0000988, GO:0001010, GO:0001011, GO:0001071, GO:0001074, GO:0001075, GO:0001077, GO:0001078, GO:0001133, GO:0001190, GO:0001200, GO:0001201, GO:0001205, GO:0001206, GO:0001227, GO:0001228, GO:0003700, GO:0003705, GO:0004879, GO:0006355, GO:0006357, GO:0038050, GO:0038052, GO:0044212, and GO:0045944.

Results

Genome-wide CLOCK-binding sites determined by ChIP-Seq

To identify genes regulated by CLOCK, binding of CLOCK to the cis-regulatory elements is good proxy. To detect CLOCK-binding sites in a genome wide-manner, I analyzed the CLOCK ChIP-Seq data (accession number: DRP001092). The number of reads of ChIP-Seq data were >20 million reads for each time point (**Table 2.1, Table 2.2**). The reads were mapped onto the mouse genome, allowing at most two mismatches. The CLOCK ChIP-Seq data using samples prepared at 2 time points (ZT 8 and 22) under LD conditions identified 5,801 CLOCK-binding sites at ZT8 (**Supplementary Table 2.1**[*]). CLOCK ChIP-Seq data using samples prepared at 8 time points (CT2, 5, 8, 11, 14, 17, 20 and 22) across the day under DD conditions identified 7,978 CLOCK-binding sites at CT8 (**Supplementary Table 2.2**). Among them, 2,400 sites of CLOCK binding were also detected in a previous study (Koike *et al*., 2012). Typically, strong peaks of CLOCK-ChIP reads were detected at the three positions in the *Dbp* locus (**Figure 2.1A**). In addition to these established sites, the CLCOK ChIP-Seq data identified many novel CLOCK-binding sites in the present study **(Figure 2.1B**). Overall, the heat maps showed that almost all the binding sites exhibited a day-night variation as well as a circadian change in terms of CLOCK occupancy (**Figure 2.1C**). These results confirmed that CLOCK rhythmically binds to DNA in a genome-wide manner.

[*] Supplementary tables are available from the following URL:

https://www.dropbox.com/sh/60pa0bqm3ptiajs/AAArN8TlECJWfd9KWbPPGL10a

Table 2.1: Mapping statistics of CLOCK ChIP-Seq data under the LD condition

Sample	Total reads	Unmapped	%Unmapped	Mapped	%Mapped	Unique	%Unique	Multi	%Multi
ZT08 CLOCK	24,814,060	1,680,980	6.8	23,133,080	93.2	18,781,183	75.7	4,351,897	17.5
ZT20 CLOCK	20,701,252	1,359,451	6.6	19,341,801	93.4	15,563,274	75.2	3,778,527	18.3
ZT08 1D4	34,903,171	2,106,676	6.0	32,796,495	94.0	26,173,449	75.0	6,623,046	19.0
ZT20 1D4	38,006,035	2,855,626	7.5	35,150,409	92.5	27,809,370	73.2	7,341,039	19.3
ZT08 input	40,149,640	1,503,935	3.7	38,645,705	96.3	30,734,244	76.5	7,911,461	19.7
ZT20 input	39,451,041	1,468,453	3.7	37,982,588	96.3	30,118,349	76.3	7,864,239	19.9

Table 2.2: Mapping statistics of CLOCK ChIP-Seq data under the DD condition

Sample	Total reads	Unmapped	% Unmapped	Mapped	% Mapped	Unique	% Unique	Multi	% Multi
CT02 CLOCK	28,316,818	1,828,663	6.5	26,488,155	93.5	20,986,937	74.1	5,501,218	19.4
CT05 CLOCK	23,028,393	1,417,093	6.2	21,611,300	93.8	17,074,439	74.1	4,536,861	19.7
CT08 CLOCK	26,802,107	2,871,361	10.7	23,930,746	89.3	18,959,795	70.7	4,970,951	18.5
CT11 CLOCK	27,007,590	2,222,955	8.2	24,784,635	91.8	19,582,742	72.5	5,201,893	19.3
CT14 CLOCK	28,707,621	1,721,628	6.0	26,985,993	94.0	21,474,636	74.8	5,511,357	19.2
CT17 CLOCK	29,896,705	3,523,487	11.8	26,373,218	88.2	20,731,864	69.3	5,641,354	18.9
CT20 CLOCK	31,648,838	2,032,882	6.4	29,615,956	93.6	23,351,120	73.8	6,264,836	19.8
CT23 CLOCK	28,866,990	2,946,493	10.2	25,920,497	89.8	20,368,507	70.6	5,551,990	19.2
CT02 1D4	32,595,544	2,626,448	8.1	29,969,096	91.9	23,463,410	72.0	6,505,686	20.0
CT05 1D4	25,418,765	1,806,255	7.1	23,612,510	92.9	18,726,553	73.7	4,885,957	19.2
CT08 1D4	25,062,236	1,852,925	7.4	23,209,311	92.6	18,169,953	72.5	5,039,358	20.1
CT11 1D4	37,015,953	2,602,564	7.0	34,413,389	93.0	26,598,399	71.9	7,814,990	21.1
CT14 1D4	26,185,744	1,936,001	7.4	24,249,743	92.6	19,230,910	73.4	5,018,833	19.2
CT17 1D4	24,683,625	3,072,692	12.4	21,610,933	87.6	16,976,003	68.8	4,634,930	18.8
CT20 1D4	30,760,715	5,472,906	17.8	25,287,809	82.2	20,295,386	66.0	4,992,423	16.2
CT23 1D4	35,855,288	9,623,901	26.8	26,231,387	73.2	20,652,349	57.6	5,579,038	15.6
CT02 input	38,132,798	1,987,798	5.2	36,145,000	94.8	27,800,164	72.9	8,344,836	21.9
CT05 input	30,638,235	2,016,149	6.6	28,622,086	93.4	22,227,293	72.5	6,394,793	20.9
CT08 input	29,162,920	1,913,317	6.6	27,249,603	93.4	21,199,983	72.7	6,049,620	20.7
CT11 input	25,952,374	1,480,520	5.7	24,471,854	94.3	18,972,182	73.1	5,499,672	21.2
CT14 input	30,571,894	1,706,418	5.6	28,865,476	94.4	22,444,683	73.4	6,420,793	21.0
CT17 input	23,097,076	1,273,377	5.5	21,823,699	94.5	17,003,154	73.6	4,820,545	20.9
CT20 input	25,783,865	1,389,550	5.4	24,394,315	94.6	18,988,165	73.6	5,406,150	21.0
CT23 input	23,219,016	1,348,074	5.8	21,870,942	94.2	17,052,688	73.4	4,818,254	20.8

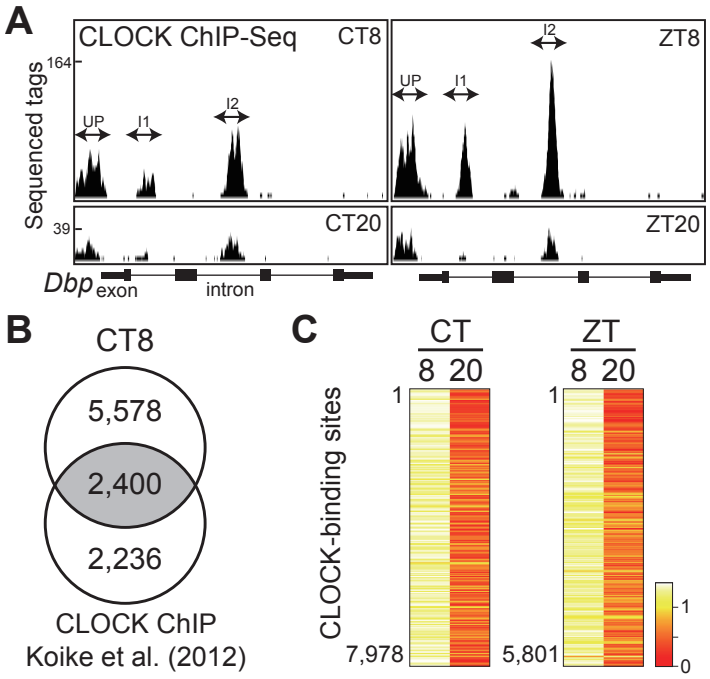

Figure 2.1: Identification of genome-wide CLOCK-binding sites

(A) The CLOCK ChIP-Seq data for the *Dbp* locus showing three rhythmic binding sites under DD (CT8, left) and LD (ZT8, right) conditions. UP: upstream. I1: first intron. I2: second intron. (B) Overlap of CLOCK-binding sites at CT8 in this study with those in a previous study (Koike *et al*., 2012). An overlap was called if CLOCK-binding sites at CT8 were within 120 bp of the peak summits in the previous study. (C) Heat maps of the sequenced reads around the identified CLOCK-binding sites under DD (CT8) and LD (ZT8) conditions. The sites were ordered by the read number for the data sets at CT8 and ZT8. For normalization, the numbers of sequenced reads were divided by the root mean square for each row in the maps.

Motif centrality analysis of ChIP-Seq

The CLOCK-BMAL1 complex binds to the canonical palindromic E-box, CACGTG (Gekakis *et al*., 1998), and to its related motifs called E′-box and E-box-like sequences (Yoo *et al*., 2005; Kiyohara *et al*., 2005; Kumaki *et al*., 2008; Ueda *et al*., 2005). These previous studies focused on selected genes, and therefore, present knowledge on the CLOCK-binding sequences might have underestimated and/or overestimated the binding motifs. High-quality ChIP-Seq data in this study made it possible to quantitatively and comprehensively investigate CLOCK-binding motifs *in vivo*. MEME is a widely used tool for determining DNA-binding motifs (Bailey and Elkan, 1994), and BMAL1-binding motifs were defined by using MEME in previous studies (Hatanaka *et al*., 2010; Rey *et al*., 2011). Here we should consider the feature of MEME, which aims at determining representative sequence motifs (**Figure 2.2E**) rather than explicitly evaluating to what extent each of the related motifs is used in a genome-wide manner.

In order to extract detailed characteristics of the CLOCK-binding motifs, I developed a bioinformatics method termed 'motif centrality analysis of ChIP-Seq' (MOCCS), which enumerates and evaluates all the significant DNA-binding motifs based on ChIP-Seq data sets. This method takes advantage of the fact that significant DNA-binding motifs of transcription factors should frequently appear around their binding sites identified by ChIP-Seq. For example, a histogram of the appearance of CACGTG or CACGCG around the CLOCK-binding sites showed a sharp peak (**Figures 2.2AC**). To quantify the sharpness of the peak, a cumulative relative frequency curve was drawn for every 6-mer (**Figures 2.2BD**), and the area under the curve (AUC) was calculated. It should be noted that when a peak becomes sharper, the AUC becomes larger. The details of MOCCS are described in the Materials and Methods section.

Repertoire of CLOCK-binding motifs

Application of MOCCS to CLOCK ChIP-Seq data revealed motifs significantly concentrated in the CLOCK-binding sites (**Figures 2.2BD, and Table 2.3**). Reasonably, the motif with the largest AUC was the canonical E-box motif CACGTG in both the CT8 and ZT8 data sets, and the second was CACGTT (AACGTG) (nucleotides mismatched with CACGTG are underlined), which is another established CLOCK-binding motif (Yoo *et al*., 2005). The other significant motifs included CACATG (CATGTG) and CACGCG (CGCGTG): the former is known to function in the promoter region of the *Dbp* gene (Kiyohara *et al*., 2008), and the latter is evolutionarily conserved in the *Per1*

locus (Kumaki *et* al., 2008). Overall, the fifth (second) position of the CACGTG sequence was less selective, and all CACGNG (CNCGTG) motifs had large AUCs. More importantly, MOCCS detected two novel potential motifs, CATGCG (CGCATG) and a palindromic motif, TACGTA.

Base on the result of MOCCS analysis, promoter assays were performed at the Fukada lab to evaluate CLOCK-BMAL1-dependent transcription via all the one-mismatch sequences and the two potential motifs with two mismatches (CATGCG and TACGTA). CLOCK and BMAL1 activated transcription was shown for the canonical E-box CACGTG and six non-canonical sequences, CACGTT, CACGCG, CACGGG, CACGAG, CACATG, and CATGCG, all of which were predicted by MOCCS. In addition, all of the one-mismatch sequences that were not predicted to be CLOCK-binding motifs by MOCCS was revealed to show no CLOCK and BMAL1 activated transcription. These observations strongly support the results of MOOCS analysis. For the two-mismatch sequence CATGCG, the Fukada lab confirmed rhythmic CLOCK-binding to CATGCG using the 30-bp genomic sequence around the CLOCK-binding sites containing CATGCG but the other motifs. This observation corroborates the MOCCS results indicating that the two-mismatch sequence CATGCG functions as a CLOCK-binding motif. In contrast, the other two-mismatch potential motif, TACGTA, showed no enhancer activity in the promoter assay. Note that TACGTA appeared far less frequently around the CLOCK-binding sites than the other motifs (**Table 2.3**).

Tandem E-boxes with 6- to 7-bp spacer were observed at the BMAL1 binding sites identified by BMAL1 ChIP-Seq (Rey *et al*., 2011). To assess whether The CLOCK-binding motifs obtained by MOCCS analysis show the similar tendency, I calculated the relative distance between the positions of the seven motifs around the CLOCK-binding sites. The CLOCK-binding motifs were found in tandem frequently with a 6- to 7-bp spacer (**Figure 2.2G**), consistent with Rey *et al*. (2011). In addition, the adjoining sequences around the CLOCK-binding motifs tended to be GC rich (data not shown), which is consistent with the CLOCK-binding region model predicted in a previous study (Kumaki *et al*., 2008). These results further support that the canonical and non-canonical E-boxes revealed by MOCCS analysis would represent the set of *bona fide* CLOCK-binding motifs.

Collectively, MOCCS revealed that the non-canonical E-box motifs CACGNG, CACGTT, and CATG[T/C]G were targeted by CLOCK in a genome-wide manner, rather than being limited exceptions. Furthermore, MOCCS reveled nucleotide selectivity relative to the canonical CACGTG motif on different positions within the E-boxes, as described above. Such detailed information on non-canonical E-boxes and nucleotide selectivity could not be extracted from the results obtained

from the same dataset by other algorithms for searching DNA-binding motifs such as MEME and POSMO (Ma *et al*., 2012) (**Figures 2.2EF**). Therefore, MOCCS is a powerful tool for determining all DNA-binding motifs from ChIP-Seq data sets.

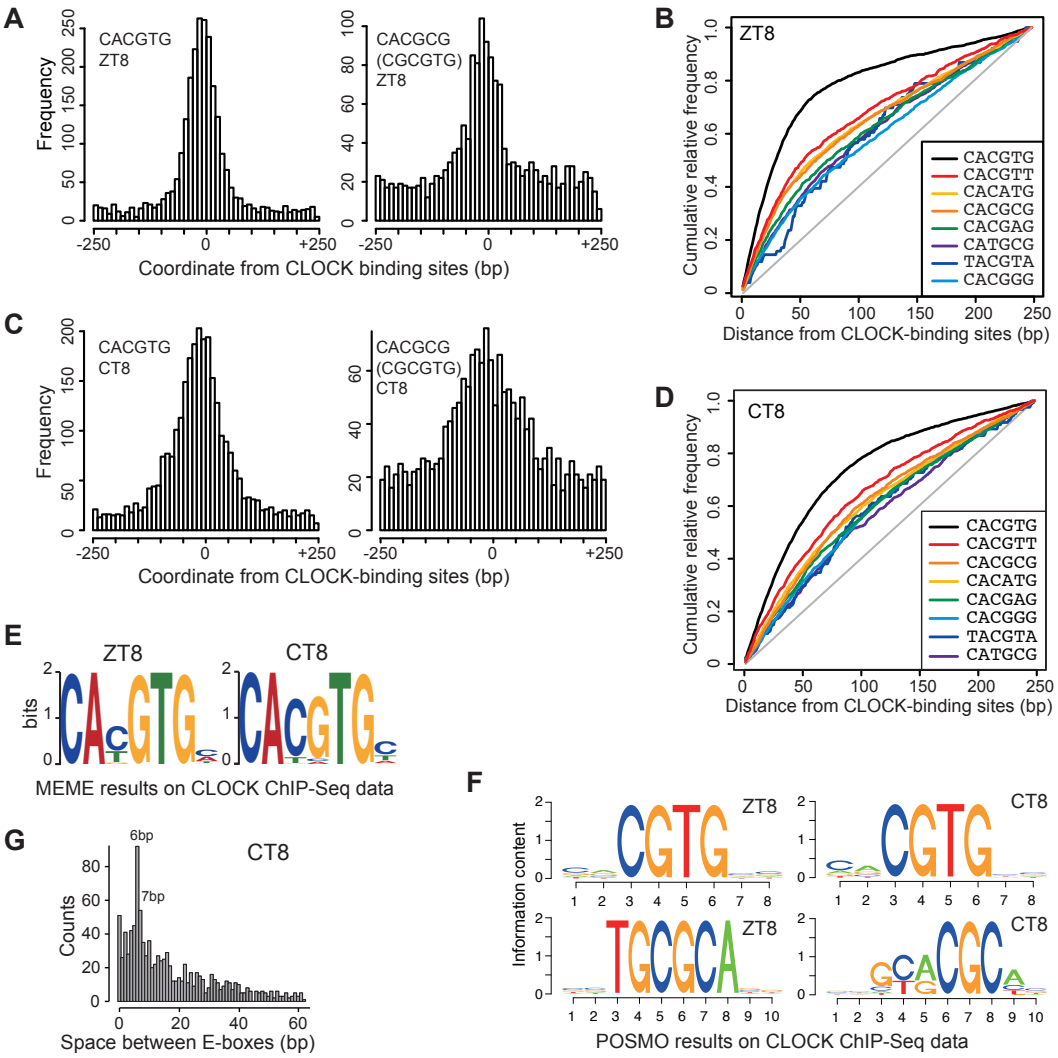

Figure 2.2: Determination of CLOCK-binding motifs by MOCCS

(A and C) Frequency distribution of the indicated sequences around the CLOCK-binding sites observed at ZT8 (A) and CT8 (B). The bin size is set for 10 bp. (B and D) Cumulative relative frequency curves of all significant motifs around the CLOCK-binding sites observed at ZT8 (B) and CT8 (D). The x axis represents absolute values of distance from the CLOCK-binding sites. (E) Overrepresented CLOCK-binding motif determined by MEME. (F) Overrepresented CLOCK-binding motif determined by POSMO. (G) Frequency histogram of the spacer lengths of tandem CLOCK-binding motifs (CACGNG, CACGTT, and CATG[C/T]G) in the window of ±40 bp around each CLOCK-binding site.

Table 2.3: List of CLOCK-binding motifs revealed by MOCCS

Motif	Normalized AUC	Count
CT8		
CACGTG	15.72	2,899
CACGTT (AACGTG)	10.27	1,181
CACGCG (CGCGTG)	7.95	1,730
CACATG (CATGTG)	7.90	3,910
CACGAG (CTCGTG)	6.62	1,233
CACGGG (CCCGTG)	6.29	1,573
TACGTA	5.91	121
CATGCG (CGCATG)	5.15	857
ZT8		
CACGTG	18.82	2,591
CACGTT (AACGTG)	11.64	910
CACATG (CATGTG)	10.10	2,873
CACGCG (CGCGTG)	9.86	1,607
CACGAG (CTCGTG)	8.14	1,034
CATGCG (CGCATG)	7.42	738
TACGTA	6.92	76
CACGGG (CCCGTG)	6.15	1,388

Rhythmic gene expression

The genome-wide rhythmic CLOCK binding is expected to be accompanied by rhythmic gene expression. To detect rhythmically expressed genes, I analyzed the mRNA-Seq (poly(A)-tailed RNA-Seq) data (accession number: DRP001092), which were derived from mouse liver at eight time points (CT2, 5, 8, 11, 14, 17, 20 and 22) under the DD condition. The sequenced reads were mapped onto the mouse genome allowing one mismatch, and this analysis yielded about 20 million mapped reads for each sample (**Table 2.4**). Among 37,314 genes, including noncoding RNAs (Ensembl, release 64), 11,926 genes were found to be expressed ones, while 1,126 genes (9.4% of expressed genes) were rhythmic in mouse liver (**Supplementary Table 2.3**). For example, a well-known rhythmically expressed gene *Dbp* was rhythmically expressed (**Figures 2.3AB**), as confirmed by the Fukada lab using qRT-PCR. Among the 1,126 rhythmic genes identified, >60% of genes were also reported to be rhythmically expressed genes in previous studies **(Figure 2.3C**) (Koike *et al*., 2012; Menet *et al*., 2012). The heat map showed a great diversity in circadian phases of the rhythmic genes (**Figure 2.3D**), indicating cooperative actions of the E-box, D-box, and RRE, as reported previously (Ueda *et al*., 2005; Ukai-Tadenuma *et al*., 2011).

To understand how strongly each gene is regulated by CLOCK, all the genes were given "ChIP scores" based on the CLOCK ChIP-Seq data at CT8 (**Supplementary Table 2.4**). The ChIP score was defined as the total number of sequenced reads that were mapped to all CLOCK-binding sites within ±10 kb from the transcription start site of each gene or in the gene body. A total of 2,234 genes with a ChIP score of >60 were then designated CLOCK targets. Comparative analysis of the ChIP-Seq and RNA-Seq data revealed 324 rhythmic genes among 1,629 CLOCK targets expressed in mouse liver (**Supplementary Table 2.4**).

In order to strengthen the results, I analyzed the RNA-Seq data using livers of *Bmal1*-KO mice and their WT littermates (The accession number: DRP001349) **(Table 2.5**). I detected genes whose expression levels were significantly changed between *Bmal1*-KO and WT mice at least at one time point (CT2, 8, 14, and 20) (**Supplementary Table 2.8**). These differentially expressed genes were 5.7 times more enriched in the rhythmic CLOCK targets than all the expressed genes, which indicates the validity of the rhythmic CLOCK targets defined in this study.

Table 2.4: Mapping statistics of the RNA-Seq data under DD condition

Sample	Total reads	Unmapped	%Unmapped	Mapped	%Mapped	Unique	%Unique	Multi	%Multi
CT02	31,589,335	2,605,777	8.2	28,983,558	91.8	23,697,620	75.0	5,285,938	16.7
CT05	30,247,944	2,555,944	8.4	27,692,000	91.6	23,074,619	76.3	4,617,381	15.3
CT08	27,419,358	3,940,458	14.4	23,478,900	85.6	19,232,009	70.1	4,246,891	15.5
CT11	35,709,409	6,855,427	19.2	28,853,982	80.8	23,810,327	66.7	5,043,655	14.1
CT14	29,527,144	3,618,354	12.3	25,908,790	87.7	21,593,143	73.1	4,315,647	14.6
CT17	31,809,803	7,233,153	22.7	24,576,650	77.3	20,095,382	63.2	4,481,268	14.1
CT20	31,842,351	3,655,704	11.5	28,186,647	88.5	23,387,453	73.4	4,799,194	15.1
CT23	28,558,329	3,633,244	12.7	24,925,085	87.3	20,709,785	72.5	4,215,300	14.8

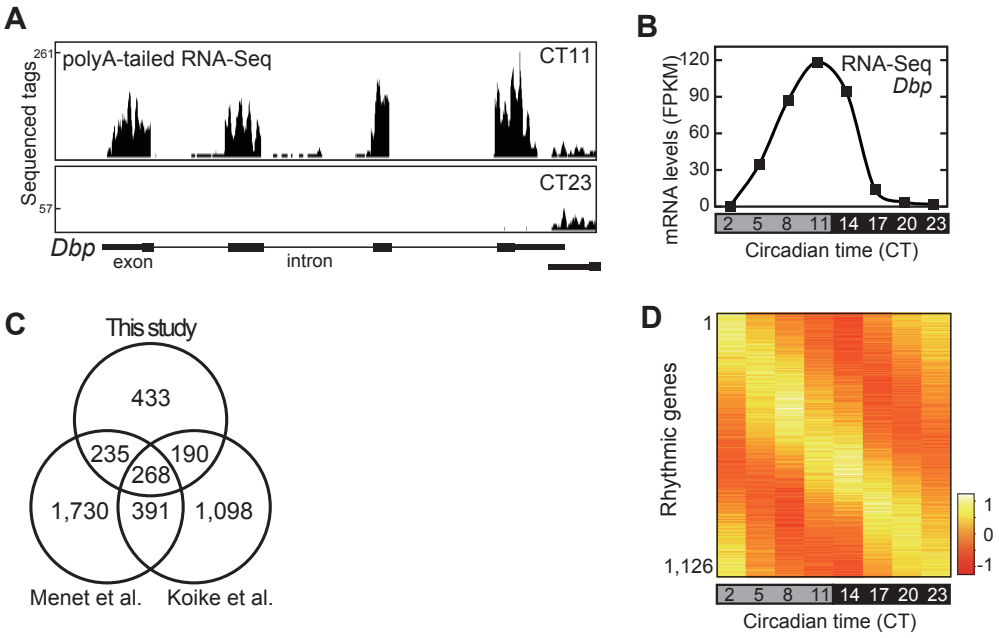

Figure 2.3: Rhythmically expressed genes revealed by mRNA-Seq

(A) RNA-Seq data at the *Dbp* locus showing a robust change in its expression level in a time-of-day-dependent manner. (B) Circadian expression profiles of *Dbp* transcript in mouse liver revealed by RNA-Seq analysis. The mRNA levels are shown as FPKM (fragments per kilobase of exon per million fragments) values. (C) Overlaps of rhythmically expressed genes in this study with those in previous RNA-Seq studies using mouse liver (Menet *et al*., 2012; Koike *et al*., 2012). The genes whose identifiers were correctly mapped to the Ensembl gene identification are compared. (D) Heat map of the expression level of each rhythmic gene in the RNA-Seq analysis. Genes were ordered by the peak phases from early subjective day to late subjective night. The FPKM values were normalized so that the mean and the variance were 0 and 1, respectively, for each row of the maps.

Table 2.5: Mapping statistics of the RNA-Seq data of *Bmal1* KO mice and their littermates under the DD condition

Sample	Total reads	Unmapped	% Unmapped	Mapped	% Mapped	Unique	% Unique	Multi	% Multi
WT CT02	54,816,216	5,450,089	9.9	49,366,127	90.1	42,958,805	78.4	6,407,322	11.7
WT CT08	52,525,606	5,807,780	11.1	46,717,826	88.9	40,732,748	77.5	5,985,078	11.4
WT CT14	69,722,484	7,317,213	10.5	62,405,271	89.5	54,051,215	77.5	8,354,056	12.0
WT CT20	45,099,148	4,914,985	10.9	40,184,163	89.1	34,839,684	77.3	5,344,479	11.9
KO CT02	61,632,346	5,645,028	9.2	55,987,318	90.8	49,605,738	80.5	6,381,580	10.4
KO CT08	66,201,196	6,313,567	9.5	59,887,629	90.5	52,704,061	79.6	7,183,568	10.9
KO CT14	61,737,692	5,297,982	8.6	56,439,710	91.4	50,061,416	81.1	6,378,294	10.3
KO CT20	55,573,622	5,592,694	10.1	49,980,928	89.9	43,942,015	79.1	6,038,913	10.9

Rhythmic alternative splicing

The alternative splicing is a key regulator of gene expression as it generates numerous transcripts from a single protein-coding gene, which largely increases the use of genetic information. Daily fluctuation of splicing of 3′-terminal intron of *Per* was reported (Majercak *et al*., 2004). Unlike conventional microarray, RNA-Seq provides opportunity to search for splicing events. Thus, I used the mRNA-Seq data to search for temporal variation in alternative splicing events.

A cassette type exon (**Figure 2.4A**) is the most frequent form (25-30%) of alternative splicing (Nagasaki *et al*., 2006). Thus, I focused on the circadian variation of cassette-type alternative splicing. Inclusion ratio of alternative exon was calculated by counting the number of sequencing reads that were mapped to exon junction, and 83 exons were identified as cassette-type exons which was rhythmically spliced (**Supplementary Table 2.5**). In particular, clear circadian rhythms were observed in the alternative splicing of *Misshapen-like Kinase 1* (*Mink1*) and *Ubiquilin 1* (*Ubqln1*) genes (**Figure 2.4B**). These rhythmic splicing should cause time-of-day variation of protein functions. MINK1 belongs to the Ste20 kinase family that has been shown to act as MAP4K, and a splicing isoform of MINK1 is known to trigger JNK pathway (Hu *et al*., 2004), which is important for the circadian clockwork (Yoshitane *et al*., 2012). On the other hand, UBQLN1 contains one ubiquitin-like domain and one ubiquitin-associated domain, and functionally associates the ubiquitination machinery with the proteasome. UBQLN1 directly interacts with Presenilin1/2, which are catalytic components in gamma-secretase enzyme complex that generates beta-amyloid (Haapasalo *et al*., 2011). Genetic variants in *Ubqln1* are reported to increase the risk for Alzheimer's disease by increasing short isoform lacking exon 8 in the brain (Bertram *et al*., 2005). In this study, alternative splicing was observed for exon 8 of *Ubqln1* in time-of day dependent manner, raising the possibility that perturbation of circadian clock increases the risk for Alzheimer's disease. These results suggest that RNA splicing would be a fundamental event that is regulated by circadian clock.

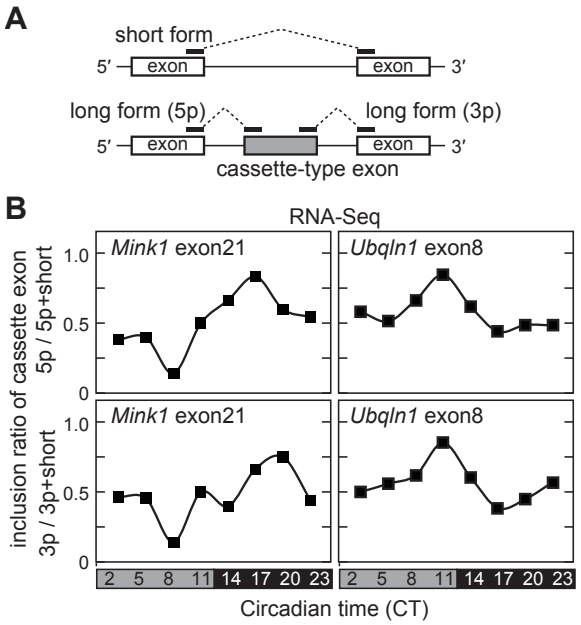

Figure 2.4: Circadian rhythms of alternative splicing events

(A) Schematic representation of cassette-type alternative splicing. Short form read represents those skipping the cassette-type exon. Long form reads (5p) and (3p) represent those mapped to 5′- and 3′- ends of the cassette-type exons, respectively. (B) Circadian variations of the inclusion ratio of Mink1 exon 21 and Ubiquilin1 exon 8 in the RNA-Seq data. Inclusion ratio of the cassette-type exon was calculated by counting the number of reads that were mapped to splice junctions, 5p or 3p, relative to the total numbers of reads of short form plus 5p (upper panels), or 3p (bottom panels), respectively.

Rhythmic small RNAs

Among CLOCK-target genes at ZT8 or CT8, 89 were pre-miRNA genes (**Figures 2.5AB; Supplementary Table 2.3**). For example, CLOCK was confirmed to target *pre-mir-148a*, *pre-mir-150*, and *pre-mir-802* genes by the promoter assay performed at the Fukada lab. However, rhythmic expression of pre-miRNA genes were not detected in the mRNA-Seq analysis, while 65 pre-miRNA genes were expressed, of which 11 were targeted by CLOCK. Yet, it is possible that expression levels of mature miRNAs show circadian variation. To address this possibility, I analyzed the small RNA-Seq data (accession number: DRP001094). The small RNA-Seq data were generated from mouse liver sampled at CT2, CT5, CT8, CT11, CT14, CT17, CT20 and CT22 under DD condition. Approximately 70-75% of 20-30 million reads in each sample were mapped onto the mouse genome with no mismatches, among which 1,141 unique sequences were identical to mature miRNAs stored in miRBase (release 18). I then summed the number of mapped reads across all time points for each miRNA and found 270 miRNAs with >100 reads (**Supplementary Table 2.6**). Among these, 84 mature miRNAs showed circadian oscillation (**Supplementary Table 2.6**). The heat map of the relative abundance revealed predominant enrichment of the rhythmic miRNAs peaking during the subjective day (**Figure 2.5C**), consistent with data from a previous study (Vollmers *et al*., 2012). Circadian profiles of typical rhythmic miRNAs are shown in **Figure 2.5D**.

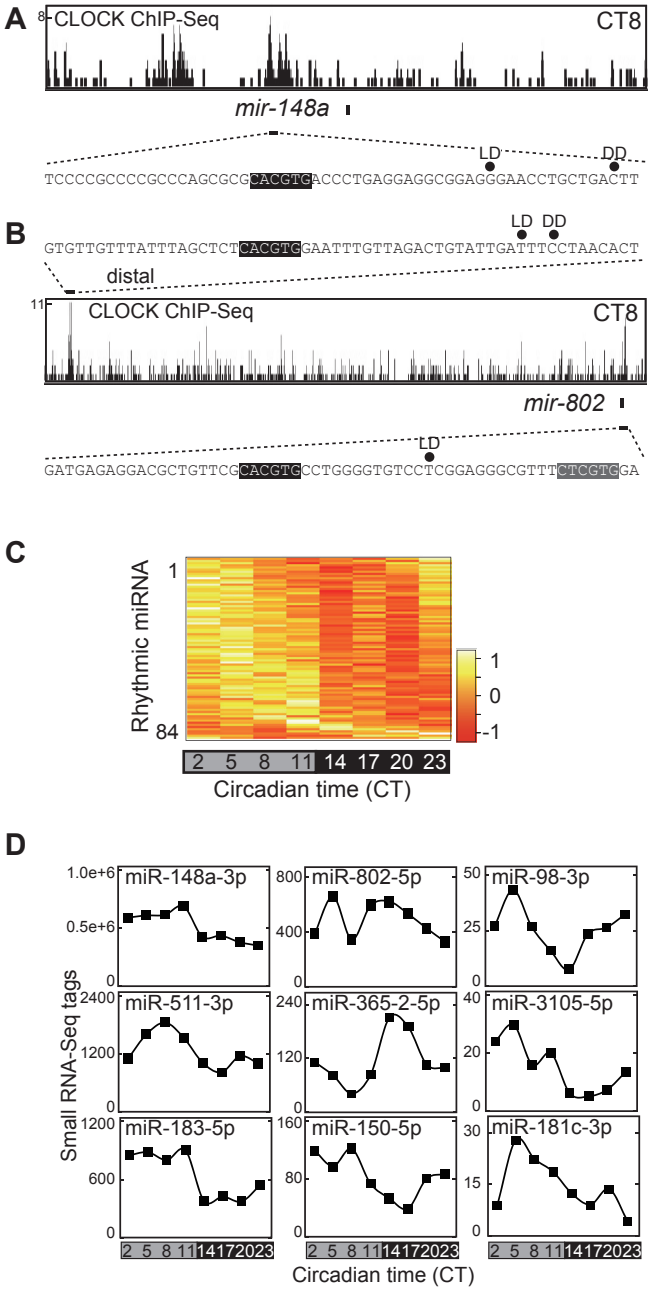

Figure 2.5: Circadian oscillation of miRNAs revealed by small RNA-Seq

(A and B) The CLOCK ChIP-Seq data around *mir-148a* (A) and *mir-802* (B) loci at CT8. LD and DD indicate peak positions of ChIP-Seq reads in ZT8 and CT8, respectively. Black and gray shading indicates the CACGTG-type E-box and one-mismatch sequences, respectively. (C) Heat map of small-RNA-Seq across the day in constant dark condition. miRNA genes were ordered by the peak phases from early subjective day to late subjective night. The numbers of sequenced reads were normalized so that the mean and the variance were 0 and 1, respectively, for each row of the maps. (D) Expression profiles of typical rhythmic miRNAs in the small-RNA-Seq analysis.

Possible functional differences among canonical and non-canonical E-boxes

In this study, I revealed the canonical and non-canonical E-boxes from CLOCK ChIP-Seq data (**Table 2.3**). Until now, functional differences between canonical/non-canonical E-boxes have been unclear owing to scarcity of known CLOCK-binding sites. Large number of the CLOCK-binding sites identified in this study provides opportunity to address this issue. To investigate the possibility of functional differences among the canonical/noncanonical E-boxes, I compared rhythmicity of their target genes and the estimated phased of the rhythmically expressed genes.

I first searched for occurrence of each motif within ±80 bp of the CLOCK-binding sites at CT8, and classified the CLOCK-binding sites according to the patterns of the motif occurrences. To clarify functional differences of each motif, I confined the following analyses to the CLOCK-binding sites of which only one kind of the motifs were found (**Table 2.6**, Unique). Next, genes associated with each CLOCK-binding sites were retrieved (**Supplementary Table 2.2**), and the number of expressed and rhythmically expressed genes were counted for each motif (**Table 2.6**). Most genes associated with the CLOCK-binding sites containing only one kind of E-boxes were expressed, consistent with CLOCK's function as a transactivator.

To assess whether different E-boxes have different efficacy of activating rhythmic expression, I compared the ratio of rhythmically expressed gene relative to expressed genes between the canonical E-box (CACGTG) and each of the other non-canonical E-boxes. Among the non-canonical E-boxes, CACTTT showed significant decrease in the ratio of rhythmically expressed genes (two-sided Fisher's exact test, p-value < 0.05 after Bonferroni correction) (**Table 2.6**). CACGTT has been well known for E′-box (Yoo et al., PNAS, 2005). Thus, this result suggests that functional differences could exist between the canonical E-box and E′-box.

In addition to rhythmicity, distributions of estimated phases for rhythmically expressed genes were compared. Some noncanonical E-boxes such as CACGGG appeared to have different phase distributions from the canonical one (**Figure 2.6**). However, difference in phase distributions of target genes between CACGTG and every non-canonical E-box was not statistically significant (Mardia-Watson-Wheeler test, p-value < 0.05 after Bonferroni correction).

Table 2.6: Canonical and non-canonical E-boxes and rhythmicity

Motif	# of CLOCK binding sites (CT8)		# of genes				*p*-value of Fisher's exact test
	All	Unique	All	Expressed	Rhythmic	%Rhythmic	
CACGTG	1805	1011	1177	1003	173	17.2	-
CACATG	1477	867	801	692	108	15.6	3.88E-01
CACGCG	696	249	397	351	45	12.8	5.27E-02
CACGGG	646	250	305	263	41	15.6	5.79E-01
CACGTT	640	335	409	369	39	10.6	2.35E-03
CACGAG	540	224	270	229	40	17.5	9.23E-01
CATGCG	316	79	113	103	9	8.7	2.53E-02

(a) Number of the CLOCK-binding sites at CT8 of which the indicated motif was found within ±80 bp.
(b) Number of the CLOCK-binding sites at CT8 of which only the indicated motif was found within ±80 bp.
(c) Genes that were associated with the CLOCK-binding sites (see Supplementary table 2.2).
(d) The *p*-values of Fisher's exact test (two-sided) to determine whether there is a nonrandom association between motif sequences and rhythmicity (rhythmically expressed or non-rhythmically expressed genes) in comparison of the canonical E-box (CACGTG) and each of the other non-canonical E-boxes.

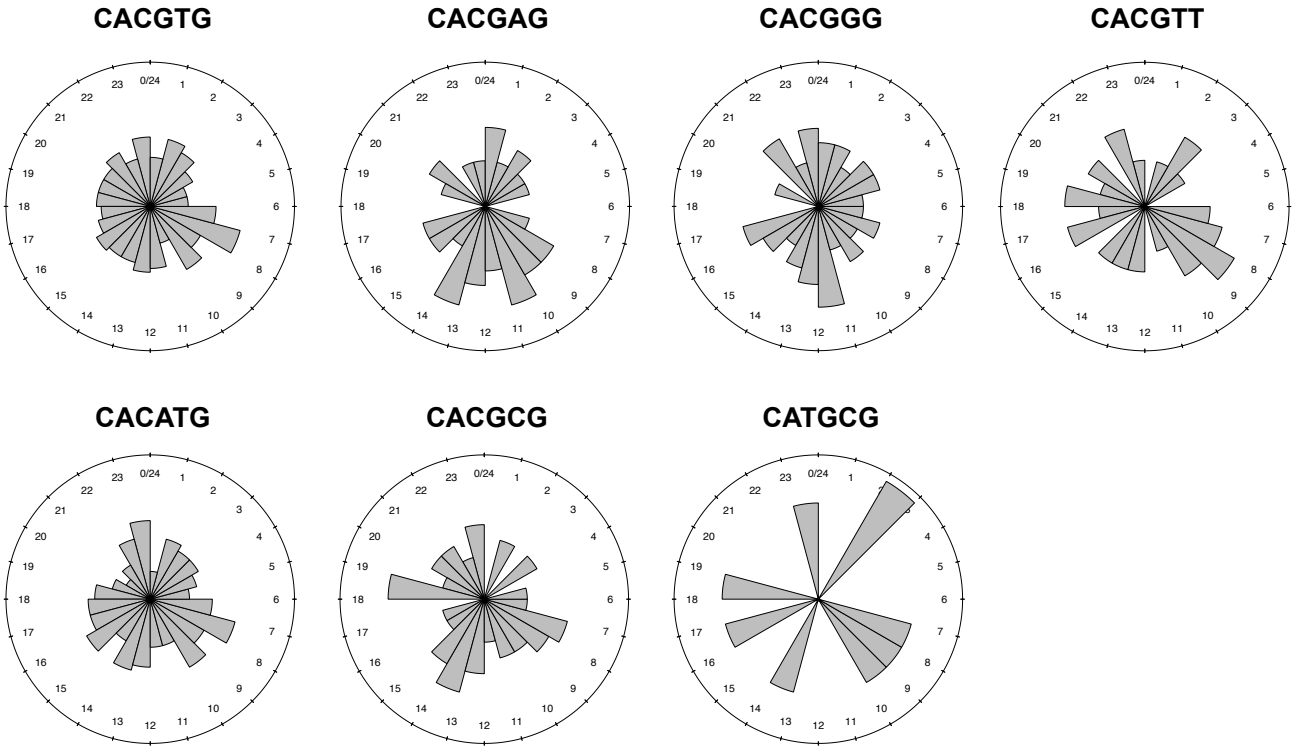

Figure 2.6: Distributions of phases of rhythmically expressed genes associated with the CLOCK-binding sites

Shown are distributions of estimated phases of rhythmically expressed genes that are associated with the CLOCK-binding sites at CT8 of which only the indicated motif was found within ±80 bp. Rose diagrams were visualized using the R package 'circular'.

Discussion

Canonical and non-canonical E-boxes

The results of MOCCS analysis demonstrates that the CLOCK-BMAL1 complex recognizes CACGNG, CACGTT, and CATG[T/C]G in a genome-wide manner. Note that, however, it might be possible that other CLOCK-binding motifs were not found owing to potential bias in immunoprecipitation efficiency of antibodies. For example, CLOCK is dynamically form protein complexes with several proteins depending on the mode of transactivation (Gustafson and Partch, 2014). Thus, some CLOCK proteins in a particular complex might be less frequently captured by the antibody used to generated the CLOCK ChIP-Seq data analyzed in this study. Nevertheless, the motifs revealed in this study were at least targeted by CLOCK proteins that were captured by the antibody. Therefore, the canonical and non-canonical E-boxes revealed by MOCCS analysis are expected to represent the comprehensive set of CLOCK-binding motifs for at least the CLOCK-binding sites identified in this study.

MOCCS and existing motif discovery tools

A transcription factor (TF) recognizes similar but different DNA sequences (Stormo, 2000). This variability of sequences targeted by a TF has encouraged developing motif discovery tools to search for a representation of the binding specificity of the TF. The two major representations are degenerate consensus sequences and position weight matrices (PWMs) (Stormo, 2000). In general, motif discovery tools calculate either of the representation.

The situation is unchanged in even the ChIP-Seq era. Until now, many ChIP-Seq experiments have been performed and they have yielded genome-wide information on occupancy of various transcription factors, as exemplified by the ENCODE project (The ENCODE Project Consortium, 2012). These new massive datasets have challenged conventional motif discovery tools because of the increased size of input sequences (Zambelli *et al*., 2013). Since ChIP-Seq experiments identify the binding sites of the transcription factor under investigation in a genome-wide manner, the input sequences provided to the motif discovery tools usually amount to thousands or ten thousands of sequences with the length of several hundred base pairs. To overcome the larger input size, many tools tailored for ChIP-Seq data have been developed to accelerate the runtime of search for DNA-binding motifs in the form of degenerate consensus sequences or PWMs (*e.g*., Sharov and Ko, 2009; Kulakovskiy *et al*., 2010; Bailey 2011; Machanick and Bailey, 2011;

Hartmann *et al*., 2013). In other words, even for the motif discovery tools tailored for ChIP-Seq data, the objective is to search for summarized representations of binding specificity of TFs.

Contrary to these tools, MOCCS aims at extracting all *k*-mers that are targeted by transcription factors, rather than obtaining summarized representation of these *k*-mers. This has become possible with increased number of TFBSs identified by ChIP-Seq. In practice, MOCCS calculates frequency distribution of each *k*-mer around TFBSs, and thereby evaluates each *k*-mer separately to see whether the *k*-mer is targeted by the transcription factor under investigation. As the aims are different from MOCCS, evaluation of MOCCS based on comparison with other existing motif discovery tools was not performed in the present study.

Future perspectives of MOCCS

In this study, I set the motif length *k* to 6 and only focused on significant motifs with at most two mismatches to the canonical E-box since interest of this study was canonical and non-canonical E-boxes. To apply MOCCS to the broader fields of ChIP-Seq studies, three future perspectives are envisioned.

The first is to extend the motif length *k*. DNA-binding motifs often have longer length than 6 bp (for example, median length of DNA-binding motifs registered in the JASPAR database is 11 bp: JASPAR CORE Vertebrate 5.0_ALPHA; Mathelier *et al*., 2014). In addition, some motifs consists of tandemly-arranged sequences enclosing non-specific spacers. However, simply extending *k* will fail to find significant motifs since longer *k*-mers occur less frequency and thus would lead to unstable calculation of AUCs. To circumvent this issue, for example, combining significant motifs with length of less than *k* to infer *k*-length motifs is one way to try.

The second is overlapping motifs. In this study, I only focused on *k*-mers with at most two mismatched to the canonical E-box. On the other hand, the primary output of MOCCS included *k*-mers overlapping with the significant motifs by $k-1$ length. To properly interpret such overlapping motifs, a method to integrate the whole results would be necessary.

The third is improvement of statistical thresholding on the AUC scores. Experimental validation showed no enhancer activity of TACGTA, one of the other two-mismatch potential motif. TACGTA appeared far less frequently around the CLOCK-binding sites than the other motifs (**Table 2.3**), suggesting that the small count of TACGTA presumably lead to the AUC score higher than the threshold by chance. To overcome this issue, more sophisticated thresholding considering the raw count should be applied in the future.

Possible mechanisms of rhythmicity in miRNA abundance

In this study, I found miRNAs showing circadian rhythms in their abundance. One possible mechanisms underlying these rhythms is transcriptional regulation of pri-miRNAs by the transcription factors with rhythmic expression. Consistently, several miRNAs were shown to be targeted by CLOCK (**Figure 2.5AB**). In addition, as discussed below, other rhythmic transcription factors might regulate rhythmic expression of miRNAs that were not directly targeted by CLOCK. Thus, analyses of ChIP-Seq data sets of other rhythmic transcription factors would help assess to what extent transcriptional regulation is important for generating rhythms in miRNA abundance.

Another possibility is involvement of the biogenesis processes of miRNAs. The biogenesis of miRNAs includes several post-transcriptional processes including processing and maturation (Krol *et al*., 2010), and thus the rhythms in those processes could be a cause of the temporal variations of miRNA abundance observed. To address this possibility, I assessed whether rhythmic expressions were detected for genes involved in the biogenesis of miRNAs. Genes involved in processing (*Drosha* and *Dgcr8*), transport (*Xpo5*), maturation (*Dicer1* and *TRBP* (*Tarbp2*)), and silencing (*GW182* (*Tnrc6a*) and *Ago2* (*Eif2c2*)) of miRNAs did not show in our RNA-seq data (**Supplementary table 2.3**). However, *Ago1* (*Eif2c1*) and *Ago4* (*Eif2c4*), which encode the AGO proteins with no silencer activity (Meister, 2013), showed rhythmic expressions with the estimated phases of 7.5 and 5.6, respectively (**Supplementary table 2.3**). Thus, trapping of miRNAs in AGO1 and AGO4 might generate rhythms in miRNA abundance. However, it is difficult to explain why approximately 70% of expressed miRNAs in the small RNA-Seq data did not show rhythms in their abundance, even in the case of miRNAs with high read numbers (**Supplementary Table 2.6**), as several studies demonstrate that miRNAs are randomly loaded to the individual AGO proteins (Dueck *et al*., 2012; Wang *et al*., 2012). On the other hand, since another study reports preferences in miRNA sorting for different AGO proteins in mammals (Burroughs *et al*., 2011), it is possible that miRNAs are sorted preferentially to different AGO proteins in a tissue-specific manner. Therefore, further studies on miRNA sorting in mammals are required to conclude whether rhythmic expressions of these *Ago* genes underlie the rhythmicity of mature miRNAs.

Indirect regulations of rhythmic gene expression by CLOCK

Comparison of CLOCK ChIP-Seq and RNA-Seq data revealed that 802 rhythmic genes were not included in the CLOCK targets, suggesting the importance of the output pathways that are mediated by several mechanisms yet regulated by CLOCK.

Such indirect regulations would be partly mediated by the actions of transcription factors that affect the expression of their target genes, as proposed in Miller *et al*. (2007). In fact, gene ontology (GO) analysis identified 250 transcription factors as CLOCK targets, some of which were rhythmically expressed in mouse liver (**Supplementary Table 2.7**). For example, the Fukada lab showed that several genes encoding transcription factors of KLF Krüppel-like factor (KLF) family were targeted by CLOCK and that those genes show rhythmic expression. Since ChIP-Seq data have accumulated in recent years, integration of ChIP-Seq data of diverse transcription factors and RNA-Seq data is expected to reveal the multistep cascades of transcription factors as the output pathways of the circadian oscillator.

Among 1084 rhythmically expressed genes, 13 genes were long non-coding RNAs (lncRNAs). For instance, the Fukada lab confirmed that 0610005C13Rik is rhythmically expressed and targeted by CLOCK. Although functions of most lncRNAs remain largely unknown, lncRNAs have been shown to inhibit or enhance gene expression through several mechanisms (Geisler and Coller, 2013). Thus, circadian expression of lncRNAs may be one of the output pathways of the circadian oscillator.

miRNAs have been predicted to target >30% of the protein-coding mRNAs (Lewis *et al*., 2005), and the contribution of miRNA expression toward regulating the circadian clockwork in several organisms has been shown (Cheng *et al*., 2007; Kadener *et al*., 2009; Kojima *et al*., 2011). miRNAs recognize 3′-UTR sequences of their targets, and thereby, protein synthesis is inhibited, or deadenylation/degradation of the target mRNA is triggered. For example, using TargetScan (release 18, Garcia *et al*. (2011)), *Clock* was predicted to be targeted by miR-148a-3p. Experiments conducted at the Fukada lab revealed that the mRNA level of *Caveolin-1* (*Cav1*), a known target of miR-802 (Lin *et al*., 2011), shows a circadian rhythm with a phase shifted largely from those of *pri-mir-802* while transcription itself shows no significant circadian variation in Pol2-ChIP analysis. Thus, it is possible that post-transcriptional regulation by miRNAs could mediate the output of the circadian oscillator.

In this study, I showed circadian oscillation of alternative splicing events. Molecular mechanism of circadian regulation in global RNA splicing remains unclear, but it is noteworthy that CLOCK binding and rhythmic transcription were observed for several genes encoding splicing factors such as polypyrimidine tract binding protein (PTBP1) (**Supplementary Tables 2.1, 2.2, and 2.3**). PTBP1 binds to pyrimidine-rich sequences represented by UCUU to regulate alternative splicing (Pérez *et al*., 1997). Interestingly, pyrimidine-rich sequences with core element UCUU are

observed around alternative cassette exons of *Mink1* and *Ubqln1* (data not shown). Consistently, McGlincy *et al*. (2012) showed that dozens of splicing factors are rhythmically expressed. They also found that circadian alternative splicing events were changed in circadian mutant mice using exon-arrays (McGlincy *et al*., 2012). Hence, splicing regulation by splicing factors that are regulated by CLOCK could be one of the output pathways of the circadian oscillator.

Conclusion

In this study, I developed MOCCS, a method to enumerate DNA-binding motifs from ChIP-Seq data. MOCCS would be useful for broader people to extract DNA-binding motifs at high resolution using ChIP-Seq data.

This study suggested two types of the output pathways of the circadian oscillation via CLOCK. One is direct regulation, *i.e*., transactivation through CLOCK binding to canonical and non-canonical E-box motifs, and the other is indirect regulations through transactivation by other transcription factors and post-transcriptional regulations involving miRNAs, lncRNAs, and alternative splicing.

Chapter 3: Positive selection on gene copy number variations in adaptive evolution

Positive selection is an evolutionary process by which an allele increases its frequency in a population. Copy number variations (CNVs) constitute a significant proportion of genomic diversity. In particular, gene copy number variations (GCNVs), which change the numbers of gene loci in genomes, can significantly alter gene functions and dosages, and thus can undergo positive selection. However, positive selection of CNVs has not been proved. One way of assessing the possibility is to search for increase or decrease of copy numbers in parallel evolution of freshwater groups of three-spined stickleback. Parallel evolution is the adaptive evolution in which the same genotypes or phenotypes are selected in different but related lineages, and provides strong evidence of positive selections. To address the possibility of positive selection on GCNVs using the system of parallel evolution of three-spined stickleback, I analyzed resequencing data of multiple individuals from freshwater and marine populations. A novel approach was devised to detect GCNVs under parallel selection from whole-genome sequencing data with low coverage by comparing two resequencing datasets. Application of the method to the resequencing data of sticklebacks revealed GCNVs that were likely under parallel selection. Many of the identified GCNVs showed increase in gene copy numbers in freshwater individuals, which is consistent with the notion that increase in gene copy number would facilitate adaptive evolution. These results suggest that contribution of GCNVs should be considered in studies on adaptive evolution.

Background

Detection of positive selection

Understanding the genetic basis of adaptive evolution is one of the major goals in evolutionary biology (Biswas and Akey, 2006; Barrett and Schluter, 2008; Barrett and Hoekstra, 2011; Kocher 2004; Prentis *et al*., 2008). When populations adapt to new environments, positive selection can increase frequencies of specific genetic variations that have greater fitness than others, sometimes

---

This chapter is in part published in the following publication:

Shotaro Hirase*, Haruka Ozaki* and Wataru Iwasaki, Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes, *BMC Genomics* **15**, 735 (2014)
(* Equal contributions)

resulting in the fixation of those variations (Biswas and Akey, 2006; Barrett and Schluter, 2008; Barrett and Hoekstra, 2011).

To detect positive selection, two major approaches have achieved significant success. One approach is molecular evolutionary analysis of protein-coding gene sequences. Comparison of the synonymous and nonsynonymous nucleotide substitution rates has been adopted by many studies to identify positive selection (Biswas and Akey, 2006; Nielsen, 2005). While this approach is applicable to only protein-coding genes that have accumulated sufficient numbers of nucleotide substitutions, the other approach targets shorter time-scale events by detecting the fixation of single nucleotide variations (SNVs) within populations (Biswas and Akey, 2006). Many SNVs were found to be associated with phenotypic variations, including cis-elemental SNVs that affect gene expression levels (e.g., Cheung *et al*., 2005). Analyses of polymorphism distributions have revealed positive selection of a number of SNVs (e.g., Akey *et al*., 2004; Carlson *et al*., 2005).

Positive selection on gene copy number variations

These approaches focused on positive selection on variations due to nucleotide substitutions. However, it has recently been revealed that copy number variations (CNVs), or gains or losses of DNA segments, constitute a significant proportion of genomic diversity (Feuk *et al*., 2006; Cridland and Thornton, 2010; DeBolt 2010; Quinlan *et al*., 2010; Brown *et al*., 2012; Handsaker *et al*., 2010). Because CNVs are known to result in significant phenotypic effects that include human diseases (McCarroll and Altshuler, 2007), they are also expected to be under positive selection. In particular, gene copy number variations (GCNVs), which change the numbers of gene loci in genomes, can significantly alter gene dosages and yield new gene functions (Kondrashov and Kondrashov, 2006; Chen *et al*., 2008). As expected, the possibility of fixation of CNVs by positive selection has been reported in several phylogenetic groups (Emerson *et al*., 2008; Gazave *et al*., 2011).

Aim of this study

Parallel evolution, which is the adaptive evolution of the same trait in related but independent lineages, can provide evidence of positive selection, because genetic drift is unlikely to produce concerted changes in independent lineages (Rundel *et al*., 2000).

The marine and freshwater phenotypes of three-spined sticklebacks (*Gasterosteus aculeatus*) are an excellent system to investigate parallel evolution (Rundel *et al*., 2000). This species inhabits

a large number of marine, estuarine, and freshwater environments in Asia, Europe, and North America. After the retreat of Pleistocene glaciers, the marine ancestors have colonized and adapted to newly created freshwater habitats over the world, showing repeated changes in the body shape, skeletal armor, trophic specialization, pigmentation, salt handling, life history, and mating preference (Bell and Foster, 1994; McKinnon and Rundle, 2002).

Previous studies revealed that this independent evolution of similar phenotypes in the freshwater groups occurred due to parallel selection on the globally shared, standing SNVs in the same genes in different freshwater populations, providing strong evidence that positive selection on these SNVs contributed to the adaptive evolution toward the freshwater environments (Colosimo *et al*., 2004; Colosimo *et al*., 2005; Jones *et al*., 2012). Recently, Feulner *et al*. (2013) reported a significant number of CNVs in a marine population of the sticklebacks. Therefore, as with SNVs, GCNVs can also be under parallel selection through the evolution of sticklebacks. To investigate this possibility, I analyzed whole-genome resequencing data from marine and freshwater groups of three-spined sticklebacks and searched for GCNVs that contributed to the parallel evolution of the three-spined sticklebacks.

Materials and Methods

Resequencing data

A resequencing dataset of 10 marine and 10 freshwater individuals was previously generated using an Illumina Genome Analyzer II (36-51 bp, single-end), which yielded approximately sixty-million million reads (approximately 2.3×) per individual (Jones *et al*., 2012, **Table 3.1**). I downloaded the data from NCBI Sequence Read Archive (SRA, Leinonen *et al*. (2011)). The accession numbers were SRX077979, SRX079119, SRX079120, SRX077981, SRX077982, SRX077990, SRX077978, SRX076627, SRX079121, SRX077983, SRX077984, SRX077986, SRX077980, SRX077988, SRX077989, SRX077987, SRX077991, SRX077992, SRX076626, SRX077985, SRX077993, and SRX077994.

The sequenced reads from each individual were mapped to the stickleback genome using the Bowtie 0.12.8 software (Langmead *et al*., 2009) (**Figure 1.1A**). The Bowtie option of '-m 1' was adopted to remove reads with multiple hits. In addition, to obtain reliable GCNVs that were not affected by the mapping parameter selection, I adopted three different values (70, 100, and 130) for the '-e' option, which designated the maximum permitted total quality values at all mismatched positions throughout a read alignment. To avoid the effects of potential PCR duplicates, if multiple reads were aligned to the same position, all of the reads except for those with the highest mapping quality were removed using SAMtools (version 0.1.18, Li *et al*. (2009)) with the command 'samtools rmdup -s'. The statistics for each mapping option are shown in **Table 3.1**.

Table 3.1: Summary of resequencing dataset of 10 marine and 10 freshwater sticklebacks and mapping statistics

Sample ID	Accession number	Phenotype	Basin	Gographic region	Number of reads	Number of reads aligned (Percentage)		
						-e 70	-e 100	-e 130
ABW	SRX077979	Freshwater	Atlantic	Iceland	65,904,900	36426325 (55.27%)	36587409 (55.52%)	36596174 (55.53%)
BIGL	SRX079119	Freshwater	Pacific	California	62,982,716	33091246 (52.54%)	34251500 (54.38%)	34800423 (55.25%)
FTC	SRX079120	Freshwater	Pacific	Washington	130,008,902	54744497 (42.11%)	54763528 (42.12%)	54762816 (42.12%)
HUTU	SRX077981, SRX077982	Freshwater	Pacific	Washington	152,468,755	54845476 (35.97%)	56089798 (36.79%)	56544376 (37.09%)
MATA	SRX077990	Freshwater	Pacific	California	66,055,414	33542818 (50.78%)	33530990 (50.76%)	33504790 (50.72%)
MUDL	SRX077978	Freshwater	Pacific	Alaska	39,467,325	21045530 (53.32%)	21183088 (53.67%)	21206196 (53.73%)
NOST	SRX076627	Freshwater	Atlantic	Norway	18,812,768	11152066 (59.28%)	11241084 (59.75%)	11285321 (59.99%)
PAXB	SRX079121	Freshwater	Pacific	British Columbia	53,647,262	30239994 (56.37%)	31054401 (57.89%)	31359584 (58.46%)
SCX	SRX077983	Freshwater	Atlantic	Germany	34,331,961	17594106 (51.25%)	17749879 (51.70%)	17779676 (51.79%)
SHEL	SRX077984	Freshwater	Atlantic	Scotland	54,599,491	27542876 (50.45%)	27692432 (50.72%)	27711276 (50.75%)
ANTL	SRX077986	Marine	Atlantic	Nova Scotia	40,210,299	14831972 (36.89%)	14997902 (37.30%)	15028743 (37.38%)
BDGB	SRX077980	Marine	Pacific	California	35,332,775	20949468 (59.29%)	21111998 (59.75%)	21142925 (59.84%)
BIGR	SRX077988	Marine	Pacific	California	36,099,888	22842437 (63.28%)	22846665 (63.29%)	22827525 (63.23%)
GJOG	SRX077989	Marine	Atlantic	Iceland	53,822,455	28877690 (53.65%)	29040512 (53.96%)	29077068 (54.02%)
GORT	SRX077987	Marine	Atlantic	Scotland	48,046,011	27148586 (56.51%)	27276293 (56.77%)	27279833 (56.78%)
JAMA	SRX077991	Marine	Pacific	Japanese	58,203,987	31192808 (53.59%)	32074859 (55.11%)	32443325 (55.74%)
JMRP	SRX077992	Marine	Atlantic	Scotland	56,314,624	20602716 (36.59%)	20608385 (36.60%)	20600818 (36.58%)
NEU	SRX076626	Marine	Atlantic	Germany	85,802,994	34304198 (39.98%)	34324465 (40.00%)	34318061 (40.00%)
RABS	SRX077985, SRX077993	Marine	Pacific	Alaska	65,798,576	32386458 (49.22%)	33019706 (50.18%)	33251296 (50.53%)
SALR	SRX077994	Marine	Pacific	British Columbia	36,887,238	18176089 (49.27%)	19420930 (52.65%)	20025564 (54.29%)

Genome sequence and gene annotation

The three-spined stickleback genome sequence (BROADS1.56) and the annotated gene models were taken from the Ensembl database (release 72, Hubbard *et al*. (2002)). The genome sequence has been generated from a line derived from a freshwater population (Bear Paw Lake, Jones *et al*., 2012).

For each GCNV likely under parallel selection, I obtained functional annotations of the gene from the Ensembl database. If the functional annotations were unavailable, BLASTX searches (Altschul *et al*., 1997) against the NCBI non-redundant protein database (nr) (Benson *et al*., 2010) were conducted with an E-value cutoff of 1e-14, and the hit with the highest bit-score and its annotated protein name was retrieved.

Analysis of resequencing data

I compared the numbers of mapped reads for each gene between the freshwater and marine groups to identify GCNVs under parallel selection (**Figure 1.1B**). If the numbers of mapped reads were significantly larger in the freshwater group, the gene would have been duplicated or multiplied specifically in the genomes of the freshwater group. If the numbers were significantly smaller, the gene would have been deleted or its copy number would have decreased.

The most 5′- and 3′- positions of each gene were retrieved from the Ensembl annotation, and the numbers of mapped reads that overlapped with the above area (i.e., any exonic or intronic region) were counted using the 'intersectBed' command in BEDTools (Quinlan and Hall, 2010). Because insufficient numbers of mapped reads may result in the detection of false GCNVs, I removed genes from the subsequent analysis if the median of the numbers of the mapped reads per 100 bp of the gene lengths was less than one, or if no reads were mapped in at least one individual resequencing data. For normalization, the numbers were divided by the total number of mapped reads across the genome for each individual. Then, I searched for GCNVs under parallel selection by detecting genes that showed significant differences in the normalized read numbers between the freshwater and marine groups using the edgeR package (Robinson *et al*., 2010) with a false discovery rate (FDR) < 0.05. I regarded genes that were significant under all of the three different mapping options ("-e 70", "-e 100", and "-e 130") as GCNVs likely under parallel selection.

To confirm that the number of identified GCNVs under parallel selection was significantly larger than that expected by chance (i.e., by genetic drift), I calculated an empirical *p*-value based on a permutation test. I randomly reallocated the 10 freshwater and 10 marine individuals into two groups 10,000 times, performed the same analyses, and obtained the null distribution of numbers of GCNVs.

SNP analysis

If the identified GCNVs involved gene duplications or multiplications, three or more different allelic sequences should be observed within the gene in each individual of each group, because three or more different allelic sequences cannot originate from a diploid genome. Thus, I examined whether three or more different allelic sequences were observed in the identified GCNVs (**Figure 1.1C**).

For each of the identified GCNVs, SNVs were called by applying the SAMtools/BCFtools pipeline (Li *et al*., 2009) to the reads that were mapped with the '-e 100' option. The SAMtools/ BCFtools pipeline was used with default parameters, except for the '-Q 30' option, to consider bases that were called with high quality only. I enumerated every pair of SNV positions that was located within the read length, i.e., 36 bp (within-read-length SNV position pairs). The numbers of different nucleotide pairs for each of the within-read-length SNV position pairs were counted, where each nucleotide pair was supported by multiple reads. Finally, I selected GCNVs that showed three or more different nucleotide pairs in at least three individuals of either group.

Microarray data analysis

Microarray data of gills of two families of pure marine and pure freshwater crosses under short and long photoperiods (Kitano *et al*., 2010) were downloaded from Center for Information Biology Gene Expression (http://cibex.nig.ac.jp) with the accession number CBX139. Two marine and freshwater datasets were treated as biological replicates. If multiple probes were mapped to one transcript, the median signal intensity of these probes was used. After removing intra-gene probes, genes with significant expression-value differences between the marine and freshwater groups were identified using the eBayes method in the limma package (Smyth, 2005). The *p*-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

Results and Discussion

GCNVs that likely contributed to the parallel evolution of three-spined sticklebacks

I downloaded whole-genome resequencing data of 10 marine and 10 freshwater individuals of three-spined sticklebacks (Jones *et al*., 2012) from NCBI Sequence Read Archive (Leinonen *et al*., 2011). Both groups consisted of individuals that were derived from diverse areas along the Pacific and Atlantic Ocean coastlines (**Table 3.1**). Thus, genetic variations that were specifically shared among individuals in the freshwater (and marine) group were likely due to parallel selection.

The coverage of the resequencing data was low (approximately 2.3× per individual as reported in Jones *et al*., 2012), making it difficult to apply conventional CNV detection tools (*e.g*., *Abyzov et al*., 2011). Thus, to increase the sensitivity of detecting GCNVs under parallel selection, I devised a novel approach that was based on a statistical method (**Figures 3.1AB**). The sequenced reads from each of the 20 individuals were mapped to the reference stickleback genome, and the numbers of the mapped reads were counted for each gene to estimate changes in their copy numbers. Genes that showed significant differences in the numbers of mapped reads between both groups were identified as GCNVs likely under parallel selection (**Figures 3.1AB**, see Materials and Methods).

Twenty-four genes showed significant differences in the numbers of mapped reads between both groups (**Figure 3.2 and Table 3.2**). Among these genes, five showed more copies in the individuals of the marine group (freshwater-decreased GCNVs) and 19 showed more copies in those of the freshwater group (freshwater-increased GCNVs). I confirmed that the number of the identified GCNVs was significantly larger than that expected by chance based on a permutation test ($p < 0.05$) for each mapping option. Collectively, these results suggested that the 24 GCNVs were likely due to parallel selection. Note that the 2.3× coverage of the resequencing data (Jones *et al*., 2012) would have led to underestimation of the numbers of GCNVs between the marine and freshwater groups. A higher sequencing coverage may result in detection of more GCNVs.

Among the identified GCNVs, *neurexophilin and PC-esterase domain family member 3* (*NXPE3*) overlapped with a region that was reported as a CNV in a marine group of three-spined sticklebacks (Feulner *et al*., 2013). In addition, the identified GCNVs included well-known multigenic families such as *sulfotransferase* (*SULT*), *NOD-like receptor* (*NLR*), *apolipoprotein L* (*APOL*), *kinesin family* (*KIF*), and *myosin heavy chain* (*MyHC*). The finding that the identified GCNVs included genes in multigenic families was consistent with the idea that GCNVs of

multigenic family genes are more likely to occur than those of single-copy genes. This is because, fatal effects due to copy-number changes of multigenic family genes tend to be less than those of single-copy genes (Nguyen *et al*., 2006). It would be notable that GCNVs were previously observed for *APOL* (Perry *et al*., 2008), *KIF* (Conrad *et al*., 2009) and *SULT* (Hebbring *et al*., 2007) in primates and for *MyHC* in fish (Ikeda *et al*., 2007).

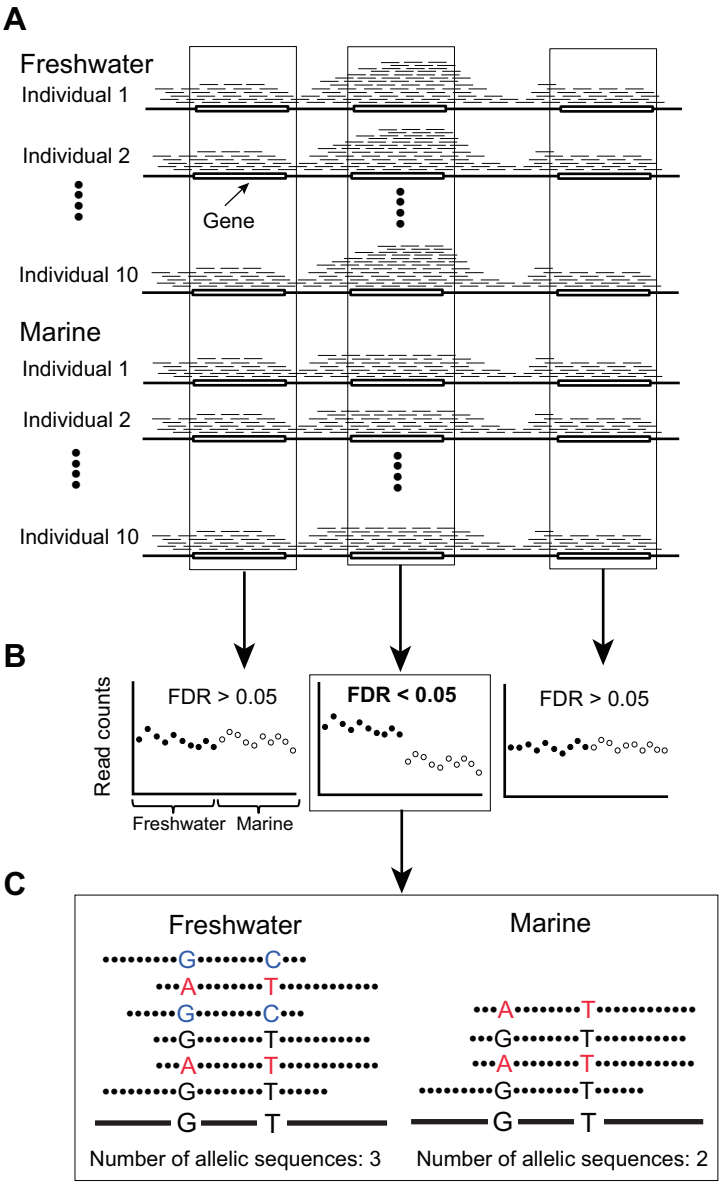

Figure 3.1: Schematic diagram of the method for identifying GCNVs likely under parallel selection
(A) Re-sequenced reads (thin lines) from each individual were mapped to the stickleback reference genome (thick lines). (B) The numbers of mapped reads that overlapped with genes were counted, and we searched for genes that showed significant differences in the normalized read numbers between the freshwater (closed circles) and marine groups (open circles) with a false discovery rate (FDR) < 0.05. Genes that showed significant differences under the three mapping options were regarded as GCNVs likely under parallel selection. (C) The number of different allelic sequences was counted for each of the identified GCNVs by enumerating every pair of SNV positions that was located within the read length. If three or more allelic sequences were observed for a gene, the GCNV involved duplications or multiplications.

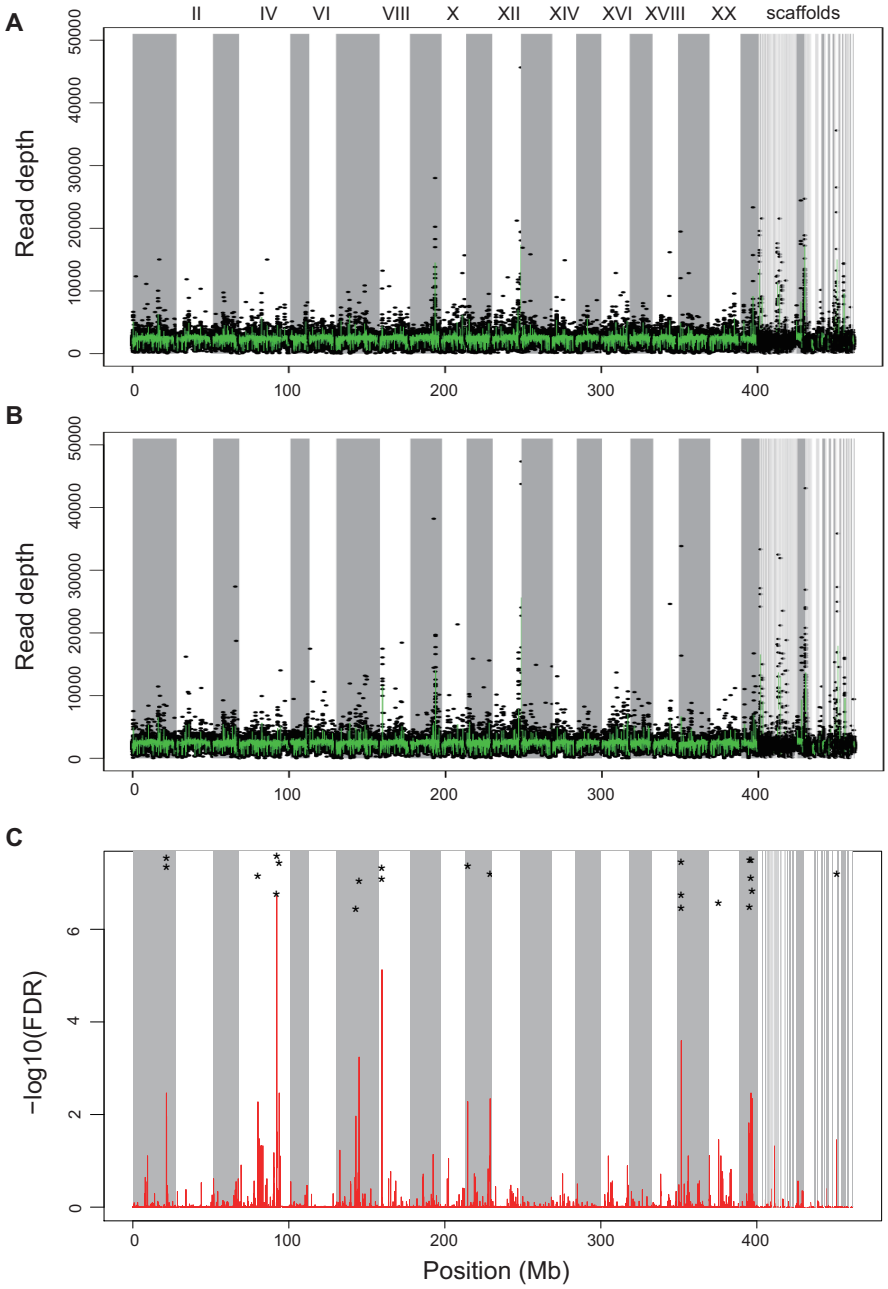

Figure 3.2: GCNVs likely under parallel selection

The normalized numbers of mapped reads per 1-Mb gene length for each gene across the genomes of the (A) freshwater and (B) marine groups. Each black point represents the number for each gene in each individual, and the green lines represent the mean values for each gene across individuals. (C) The false discovery rate of the EdgeR analysis on the differences in the numbers of mapped reads between the freshwater and marine groups for each gene. Asterisks indicate the positions of the GCNVs under parallel selection (FDR < 0.05).

Table 3.2: Gene copy number variations likely under parallel selection

Ensembl gene ID	Genomic location			Group having more copies	In divergent regions [21]	Gene annotation
	Linkage group	Start	End			
ENSGACG00000014268	groupI	21,543,442	21,565,537	Freshwater	Yes	Tensin 1 (TNS1)
ENSGACG00000014289	groupI	21,600,545	21,614,802	Freshwater	Yes	Serine/threonine kinase 11 interacting protein (STK11IP)
ENSGACG00000018214	groupIV	11,925,723	11,934,224	Freshwater	No	Kinesin family member 3A (KIF3A)
ENSGACG00000019313	groupIV	23,928,955	23,953,125	Freshwater	No	Tubulin tyrosine ligase-like family member 12 (TTLL12)
ENSGACG00000019321	groupIV	23,968,608	23,982,358	Freshwater	Yes	Sulfotransferase family 4A member 1 (SULT4A1)
ENSGACG00000020171	groupVII	12,721,951	12,727,083	Freshwater	No	Protein phosphatase 1 regulatory (inhibitor) subunit 14A (PPP1R14A)
ENSGACG00000014553	groupXI	15,607,308	15,613,431	Freshwater	No	Apolipoprotein L 2 (APOL2)
ENSGACG00000002886	groupXIX	2,446,925	2,473,806	Freshwater	Yes	NLR family CARD domain containing 5 (NLRC5)
ENSGACG00000002902	groupXIX	2,484,537	2,497,605	Freshwater	Yes	*Myosin heavy chain (MyHC)
ENSGACG00000002933	groupXIX	2,501,529	2,511,962	Freshwater	Yes	*Myosin heavy chain (MyHC)
ENSGACG00000006397	groupXX	6,176,973	6,190,798	Freshwater	No	Dopa decarboxylase (aromatic L-amino acid decarboxylase)(DDC)
ENSGACG00000002551	groupXXI	5,808,646	5,870,440	Freshwater	No	*Rab effector MyRIP-like (MYRIP)
ENSGACG00000002682	groupXXI	6,189,464	6,240,135	Freshwarer	No	Neuropilin (NRP) and tolloid (TLL)-like 1 (NETO1)
ENSGACG00000002744	groupXXI	6,534,938	6,558,550	Freshwater	No	Junctophilin 1 (JPH1)
ENSGACG00000002857	groupXXI	7,179,938	7,191,684	Freshwater	No	Carboxypeptidase A6 (CPA6)
ENSGACG00000002913	groupXXI	7,252,896	7,262,425	Freshwater	No	Minichromosome maintenance domain containing 2 (MCMDC2)
ENSGACG00000002918	groupXXI	7,255,256	7,257,350	Freshwater	No	*Unknown
ENSGACG00000003408	groupXXI	7,994,019	7,996,973	Freshwater	No	*Neoverrucotoxin
ENSGACG00000015099	scaffold_68	405,524	407,382	Freshwater	No	LSM14B SCD6 homolog B (S. cerevisiae) (LSM14B)
ENSGACG00000019508	groupIV	25,553,051	25,563,391	Marine	No	Neurexophilin and PC-esterase domain family member 3 (NXPE3)
ENSGACG00000020238	groupVII	14,778,775	14,788,878	Marine	No	*Gap-Pol polyprotein-like
ENSGACG00000003374	groupVIII	1,526,335	1,528,158	Marine	No	*Unknown
ENSGACG00000003379	groupVIII	1,528,722	1,530,746	Marine	No	*Unknown
ENSGACG00000005313	groupXI	1,204,843	1,206,464	Marine	No	*Heat shock protein (HSP)

*Gene annotations were based on BlastX search if Ensembl annotations were unavailable.

Segmental duplications/multiplications or deletions behind the identified GCNVs

An important characteristic of the 24 GCNVs likely under parallel selection was that they frequently appeared at close locations on the genomes (**Figure 3.2**). This observation implied that those GCNVs would have resulted from segmental duplications/multiplications or deletions of genomic regions that contained multiple genes (i.e., gene clusters). **Figure 3.3** represents the ratios of the numbers of reads that were mapped to genes in and around the gene clusters in the linkage groups VIII and XIX, which were suspected to have experienced segmental duplications or deletions. This observation was consistent with a previous study that reported that CNVs sometimes involve segmental duplications (Gazave *et al*., 2011).

Next, I compared the locations of the 24 GCNVs with divergent regions that were designated by Jones *et al*. (2012), because a previous study reported that many CNVs in primates overlapped with genes under positive selection (Gokcumen *et al*., 2011). The divergent regions were three-spined stickleback genomic regions whose sequences showed signs of parallel evolution of nucleotide variations between the marine and freshwater groups. The aforementioned gene cluster in the linkage group XIX overlapped with the divergent regions, suggesting that both nucleotide sequences and copy numbers of the genes in this region would have been under parallel selection during adaptation to the freshwater environment. However, most of the GCNVs did not overlap with the divergent regions, which suggested that their copy numbers, but not sequences, would have been under parallel selection (**Table 3.2**).

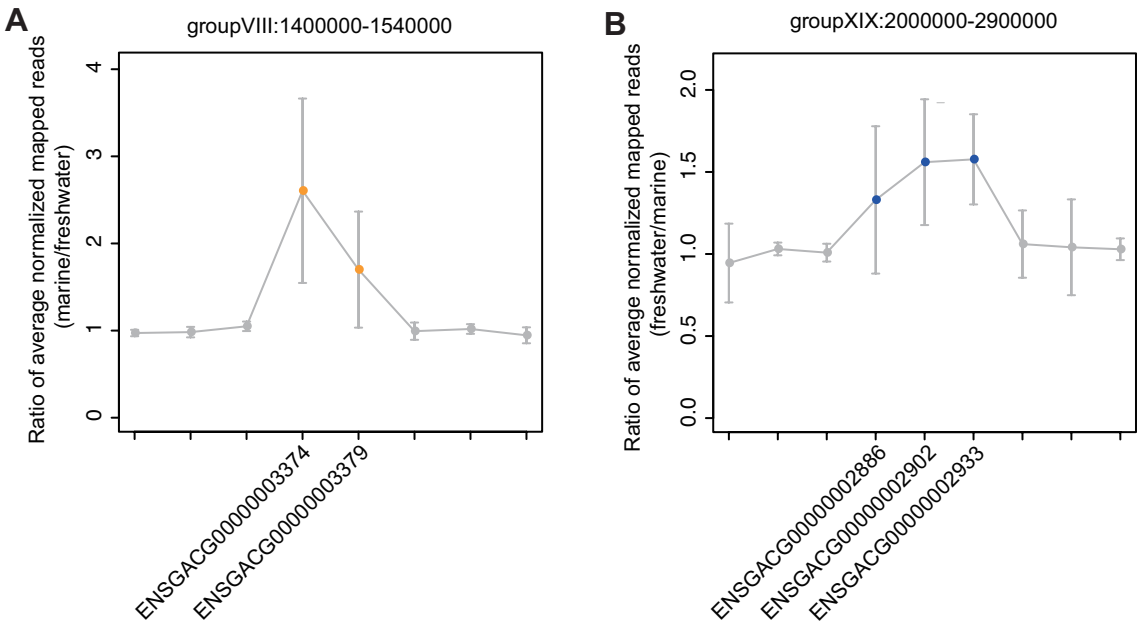

Figure 3.3: Segmental duplications/multiplications or deletions underlying the clusters of GCNVs likely under parallel selection

Gene clusters that included GCNVs likely under parallel selection located in the linkage groups (A) VIII and (B) XIX are shown with three genes upstream or downstream. Each point represents the ratio of the average of the normalized numbers of the mapped reads between the two groups. The identified GCNVs with more copies in the marine and freshwater groups are colored by orange and blue, respectively. Genes were excluded from visualization if the median of the numbers of mapped reads per 100 bp of the gene length was less than one or if no reads were mapped in at least one individual. The error bars indicate standard deviations of the ratios that were calculated for pairs of freshwater and marine groups derived from the same geographic regions. (If multiple samples were derived from the same geographic region for either group, the average of the normalized number of reads was used for the calculation.)

Larger gene copy numbers in the derivative, freshwater phenotype

Among the 24 GCNVs likely under parallel selection, larger gene copy numbers were more frequently associated with the freshwater group (19 out of 24, **Table 3.2**). This was consistent with the fact that the freshwater phenotype is derivative, because increase, rather than decrease, in gene copy numbers is expected to facilitate adaptation to new environments by introducing new physiology and morphology to the organism (Hoffmann and Willi, 2008). For example, Chen *et al*. (2008) suggested that duplications of protein coding genes contributed to the physiological fitness of Antarctic notothenioids in freezing polar conditions. In particular, the freshwater-increased GCNVs included two genes involved in the inflammatory response (*APOL2*, *NLRC5*) and two genes that were homologous to *MyHC* (ENSGACG00000002902, ENSGACG00000002933). A previous study showed parallel divergences between littoral and pelagic phenotype pairs of three-spined stickleback MHC genes, which are key genes in the immune system and would be associated with parasite communities in each habitat (Scharsack *et al*., 2007). Various types of myosin genes were reported to have appeared during the evolution of teleost fish, and those variations were supposed to have contributed to the adaptation to variable aquatic conditions (Ikeda *et al*., 2007). Thus, I expect that those GCNVs would have played important roles in adaptation to the freshwater environment.

The larger gene copy numbers in the freshwater group could be due to the choice of the reference genome sequence. I used the reference genome that was generated from a freshwater lineage, thus the mapping efficiency of the sequencing data of the marine group might be lower for genes that accumulated many SNVs between the marine and freshwater groups. To examine whether the detected GCNVs were derived from the mapping efficiency bias toward the freshwater group, I investigated the frequencies of SNVs of the 19 freshwater-increased GCNVs using reads that were mapped with the '-e 100' option. The most divergent gene was ENSGACG00000015099, which contained an average of 1.02 SNVs per 1 kb along the gene body in the marine group. This frequency was insufficient to produce the observed differences in the numbers of mapped reads. Therefore, the mapping efficiency bias was unlikely to explain the large number of the freshwater-increased GCNVs.

GCNVs that were likely due to duplication or multiplication

To confirm whether the detected GCNVs under parallel selection were due to duplications or multiplications in the freshwater group, I counted the numbers of different allelic sequences within the regions of the GCNVs (**Figure 3.1C**). Two freshwater-increased GCNVs

(ENSGACG00000003408 and *APOL2*) (**Figures 3.4AB**) were strongly predicted to be such GCNVs, because they were supported by at least two within-read-length SNV position pairs in three individuals of the freshwater group (**Tables 3.2 and 3.3**). Read depths along the genomic coordinates were not stable probably due to sequencing biases, thus their differences were clearly observed in the regions with large read depths. It was notable that the read depths in the intronic regions of *APOL2* of the freshwater group were higher than those of the marine group (**Figure 3.4B**), suggesting that this gene was recently duplicated with their intronic sequences. In addition, multiple copies of one freshwater-decreased GCNV (ENSGACG00000003374) (**Figure 3.4C**) were predicted to exist on the genomes of the marine group by the same analysis on the marine group. Another freshwater-decreased GCNV (*NXPE3*) was also supported by at least one within-read-length SNV position pair in three individuals of the marine group (**Tables 3.2 and 3.3**). The copy numbers of these two genes (ENSGACG00000003374 and *NXPE3*) would have decreased during the adaptation to the freshwater environment.

The APOL2 gene is a member of the apolipoprotein L gene family. This gene family is involved in pathogen immunity and was previously reported to have been under positive selection in primates (Smith and Malik, 2009). Another previous study found copy number differences in the APOL1 gene between human and chimpanzee and suggested that these differences were involved in the adaptive phenotype differentiation of the inflammatory response (Perry *et al*., 2008). The duplications or multiplications of APOL2 might have contributed to adaption of the immune system to the freshwater environment. For ENSGACG00000003408, I conducted BLASTX searches against NCBI nr database because no functional descriptions were available in the Ensembl database. The best hit for this gene was a neoverrucotoxin subunit alpha-like gene of *Oreochromis niloticus* with E-value = 0.0 (Accession numbers of the hits were XP_003449498, XP_003449506, and XP_003449483). This gene was reported to be overexpressed in the brooding tissue of pregnant specimens of a species in genus Syngnathus (Small *et al*., 2013), which belongs to the same order as the three-spined stickleback does. The duplications or multiplications of ENSGACG00000003408 might have had roles in pregnancy functions in the freshwater environment. I could not obtain any hit for ENSGACG00000003374. A previous study reported GCNVs of *NXPE3* within marine populations (Feulner *et al*., 2013). *NXPH3* is a neuropeptide-like molecule that functions in brain (Beglopoulos *et al*., 2005), and neuropeptides were suggested to control migratory behaviors (Mueller *et al*., 2011). The decrease of the *NXPE3* copy numbers in the freshwater group might have been associated with their anadromous behavior (Bell and Foster, 2000).

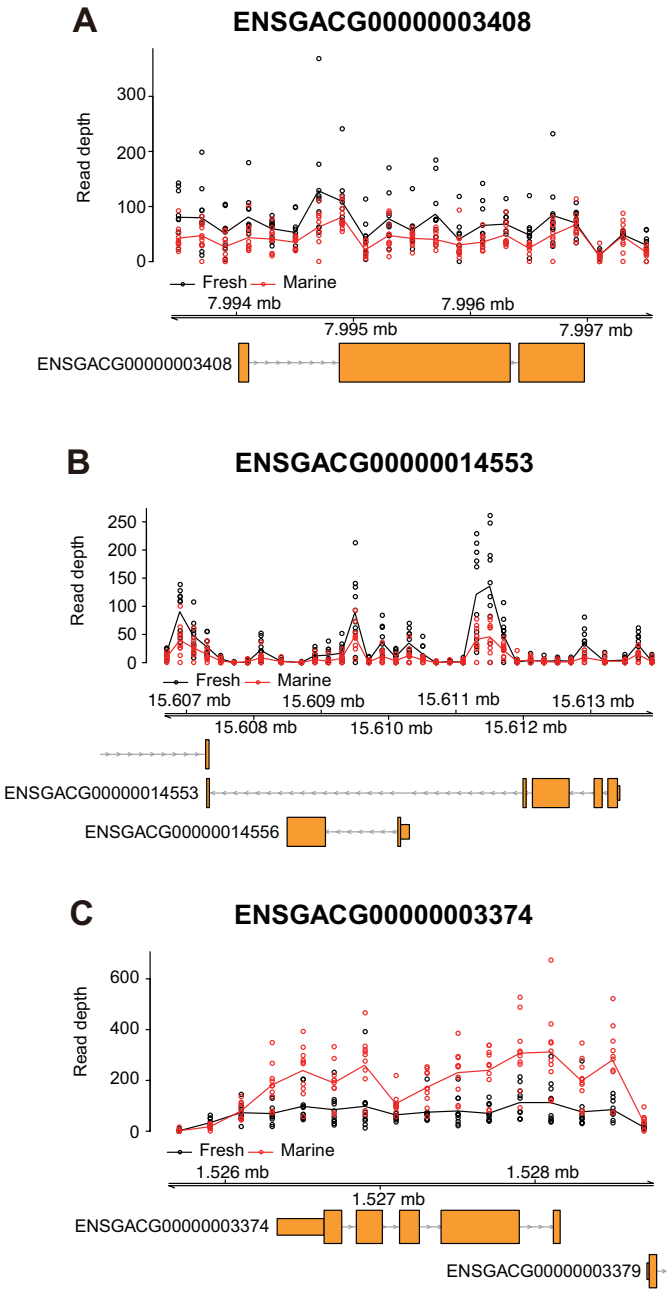

Figure 3.4: Numbers of mapped reads in two freshwater-increased and one freshwater-decreased GCNVs

Each point and line represent the normalized numbers and average normalized numbers, respectively, of the mapped reads per 200-bp non-overlapping window for 10 freshwater (black) and 10 marine (red) individuals. (A and B) Two freshwater-increased and (C) one freshwater-decreased GCNVs that were confirmed by three or more different allelic sequences, are shown. Gene models are shown at the bottom of each panel.

Table 3.3: Numbers of SNV pairs in which three or more haplotypes were observed based on '-e 100' mapping condition

Ensembl gene ID	Group having more copies	Freshwater									Marine									
		ABW	BIGL	FTC	HUTU	MATA	MUDL	NOST	PAXB	SCX	SHEL	ANTL	BDGB	BIGR	GJOG	GORT	JAMA	JMRP	NEU	RABS	SALR
ENSGACG00000002551	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000002682	Freshwarer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSGACG00000002744	Freshwater	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
ENSGACG00000002857	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000002886	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000002902	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000002913	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000002918	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000002933	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000003408	Freshwater	0	0	1	6	1	0	0	0	3	0	0	0	2	0	0	0	0	0	0	0
ENSGACG00000006397	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000014268	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000014289	Freshwater	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000014553	Freshwater	0	4	3	1	5	0	0	0	3	0	0	0	0	0	0	1	0	0	0	0
ENSGACG00000015099	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000018214	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000019313	Freshwater	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000019321	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000020171	Freshwater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000003374	Marine	1	0	0	0	0	0	1	3	0	1	0	0	0	10	4	6	1	3	9	1
ENSGACG00000003379	Marine	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSGACG00000005313	Marine	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSGACG00000019508	Marine	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0
ENSGACG00000020238	Marine	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0

Differential expressions of genes between the two environments

If the two strongly supported freshwater-increased GCNVs actually contributed to the parallel evolution of the three-spined sticklebacks, the amount of transcription products of these genes should be important for the adaptation. Thus, I analyzed microarray data of gills of three-spined sticklebacks in marine and freshwater groups under the short and long photoperiod conditions (Kitano *et al*., 2010), and evaluated whether these two genes showed significant differential expressions between the two groups. As expected, the gene expression values of *APOL2* and ENSGACG00000003408 were higher in the freshwater group than those in the marine group highly significantly ($p < 0.005$ after Bonferroni correction) under the short photoperiod condition (**Table 3.4**). The short photoperiod condition resembled winter, thus these genes might have contributed to the fitness though the overwintering survival (Barrett *et al*., 2008).

In addition to the above two freshwater-increased GCNVs, another freshwater-increased GCNV, ENSGACG00000002551, showed higher expression in freshwater than in marine groups under the short-photoperiod condition (**Table 3.4**). Although the analysis of the numbers of different allelic sequences in the present study did not detect duplication for the GCNV, it is possible that the actually duplicated GCNVs were overlooked owing to low coverage of the resequencing data. Thus, ENSGACG00000002551 would be the important candidate for future studies to investigate GCNVs.

The observed increase in expression levels of freshwater-increased GCNVs in the freshwater group might possible be due to a potential bias toward higher expression levels in freshwater groups in the microarray data used in the present study. To assess this possibility, I searched for differentially expressed genes between marine and freshwater groups irrespective of the genes being GCNVs identified in this study. Of 22913 genes, 374 genes showed higher expression in marine individuals and 385 genes showed higher expression in freshwater individuals (adjusted p-value < 0.05) (**Figure 3.5**). This indicates that differences in gene expression levels were not biased toward higher expression in freshwater groups, further supporting that the increased abundance of mRNAs of GCNVs might have contributed to adaptation to the freshwater environments.

Table 3.4: GCNVs that showed higher expression in freshwater than marine groups in gills under the short-photoperiod condition

Ensembl gene ID	Ensembl transcript ID	log2FC	*p*-value	Adjusted *p*-	Duplication detected
ENSGACG00000014553	ENSGACT00000019239	-1.3725	5.255E-04	3.227E-02	Yes
ENSGACG00000003408	ENSGACT00000004466	-1.2225	5.706E-04	3.329E-02	Yes
ENSGACG00000002551	ENSGACT00000003350	-1.1700	8.343E-04	3.787E-02	No

(a) Logarithms 2 of fold changes of gene expression levels in marine groups over those of freshwater groups.
(b) The p-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

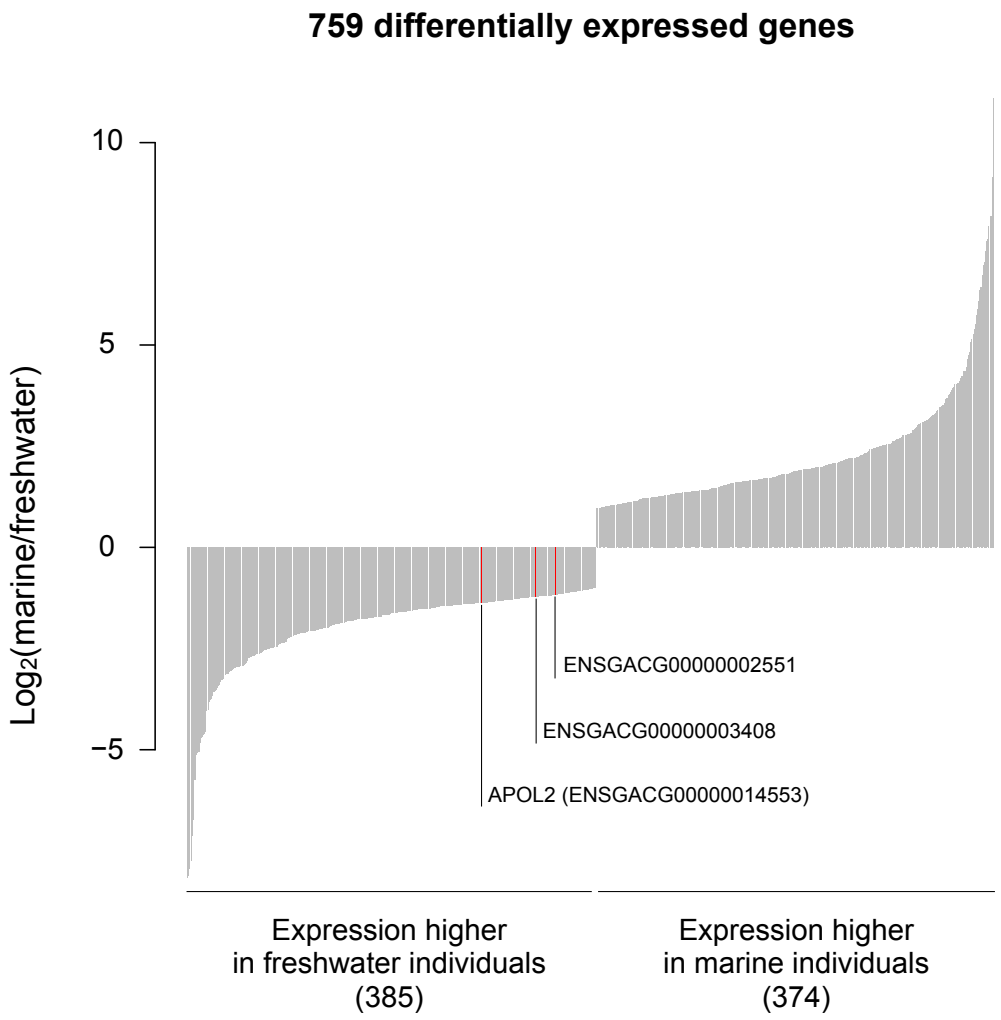

Figure 3.5: Comparison of differentially expressed genes between marine and freshwater groups by microarray analysis

The 759 genes with differential expression in the gill between marine and freshwater groups under the short photoperiod are shown (adjusted *p*-value < 0.05). The y-axis represents log2 of fold changes of expression levels (marine/freshwater). The positive and negative values of the y-axis indicate higher expression in marine and freshwater groups, respectively. The x-axis represents genes ordered by the fold changes. Red bars indicate GCNVs for which differential expressions between marine and freshwater groups were detected.

Conclusion

In this study, I showed the possibility that GCNVs underwent positive selection in the parallel evolution of the three-spined sticklebacks and had a role in the adaptation to the freshwater environment. It would be notable that many CNVs were found in a marine population of three-spined sticklebacks (Feulner *et al*., 2013), which suggests the existence of globally shared, standing CNVs that can contribute to the parallel evolution within natural population. These results suggest that the contribution of GCNVs should be considered in studies on adaptive evolution of diverse species.

References

Chapter 1: General introduction

Francis Crick. (1970). Central dogma of molecular biology. *Nature* **227**, 561-563.

Theodosius Dobzhansky. (1973). Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **35,** 125-129.

Robert D. Fleischmann, *et al.* (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512.

François Jacob. (1998). Of Flies, Mice, and Men. (Giselle Weiss, Trans.) New York, Harvard University Press.

Chapter 2: CLOCK-controlled regulations of circadian rhythms through canonical and non-canonical E-Boxes

T L Bailey and C Elkan. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28-36.

Timothy L. Bailey. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653-1659.

Joseph Bass and Joseph S Takahashi. (2010). Circadian integration of metabolism and energetics. *Science* **330**, 1349-1354.

Lars Bertram, Mikko Hiltunen, Michele Parkinson, Martin Ingelsson, Christoph Lange, Karunya Ramasamy, Kristina Mullin, Rashmi Menon, Andrew J Sampson, Monica Y Hsiao, Kathryn J Elliott, Gonül Velicelebi, Thomas Moscarillo, Bradley T Hyman, Steven L Wagner, K David Becker, Deborah Blacker, and Rudolph E Tanzi. (2005). Family-based association between Alzheimer's disease and variants in UBQLN1. *N. Engl. J. Med.* **352**, 884-894.

Steven A. Brown, Elzbieta Kowalska, and Robert Dallmann. (2012). (Re)inventing the Circadian Feedback Loop. *Dev. Cell* **22**, 477-487.

Alexander Maxwell Burroughs, Yoshinari Ando, Michiel Jan Laurens de Hoon, Yasuhiro Tomaru, Harukazu Suzuki, Yoshihide Hayashizaki, and Carsten Olivier Daub. (2011). Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA Biol.* **8**, 158-177.

Patricia P. Chan and Todd M. Lowe. (2009). GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93-D97.

Hai Ying M Cheng, Joseph W. Papp, Olga Varlamova, Heather Dziema, Brandon Russell, John P. Curfman, Takanobu Nakazawa, Kimiko Shimizu, Hitoshi Okamura, Soren Impey, and Karl Obrietan. (2007). microRNA Modulation of Circadian-Clock Period and Entrainment. *Neuron* **54**, 813-829.

Anne Dueck, Christian Ziegler, Alexander Eichner, Eugene Berezikov, and Gunter Meister. (2012). MicroRNAs associated with the different human Argonaute proteins. *Nucleic Acids Res.* **40**, 9850-9862.

S. Falcon and R. Gentleman. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258.

David M Garcia, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.* **18**, 1139-1146.

Sarah Geisler and Jeff Coller. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14**, 699-712.

N Gekakis, D Staknis, H B Nguyen, F C Davis, L D Wilsbacher, D P King, J S Takahashi, and C J Weitz. (1998). Role of the CLOCK protein in the mammalian circadian mechanism. *Science* **280**, 1564-1569.

Yuchun Guo, Georgios Papachristoudis, Robert C. Altshuler, Georg K. Gerber, Tommi S. Jaakkola, David K. Gifford, and Shaun Mahony. (2010). Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26**, 3028-3034.

Chelsea L. Gustafson and Carrie L. Partch. (2015). Emerging Models for the Molecular Basis of Mammalian Circadian Timing. *Biochemistry* **54**, 134-149.

Annakaisa Haapasalo, Jayashree Viswanathan, Kaisa Ma Kurkinen, Lars Bertram, Hilkka Soininen, Nico P Dantuma, Rudolph E Tanzi, and Mikko Hiltunen. (2011). Involvement of ubiquilin-1 transcript variants in protein degradation and accumulation. *Commun. Integr. Biol.* **4**, 428-32.

Holger Hartmann, Eckhart W Guthöhrlein, Matthias Siebert, Sebastian Luehr, and Johannes Söding. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.* **23**, 181-194.

Fumiyuki Hatanaka, Chiaki Matsubara, Jihwan Myung, Takashi Yoritaka, Naoko Kamimura, Shuichi Tsutsumi, Akinori Kanai, Yutaka Suzuki, Paolo Sassone-Corsi, Hiroyuki Aburatani, Sumio Sugano, and Toru Takumi. (2010). Genome-wide profiling of the core clock protein BMAL1 targets reveals a strict relationship with metabolism. *Mol. Cell. Biol.* **30**, 5636-5648.

Yuanming Hu, Cindy Leo, Simon Yu, Betty C B Huang, Hank Wang, Mary Shen, Ying Luo, Sarkiz Daniel Issakani, Donald G. Payan, and Xiang Xu. (2004). Identification and functional characterization of a novel human Misshapen/Nck interacting kinase-related kinase, hMINKβ. *J. Biol. Chem.* **279**, 54387-54397.

Michael E. Hughes, Luciano DiTacchio, Kevin R. Hayes, Christopher Vollmers, S. Pulivarthy, Julie E. Baggs, Satchidananda Panda, and John B. Hogenesch. (2009). Harmonics of circadian gene transcription in mammals. *PLoS Genet.* **5**, e1000442.

Sebastian Kadener, Jerome S. Menet, Ken Sugino, Michael D. Horwich, Uri Weissbein, Pipat Nawathean, Vasia V. Vagin, Phillip D. Zamore, Sacha B. Nelson, and Michael Rosbash. (2009). A role for microRNAs in the Drosophila circadian clock. *Genes Dev.* **23**, 2179-2191.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Yota B. Kiyohara, Keigo Nishii, Maki Ukai-Tadenuma, Hiroki R. Ueda, Yasuo Uchiyama, and Kazuhiro Yagita. (2008). Detection of a circadian enhancer in the mDbp promoter using prokaryotic transposon vector-based strategy. *Nucleic Acids Res.* **36**, e23.

N. Koike, S.-H. Yoo, H.-C. Huang, V. Kumar, C. Lee, T.-K. Kim, and J. S. Takahashi. (2012). Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. *Science* **338**, 349-354.

Shihoko Kojima, Danielle L Shingle, and Carla B Green. (2011). Post-transcriptional control of circadian rhythms. *J. Cell Sci.* **124**, 311-320.

Ana Kozomara and Sam Griffiths-Jones. (2011). MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152-157.

Jacek Krol, Inga Loedige, and Witold Filipowicz. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* **11**, 597-610.

I. V. Kulakovskiy, V. a. Boeva, a. V. Favorov, and V. J. Makeev. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* **26**, 2622-2623.

Yuichi Kumaki, Maki Ukai-Tadenuma, Ken-ichiro D Uno, Junko Nishio, Koh-hei Masumoto, Mamoru Nagano, Takashi Komori, Yasufumi Shigeyoshi, John B Hogenesch, and Hiroki R Ueda. (2008). Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14946-14951.

B Langmead, C Trapnell, M Pop, and SL Salzberg. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10**, R25.

Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20.

Heng Li and Richard Durbin. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Dao-Hong Lin, Peng Yue, Chunyang Pan, Peng Sun, and Wen-Hui Wang. (2011). MicroRNA 802 stimulates ROMK channels by suppressing caveolin-1. *J. Am. Soc. Nephrol*. **22**, 1087-1098.

Phillip L Lowrey and Joseph S Takahashi. (2004). Mammalian circadian biology: elucidating genome-wide levels of temporal organization. *Annu. Rev. Genomics Hum. Genet*. **5**, 407-441.

Xiaotu Ma, Ashwinikumar Kulkarni, Zhihua Zhang, Zhenyu Xuan, Robert Serfling, and Michael Q. Zhang. (2012). A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res*. **40**, e50.

Philip Machanick and Timothy L. Bailey. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-1697.

John Majercak, Wen-Feng Chen, and Isaac Edery. (2004). Splicing of the period gene 3'-terminal intron is regulated by light, circadian clock factors, and phospholipase C. *Mol. Cell. Biol*. **24**, 3359-3372.

Marcel Martin. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10.

Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, François Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih Yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. (2014). JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*. **42**, D142-147.

Nicholas J McGlincy, Amandine Valomon, Johanna E Chesham, Elizabeth S Maywood, Michael H Hastings, and Jernej Ule. (2012). Regulation of alternative splicing by the circadian clock and food related cues. *Genome Biol*. **13**, R54.

Gunter Meister. (2013). Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet*. **14**, 447-59.

Jerome S. Menet, Joseph Rodriguez, Katharine C. Abruzzi, and Michael Rosbash. (2012). Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *Elife* **1**, :e00011.

Brooke H Miller, Erin L McDearmon, Satchidananda Panda, Kevin R Hayes, Jie Zhang, Jessica L Andrews, Marina P Antoch, John R Walker, Karyn A Esser, John B Hogenesch, and Joseph S Takahashi. (2007). Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation. *Proc. Natl. Acad. Sci. U. S. A*. **104**, 3342-3347.

Hideki Nagasaki, Masanori Arita, Tatsuya Nishizawa, Makiko Suwa, and Osamu Gotoh. (2006). Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics* **22**, 1211-1216.

I Pérez, C H Lin, J G McAfee, and J G Patton. (1997). Mutation of PTB binding sites causes misregulation of alternative 3' splice site selection in vivo. *RNA* **3**, 764-778.

Guillaume Rey, François Cesbron, Jacques Rougemont, Hans Reinke, Michael Brunner, and Felix Naef. (2011). Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biol*. **9**, e1000595.

Jürgen A Ripperger and Ueli Schibler. (2006). Rhythmic CLOCK-BMAL1 binding to multiple E-box motifs drives circadian Dbp transcription and chromatin transitions. *Nat. Genet*. **38**, 369-374.

Alexei A. Sharov and Minoru S H Ko. (2009). Exhaustive search for over-represented DNA sequence motifs with cisfinder. *DNA Res*. **16**, 261-273.

G D Stormo. (2000). DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23.

Joseph S Takahashi, Hee-Kyung Hong, Caroline H Ko, and Erin L McDearmon. (2008). The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nat. Rev. Genet.* **9**, 764-775.

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178-192.

Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46-53.

Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 516-520.

Hiroki R Ueda, Satoko Hayashi, Wenbin Chen, Motoaki Sano, Masayuki Machida, Yasufumi Shigeyoshi, Masamitsu Iino, and Seiichi Hashimoto. (2005). System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nat. Genet.* **37**, 187-192.

Maki Ukai-Tadenuma, Rikuhiro G Yamada, Haiyan Xu, Jürgen a Ripperger, Andrew C Liu, and Hiroki R Ueda. (2011). Delay in feedback repression by cryptochrome 1 is required for circadian clock function. *Cell* **144**, 268-81.

Hideki Ukai and Hiroki R Ueda. (2010). Systems biology of mammalian circadian clocks. *Annu. Rev. Physiol.* **72**, 579-603.

Christopher Vollmers, Robert J. Schmitz, Jason Nathanson, Gene Yeo, Joseph R. Ecker, and Satchidananda Panda. (2012). Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. *Cell Metab.* **16**, 833-845.

Dongmei Wang, Zhaojie Zhang, Evan O'Loughlin, Thomas Lee, Stephane Houel, Dónal O'Carroll, Alexander Tarakhovsky, Natalie G. Ahn, and Rui Yi. (2012). Quantitative functions of argonaute proteins in mammalian development. *Genes Dev.* **26**, 693-704.

K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. a. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, 1-14.

Seung-Hee Yoo, Caroline H Ko, Phillip L Lowrey, Ethan D Buhr, Eun-joo Song, Suhwan Chang, Ook Joon Yoo, Shin Yamazaki, Choogon Lee, and Joseph S Takahashi. (2005). A noncanonical E-box enhancer drives mouse Period2 circadian oscillations in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2608-13.

Hikari Yoshitane, Sato Honma, Kiyomichi Imamura, Hiroto Nakajima, Shin-ya Nishide, Daisuke Ono, Hiroshi Kiyota, Naoya Shinozaki, Hirokazu Matsuki, Naoya Wada, Hirofumi Doi, Toshiyuki Hamada, Ken-ichi Honma, and Yoshitaka Fukada. (2012). JNK regulates the photic response of the mammalian circadian clock. *EMBO Rep.* **13**, 455-461.

Federico Zambelli, Graziano Pesole, and Giulio Pavesi. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief. Bioinform.* **14**, 225-37.

Chapter 3: Positive selection on gene copy number variations in adaptive evolution

Alexej Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974-84.

Joshua M. Akey, Michael A. Eberle, Mark J. Rieder, Christopher S. Carlson, Mark D. Shriver, Deborah A. Nickerson, and Leonid Kruglyak. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*. **2**, e286.

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389-3402.

Rowan D H Barrett, Sean M Rogers, and Dolph Schluter. (2008). Natural selection on a major armor gene in threespine stickleback. *Science* **322**, 255-257.

Rowan D. H. Barrett and Hopi E. Hoekstra. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet*. **12**, 767-780.

Michael A Bell and Susan A Foster. (1994). The Evolutionary Biology of the Threespine Stickleback. Oxford University Press.

Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. (2010). GenBank. *Nucleic Acids Res*. **38**, D46-D51.

Shameek Biswas and Joshua M. Akey. (2006). Genomic insights into positive selection. *Trends Genet*. **22**, 437-446.

K. H. Brown, K. P. Dobrinski, A. S. Lee, O. Gokcumen, R. E. Mills, X. Shi, W. W. S. Chong, J. Y. H. Chen, P. Yoo, S. David, S. M. Peterson, T. Raj, K. W. Choy, B. E. Stranger, R. E. Williamson, L. I. Zon, J. L. Freeman, and C. Lee. (2012). Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proc. Natl. Acad. Sci*. **109**, 529-534.

Zuozhou Chen, C-H Christina Cheng, Junfang Zhang, Lixue Cao, Lei Chen, Longhai Zhou, Yudong Jin, Hua Ye, Cheng Deng, Zhonghua Dai, Qianghua Xu, Peng Hu, Shouhong Sun, Yu Shen, and Liangbiao Chen. (2008). Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci. U. S. A*. **105**, 12944-12949.

V G Cheung, R S Spielman, K G Ewens, T M Weber, M Morley, and J T Burdick. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365-1369.

Pamela F Colosimo, Kim E Hosemann, Sarita Balabhadra, Guadalupe Villarreal, Mark Dickson, Jane Grimwood, Jeremy Schmutz, Richard M Myers, Dolph Schluter, and David M Kingsley. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**, 1928-1933.

Pamela F. Colosimo, Catherine L. Peichel, Kirsten Nereng, Benjamin K. Blackman, Michael D. Shapiro, Dolph Schluter, and David M. Kingsley. (2004). The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biol*. **2**, e109.

Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun Hwa Ihm, Kati Kristiansson, Daniel G Macarthur, Jeffrey R Macdonald, Ifejinelo Onyiah, Andy Wing Chun Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, Chris Tyler-Smith, Nigel P Carter, Charles Lee, Stephen W Scherer, and Matthew E Hurles. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712.

Seth Debolt. (2010). Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales. *Genome Biol. Evol*. **2**, 441-453.

J J Emerson, Margarida Cardoso-Moreira, Justin O Borevitz, and Manyuan Long. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**, 1629-1631.

Lars Feuk, Lars Feuk, Andrew R Carson, Andrew R Carson, Stephen W Scherer, and Stephen W Scherer. (2006). Structural variation in the human genome. *Nat. Rev. Genet*. **7**, 85-97.

Philine G D Feulner, Frédéric J J Chain, Mahesh Panchal, Christophe Eizaguirre, Martin Kalbe, Tobias L. Lenz, Marvin Mundry, Irene E. Samonte, Monika Stoll, Manfred Milinski, Thorsten B H Reusch, and Erich Bornberg-Bauer. (2013).

Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Mol. Ecol.* **22**, 635-649.

Elodie Gazave, Fleur Darré, Carlos Morcillo-Suarez, Natalia Petit-Marty, Angel Carreño, Urko M Marigorta, Oliver a Ryder, Antoine Blancher, Mariano Rocchi, Elena Bosch, Carl Baker, Tomàs Marquès-Bonet, Evan E Eichler, and Arcadi Navarro. (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res.* **21**, 1626-1639.

Robert E Handsaker, Joshua M Korn, James Nemesh, and Steven A McCarroll. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269-276.

Scott J. Hebbring, Araba A. Adjei, Janel L. Baer, Gregory D. Jenkins, Jianping Zhang, Julie M. Cunningham, Daniel J. Schaid, Richard M. Weinshilboum, and Stephen N. Thibodeau. (2007). Human SULT1A1 gene: Copy number differences and functional implications. *Hum. Mol. Genet.* **16**, 463-470.

Ary a Hoffmann and Yvonne Willi. (2008). Detecting genetic responses to environmental change. *Nat. Rev. Genet.* **9**, 421-432.

T. Hubbard. (2002). The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38-41.

Felicity C. Jones, Manfred G. Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, Mono Pirun, Michael C. Zody, Simon White, Ewan Birney, Stephen Searle, Jeremy Schmutz, Jane Grimwood, Mark C. Dickson, Richard M. Myers, Craig T. Miller, Brian R. Summers, Anne K. Knecht, Shannon D. Brady, Haili Zhang, Alex a. Pollen, Timothy Howes, Chris Amemiya, Jen Baldwin, Toby Bloom, David B. Jaffe, Robert Nicol, Jane Wilkinson, Eric S. Lander, Federica Di Palma, Kerstin Lindblad-Toh, and David M. Kingsley. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61.

Jun Kitano, S.C. Lema, J.A. Luckenbach, Seiichi Mori, Yui Kawagishi, Makoto Kusakabe, Penny Swanson, and C.L. Peichel. (2010). Adaptive divergence in the thyroid hormone signaling pathway in the stickleback radiation. *Curr. Biol.* **20**, 2124-2130.

Thomas D Kocher. (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5**, 288-298.

Fyodor A. Kondrashov and Alexey S. Kondrashov. (2006). Role of selection in fixation of gene duplications. *J. Theor. Biol.* **239**, 141-151.

Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. (2011). The sequence read archive. *Nucleic Acids Res.* **39**, D54-D56.

Jakob C Mueller, Francisco Pulido, and Bart Kempenaers. (2011). Identification of a gene associated with avian migratory behaviour. *Proc. R. Soc. B* **278**, 2848-2856.

Duc Quang Nguyen, Caleb Webber, and Chris P. Ponting. (2006). Bias of selection on human copy-number variants. *PLoS Genet.* **2**, 198-207.

Rasmus Nielsen. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197-218.

George H. Perry, Fengtang Yang, Tomas Marques-Bonet, Carly Murphy, Tomas Fitzgerald, Arthur S. Lee, Courtney Hyland, Anne C. Stone, Matthew E. Hurles, Chris Tyler-Smith, Evan E. Eichler, Nigel P. Carter, Charles Lee, and Richard Redon. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Res.* **18**, 1698-1710.

Peter J. Prentis, John R U Wilson, Eleanor E. Dormontt, David M. Richardson, and Andrew J. Lowe. (2008). Adaptive evolution in invasive species. *Trends Plant Sci.* **13**, 288-294.

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

M D Robinson, D J McCarthy, and G K Smyth. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140.

H D Rundle, L Nagel, J Wenrick Boughman, and D Schluter. (2000). Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**, 306-308.

Jörn P Scharsack, Martin Kalbe, Chris Harrod, and Gisep Rauch. (2007). Habitat-specific adaptation of immune responses of stickleback (Gasterosteus aculeatus) lake and river ecotypes. *Proc. R. Soc. B* **274**, 1523-1532.

Clayton M. Small, April D. Harlin-Cognato, and Adam G. Jones. (2013). Functional similarity and molecular divergence of a novel reproductive transcriptome in two male-pregnant Syngnathus pipefish species. *Ecol. Evol.* **3**, 4092-4108.

Eric E. Smith and Harmit S. Malik. (2009). The apolipoprotein L family of programmed cell death and immunity genes rapidly evolved in primates at discrete sites of host-pathogen interactions. *Genome Res.* **19**, 850-858.

G Smyth. (2005). limma: Linear Models for Microarray Data. In *Bioinforma. Comput. Biol. Solut. Using R Bioconductor*.

B Langmead, C Trapnell, M Pop, and SL Salzberg. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

R. D H Barrett and Dolph Schluter. (2008). Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38-44.

Vassilios Beglopoulos, Monique Montag-Sallaz, Astrid Rohlmann, Kerstin Piechotta, Mohiuddin Ahmad, Dirk Montag, and Markus Missler. (2005). Neurexophilin 3 is highly localized in cortical and cerebellar regions and is functionally important for sensorimotor gating and motor coordination. *Mol. Cell. Biol.* **25**, 7278-7288.

Christopher S. Carlson, Daryl J. Thomas, Michael A. Eberle, Johanna E. Swanson, Robert J. Livingston, Mark J. Rieder, and Deborah A. Nickerson. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15**, 1553-1565.

Julie M. Cridland and Kevin R. Thornton. (2010). Validation of rearrangement break points identified by paired-end sequencing in natural populations of Drosophila melanogaster. *Genome Biol. Evol.* **2**, 83-101.

Omer Gokcumen, Paul L Babb, Rebecca C Iskow, Qihui Zhu, Xinghua Shi, Ryan E Mills, Iuliana Ionita-Laza, Eric J Vallender, Andrew G Clark, Welkin E Johnson, and Charles Lee. (2011). Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol.* **12**, R52.

Daisuke Ikeda, Yosuke Ono, Phil Snell, Yvonne J K Edwards, Greg Elgar, and Shugo Watabe. (2007). Divergent evolution of the myosin heavy chain gene family in fish and tetrapods: evidence from comparative genomic analysis. *Physiol. Genomics* **32**, 1-15.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Steven A McCarroll and David M Altshuler. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37-S42.

J S McKinnon and H D Rundle. (2002). Speciation in nature:the threespine stickleback model systems. *Trends Ecol. Evol.* **17**, 480-488.

Aaron R. Quinlan, Royden a. Clark, Svetlana Sokolova, Mitchell L. Leibowitz, Yujun Zhang, Matthew E. Hurles, Joshua C. Mell, and Ira M. Hall. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* **20**, 623-635.