

UNDERSTANDING THE GENETIC BASIS OF NATURAL  
VARIATION IN THE REGULATION OF CIRCADIAN CLOCK OF  
*NEUROSPORA CRASSA*

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

by

Tae Sung Kim

January 2009

© 2009 Tae Sung Kim

UNDERSTANDING THE GENETIC BASIS OF NATURAL  
VARIATION IN THE REGULATION OF CIRCADIAN CLOCK OF  
*NEUROSPORA CRASSA*

Tae Sung Kim, Ph. D.

Cornell University 2009

Circadian clock has been found in all forms of life from bacteria to humans. Its biological function is thought to provide organisms with time keeping ability, which enables organisms to control their behavioral, physiological and cellular activities efficiently on daily basis environmental changes.

Over the past four decades, *Neurospora crassa* has been developed as a model organism for the study of circadian clocks. However, despite the intensive molecular characterizations of the Neurospora circadian clock, our understanding of this system is far from comprehensive.

Quantitative Trait Loci (QTL) analyses, using natural strains, have been successfully utilized over the past decade to dissect complex traits down to a naturally occurring polymorphism that is relevant to phenotypic variations. The high quality genomic sequence and sophisticated molecular biology tools, in combination with the QTL analysis, may make it possible to increase the understanding of mechanisms of circadian regulation and may also provide insights into the biological role of the circadian clock, especially in the process of adapting to local environments, a topic that is somewhat overlooked in current research.

In this work, I have explored an alternative strategy to uncover new perspectives in the Neurospora circadian clock. My research has laid the groundwork for QTL analysis and has demonstrated QTL analysis of the clock phenotypes, period and entrained phase using natural populations.

In chapter II, I describe the computational, statistical and genetic analyses performed to evaluate the marker potential of *Neurospora* simple sequence repeat (SSR) and to investigate the biological role of the SSR

In chapter III, I describe the research regarding the development of two important bioinformatic tools which include 1) a genetic marker management system which facilitates QTL analysis and subsequent positional cloning steps, and 2) an automatic image processing system for the *Neurospora* circadian clock phenotype.

Lastly, in chapter IV, I describe the results of QTL analysis for the two clock phenotypes (period, phase) in three natural F1 populations using two independent statistical methods. Subsequently, I confirmed the QTL effects of one of those in the BC4 generation which were predicted from the F1 populations by constructing near isogenic lines (NIL).

## BIOGRAPHICAL SKETCH

Tae Sung Kim was born as the first son of Bong Lim Park and Hyeung Jin Kim on January 17<sup>th</sup>, 1973. He grew up with his parents and sisters, Sung Hee Kim, Sung Yeun Kim and Sung Min Kim on Je-ju island in Korea, blessed with a diverse landscape and a subtropical climate. From 1992 to 1999, he studied at Kyung Hee University, Korea, attaining his bachelor degree after majoring in horticultural biotechnology at College of Life Science. He continued to his study in the graduate program of Plant Science at Seoul National University and completed a Master in Science degree in 2001. While in graduate school, he worked as a research assistant in the Vegetable Crops Breeding Lab, studying the genetic transformation of the insecticidal Bt gene (*cry1Ac*) into hot peppers (*Capcicum annum*), one of the major vegetable crops in Korea. After graduating from Seoul National University, he moved to the United States in 2002 to study plant genetics in the Department of Plant Breeding and Genetics at Cornell University. Under the direction of Dr. Walter DeJong in this department, Tae Sung developed a strong interest in the evolutionary mechanisms of allelic variation in complex traits. In 2004, his passion for the topic led him to change his field of study to Neurospora circadian clock variation under the guidance of Kwangwon Lee in the Department of Plant Pathology and Plant-Microbe Biology.

## ACKNOWLEDGMENTS

First of all, I would truly like to thank God, my Lord, for guiding me up to this point. I would also like to acknowledge my major advisor and committee members, Dr. Kwangwon Lee, Micheal Milgroom and Susan McCouch for their sincere and wonderful academic guidance during my Ph.D program. I would like to extend a special thanks to the chairperson of my academic program, Dr. Kwangwon Lee, for supporting me financially and reminding me of the reason why I need to continue this journey. Thanks to all of my committee members, I have been able to appreciate the joys of learning something new. I also appreciate the help and dedication I have received from my collaborators, Dr. James G. Booth, Hugh G. Gauch, Jr., Dr. Qi Sun , Dr. Jason Mezey, Benjamin A. Logsdon, Noah Whitman, Jongsun Park and Dr. Yong-Hwan Lee. I also am thankful to my current and previous lab members, especially, Sohyun Park, Jung Il Bae, Anna Yarusskaya and Lisa Huang for their friendship and assistance in my lab work. I am grateful to all my friends, colleagues and faculty members who helped me at some point of my Ph.D. Program in the Departments of Plant Pathology & Plant-Microbe Biology and Plant Breeding & Genetics.

Finally, I always appreciate the continuous support of my family here in Ithaca, my wife Min A Choi and lovely daughter Sau Yoon, as well as my family in Korea, Hyueng Jin Kim, Bong Lim Park and my sisters.

## TABLE OF CONTENTS

<b>BIOGRAPHICAL SKETCH</b>	iii
<b>ACKNOWLEDGMENTS</b>	iv
<b>LIST OF FIGURES</b>	ix
<b>LIST OF TABLES</b>	xii
<b>Chapter 1: GENERAL INTRODUCTION</b>	1
Intrinsic properties of circadian clock	1
Conceptual structure of circadian clock	1
Mechanism of autoregulatory cellular oscillation	2
Biological role of circadian clock	4
<i>Neurospora crassa</i> as a model system in the chronobiology	6
Circadian clock study in <i>Neurospora crassa</i> .	8
Summary of the chapters	11
Reference	13
<b>Chapter 2: SIMPLE SEQUENCE REPEATS IN NEUROSPORA</b>	22
<b>CRASSA: DISTRIBUTION, POLYMORPHISM AND</b>	
<b>EVOLUTIONARY INFERENCE</b>	
Introduction; Simple Sequence Repeats in <i>Neurospora crassa</i>	22
Method; SSR analysis in the Neurospora genome sequence	24
Method; Statistical frame work to infer SSR genesis rate in chromosomes and genomic locations	24
Method; SSR marker development from Neurospora genome	25
Method; linkage mapping analysis of Neurospora genome	27
Experiment and Result; Genome-wide distribution of SSRs by the SSR unit size	27

## **Chapter 2 (continued)**

Experiment and Result; SSR genesis rate in chromosomes and genomic locations	34
Experiment and Result; Characterizations of size polymorphism in SSRs among natural accessions	39
Experiment and Result; The statistical inference for evolutionary forces of size variation of SSRs	42
Experiment and Result; Genetic map construction	55
Discussion; Distribution of SSRs in the sequenced <i>N. crassa</i> genome	57
Discussion; The potential role of AAR encoded by tri-nt SSRs	63
Discussion; Evolutionary inference of SSR variations in <i>N. crassa</i>	65
Discussion; The marker potential of Neurospora SSR	66
Discussion; SSR based genetic map construction for <i>N. crassa</i> genome	68
Reference	70

## **Chapter 3: TOOL DEVELOPMENTS TO FACILITATE QTL ANALYSIS AND SUBSEQUENT POSITIONAL CLONING** 82

### **PROCEDURES**

Introduction; The need for tool development for the efficient QTL analysis	82
Introduction; MoMMS: A <u>Molecular</u> <u>Marker</u> <u>Management</u> <u>System</u>	83
Introduction; Circamate, the high through-put image processing system for the circadian clock analysis	86
Experiment and Result; Establishment of a phase evaluation function which is compatible to cycling condition	88
Method; Implementations of MoMMS	91

**Chapter 3 (continued)**

Experiment and Result; Applications of MoMMS in QTL analysis and subsequent positional cloning	91
Reference	97

**Chapter 4: QUANTITATIVE TRAIT LOCI FOR THE CIRCADIAN 99****CLOCK IN *NEUROSPORA CRASSA***

Introduction: Quantitative Trait Loci (QTL) approach for Neurospora	99
Clock study	
Method; Strains and growth conditions	104
Method; Race tube experiment	104
Method; Genotyping and genetic map construction	105
Method; QTL analysis	106
Method; The Near isogenic line (NIL) construction	108
Experiments and Results; The phenotypic analysis of Neurospora circadian rhythm (Period and phase)	110
Experiments and Results; Comparisons of two QTL methods	113
(CIM vs BMQ)	
Experiments and Results; Comparisons with the previous clock study	122
Experiment and Results; Further characterizations of the QTL effect	132
Discussion; Phenotypic variations of circadian rhythm in F1 mapping population	137
Discussion; Advantages of haploid organism in QTL analysis	137
Discussion; Comparisons of two different QTL methods (CIM vs BMQ)	139
Discussion; Important considerations in the application of QTL approach to Neurospora circadian clock study	140

**Chapter 4 (continued)**

Discussion; Neurospora clock QTLs	141
Discussion; Functional relationships between clock phenotypes (period vs phase)	142
Reference	144

## LIST OF FIGURES

Figure 1-1.	Schematic diagram of the operation of circadian clock to keep tract of local times	2
Figure 1-2.	Regulation of circadian clock under FRC	3
Figure 1-3.	The rhythmic tproduction of asexual spores in <i>N. crassa</i>	7
Figure 1-4.	The image of the race tube assay	8
Figure 1-5.	The schematic diagram of circadian clock regulation under free running condition	9
Figure 2-1.	Genome-wide distribution of relative abundance of SSRs by the unit size	29
Figure 2-2.	Genome-wide distribution of relative abundance of SSRs by the SSR types in different SSR unit number	30
Figure 2-3.	The predicted and observed frequencies of amino acid repeats encoded by tri-nucleotide SSRs	32
Figure 2-4.	The distributions of the repeat lengths of the different types of amino acid repeats encoded by tri-nucleotide SSRs	33
Figure 2-5.	The distribution SSRs in different chromosomes in the <i>N. crassa</i> genome	35
Figure 2-6.	The polymorphic information content analysis from 129 SSR loci	49
Figure 2-7.	Comparison of PIC values of the AC/CA SSR type in two different population sizes	50
Figure 2-8.	Three hypotheses for the size variation of SSRs	51
Figure 2-9.	Genetic linkage maps of three mapping populations	60
Figure 2-10.	The distribution of the randomly selected SSRs by	67

Figure 2-10 (continued)

	the unit number	
Figure 3-1.	Procedures to discover a gene/genes from QTL analysis	83
Figure 3-2.	Graphical summary in the operational processes of MoMMS	85
Figure 3-3.	The three analysis steps by Cricamate	87
Figure 3-4.	The comparison of the conidiation rhythm between free running condition (A) and entraining condition to 24 hour cycle (B)	88
Figure 3-5.	Comparison of phase inference from between the normalized and original image	90
Figure 3-6.	Flow chart in the operational process of the updated Circamate	92
Figure 3-7.	Simple sequence repeat (SSR) marker developments of <i>Neurospora crassa</i> genome using MoMMS	94
Figure 3-8.	Finding candidate genes for QTL in the <i>N. crassa</i> genome	96
Figure 4-1.	Sechmatic diagram of <i>Neurospora</i> QTL analysis	102
Figure 4-2.	The flow chart that describes the process for the characterizations QTLs defined in F1 mapping population	103
Figure 4-3.	Strategy for the QTL confirmation and the subsequent high resolution mapping step	109
Figure 4-4.	Graphical summary of the collection site of <i>N. crassa</i> accessions used as parental stains in crosses	112
Figure 4-5.	The circadian period variation in the F1 populations of our study	115
Figure 4-6.	The entrained phase variation under 12:12 light/dark condition in the F1 populations of our study	116
Figure 4-7.	Scatter plot analysis between period and phase in N2, N4 and N6	117

Figure 4-8.	Distributions of PPT for neutral markers	119
Figure 4-9.	PPT for three simulated QTLs	120
Figure 4-10.	Summary of Bayesian QTL analysis (BMQ, CIM)	121
Figure 4-11.	Distribution of PPTs of QTLs in BMQ analysis	122
Figure 4-12.	The graphical description of composite interval mapping (CIM) analysis in period length under free running condition in N4 (A) and N6 (B) populations	127
Figure 4-13.	The graphical description of composite interval mapping (CIM) analysis in the phase under the 12:12 LD cycle in N2(A), N4 (B) and N6 (C) population	128
Figure 4-14.	Comparison of mapped loci between Neurospora physical map (A) and linkage maps derived from N6 (B), N4 (C), N2 (D) cross respectively	129
Figure 4-15.	Venn diagram analysis of period (A) and phase QTL (B) among populations	130
Figure 4-16.	Venn diagram analysis of between period and phase in population specific (A and B) and all three populations (C)	131
Figure 4-17.	Comparison of phase phenotype between 13 individuals of BC4F1 with a targeted allele and those of recurrent and donor parents	135
Figure 4-18.	CIM analysis to confirm the QTL effect of N6Cbpha6 at BC4F1	136
Figure 4-19.	Phylogenetic analysis of Neurospora natural strains collected from diverse geographic region	138
Figure 4-20.	Graphical summary of our QTL study for Neurospora circadian clock phenotypes	143

## LIST OF TABLES

Table 2-1. Relative abundance of SSR types by functional genome regions in <i>N. crassa</i>	28
Table 2-2. Sequential analysis of deviance of log linear model for all 19 SSR types	36
Table 2-3. Sequential analysis of deviance of log linear model for 8 mono/di-nt SSR types, intergenic region only	38
Table 2-4. Sequential analysis of deviance of log linear model for 11 tri-nt SSR types	39
Table 2-5. Comparison of log abundance rates in genic and intergenic regions based on a Poisson model	40
Table 2-6. The physical location and PIC values of 131 SSR loci in the <i>Neurospora crassa</i> genome	43
Table 2-7. The list of 33 SSR loci for statistical analyses	52
Table 2-8. Line-cross populations from <i>N. crassa</i> accessions	53
Table 2-9. The marker quality of the mapped SSR loci	58
Table 4-1. <i>N. crassa</i> accessions used as parental stains in crosses	111
Table 4-2. Phenotypic variation in period length and phase in N2, N4 and N6 population	114
Table 4-3. Summary of the additive QTLs in circadian properties that are segregated in three different population formed by <i>N. crassa</i> natural accessions using Bayesian QTL analysis	123
Table 4-4. The information of backcross parent and tested marker	132
Table 4-5. The steps required at each backcross (BC) generation in the NIL Construction	133

# CHAPTER 1

## GENERAL INTRODUCTION

### **Intrinsic properties of circadian clock**

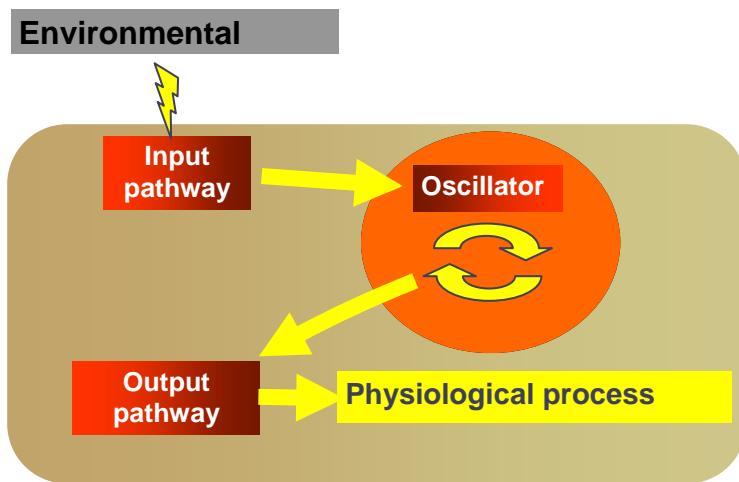
Biological rhythms with about a 24 hour period, called circadian rhythms, have been found in all forms of life from bacteria to humans [1, 2]. Since the circadian rhythm is tightly controlled by an intercellular pacemaker or circadian clock, the rhythmic patterns thus reflect the intrinsic properties of the circadian clock. In addition, the rhythm should maintain a self sustained endogenous oscillation under constant environment or free running condition (FRC) (self sustaining nature) and its period in the FRC is about 24 hours. It should also be able to be reset and synchronized by environmental signals, primarily by light and temperature cycles (entrainment). The circadian rhythm's operation is relatively stable at different temperatures if it is within a physiological temperature range (temperature compensation) [1-4].

It is thought that those intrinsic clock functions provide organisms with time keeping ability to anticipate and deal with daily basis external environmental changes effectively [1-7], which enable organisms to controlling their behavioral, physiological and cellular activity at the right time of day [8].

### **Conceptual structure of circadian clock**

The circadian clock is thought to be comprised of at least three different functional mechanisms or pathways, which include the input, the circadian oscillator, and the output pathways. The input pathway perceives environmental cues, such as light and temperature signals. Temporal information is then transferred to the circadian oscillator where the endogenous rhythm of the circadian oscillator is

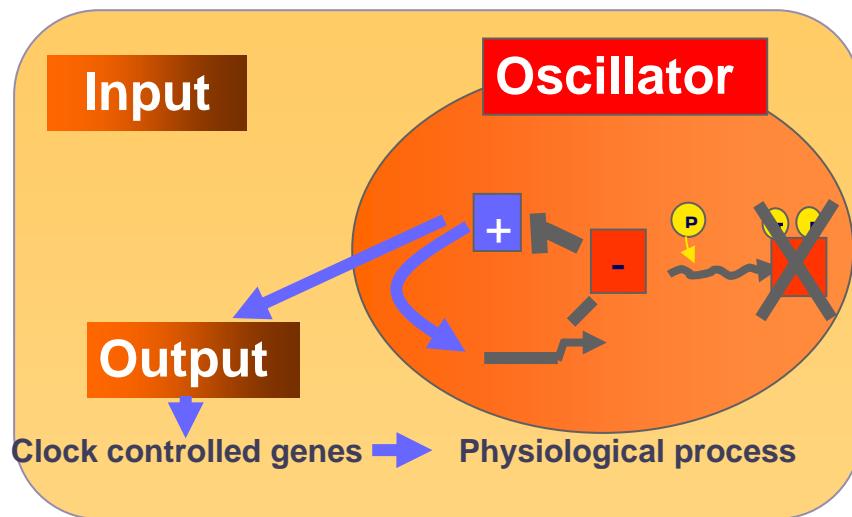
modified by temporal information. Eventually, the information reaches the output pathways to regulate the expression of target genes, which are related to various physiological processes (Figure 1-1). However, much remains to be elucidated in this process, especially about how the input, oscillator and output pathways can communicate temporal information among each other.



**Figure 1-1. Schematic diagram of the operation of the circadian clock to keep track of local times.**

### Mechanism of autoregulatory cellular oscillation

In the circadian system, the most well characterized part is the circadian oscillator. Under the FRC, the output phenotype of the circadian rhythm, called period, exclusively reflects endogenous regulation of the circadian oscillator (Figure 1-2).



**Figure 1-2. Regulation of the circadian clock under FRC.**

A forward genetic approach has been applied to find components associated with the endogenous regulation of the cellular oscillator by focusing on mutants with altered periods or arrhythmic phenotypes [9]. The cloning and molecular characterization of genes, especially the Drosophila period (*per*) gene using this approach became a foundation of the current paradigm of the cellular oscillation of the circadian clock [10, 11]. The *per* gene encodes a novel protein of unknown function that accumulates in the nucleus. A mutation in the *per* gene affects the circadian rhythm [11]. Hardin et al (1990) show that the levels of *per* gene product undergo circadian oscillation at the transcription and translational level, but *per* mRNA level is inhibited by the enrichment of its gene product (PER protein). This suggests a hypothesis where an autoregulatory negative feedback loop underlies the self sustaining rhythm of the cellular oscillator [11]. Based on the model proposed, other components related to the negative feedback loop began to be characterized in this organism and other model systems as well [3, 11-13].

The transcriptional/translational negative feedback loop is a fundamental mechanism underlying the autonomous cellular oscillator in all eukaryotes studied to date [2, 10, 13], where the positive elements of the loop initiate the transcription of a gene/genes that encode the negative element. As the concentration of the gene product of the negative element increases, it interacts with the positive elements by regulating the phosphorylation status of the positive elements, which in turn inhibits the gene transcription of the negative element[14]. However, phosphorylation-mediated degradation of the negative elements by SCF type E3 ubiquitinase complete the oscillation [15]. The degradation of the negative element leads to reactivation of the positive elements, which allows a new cycle. In addition, at different times of day in the cytoplasm, the negative element activates the expression of the positive elements to form positive feedback. This positive feedback mechanism, which interlocks with negative-feedback loops is important for maintaining the stability and robustness of the cellular oscillator[16]. Interestingly, the transcriptional/translational negative feedback mechanism seems to be universal in eukaryotes, but the components in the pathway are greatly different among organisms. For example, the clock gene sequences share no or little homology across taxa except functional domains [2, 13, 17]. Thus, it is speculated that the common feature in the clock operation shared by eukaryotes may be the result of the convergent evolution [18].

### **Biological role of circadian clock**

One of the crucial roles of the circadian clock is to synchronize organisms to local time in order to control the behavioral, physiological, and cellular activities in response to daily or seasonal environmental changes. Most popular examples regarding circadian clock-mediated activities include many important biological process; nitrogen fixation in cyanobacteria [19], scent emission and stability of

photosynthesis in plants [20, 21], conidiation in *Neurospora crassa* [22], olfactory responses of *Drosophila melanogaster* [23], luteinizing hormone levels in birds [24], and wheel running activity in hamsters [25]. Thus, it is reasonable to think that the proper regulation of circadian clock may provide a fitness advantage or have adaptive significance to organisms [8, 21, 26]. Nevertheless, justifying this statement is not as simple as it seems [18] because fitness itself is not easily defined.

A classic definition of fitness in evolutionary biology is “a measure of reproductive success and the passing on of genes.[18]” Many factors like longevity, survival rate, growth, and development are somewhat related to the fitness, but these ancillary factors may not be direct measures of the fitness[18]. However, in many cases, researchers fell short of finding a correlation between circadian clock function and these ancillary factors and failed to demonstrate the fitness advantage of the circadian clock.

To test the fitness advantage hypothesis appropriately, three approaches can be proposed as addressed in the review paper from Johnson (2005). The first was executed by DeCoursey et al (2000). In this approach, the circadian clock phenotype is manipulated (for example, compromise circadian clock phenotype by a clock gene knock out or by a surgical control) and then the overall effect is evaluated in natural conditions [27]. In the second approach different genotypes in clock properties (for example: wild type versus clock mutant) can be competed in various environments under a laboratory setting [28]. The third approach compares genetic variation between or among natural strains adapted in different environments using Quantitative Trait Loci (QTL) analyses or association studies [29-31].

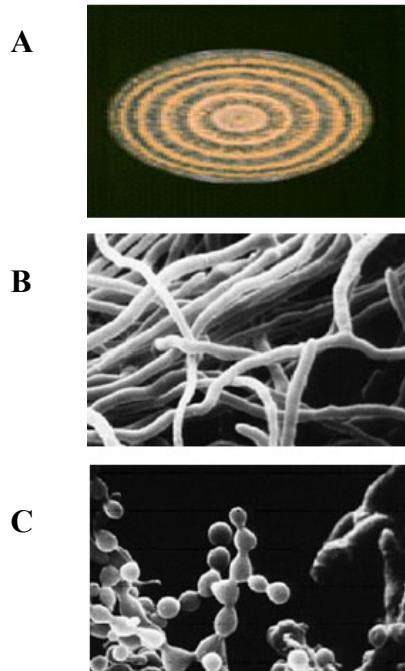
In general, a criticism of the first (competition experiment in a laboratory setting) and second (manipulation of circadian clock phenotype) approaches is that these settings are artificial. Here are the reasons; in the first approach, tested in natural

conditions, a mutant may not be a good subject to use because the overall conclusion may be drawn from other effects, rather than from the circadian clock function, via pleiotropy effect [21, 32]. For the second approach, in many cases, the laboratory setting may not be precise enough to simulate the selective pressures occurring in natural environments due to technical limitations if the environment is too complex[33]. However, some elaborate laboratory settings have been successfully applied to make strong inferences about natural phenomena, since controlled experiments under the laboratory environments can erase unnecessary noises that may distract from the logic of the investigation [34-36]. In this context, the third approach, which tested the hypothesis through QTL or association studies using natural strains under elaborated laboratory setting, seems to be the most promising approach to test the role of the circadian clock in adaptive processes.

### ***Neurospora crassa* as a model system in the chronobiology**

The ascomycete filamentous fungus *Neurospora crassa*, which is called red bread mold, has served as one of the leading model systems in circadian clock studies. In this fungal species, 10,620 genes are documented by a >16-fold sequence coverage and the availability of information is supported by sophisticated web based bioinformatics tools; targeted gene replacements are now possible[37]; a complete gene knock out project is under way [37, 38]; inexpensive whole genome microarrays are available and facilitate transcriptional profiling[39]. In addition to the genetic tractability and many resources, one of the reasons that *N. crassa* has become a model system in chronobiology is the obvious output phenotype of the circadian clock, which is the rhythmicity of the production of asexual spores (conidiation) as shown in Figure 1-3 [22]. *N. crassa* creates a banding phenotype which results from the changes in

developmental stages between the asexual and vegetative stages in a circadian clock manner.

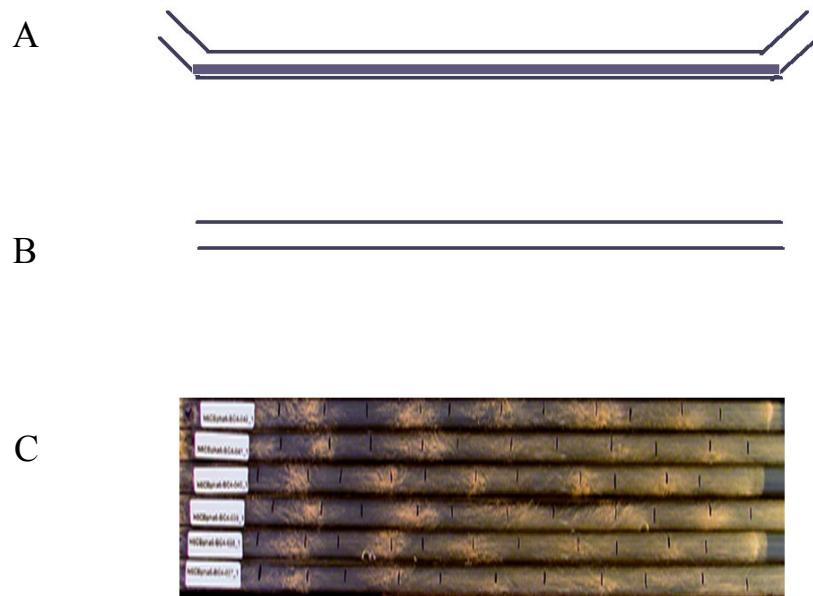


**Figure 1-3. The rhythmic production of asexual spores in *N. crassa*.**

A. The rhythmic pattern of conidiation shown from the culture of *N. crassa* at Petri dish under free running condition. B. Microscopic view on the vegetative growth, which is dark ring part at (A). C. Microscopic view of the conidiation, which represent bright ring portion in (A). This figure is obtained from the web page of Fungal Genetics Stock Center (<http://www.fgsc.net/>).

The most common assay to measure the *Neurospora* clock is the race tube assay. As the result of circadian clock-controlled asexual development, the banding phenotype can be interpreted more precisely in a long glass tube or “race tube” since the growth rate is unified under this environment (Figure 1-4 B and C) [40]; after a segment of mycelia of a *Neurospora* strain in one end is inoculated, one can measure a circadian clock property of a strain without any sophisticated instrument (Figure 1-4 B

and C). This easily detected clock phenotype made the *Neurospora* circadian rhythm an attractive system for genetic studies.



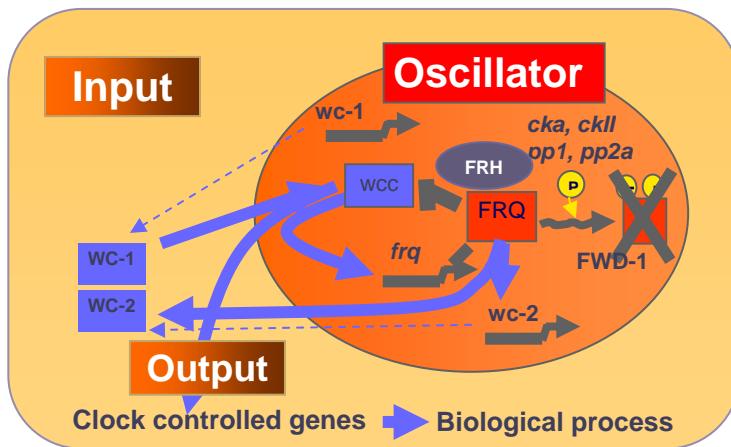
**Figure 1-4. Image of the race tube assay.**

A. The individual race tube image at side view. Minimal media with agar is used to fill in the bottom. B. The individual race tube image from top view. The mycelial segment is inoculated in one end. C. Race tube image of a 6 pack where mycelial segments of *N.crassa* are inoculated and cultured for 5 days.

#### Circadian clock study in *Neurospora crassa*.

The availability of powerful genetic analysis tools and the easily assayable clock phenotype in *Neurospora* has made the system one of the most successful model organisms to dissect the circadian clock by forward genetics approaches [2, 9, 40]. Mutant screens for clock genes have focused on mutants with altered period or arrhythmic phenotypes caused by a single mutation inherited through Mendelian

segregation [9]. The most interesting gene discovered in these mutant screenings is frequency (*frq*), which when mutated, leads to strains with long period, short period or arrhythmic phenotype [2, 9, 40]. This finding led to the proposal that a single gene could function as a “state variable” for the circadian oscillator [41]. Cloning and characterizing the *frq* gene significantly advanced molecular understanding of eukaryotic circadian oscillators (Figure 1-5) [4, 12].



**Figure 1-5. Schematic diagram of circadian clock regulation under free running conditions.**

For about two decades, great effort has been made to understand mechanism of circadian oscillators as summarized in Figure 1-5. Central components of the transcriptional/translational negative feedback loop in *Neurospora* include two negative elements, the proteins FREQUENCY (FRQ) and FRQ-RELATED HELICASE (FRH), and two positive elements, WHITE COLLAR-1 (WC-1) and WHITE COLLAR-2 (WC-2), which form a hetero duplex, WHITE COLLAR COMPLEX (WCC) through Per-Ant-Sim domain (PAS) [2, 4, 12, 38]. In addition to these core components, there are a number of proteins with supporting roles: the

various kinases (Casein kinase I (CKI) and II (CKII)), phosphatases (Protein phosphatase 1(PP1), protein phosphatase 2a (PP2a)) that regulate phosphorylation status of positive element and negative elements during time of a day[12]. The SCF type E3 ubiquitinase, FWD-1, is responsible for the termination of a cycle by degrading the negative element based on the phosphorylation status of the negative element, FRQ[15].

At subjective dawn, the WCC is already bound to the Clock Box (C-box) in the *frq* promoter, a region containing two GATG repeats, and is actively initiating *frq* transcription. The FRQ is translated with little lag, dimerizes, and assembles with FRH into the FFC in the cytoplasm, and moves to the nucleus[38]. FRQ, in the FFC, participates in several pleotropic actions, which affect the kinetics of the circadian cycle. The first and dominant action is that the FFC acts to reduce the activity of the WCC by regulating the phosphorylation status of one or both elements of the WCC with the help of CKI and CKII, which make the transcription of *frq* less active [14, 38]. At different times of day in the cytoplasm, the FFC activates the expression of the positive elements to form positive feedback. This positive feedback mechanism interlocks negative-feedback loops and is important for maintaining the stability and robustness of the cellular oscillator[16].

FRQ disappears through phosphorylation-mediated proteasomal degradation[15]. As soon as FRQ is translated, it begins to be phosphorylated by CKI and CKII[12]. The kinetics of FRQ phosphorylation are likely to be the most critical feature determining the period length of the clock; point mutations in single phosphorylated residues can shift the period from 22 to 35 h [4]. Phosphorylation of FRQ promotes its interaction with the WD-40 domain of FWD-1, SCF-type ubiquitin (E3) ligase, which promotes the precipitous turnover of FRQ around the middle of the subjective night, resulting in the completion of a cycle[15]. When FRQ disappears, the

phosphorylation-mediated inactivation of the WCC is reversed presumably by PP2A and the expression of *frq* by WCC launch a new cycle [2, 4, 12, 38].

Despite the intensive molecular characterizations of *frq* and other known clock genes in *N. crassa*, there is still no comprehensive understanding of the Neurospora circadian clocks. Our understanding of circadian regulation is mainly focused on the circadian clock regulation under FRC, and most of the genetic screening done previously was focused on identifying period determinants. Thus, just a handful of genetic loci that are responsible for other clock properties, such as entrained phase or temperature compensation, have been characterized. With advances in our understanding of the molecular structure of Neurospora clocks, the circadian clock is more tightly linked to other cellular machineries than previously speculated [1]. However, we still don't have a clear picture of the biological role of the circadian clock, for example its adaptive significance. Lastly, our understanding of the cellular oscillator is incomplete. It has been suggested that the *frq*-based oscillator might not be the only oscillator. For example *frq*-less oscillators [42] coupled to other oscillators in a cell have been proposed [43]. Currently, we know very little about the genetic basis of these loosely defined oscillators [4]. Forward genetics may be useful to make a breakthrough, but the conventional forward genetics approach may not as helpful as proposed. That is because it can uncover neither genetic loci with subtle clock phenotypes nor those associated with essential cellular functions [44, 45]. Thus, we explored an alternative strategy for detecting novel genetic loci affecting the Neurospora circadian clock.

## **Summary of the chapters**

In the second chapter, the results of investigation in the distribution and size variability of SSRs across the *N. crassa* genome are described. This work was

primarily dedicated to establishing a genetic marker system. I had four specific questions in the project. 1) Is the distribution of SSRs random or not in the *N. crassa* genome? If it is not random, what factors could explain this? 2) What are the biological functions of SSRs? 3) What are the forces causing size variation in SSRs? 4) Can we use SSRs for population studies in intra-species populations as previously suggested?

In the third chapter, I describe the research leading to the development of two important tools that facilitate QTL analysis and subsequent positional cloning. The first tool we developed is MoMMs (Molecular Marker Management System). This software assists with SSR marker development and integrates all relevant bioinformatic and experimental data derived from the polymorphic SSR markers to one database. In addition, this program can visualize the physical location of the SSR markers at the chromosome and nucleotide level to facilitate genetic mapping and subsequent positional cloning procedures. The second tool, I involved an update of the phase evaluation function in Circamate, which is a high through-put image processing software for circadian rhythm analysis.

Lastly, in the fourth chapter, the results of QTL analysis, followed by QTL characterization are described. The QTL study is performed for the two clock phenotypes, period and entrained phase, using natural populations. In an effort to efficiently find natural genetic variations affecting the clock phenotype, we employed two QTL analyses, composite interval mapping (CIM) and Bayesian multiple QTL (BMQ) analysis in three independent mapping populations derived from natural accessions that were collected from geographically isolated areas [59]. Subsequently, I confirmed the QTL effect in the BC4 generation based on prediction in the F1 generation, by constructing near isogenic line (NIL)

## REFERENCES

1. Bell-Pedersen D, Crosthwaite SK, Lakin-Thomas PL, Merrow M, Okland M: **The Neurospora circadian clock: simple or complex?** *Philos Trans R Soc Lond B Biol Sci* 2001, **356**(1415):1697-1709.
2. Dunlap JC, Loros JJ: **The neurospora circadian system.** *J Biol Rhythms* 2004, **19**(5):414-424.
3. Bell-Pedersen D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, Zoran MJ: **Circadian rhythms from multiple oscillators: lessons from diverse organisms.** *Nat Rev Genet* 2005, **6**(7):544-556.
4. Dunlap JC, Loros JJ: **How fungi keep time: circadian system in Neurospora and other fungi.** *Curr Opin Microbiol* 2006, **9**(6):579-587.
5. Edwards KD, Anderson PE, Hall A, Salathia NS, Locke JC, Lynn JR, Straume M, Smith JQ, Millar AJ: **FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock.** *Plant Cell* 2006, **18**(3):639-650.
6. Elvin M, Loros JJ, Dunlap JC, Heintzen C: **The PAS/LOV protein VIVID supports a rapidly dampened daytime oscillator that facilitates entrainment of the Neurospora circadian clock.** *Genes Dev* 2005, **19**(21):2593-2605.

7. Heintzen C, Loros JJ, Dunlap JC: **The PAS protein VIVID defines a clock-associated feedback loop that represses light input, modulates gating, and regulates clock resetting.** *Cell* 2001, **104**(3):453-464.
8. Vitalini MW, de Paula RM, Park WD, Bell-Pedersen D: **The rhythms of life: circadian output pathways in Neurospora.** *J Biol Rhythms* 2006, **21**(6):432-444.
9. Feldman JF, Hoyle MN: **Isolation of circadian clock mutants of Neurospora crassa.** *Genetics* 1973, **75**(4):605-613.
10. McClung CR: **Plant circadian rhythms.** *Plant Cell* 2006, **18**(4):792-803.
11. Hardin PE, Hall JC, Rosbash M: **Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels.** *Nature* 1990, **343**(6258):536-540.
12. Liu Y, Bell-Pedersen D: **Circadian rhythms in Neurospora crassa and other filamentous fungi.** *Eukaryot Cell* 2006, **5**(8):1184-1193.
13. King DP, Takahashi JS: **Molecular genetics of circadian rhythms in mammals.** *Annu Rev Neurosci* 2000, **23**:713-742.
14. Brunner M, Schafmeier T: **Transcriptional and post-transcriptional regulation of the circadian clock of cyanobacteria and Neurospora.** *Genes Dev* 2006, **20**(9):1061-1074.

15. He Q, Cheng P, Liu Y: **The COP9 signalosome regulates the Neurospora circadian clock by controlling the stability of the SCFFWD-1 complex.** *Genes Dev* 2005, **19**(13):1518-1531.
16. Lee K, Loros JJ, Dunlap JC: **Interconnected feedback loops in the Neurospora circadian system.** *Science* 2000, **289**(5476):107-110.
17. Hardin PE: **Transcription regulation within the circadian clock: the E-box and beyond.** *J Biol Rhythms* 2004, **19**(5):348-360.
18. Johnson CH: **Testing the adaptive value of circadian systems.** *Methods Enzymol* 2005, **393**:818-837.
19. Johnson CH, Golden SS, Ishiura M, Kondo T: **Circadian clocks in prokaryotes.** *Mol Microbiol* 1996, **21**(1):5-11.
20. Kolosova N, Gorenstein N, Kish CM, Dudareva N: **Regulation of circadian methyl benzoate emission in diurnally and nocturnally emitting plants.** *Plant Cell* 2001, **13**(10):2333-2347.
21. Dodd AN, Salathia N, Hall A, Kevei E, Toth R, Nagy F, Hibberd JM, Millar AJ, Webb AA: **Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage.** *Science* 2005, **309**(5734):630-633.

22. Pittendrigh CS, Bruce SG, Rosensweig NS, Rubin ML: **Growth patterns in Neurospora: A biological clock in Neurospora.** *Nature* 1959, **184**(4681):169-170.
23. Krishnan B, Dryer SE, Hardin PE: **Circadian rhythms in olfactory responses of Drosophila melanogaster.** *Nature* 1999, **400**(6742):375-378.
24. Follett BK, Mattocks PW, Jr., Farner DS: **Circadian function in the photoperiodic induction of gonadotropin secretion in the white-crowned sparrow, Zonotrichia leucophrys gambelii.** *Proc Natl Acad Sci U S A* 1974, **71**(5):1666-1669.
25. Ralph MR, Menaker M: **A mutation of the circadian system in golden hamsters.** *Science* 1988, **241**(4870):1225-1227.
26. Johnson CH: **Endogenous timekeepers in photosynthetic organisms.** *Annu Rev Physiol* 2001, **63**:695-728.
27. DeCoursey PJ, Walker JK, Smith SA: **A circadian pacemaker in free-living chipmunks: essential for survival?** *J Comp Physiol [A]* 2000, **186**(2):169-180.
28. Ouyang Y, Andersson CR, Kondo T, Golden SS, Johnson CH: **Resonating circadian clocks enhance fitness in cyanobacteria.** *Proc Natl Acad Sci U S A* 1998, **95**(15):8660-8664.

29. Gottlieb DJ, O'Connor GT, Wilk JB: **Genome-wide association of sleep and circadian phenotypes.** *BMC Med Genet* 2007, **8 Suppl 1**:S9.
30. Michael TP, Salome PA, Yu HJ, Spencer TR, Sharp EL, McPeek MA, Alonso JM, Ecker JR, McClung CR: **Enhanced fitness conferred by naturally occurring variation in the circadian clock.** *Science* 2003, **302**(5647):1049-1053.
31. Costa R, Peixoto AA, Barbujani G, Kyriacou CP: **A latitudinal cline in a Drosophila clock gene.** *Proc Biol Sci* 1992, **250**(1327):43-49.
32. Beaver LM, Rush BL, Gvakharia BO, Giebultowicz JM: **Noncircadian regulation and function of clock genes period and timeless in oogenesis of Drosophila melanogaster.** *J Biol Rhythms* 2003, **18**(6):463-472.
33. Klarsfeld A, Rouyer F: **Effects of circadian mutations and LD periodicity on the life span of Drosophila melanogaster.** *J Biol Rhythms* 1998, **13**(6):471-478.
34. Balasubramanian S, Sureshkumar S, Agrawal M, Michael TP, Wessinger C, Maloof JN, Clark R, Warthmann N, Chory J, Weigel D: **The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana.** *Nat Genet* 2006, **38**(6):711-715.

35. Nozue K, Covington MF, Duek PD, Lorrain S, Fankhauser C, Harmer SL, Maloof JN: **Rhythmic growth explained by coincidence between internal and external cues.** *Nature* 2007, **448**(7151):358-361.
36. Filiault DL, Wessinger CA, Dinneny JR, Lutes J, Borevitz JO, Weigel D, Chory J, Maloof JN: **Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light.** *Proc Natl Acad Sci U S A* 2008, **105**(8):3157-3162.
37. Colot HV, Park G, Turner GE, Ringelberg C, Crew CM, Litvinkova L, Weiss RL, Borkovich KA, Dunlap JC: **A high-throughput gene knockout procedure for Neurospora reveals functions for multiple transcription factors.** *Proc Natl Acad Sci U S A* 2006, **103**(27):10352-10357.
38. Dunlap JC: **Proteins in the Neurospora circadian clockworks.** *J Biol Chem* 2006, **281**(39):28489-28493.
39. Kasuga T, Townsend JP, Tian C, Gilbert LB, Mannhaupt G, Taylor JW, Glass NL: **Long-oligomer microarray profiling in Neurospora crassa reveals the transcriptional program underlying biochemical and physiological events of conidial germination.** *Nucleic Acids Res* 2005, **33**(20):6469-6485.
40. Loros JJ, Dunlap JC: **Genetic and molecular analysis of circadian rhythms in Neurospora.** *Annu Rev Physiol* 2001, **63**:757-794.

41. Aronson BD, Johnson KA, Dunlap JC: **Circadian clock locus frequency: protein encoded by a single open reading frame defines period length and temperature compensation.** *Proc Natl Acad Sci U S A* 1994, **91**(16):7683-7687.
42. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome.** *Science* 2001, **291**(5507):1304-1351.
43. Lakin-Thomas PL: **Transcriptional feedback oscillators: maybe, maybe not.** *J Biol Rhythms* 2006, **21**(2):83-92.
44. He Q, Shu H, Cheng P, Chen S, Wang L, Liu Y: **Light-independent phosphorylation of WHITE COLLAR-1 regulates its function in the Neurospora circadian negative feedback loop.** *J Biol Chem* 2005, **280**(17):17526-17532.
45. Mackay TF: **The genetic architecture of quantitative traits.** *Annu Rev Genet* 2001, **35**:303-339.
46. Alonso-Blanco C, Koornneef M: **Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics.** *Trends Plant Sci* 2000, **5**(1):22-29.
47. Shimomura K, Low-Zeddies SS, King DP, Steeves TD, Whiteley A, Kushla J, Zemenides PD, Lin A, Vitaterna MH, Churchill GA *et al*: **Genome-wide**

- epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice.** *Genome Res* 2001, **11**(6):959-980.
48. Doerge RW: **Mapping and analysis of quantitative trait loci in experimental populations.** *Nat Rev Genet* 2002, **3**(1):43-52.
49. Darrah C, Taylor BL, Edwards KD, Brown PE, Hall A, McWatters HG: **Analysis of phase of LUCIFERASE expression reveals novel circadian quantitative trait loci in Arabidopsis.** *Plant Physiol* 2006, **140**(4):1464-1474.
50. Edwards KD, Lynn JR, Gyula P, Nagy F, Millar AJ: **Natural allelic variation in the temperature-compensation mechanisms of the *Arabidopsis thaliana* circadian clock.** *Genetics* 2005, **170**(1):387-400.
51. Suzuki T, Ishikawa A, Yoshimura T, Namikawa T, Abe H, Honma S, Honma K, Ebihara S: **Quantitative trait locus analysis of abnormal circadian period in CS mice.** *Mamm Genome* 2001, **12**(4):272-277.
52. Swarup K, Alonso-Blanco C, Lynn JR, Michaels SD, Amasino RM, Koornneef M, Millar AJ: **Natural allelic variation identifies new genes in the *Arabidopsis* circadian system.** *Plant J* 1999, **20**(1):67-77.
53. Sargent ML, Woodward DO: **Genetic determinants of circadian rhythmicity in *Neurospora*.** *J Bacteriol* 1969, **97**(2):861-866.

54. Sargent ML, Kaltenborn SH: **Effects of Medium Composition and Carbon Dioxide on Circadian Conidiation in Neurospora.** *Plant Physiol* 1972, **50**(1):171-175.
55. Park S, Lee K: **Inverted race tube assay for circadian clock studies of the Neurospora accessions.** *Fungal Genet News* 2004, **51**:12-14.
56. Belden WJ, Larrondo LF, Froehlich AC, Shi M, Chen CH, Loros JJ, Dunlap JC: **The band mutation in Neurospora crassa is a dominant allele of ras-1 implicating RAS signaling in circadian output.** *Genes Dev* 2007, **21**(12):1494-1505.
57. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*: **The genome sequence of the filamentous fungus Neurospora crassa.** *Nature* 2003, **422**(6934):859-868.
58. Davis RH, Perkins DD: **Timeline: Neurospora: a model of model microbes.** *Nat Rev Genet* 2002, **3**(5):397-403.
59. Turner BC, Perkins DD, Fairfield A: **Neurospora from natural populations: a global study.** *Fungal Genet Biol* 2001, **32**(2):67-92.

## CHAPTER 2

### SIMPLE SEQUENCE REPEATS IN *NEUROSPORA CRASSA*: DISTRIBUTION, POLYMORPHISM AND EVOLUTIONARY INFERENCE

#### **Introduction; Simple Sequence Repeats in *Neurospora crassa***

Simple sequence repeats (SSRs) refer to the sequences that are one to six-nucleotides (nt) repeated in tandem in a genome. SSRs have many advantageous features for various biological studies: SSRs are ubiquitous and abundant in a genome, highly variable and suitable for high-throughput applications [1 , 2-8]. In addition to practical usages of SSRs for biological studies, the SSRs have also been under the intense scrutiny of researchers to elucidate the evolution of genomes: (1) why are they ubiquitously present in a genome, (2) how do they arise, (3) why are they unusually polymorphic, and (4) what are their biological or structural functions are [1, 9]? The evolutionary dynamics of SSRs have been actively discussed and hypotheses for experimental confirmation have been reviewed in the recent literature [1, 9-11].

The growing numbers of completed genome sequences in eukaryotic organisms from fungi to human have greatly assisted understanding SSRs at the genome-wide level. One obvious observation from the genome-wide studies was that the distribution of SSRs in the genome was not random in several respects: tri-nt and hexa-nt SSRs in coding regions were the dominant SSR types; other SSR repeat types (except tri-nt or hexa-nt SSRs) were found in excess in the non-coding regions of the genome but were rare in coding regions; differential distribution in terms of abundance of SSRs was observed in between intronic and intergenic regions 5' and 3' UTRs, and different chromosomes; and lastly, different species have different frequencies of SSR types and repeat units [2, 10, 12, 13]. The current experimental and observational evidence suggests that these non-random distributions of SSRs, both in coding and non-coding

regions, may be associated with a functional significance, which presumably results in adaptive advantages [9, 14-20]. Two alternative hypotheses were suggested to explain the genesis of SSRs. These hypotheses propose that SSRs originate either spontaneously from/within unique sequences (*de novo* genesis) or that they are brought about in a primal form into a receptive genomic location by mobile elements (adoptive genesis). These two hypotheses are both adequate for explaining the ubiquitous distribution of SSRs. However, there remains much to be understood to elucidate which one is right and how the non-random distribution of SSRs has emerged in the eukaryotic genome [1, 9-11]. *N. crassa* has been well characterized for its diverse genome defence mechanisms that inactivate genetic mobile elements and gene duplication across the genome except in some restricted regions close to telomeres and centromeres [21-24]. Thus, we reasoned that characterizing the SSR distribution in the *N. crassa* genome would provide a unique opportunity to explore the non-random distribution of SSRs shaped by the *de novo* genesis in the eukaryotic genome.

In this report, we investigate the distribution and size variability of SSRs across the *N. crassa* genome. We had four specific questions in mind. 1) Is the distribution of SSRs random or not in the *N. crassa* genome? If it is not random, what factors could explain this? 2) What are the biological functions of SSRs? 3) What are the forces causing the size variation of SSRs? 4) Could we use SSRs for population studies in intra-species populations as previously suggested [8]?

Our data on the distribution and size variation of SSRs in the *N. crassa* genome reveal both similarities and uniqueness in composition and distribution patterns in comparison to the other eukaryotic genomes, including other sequenced fungal organisms. We discuss the potential forces for shaping non-random distribution and size variation of SSRs, and biological implications of size variations of SSRs in the *N. crassa* genome.

### **Method; SSR analysis in the *Neurospora* genome sequence**

The 39.2 Mb *Neurospora* genome sequence, release 7, was downloaded from the Broad Institute website, <http://www.broad.mit.edu/annotation/genome/neurospora/Home.html>, and was analyzed to identify SSRs. We utilized the “tandem repeat finder” program[25]. We used stringent cut-off parameters as follows: matching weight = 2, mismatching penalty = 7, indel penalty = 7, match probability = 80, indel probability = 10, minimum alignment score to report = 50, and maximum period size to report = 6. From the analysis, we selected 2749 SSRs and subsequently categorized the SSRs by unit size and repeat motif in different genomic locations. In our study, the genomic location categories were intergenic and genic (exon and intron) regions. Each of the SSRs was considered as unique and was subsequently classified according to theoretically possible combinations in each SSR. For example,  $(AC)_n$  is equivalent to  $(CA)_n$ ,  $(TG)_n$ , and  $(GT)_n$ , while  $(AGC)_n$  is equivalent to  $(GCA)_n$ ,  $(CAG)_n$ ,  $(CTG)_n$ , and  $(TGC)_n$ . Lastly, we determined the abundance of each SSR motif and unit size in the different genomic regions by normalizing the size of the corresponding genomic region. To describe the abundance of SSRs in different genomic region, we chose to use the “relative abundance”, which is calculated by dividing the number of SSRs by mega base-pair (MB) of sequences in our analyses.

### **Method; Statistical frame work to infer SSR genesis rate in chromosomes and genomic locations**

In the statistical analysis, Y is denoted as a frequency and L the associated length, then the (relative) abundance is given by  $A=Y/L$ . Our modeling strategy incorporates the frequency and length information by assuming that the counts are Poisson random variables with expected values proportional to their associated lengths. If E denotes the expected value of a count, then the expected rate is  $R=E/L$ . A log-linear model is used

to describe how these rates vary as a function of SST type, genomic region category, and chromosome. The fit of a particular model can be assessed by comparing the empirical abundance values, A, to maximum likelihood estimates of expected rates which satisfy the model assumptions, using the model deviance statistic,

$$D = 2 \sum Y \log \frac{Y}{\hat{E}}$$

where the summation is over all 266 cells in the  $19 \times 7 \times 2$  contingency table. Our most general model has the form,

$$\log R = \alpha + T + C + G + T \times G,$$

or equivalently,

$$\log E = \log L + \alpha + T + C + G + T \times G,$$

where T denotes the main effect of SSR type, C denotes the effect of chromosome, G is the main effect of genomic location category, and TxG allows for type by region interaction, that is, differential type effects by genomic category.

### **Method; SSR marker development from *Neurospora* genome**

To characterize the overall pattern of polymorphism of the SSRs in the *Neurospora* genome, we strived to select SSRs randomly from the *Neurospora* genome. We divided the genome into 250 kb windows and selected SSRs randomly within each window. A total of 164 SSR loci consisting of di- to hexa-SSRs with various sequence motifs were chosen for further analysis. The scatter plot of the selected markers showed that they were evenly distributed in the genome (Data not shown). With the SSRs selected, we designed oligos using Primer3 software (Whitehead Institute for Biomedical Research, Boston, USA) in flanking sequence to amplify the targeted SSR loci. The range of the annealing temperatures in each primer set was between 50 °C and 60 °C and the primer pairs yielded amplification products between 100 and 350 bp.

For semi-automated genotyping analysis, the 5' M13 sequence was attached to a forward primer in order to incorporate a florescent dye into the PCR product.

Fluorescent dye labelled M13 forward primer and a marker specific reverse primer were used to generate fluorescent-labelled PCR product as previously described [30]. The composition of the PCR master mix was prepared as described in Cho et al [31], and the PCR profile was modified from Schuelke as follows [30]. The basic profile was: 5 min at 94°C, 30 cycles of 30 sec at 94°C, 45 sec at 55°C, 1 min at 72°C, and 25 cycles of 15 sec at 94 °C, 30 sec at 53°C, 1 min at 72°C, and 10 min at 72°C for final extension.

Fluorescent-labelled PCR products for SSR loci were multiplexed with regard to each molecular weight and fluorescent dye. Each multiplexing set of primers was called a panel. One panel consisted of 12-15 SSR marker sets. The multiplexed PCR products were analyzed by an ABI 3730 (Applied Biosystems) according to the manufacturer's instructions. Allele sizes of SSR loci were determined using Genemapper3.0® (Applied Biosystems).

The measurement of the allelic diversity or polymorphism information content (PIC) value was first described by Botstein et al [32] and modified by Anderson et al. [33]. PIC was defined as the probability that two randomly chosen copies of gene will be different alleles different within a population. The formula for the PIC value applied in our study was as follows:

$$PIC_i = 1 - \sum_{j=1}^n P_{ij}^2$$

where  $P_{ij}$  represents the frequency of the  $j$ th allele for marker  $i$ , and summation extends over  $n$  alleles. The allelic polymorphism of the 162 SSR markers in the seven natural accessions, FGSC#2223, FGSC#4825, FGSC#4720, FGSC#4715, FGSC#3223, FGSC#4724, and FGSC#2478, were calculated following the formula. The genome structure of seven *N. crassa* strains are divergent and not related among each other (unrooted tree analysis, minimum pair-wise dissimilarity= 0.91).

### **Method; linkage mapping analysis of *Neurospora* genome**

The 564 F1 progenies (188 F1 haploid progeny from each line-cross, Table 2-6) were genotyped to determine the linkage maps for each cross. Genetic linkage maps of each population were constructed using two different algorithms, Map Manger QTX v. 0.3 [38] and GMENDEL v.3.0 [39] with the Kosambi mapping function [40]. Using Map manager, the initial linkage grouping was performed using the Double Haploid option with a threshold level of  $P= 0.001$ . Subsequently, Monte Carlo simulation with 500 iterations was used to test the marker locus order generated by GMENDEL.

### **Experiment and Result; Genome-wide distribution of SSRs by the SSR unit size**

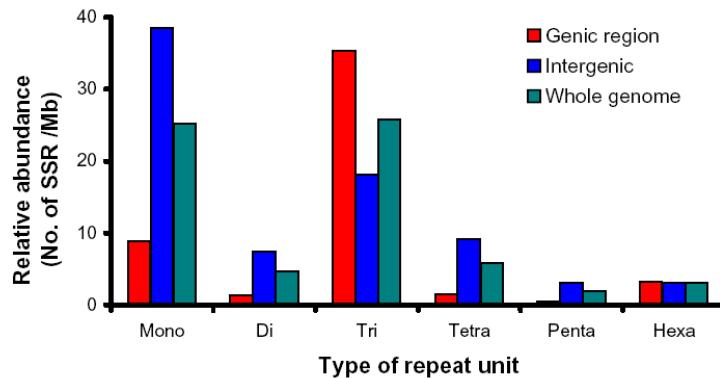
In order to systematically characterize the distribution of SSRs, we surveyed all of the SSRs in the *N. crassa* genome. With our filter conditions (Experiment and Result; SSR analysis in the *Neurospora* genome sequence), we identified 2749 SSRs, which all the information is available at

<http://ncssr.genesis.plantpath.cornell.edu/ssr.php>. SSRs were present equally in the genic and intergenic regions in the *N. crassa* genome; 51% in the genic region and 49% in the intergenic region (Figure 2-1 and Table 2-1). Tri-nt SSRs were the most abundant SSRs overall (Figure 2-1 and Table 2-1). SSRs in different repeat units show differential or non-random distributions in the different genomic locations. It is noteworthy that tri-nt SSRs were the most abundant SSR type in the genic region, whereas, mono-nt SSRs were the most abundant SSR type in the intergenic region (Figure 2-1). In an attempt to analyze the differential distribution of SSRs more clearly, we characterized the distribution of SSR types in each repeat unit across genomic locations.

**Table 2-1. Relative abundance of SSR types by functional genome regions in *N. crassa*.**

SSR types	Intergenic region (21.7 Mb)			Genic region (17.4 Mb)			Total (39.2 Mb)	
	Count	RAa	Count	RAa	Count	RA <sup>a</sup>	Count	RAa
Mono	835	38.4	3	0.2	152	56.3	990	25.3
Di	167	7.7	1	0.1	22	8.1	191	4.9
Tri	414	19.1	591	40.2	64	23.7	1084	27.7
Tetra	213	9.8	1	0.1	22	8.1	243	6.2
Penta	73	3.4	1	0.1	7	2.6	82	2.1
Hexa	84	3.9	64	4.4	7	2.6	159	4.1

<sup>a</sup>Relative abundance = number of SSR / chromosome size in mega base (MB).



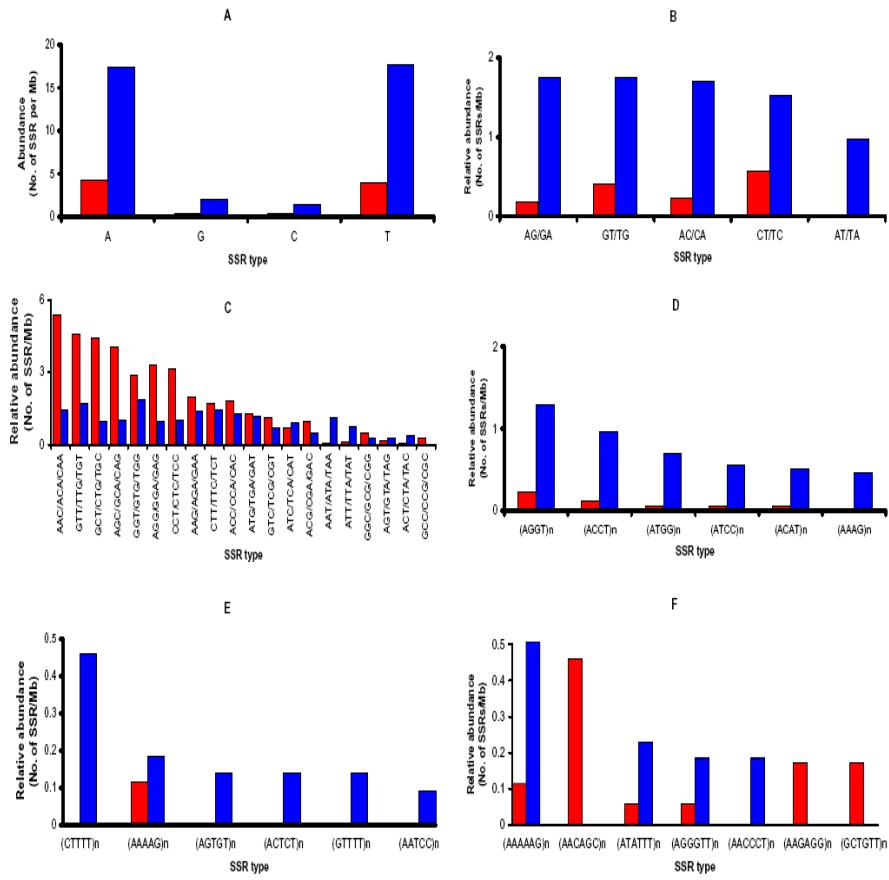
**Figure 2-1. Genome-wide distribution of relative abundance of SSRs by the unit size.**

The relative abundance was calculated by the number of SSR type / mega base-pair (MB). Each bar represents the relative abundance of SSR type in different genome locations; genic (red bar), intergenic (blue bar), and whole genome (green bar), which was calculated by (genic + intergenic) / 2. The x-axis represents SSRs that have different SSR units and the y-axis represents the relative abundance of each SSR type.

Mono-nt SSR was the second largest class by repeat unit, representing 36% of the total SSRs in *N. crassa*. Mono-nt SSRs were distributed preferentially in the intergenic and intronic regions and were rare in the exonic region (Fig. 2-1 and Table 2-1). The relative abundances of mono-nt SSRs in intergenic, intronic and exonic regions were 38.4, 56.3 and 0.2 per Mb, respectively. Among the possible four types of mono-nt repeats (poly-A, -T, -G and -C), poly-A and -T were the predominant forms: 11.4 poly-A per Mb and 11.3 poly-T per Mb in the genome (Figure 2-2).

Unlike other organisms, the di-nt SSRs were a minor class SSR type in the *N. crassa* genome (Fig. 2-1 and Table 2-1) [11, 26]. But it was consistent that the di-nt SSRs were preferentially distributed in the nongenic region (Figure 2-2) as found in other organisms [11]: 88% of di-nt SSRs present in the intergenic region and 12% in the

genic region (0.5% exon and 11.5% intron) (Figure 2-2 and Table 2-1). AG/GA, GT/TG, AC/CA are the most abundant SSR types in di-nt SSRs (about 1 SSR per Mb in each case). The relative abundance of AT/TA was about half that of AG/GA, GT/TG, and AC/CA. No GC/CG SSR type was identified in our analysis.

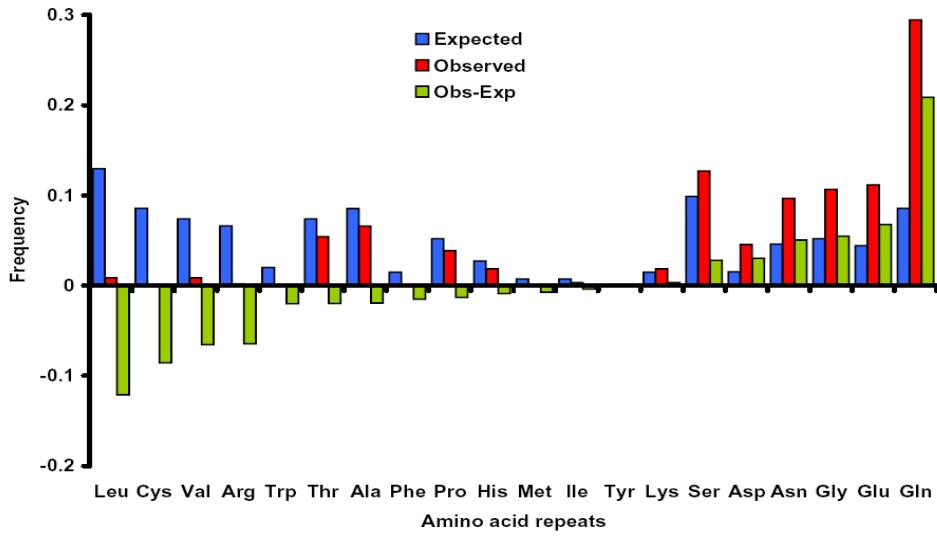


**Figure 2-2. Genome-wide distribution of relative abundance of SSRs by the SSR types in different SSR unit number.**

The relative abundances of SSRs are presented by the genome region: genic region (red bar) and non-genic region (blue bar). Each panel represents a different SSR type: mono-nucleotide SSR (A), di-nucleotide SSR (B), tri-nucleotide SSR (C), tetra-nucleotide SSR (D), penta-nucleotide SSR (E), and hexa-nucleotide SSR (F). From tetra-nucleotide SSR (D), the possible repeats of each microsatellite type are shown as (nucleotide sequence) $n$ . For example, (AGGT) $n$  stands for AGGT/GGTA/GTAG/TAGG repeats. The x-axis represents the different sequence types and the y-axis represents relative abundance where the observed count of SSRs in each category is divided by megabase of sequence.

The tri-nt repeat was the most abundant SSR in terms of unit number: 39.4 % of the total SSRs (1084 out of 2749) (Figure 2-1 and Table 2-1). The relative abundance of tri-nt SSRs in the exonic region was approximately two-fold higher than in the intergenic region (40.2 per Mb vs. 19.1 per Mb) (Figure 2-1 and Table 2-1). Among the tri-nt SSRs, AAC/ACA/CAA was the most abundant (Table 2-1). We also found that some tri-nt SSR types were not randomly distributed in the genome. For example, a group of SSR types, AAC/ACA/CAA, GCT/CTG/TGC, AGC/GCA/CAG, was preferentially located in the exonic region (Figure 2-2C). And AAT/ATA/TAA was exclusively located in the intergenic region (Figure 2-2C). The tri-nt SSRs in the exonic region are translated into amino-acid repeats, which possibly contribute to the biological function of the protein.

We investigated the frequency of the amino-acid repeats encoded by the tri-nt repeats in the exon (Figure 2-3). The frequency was measured based on the encoded amino-acid repeats that are composed of at least 5 repeats of a single amino acid without any interruption. To see if there is a bias in the distribution of amino acid repeats (AAR) encoded by the tri-nt SSRs in the exonic region, we compared the expected and observed frequencies of the encoded AAR (Figure 2-3). To estimate predicted predicted amino acid repeats encoded by exonic tri-nt SSRs. we generated the predicted amino acid sequences with an assumption that exonic tri-nt SSR sequences had an equal chance to be translated in all the possible reading frames of the tri-nt repeats. For example, SSR sequences GCTGCTGCTGCTGCTGCT can be translated in three different frames: 1) GCT GCT GCT GCT GCT, which will be translated into Ala-Ala-Ala-Ala-Ala, 2) CTG CTG CTG CTG CTG, which will be translated into Leu-Leu-Leu-Leu-Leu, and 3) TGC TGC TGC TGC TGC TGC, which will be translated into Cys-Cys-Cys-Cys-Cys. Only one of the three possible reading frames would be used to generate the “observed” amino acid repeats.

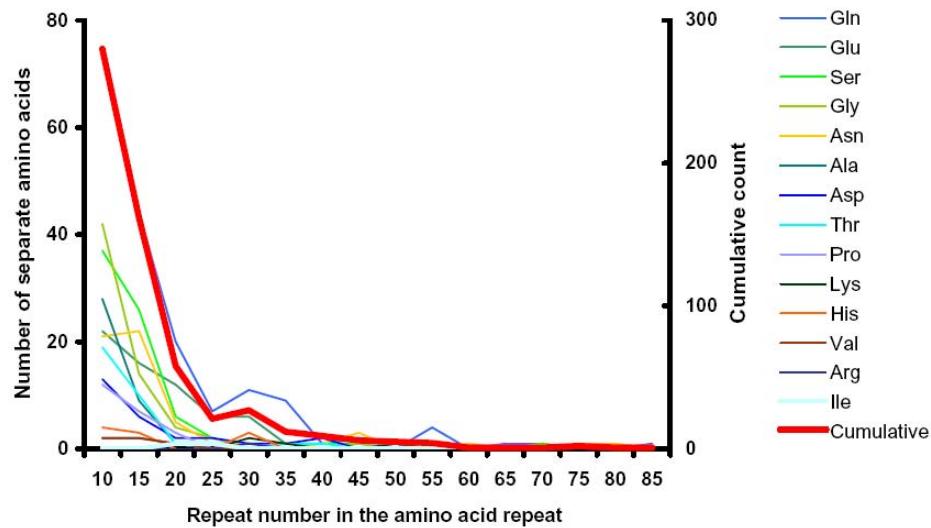


**Figure 2-3. The predicted and observed frequencies of amino acid repeats encoded by tri-nucleotide SSRs.**

Predicted (blue bars) and observed (red bars) frequencies for each amino acid repeat are presented. Green bars represent the differences between the predicted and observed frequencies for each amino acid repeat. If the expected frequency is higher than the observed frequency, the green bar is drawn below the x-axis. If the observed frequency is higher than the expected frequency, the green bar is drawn above the x-axis. Please see Methods and main text for more detailed description.

Among the AAR, three AAR accounted for 50% of the total: Glutamine (Gln), 174 repeats, 29%; Serine (Ser), 75 repeats, 12.9%; and Glycine (Gly), 66 repeats, 11.1%. Interestingly, some AAR are present far more abundantly than the expected frequency in the exonic region ( $p < 0.001$ ). These amino acids are Gln, Glutamic acid (Glu), and Asparagine (Asn), Gly. On the other hand, another group of amino acids, Cysteine (Cys), Tryptophan (Trp), Arginine (Arg), Leucine (Leu), and Valine (Val), are observed at less than expected frequencies (Fig. 2-3). The longest AAR encoded by tri-nucleotide SSRs was observed for Gln with 81 repeats. Generally, the proportion of amino acid repeats exponentially decreases as the number of repeat units increase in all types of

AAR (Fig. 2-4). This suggests that there could be functional adverse effects when an AAR becomes too large. To characterize the potential biological effects of the size variation of AAR, we grouped the proteins containing AAR using gene ontology (GO). This showed that the proteins containing the AAR that prevail in the *N. crassa* genome are involved in important biological functions in sustaining life, including physiological process (GO ID: 007582), binding (GO ID:0005488), and catalytic function (GO ID:0003824). Small modifications of these genes could trigger large effects in downstream pathways.



**Figure 2-4. The distributions of the repeat lengths of the different types of amino acid repeats encoded by tri-nucleotide SSRs.**

The x-axis represents the length of amino acid repeat and the primary y-axis (left side) is the observed count of individual amino acid repeats in a given amino acid motif (thin lines) and the secondary y-axis (right side) is the observed count of all amino acid repeats combined (thick red line).

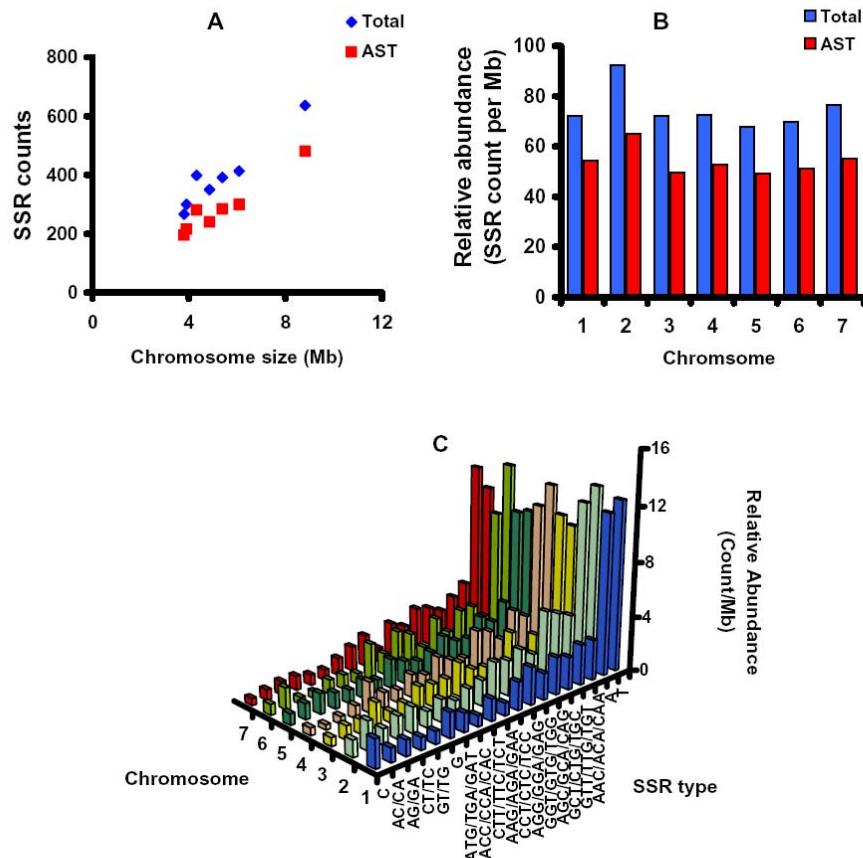
Tetra-nt SSRs were predominantly distributed in the nongenic regions (Fig. 2-2, D). The two most frequent tetra-nt SSRs were (TAGG)n and (ACCT)n, representing 1.58 and 1.07 repeats per Mb respectively in the genome (Fig. 2-2, D). Penta-nt SSRs

were also predominantly distributed in the nongenic regions (Fig. 2-2, E). The most abundant penta-nt SSR was (CTTTT)n. The relative abundance of hexa-nt SSR in the intergenic region was slightly higher than those SSRs in the exonic region: 2.14 vs 1.74 SSRs/Mb, respectively. The most common amino acid repeat encoding hexa-nt SSRs was Gln-Gln repeats (13.8% of the total amino acids encoded by hexa-nt SSRs), which is the same as Gln repeats by tri-nt SSR. The second most common AAR encoded by hexa-nt SSRs was Gly-Ser repeats and Glu-Lys repeats: 6.1 % each in hexa-nt SSRs.

### **Experiment and Result; SSR genesis rate in chromosomes and genomic locations**

The apparently non-random distribution (Figure 2-2) prompted us to further characterize the distribution of SSRs in each chromosome and different genome locations. In this analysis, our goal was to test the *de novo* genesis model in detail: what are the potential parameters that cause the genesis rate of SSRs in the *N. crassa* genome? If the genesis of SSRs (birth of SSRs) in the genome is random, one could interpret that the high abundance of SSRs as the high occurrence rate of SSRs [27].

In general, the number of SSRs increases with the size of the chromosomes except for chromosome 2 (linkage group II) (Figure 2-5 A, B). The average abundance of SSRs in chromosome 2 is significantly higher than the other chromosomes. Among 963 SSR types that we identified, only 19 different SSR types were present at least more than once per Mb (Table 2-2). Thus, we classified these SSR types as abundant SSR types (AST). Only mono-, di-, and tri-nt SSRs are included in the AST (Table 2-2). About 71% of the total *N. crassa* SSRs (1,990) belong to one of the 19 AST and the relative abundance of AST reflects the relative abundance of the total SSRs among chromosomes (Figure 2-5B). Moreover, the high copy number of AST allows us to perform statistical tests to characterize the chromosomal distribution of different SSR types.



**Figure 2-5. The distribution SSRs in different chromosomes in the *N. crassa* genome.** A. The scatter plot for cumulative abundance of SSRs (y-axis) and size of chromosome (x-axis). Red squares represent the 19 abundant SSR types (AST) and blue diamonds represent all 963 SSR types (Total). B. Relative abundance of SSRs in different chromosomes. Red bars represent AST and blue bars represent Total. C. The relative abundances of total SSRs are presented by the chromosome and by the SSR type.

**Table 2-2. Sequential analysis of deviance of log linear model for all 19 SSR types.**

Effect	Number of Parameters	Change in		Residual
		Deviance	Residual DF	Deviance
<b>Null (<math>\alpha</math> only)</b>	<b>1</b>		<b>265</b>	<b>2396.08</b>
<b>SSR type (T)</b>	<b>18</b>	<b>1504.79</b>	<b>247</b>	<b>891.30</b>
<b>Chromosome (C)</b>	<b>6</b>	<b>14.53</b>	<b>241</b>	<b>876.77</b>
<b>Genomic location (G)</b>	<b>1</b>	<b>43.98</b>	<b>240</b>	<b>832.79</b>
<b>Interaction (T x G)</b>	<b>18</b>	<b>604.00</b>	<b>222</b>	<b>228.79</b>

In addition to the variation in the distribution of different SSR types in different functional regions (Figure 2-2), the relative abundance of SSR types (SSR counts per Mb) also appears to be variable across the chromosomes (Figure 2-5C). Thus, the data suggest that the occurrence rate of a SSR type may depend on both chromosome and functional region. To statistically validate these apparent differences, we performed an analysis of SSR abundance for the 19 AST using a Poisson log linear model as shown in the previous section [28].

The probabilistic motivation for the Poisson model is that random occurrence of an SSR in the genome is synonymous with SSR “events” occurring according to a Poisson process, when traversing the genome from one end to the other. The data presented in Figure 2-2 and Figure 2-5 indicates substantial variability in abundance rates among types in each chromosome and genomic location (Figure 2-2 and 2-5). There was no a priori reason to expect variation in abundance between chromosomes. However, the log-linear modelling approach allowed multiple factors to be examined simultaneously in a unified statistical framework. Thus, we analyzed three factors in our analysis: chromosome, SSR type, and genomic location (genic vs. intergenic). Our

analysis was based on the data for the 19 AST, with a cumulative total of 1990 SSR occurrences. For the purpose of our statistical analysis, the data were summarized as two 19 by 7 contingency tables (one for genic and one for intergenic regions) giving the frequencies of the 19 SSR types on each chromosome [29]. The abundance for each SSR type/chromosome/region combination was defined as the number of SSRs divided by the length of the relevant region on the chromosome in Mb. Our statistical model assumes that the effects are all additive on a log rate scale, and therefore multiplicative on the rate scale. The goodness-of-fit of this model is summarized in the analysis of deviance decomposition given in Table 2-2. In particular, the residual deviance for the full model, 228.79 with 222 degrees of freedom, indicates a good overall fit. In addition, all of the factors in the model, including the chromosome main effects, are statistically significant. In particular, adding SSR type as an explanatory factor to the null model reduces the residual deviance by over 1500, which is clearly statistically significant ( $p<0.0001$  when compared to a chi-squared distribution with 18 degrees of freedom). Thus, abundance is clearly not uniform over SSR types. There is also a modest, but statistically significant, chromosome main effect ( $\chi^2=14.53$  with  $df=6$ ,  $p=0.02$ ). The cause of the significant chromosome effect was the higher overall SSR abundance on chromosome number 2 relative to all other chromosomes.

The statistical significance of the SSR type/genomic location interaction was partially explained by the fact that the 8 mono- and di-nt SSR types (among the 19 AST) were almost non-existent in the genic region, whereas the 11 tri-nt SSR types combined are approximately equally abundant in the genic and intergenic regions. For this reason we considered separate fits of the log linear model to the mono/di-nt SSR data and the tri-nt data, with the genomic category factor omitted from the model in the mono/di-nt data case. The sequential deviance decompositions for the two data sets are reported in Tables 2-3 and 2-4.

**Table 2-3. Sequential analysis of deviance of log linear model for 8 mono/di-nt SSR types, intergenic region only.**

Effect	Number of Parameters	Change in	Residual DF	Residual
		Deviance		Deviance
Null ( $\alpha$ only)	1		55	1130.89
SSR type (T)	7	1085.14	48	45.75
Chromosome (C)	6	8.89	42	36.86

Table 2-3 indicates a significant SSR type effect as observed in Table 2, but no strong evidence of differences between abundance rates of mono/di-nt SSR types among chromosomes. The statistical significance of the SSR type factor was a consequence of the large differences in the empirical abundance rates, which strongly suggest that the genesis rates of different SSR types were not uniform. The residual deviance after dropping chromosome as a factor in the model was 45.75 with 48 degrees of freedom, indicating a good fit for the Poisson model with SSR dependent abundance rates.

The story for the tri-nt SSR counts was more complex (Table 2-4). Dropping chromosome from the model increases the residual deviance by a statistically insignificant 10.38. Examination of the chromosome coefficients does indicate a significantly higher abundance value on chromosome number 2. This may be just a statistical anomaly or may indicate the existence of a differential SSR genesis rate in chromosome 2. The residual deviance for the reduced model was 142.69 with 136 degrees of freedom, again indicating that the Poisson variation model was reasonable. However, not only do the abundance rates vary by SSR type, but the differences depend upon the genomic location category (intergenic/genic). While abundance was generally higher in the genic region, the pattern was not uniform across all 11 tri-nt types. In some

cases there was no significant difference between genic and intergenic regions (Table 2-5).

**Table 2-4. Sequential analysis of deviance of log linear model for 11 tri-nt SSR types.**

Effect	Number of Parameters	Change in Deviance	Residual DF	Residual
				Deviance
<b>Null (<math>\alpha</math> only)</b>	1		153	<b>409.70</b>
<b>SSR type (T)</b>	10	<b>68.05</b>	143	<b>341.65</b>
<b>Chromosome (C)</b>	6	<b>10.37</b>	137	<b>331.28</b>
<b>Genomic category (G)</b>	1	<b>150.92</b>	136	<b>180.00</b>
<b>Interaction (T x G)</b>	10	<b>47.68</b>	126	<b>132.31</b>

### **Experiment and Result; Characterization of size polymorphism in SSRs among natural accessions**

Next, we tested if the size variations of SSRs among natural accessions suggest evolutionary forces for the cause of SSR size variations. We scanned the genome using a 250 kb window and randomly selected a SSR within each window. Mono-nt SSRs are not easy to accurately assay for their repeat number and could mislead our analysis [34], so they were eliminated from this analysis. Of the 1759 SSRs (after removing mono-nt SSRs), we selected 162 SSRs for further analysis as addressed in the previous section. We analyzed the characteristics of the 162 selected SSRs and found that their distribution, repeat units, and frequencies were comparable to those in the complete genome-wide collection [29].

**Table 2-5. Comparison of log abundance rates in genic and intergenic regions based on a Poisson model.**

SSR	Genic	Intergenic	Diff.	Std.err	Z-value	P-value
AAC/ACA/CAA	1.689	0.335	1.354	0.216	6.280	0.000***
AAG/AGA/GAA	0.683	0.370	0.312	0.253	1.236	0.216
ACC/CCA/CAC	0.590	0.335	0.255	0.261	0.979	0.328
AGC/GCA/CAG	1.405	0.094	1.311	0.244	5.363	0.000***
AGG/GGA/GAG	1.182	0.047	1.134	0.256	4.433	0.000***
ATG/TGA/GAT	0.247	0.222	0.026	0.292	0.088	0.930
CCT/CTC/TCC	1.164	0.047	1.116	0.257	4.352	0.000***
CTT/TTC/TCT	0.557	0.437	0.121	0.256	0.471	0.638
GCT/CTG/TGC	1.474	0.047	1.426	0.247	5.778	0.000***
GGT/GTG/TGG	1.068	0.640	0.428	0.215	1.988	0.047*
GTT/TTG/TGT	1.538	0.558	0.980	0.202	4.836	0.000***

\* indicates significance at the 5% level, \*\* 1%, and \*\*\* 0.1%

To test size polymorphism of the 162 SSRs, primers were designed and used to screen the length polymorphism in a SSR locus with 7 natural accessions. Subsequently we accessed the size variability of SSRs represented by the polymorphic index content (PIC). The PIC value 0 represents no polymorphism among alleles and the PIC value 1 represents the most complete polymorphism. Of the 162 SSR loci, 33 SSR loci were eliminated from further characterization due to PCR failure or ambiguous results. We calculated the PIC scores for the remaining 129 SSR loci. The range of the PIC scores

spans from 0.63 to 0.86. All the results of the polymorphism analysis for the 129 sampled SSR loci are described in Table 2-6 [29]. In this analysis, we considered two different parameters, physical characteristics of SSRs (repeat number, type, and length) and genome location of SSRs (chromosomes and genic vs. intergenic). First, we grouped these experimentally characterized SSR loci to test if the distributions of PIC scores are associated with different physical characteristics of the SSRs. There were no significant differences in the mean values of PIC scores among repeat units or SSR types ( $p=0.86$  and  $p=0.84$  respectively, using one-way ANOVA) (Figure 2-6A and 6B), and there was no significant correlation between PIC and repeat number ( $p=0.4$ ) (Figure 2-6C). Second, we compared PIC scores of 129 SSRs in two functional regions (genic vs. intergenic) (Figure 2-6D) and the seven chromosomes (Figure 2-6E). We also found that there were no significant differences in the distribution of PIC scores in different functional genome regions (genic vs. intergenic) ( $p = 0.2$ , Figure 2-6D) and chromosomes ( $p= 0.94$ , Figure 2-6E). Thus, these data suggest that there was no systematic difference in terms of the variations of PIC values among different physical characteristics of the SSRs tested here, or across functional regions, or chromosomes. Finally, we also compared the PIC value distributions of the same SSR type (AAC/ACA/CAA) in 20 different loci at different functional genome locations and found that there were no significant differences in PIC scores between genic and intergenic regions ( $p = 0.84$ , Figure 2-6F). These results suggest that there is no apparent bias in SSR genesis rates in 1) the physical characteristics of SSRs, 2) genomic locations, and 3) chromosomes. It is worth noting that the SSR size variability of *N. crassa* that is estimated from our study is relatively high, in comparison to other organisms [5, 6], with an average PIC score = 0.8.

We were concerned that the PIC values calculated from seven accessions might not reflect the true PIC values among all accessions in nature. To test this, we randomly chose 32 strains from the collection of natural accessions in the Fugal Genetics Stock

Center (FGSC, Kansas) and analyzed the PIC score of one SSR type, AC/CA, at three randomly chosen different loci. The PIC values of the AC/CA SSR type with two different population sizes, 7 vs. 32, were not significantly different (two sample t-test,  $p = 0.19$ , Figure 2-7).

### **Experiment and Result; Statistical inference for evolutionary forces of size variation of SSRs**

We thought that the size variation of SSRs in seven different accessions could provide some insights in terms of the occurrence of size variations in nature. We hypothesized three simple scenarios regarding the scope of evolutionary forces in SSR size variation for statistical tests: 1) Hypothesis #1 (genome-wide effect), the sizes of the SSRs in the genome are either longer or shorter for a given strain in comparison to those in other strains, 2) Hypothesis #2 (local effect), the sizes of some SSRs are either significantly shorter or longer than other SSRs, 3) Hypothesis #3, there could be both significant differences among strains and within a strain (Figure 2-8).

To test these hypotheses, we used 33 SSRs (Table 2-7) that had no missing data in seven strains (Table 2-8) [29]. The distribution of repeat numbers across all seven strains and 33 markers is considerably right-skewed.

**Table 2-6. The physical location and PIC values of 131 SSR loci in the *Neurospora crassa* genome.**

No	Name	Chromo-some	Contig	Location	Length	Unit	unit number	Repeat Number	Allele number <sup>a</sup>	Allele total tested <sup>b</sup>	PIC <sup>c</sup>
1	MN001	3	1	221811 ~ 221875	65	CCT	3	21.7	4	7	0.69
2	MN003	3	1	730845 ~ 730888	44	AGC	3	14.7	5	7	0.78
3	MN007	3	1	1500560 ~ 1500605	46	TAC A	4	11.5	5	6	0.78
4	MN008	1	2	277495 ~ 277591	97	GTT	3	32.3	5	7	0.78
5	MN009	1	2	380469 ~ 380523	55	TTC C	4	13.8	5	5	0.80
6	MN010	1	2	524605 ~ 524640	36	CTT C	4	9	3	5	0.64
7	MN011	1	2	882110 ~ 882179	70	CAA CAC	6	11.7	6	7	0.82
8	MN014	1	2	1588956 ~ 1588998	43	CAA	3	14.3	4	4	0.75
9	MN015	1	3	232922 ~ 232968	47	CAG	3	15.7	7	7	0.86
10	MN016	1	3	343335 ~ 343379	45	GAC T	4	11.3	5	7	0.78
11	MN017	1	3	652024 ~ 652064	41	ACG	3	13.7	5	6	0.78
12	MN018	1	3	896071 ~ 896119	49	AC	2	25	6	7	0.82
13	MN019	1	3	1247202 ~ 1247252	51	AGG	3	17	6	6	0.83
14	MN023	6	4	972152 ~ 972189	38	TGG	3	12.7	5	6	0.78
15	MN024	6	4	1022466 ~ 1022511	46	TTG G	4	11	4	5	0.72
16	MN026	2	5	424445 ~ 424483	39	TGC	3	13	6	7	0.82
17	MN027	2	5	523288 ~ 523331	44	TCT	3	14.7	5	7	0.73
18	MN028	2	5	833653 ~ 833728	76	GTT	3	25.3	7	7	0.86
19	MN029	1	6	66556 ~ 66597	42	ACA	3	14	3	4	0.63

**Table 2-6 (continued)**

No	Name	Chromo-some	Contig	Location	Length	Unit	unit number	Repeat Number	Allele number <sup>a</sup>	Allele total tested <sup>b</sup>	PIC <sup>c</sup>
20	MN0 30	1	6	311750 ~ 311800	51	GCA	3	17	5	5	0.80
21	MN0 32	1	7	70610 ~ 70651	42	AAG A	4	10.5	4	5	0.72
22	MN0 33	1	7	414141 ~ 414186	46	TCT	3	15.3	6	7	0.82
23	MN0 34	1	7	550900 ~ 550949	50	GAT	3	16.7	6	7	0.82
24	MN0 35	1	7	972109 ~ 972173	65	CAG CAA	6	10.8	3	3	0.67
25	MN0 36	2	8	225775 ~ 225814	40	GGT	3	13.3	5	6	0.78
26	MN0 37	2	8	445938 ~ 445986	49	CAT	3	16.3	7	7	0.86
27	MN0 38	2	8	590238 ~ 590281	44	TGT	3	14.7	4	5	0.72
28	MN0 39	2	8	860683 ~ 860717	35	TAG G	4	8.8	3	3	0.67
29	MN0 41	1	9	261776 ~ 261829	54	ACA T	4	13.5	7	7	0.86
30	MN0 42	1	9	679283 ~ 679345	63	ACA	3	21	5	5	0.80
31	MN0 45	7	10	252056 ~ 252120	65	TG	2	32.5	3	3	0.67
32	MN0 46	7	10	624049 ~ 624091	43	CTG	3	14.3	6	7	0.82
33	MN0 47	7	10	902607 ~ 902748	142	AGG T	4	35.5	6	7	0.82
34	MN0 48	5	11	160124 ~ 160173	50	CA	2	26	6	7	0.82
35	MN0 49	5	11	320460 ~ 320499	40	GCC A	4	10	4	4	0.75
36	MN0 51	5	11	764835 ~ 764878	44	CTC	3	14.7	7	7	0.86
37	MN0 52	6	12	79329 ~ 79382	54	TC	2	27.5	3	4	0.63
38	MN0 53	6	12	422999 ~ 423051	53	GGT A	4	13.3	6	7	0.82
39	MN0 54	6	12	634162 ~ 634199	38	AGC	3	12.7	6	7	0.82
40	MN0 57	5	13	264664 ~ 264705	42	AGG T	4	10.3	5	5	0.80
41	MN0 58	5	13	521169 ~ 521206	38	ACC	3	12.7	6	7	0.82
42	MN0 59	5	13	831848 ~ 831892	45	CAT	3	15	6	7	0.82

**Table 2-6 (continued)**

No	Name	Chromo-some	Contig	Location	Length	Unit	unit number	Repeat Number	Allele number <sup>a</sup>	Allele total tested <sup>b</sup>	PIC <sup>c</sup>
43	MN0 60	5	14	166889 ~ 166952	64	TGT AG	5	12.8	4	4	0.75
44	MN0 61	5	14	275448 ~ 275544	97	AC	2	48.5	4	6	0.80
45	MN0 62	5	14	731790 ~ 731824	35	GA G	3	11.7	5	7	0.73
46	MN0 65	6	16	147716 ~ 147751	36	AC	2	18	5	6	0.78
47	MN0 66	6	16	385833 ~ 385870	38	TGG	3	12.7	4	4	0.75
48	MN0 67	6	16	501791 ~ 501828	38	ACA	3	12.7	6	7	0.82
49	MN0 68	3	17	323626 ~ 323662	37	GTC	3	12.3	7	7	0.86
50	MN0 72	4	19	481576 ~ 481610	35	GGT	3	11.7	6	6	0.83
51	MN0 73	4	19	534736 ~ 534787	52	ATA C	4	13	5	5	0.80
52	MN0 74	4	20	204217 ~ 204261	45	AAC	3	14.7	5	7	0.78
53	MN0 75	4	20	256193 ~ 256291	99	TCA CCA	6	16.5	6	7	0.82
54	MN0 76	4	20	505127 ~ 505208	82	TGT	3	27.3	5	5	0.80
55	MN0 77	7	21	275287 ~ 275330	44	GA	2	22	5	5	0.80
56	MN0 78	7	21	615080 ~ 615259	180	ACA	3	60	4	6	0.72
57	MN0 79	6	22	128438 ~ 128527	90	CCT A	4	22.5	6	7	0.82
58	MN0 80	6	22	536453 ~ 536488	36	GA AA	4	9	7	7	0.86
59	MN0 81	7	23	194233 ~ 194352	120	TCA	3	40	3	3	0.67
60	MN0 82	7	23	343505 ~ 343587	83	GTA	3	27.7	4	4	0.75
61	MN0 84	3	25	25541 ~ 25581	41	TTC	3	13.7	5	7	0.78
62	MN0 86	4	26	171173 ~ 171211	39	GTT	3	13	5	5	0.80
63	MN0 87	4	26	340105 ~ 340154	50	GG A	3	16.7	3	3	0.67
64	MN0 89	3	27	400773 ~ 400817	45	GAT	3	15	6	7	0.82
65	MN0 90	4	28	138649 ~ 138692	44	TGT	3	14.7	7	7	0.86
66	MN0 92	1	29	13096 ~ 13146	51	CTC	3	17.3	3	3	0.67

**Table 2-6 (continued)**

No	Name	Chromo-some	Contig	Location	Length	Unit	unit number	Repeat Number	Allele number <sup>a</sup>	Allele total tested <sup>b</sup>	PIC <sup>c</sup>
67	MN094	2	30	255769 ~ 255808	40	TTG	3	13.3	5	5	0.80
68	MN095	7	32	24674 ~ 24717	44	CTC	3	14.7	4	4	0.75
69	MN096	7	32	262267 ~ 262346	80	GG GAA A	6	13.3	4	4	0.75
70	MN104	3	1	838136 ~ 838171	36	ACT G	4	9	4	4	0.75
71	MN108	3	1	1452588 ~ 1452625	38	TGA	3	12.7	4	6	0.72
72	MN112	1	2	700491 ~ 700535	45	CAC	3	15	5	7	0.78
73	MN114	1	2	1056415 ~ 1056474	60	AA G	3	20	4	4	0.75
74	MN116	1	2	1648807 ~ 1648894	88	TCT	3	29	3	3	0.67
75	MN117	1	3	711452 ~ 711509	58	AA AG	4	14.3	4	4	0.75
76	MN119	6	4	293442 ~ 293472	31	GTG	3	10.3	4	6	0.72
77	MN121	6	4	795125 ~ 795174	50	GA A	3	16.7	5	7	0.78
78	MN127	2	5	914661 ~ 914699	39	AC	2	19.5	3	5	0.64
79	MN128	1	6	200004 ~ 200161	158	CAA	3	52.7	5	7	0.78
80	MN129	1	7	636597 ~ 636632	36	AC	2	18	5	7	0.64
81	MN131	1	7	823852 ~ 823909	58	AA GC	4	14.5	3	4	0.63
82	MN132	2	8	602669 ~ 602703	35	CAG C	4	8.8	4	5	0.72
83	MN136	1	9	478497 ~ 478537	41	CGT T	4	10.3	6	7	0.82
84	MN142	7	10	722692 ~ 722745	54	CT	2	27	3	3	0.67
85	MN150	5	13	361740 ~ 361779	40	TTC	3	13.3	5	7	0.78
86	MN153	5	14	449098 ~ 449139	42	GCT	3	14	5	5	0.80
87	MN154	5	14	564948 ~ 565014	67	AG	2	33.5	4	4	0.75
88	MN157	6	16	245477 ~ 245519	43	GT	2	21.5	5	7	0.78
89	MN162	4	19	270621 ~ 270662	42	CAA	3	14	6	6	0.83

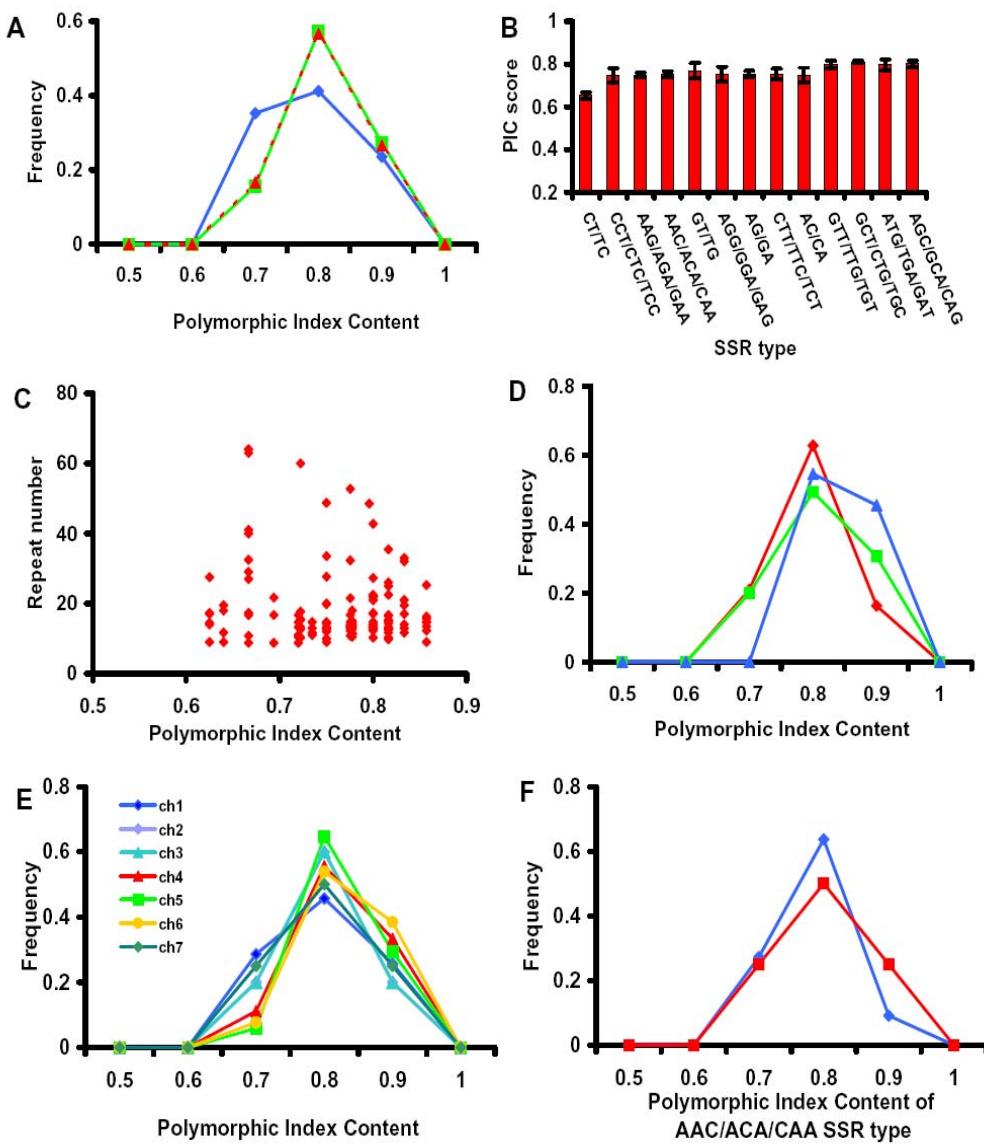
**Table 2-6 (continued)**

No	Name	Chromo-some	Contig	Location	Length	Unit	unit number	Repeat Number	Allele number <sup>a</sup>	Allele total tested <sup>b</sup>	PIC <sup>c</sup>
90	MN1 64	4	19	439820 ~ 439856	37	CAA	3	12.3	5	5	0.80
91	MN1 67	4	20	621911 ~ 621950	40	AA G	3	13.3	4	6	0.72
92	MN1 68	7	21	157800 ~ 157838	39	TGA	3	13	6	6	0.83
93	MN1 70	7	23	430962 ~ 431003	42	AGC	3	14	5	6	0.78
	MN1			290074 ~		GGT					
94	71	5	24	290135	62	A	4	15.5	4	6	0.72
95	MN1 73	3	25	289671 ~ 289709	39	AGC	3	13	5	6	0.72
96	MN1 78	3	27	153363 ~ 153415	53	TGC C	4	13.3	4	7	0.78
97	MN1 79	3	27	238013 ~ 238056	44	AAC	3	14.7	5	5	0.72
98	MN1 82	4	28	141787 ~ 141849	63	CTC	3	21	6	5	0.80
99	MN1 84	1	29	13096 ~ 13146	51	CTC	3	17.3	5	6	0.83
100	MN1 86	7	32	351411 ~ 351449	39	CAG G	4	9.8	6	5	0.80
101	MN1 88	2	33	341828 ~ 341953	126	TC	2	63	3	7	0.82
102	MN1 91	4	35	134360 ~ 134438	79	GG AT	4	19.8	4	3	0.67
103	MN1 92	4	35	224393 ~ 224431	39	TG	2	19.5	6	4	0.75
104	MN1 94	4	36	260119 ~ 260171	53	GG A	3	17.7	4	6	0.83
105	MN1 96	5	37	223562 ~ 223611	50	GCA	3	16.7	4	6	0.72
106	MN1 97	1	38	157604 ~ 157643	40	AA GA	4	10	6	5	0.72
107	MN1 99	1	39	137780 ~ 137824	45	TG	2	22.5	5	7	0.82
108	MN2 01	3	40	218684 ~ 218749	66	AG	2	33	7	5	0.80
109	MN2 03	5	41	229996 ~ 230091	96	ACA	3	32	7	7	0.83
	MN2			246781 ~							
110	05	3	42	246972	192	AAC	3	64	3	3	0.67
	MN2			116121 ~							
111	08	2	44	116161	41	TGT	3	13.7	5	7	0.78
112	MN2 13	5	46	132777 ~ 132812	36	CAA	3	12	4	7	0.78

**Table 2-6 (continued)**

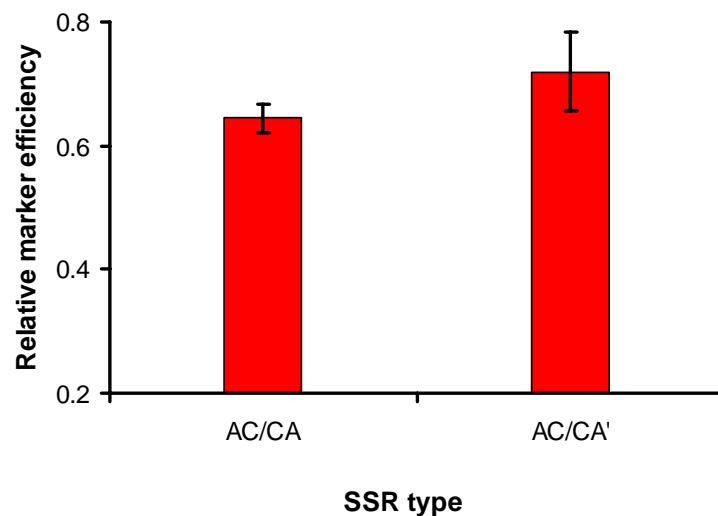
No	Name	Chrom- osome	Contig	Location	Length	Unit	unit number	Repeat Number	Allele numbera	Allele total tested <sup>b</sup>	PIC <sup>c</sup>
113	MN2 15	4	47	84629 ~ 84670	42	AGC A	4	10.8	5	4	0.75
114	MN2 20	4	51	173992 ~ 174041	50	CAA	3	16.7	4	7	0.78
115	MN2 25	2	54	210836 ~ 210872	37	ACA	3	12.3	5	7	0.69
116	MN2 27	1	56	88744 ~ 88785	42	CTT G	4	10.5	5	7	0.78
117	MN2 29	2	57	104129 ~ 104172	44	CCG A	4	11	5	6	0.78
118	MN2 31	1	58	110853 ~ 110887	35	GAC C	4	8.8	4	7	0.73
119	MN2 34	1	62	8763 ~ 8813	51	CCT	3	17	3	7	0.69
120	MN2 36	5	63	115653 ~ 115704	52	CAA	3	17.3	3	4	0.63
121	MN2 39	1	65	38240 ~ 38278	39	GTG	3	13	5	4	0.63
122	MN2 40	7	66	71475 ~ 71602	128	AAC	3	42.7	5	7	0.73
123	MN2 41	2	68	34772 ~ 34815	44	GA	2	22	5	5	0.80
124	MN2 42	3	69	83662 ~ 83807	146	AG A	3	48.7	4	5	0.80
	MN2			67709 ~		TGT					
125	43	1	70	67744	36	A	4	9	3	4	0.63
	MN2			2479 ~							
126	45	3	74	2513	35	TAT	3	11.7	3	5	0.64
	MN2			24819 ~							
127	46	7	75	24941	123	AAC	3	41	3	3	0.67
	MN2			46844 ~							
128	48	7	78	46906	63	ACA	3	21	5	5	0.80
	MN2			60095 ~							
129	49	4	79	60140	46	AAC	3	15.3	6	4	0.72
	MN2			46844 ~							
128	48	7	78	46906	63	ACA	3	21	5	5	0.80
	MN2			60095 ~							
129	49	4	79	60140	46	AAC	3	15.3	6	4	0.72

<sup>a</sup>, The allele number refer to the number of different allele in a given SSR locus.<sup>b</sup>, The total allele number refer to the tested allele from the possible 7 alleles in a SSR locus.<sup>c</sup>, PIC stands for polymorphism information contents (see method)<sup>d</sup>, NA stands for Not acquired

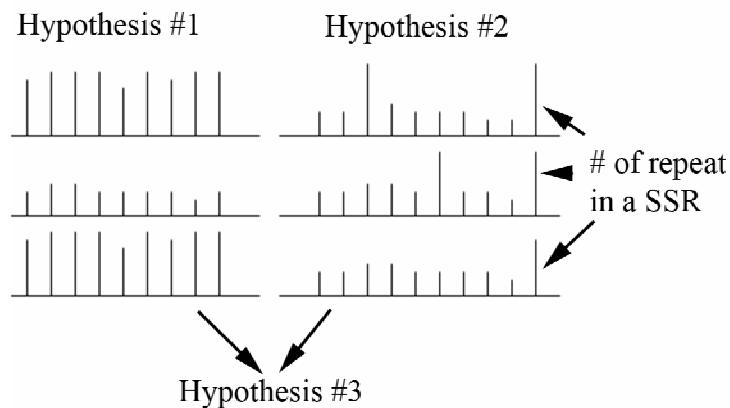


**Figure 2-6. The polymorphic information content analysis from 129 SSR loci.**

A. The frequencies of the PIC values of 129 SSR loci are displayed by the SSR type: di-nucleotide SSR (blue diamond), tri-nucleotide SSR (green square), and tetra-nucleotide SSR (red triangle). B. The comparison of PIC scores in the sampled SSRs of different SSR types. The error bars show the standard error of the mean. C. The scatter plot for PIC score and the repeat number among the sampled SSRs. The frequencies of the PIC values of 129 SSR loci are displayed by the genome region (D), exon (red diamond), intron (blue triangle) and intergenic region (green square), and by the chromosome (E). F. PIC score distribution of one SSR type, AAC/ACA/CAA, located either in genic (blue diamond) or intergenic (red square) regions.



**Figure 2-7. Comparison of PIC values of the AC/CA SSR type in two different population sizes.**



**Figure 2-8. Three hypotheses for the size variation of SSRs**

This skewness was largely removed by a (natural) log transformation. In the following analysis we attempted to isolate the sources of variation in the log transformed repeat numbers by taking into account the strain, chromosome, genome regions (genic vs. intergenic), and SSR type. It is worth noting here that, unlike in the analysis of the SSR counts, there is no particular reason that the repeat number should have a Poisson distribution. Accordingly, our analysis of the repeat numbers uses classical linear models with the natural logarithm of the repeat number as the response variable.

Comparison of the averages of repeat numbers for the seven strains using one-way ANOVA indicates significant differences among the strains. Pairwise comparisons indicate that strain FGSC#2489 has a significantly higher average of repeat numbers than the other six strains ( $P < 0.05$  for all pairwise comparisons with strain FGSC#2489).

**Table 2-7. The list of 33 SSR loci for statistical analyses**

Marker number	genomic region (genic or intergenic)	genomic region (exonic,intron and intergenic)	SSR type
MN001	genic	Exon	cct/ctc/tcc
MN003	Intergenic	Intergenic	agc/gca/cag
MN011	Intergenic	Intergenic	caacac/aacacc/acacca/caccaa/accaac/ccaaca
MN018	Intergenic	Intergenic	ac/ca
MN026	Intergenic	Intergenic	tgc/gct/ctg
MN027	genic	Exon	tct/ctt/ttc
MN028	genic	Intron	gtt/ttg/tgt
MN033	Intergenic	Intergenic	tct/ctt/ttc
MN034	Intergenic	Intergenic	gat/atg/tga
MN037	genic	Intron	cat/atc/tca
MN041	Intergenic	Intergenic	acat/cata/atac/taca
MN046	genic	Intron	ctg/tgc/gct
MN048	Intergenic	Intergenic	ac/ca
MN053	Intergenic	Intergenic	ggta/gtag/tagg/aggt
MN054	genic	Exon	agc/gca/cag
MN059	Intergenic	Intergenic	cat/atc/tca
MN067	Intergenic	Intergenic	aac/aca/caa
MN068	Intergenic	Intergenic	gtc/tcg/cgt
MN121	genic	Exon	gaa/aag/aga
MN125	genic	Exon	aac/aca/caa
MN129	Intergenic	Intergenic	ac/ca
MN136	non genic	Intergenic	cgtt/gtta/ttag/tagt
MN150	genic	Exon	tct/ctt/ttc
MN157	genic	Intron	gt/tg
MN173	Intergenic	Intergenic	agc/gca/cag
MN201	genic	Intron	ag/ga
MN203	genic	Exon	aac/aca/caa
MN215	Intergenic	Intergenic	agca/gcaa/caag/aagc
MN220	genic	Exon	aac/aca/caa
MN225	genic	Exon	aac/aca/caa
MN229	Intergenic	Intergenic	ggca/gcag/cagg/aggc
MN239	genic	Exon	gtt/ttg/tgt
MN250	genic	Intron	ctg/tgc/gct

**Table 2-8. Line-cross populations from *N. crassa* accessions.**

Cross number	Parents*	Mating type	Origin of collection	Polymorphic SSRs
N2	3223	<i>mat A</i>	Louisiana, U.S.A.	
	4724	<i>mat a</i>	Penang, Malaysia	74
N4	4720	<i>mat A</i>	India	
	4715	<i>mat a</i>	Haiti	94
N6	4825	<i>mat A</i>	Tiassalel, Ivory Coast	
	2223	<i>mat a</i>	Iowa, U.S.A.	91

\* Fungal Genetics Stock Number

The strain FGSC#2489 is the sequenced standard laboratory strain that has been developed through an extensive backcrossing in the laboratory [35]. We are tempted to speculate that the systematic difference in the repeat numbers of SSRs between FGSC#2489 and other natural accessions could be a result of repeated selection in the laboratory environment. This supports our hypothesis #1 that there was a strain specific (genome-wide) effect in SSR size variation. There are also significant differences among the six natural strains. However, after Bonferroni adjustment of the pairwise P-values, the only significant differences were strains FGSC#3223 and FGSC#2489 having higher average repeat numbers than strains FGSC#4720 and FGSC#4724. Strains FGSC#4825, FGSC#2223, and FGSC#4715 have average repeat numbers in between these two pairs, none being significantly different from either extreme after Bonferroni adjustment.

In an attempt to more carefully analyze the size variation, we performed two-way ANOVA analyses: 1) strain by functional regions, 2) strain by chromosome, and 3) strain by SSR type. Analysis of the means by strain and region shows no significant interaction ( $P=0.36$ ) and no genic region main effect ( $P=0.07$ ). Analyses of the means by strain and chromosome shows no significant interaction ( $P>0.99$ ) and no significant

chromosome main effects ( $P < 0.30$ ), but a significant strain effect ( $P < 0.0001$ ). The estimated strain effects have a similar pattern as in the one-way analysis. Thus, this result also supports the hypothesis that there was a genome-wide regulation of SSR repeat number.

Since each marker occurs exactly once in each strain, it was not possible to conduct a global test for strain by marker interaction. However, a singular-value decomposition of the 33 by 7 interaction matrix,  $M$  [36], reveals that there are two dominant components that account for more than 60% of the residual variation after accounting for strain and marker main effects. Thus, we consider a linear model of the form,

$$Y_{ij} = \mu + \beta_i^S + \beta_j^M + \lambda_1 u_{1i} v_{1j} + \lambda_2 u_{2i} v_{2j} + \varepsilon_{ij},$$

where  $Y_{ij}$  is the log repeat number for marker  $j$  on strain  $i$ ,  $\mu$  was the overall mean,  $\beta_i^S$  is the main effect of strain  $i$ , and  $\beta_j^M$  is the main effect of marker  $j$ . The vectors  $u_k$  and  $v_k$ ,  $k = 1, 2$ , are the unit eigenvectors corresponding to the largest two eigenvalues of the matrices,  $MM'$  and  $M'M$ , respectively. This is an additive main effects and multiplicative interaction (AMMI) model that has been widely used in the analysis of agricultural yield trials [37]. The  $\varepsilon_{ij}$  terms account for residual variation (interaction) not explained by the multiplicative component. The least squares estimate of the parameter,  $\lambda_k$ , is equal to the singular value associated with the eigenvectors,  $u_k$  and  $v_k$ , in the singular value decomposition of the interaction matrix,  $M$ . The square of this singular value is equal to the sum of squares explained by the multiplicative interaction component in the ANOVA decomposition for this model as summarized in Table 6.

Examination of the first eigenvector for markers reveals a very large positive loading on marker 48. On the other hand, the first eigenvector for strain was essentially a contrast between two groups of strains, one group including FGSC#4720 and FGSC#4715 and other group including FGSC#4825, FGSC#2223, FGSC#4724, FGSC#3223, and

FGSC#2489. Thus, one source of the strain by marker interaction appears to be caused by the extremely high repeat numbers for marker 48 in strains FGSC#4720 and FGSC#4715, relative to the other five strains. The second eigenvector for markers has a dominant positive loading on marker 201 and a dominant negative loading on marker 1. In this case, the corresponding eigenvector for strain was a contrast between the pair of strains, FGSC#4724 and FGSC#3223, and the remaining five strains. Thus, a second source of interaction appears to be due to the contrast between these two groups of strains with respect to the difference in repeat numbers between markers 1 and 201, relative to this contrast for any other pair of markers. These data support our hypothesis #3 that there are variations in SSR repeat numbers that genome-wide effects alone cannot explain.

An alternative simple analysis is to look for markers that have highly variable (log) repeat numbers across the seven strains. Under the assumption that the seven (log) repeat numbers for a particular marker are a random sample from a normal distribution, the sample variance is proportional to a chi-squared statistic with 6 degrees-of-freedom. Specifically,  $6s^2 / \sigma^2 \sim \chi^2(6)$ , where  $\sigma^2$  is the unknown true variance. Using the chi-squared reference distribution with  $\sigma^2$  replaced by the median sample variance from the 33 markers, we found 4 markers with significantly large sample variances ( $P < 0.005$ ). In order of increasing variance these are markers 201, 34, 48 and 1. Thus, three of the four markers with the largest sample variances are the ones found using the ANOVA methods. Based on these data, we concluded that there are genome-wide, chromosomal, and local effects in size variation of SSRs.

### **Experiment and Result; Genetic map construction**

It was suggested that SSRs could be useful molecular markers for genetic analysis in intra-species populations due to the hypervariability of SSRs [8]. Earlier in this paper we also confirmed this high variability of SSRs in *N. crassa*. In addition to

hypervariability, a useful genetic marker should show stable inheritance. Thus, we wanted to examine the stability of the SSRs as genetic markers by constructing linkage maps from intra-species populations generated by crossing *N. crassa* natural accessions (Table 2-8). We also reasoned that the polymorphic SSR markers could provide a means of detecting chromosome rearrangement if there was a significant chromosome rearrangement among accessions. We found that 140 SSR markers out of the 162 (86.4%) exhibited polymorphisms of either co-dominant or presence/absence types in at least one pair of the mapping parents.

The availabilities of the polymorphic markers are predominately population specific and the mapped SSR loci varied in the different mapping populations (Table 2-9). Of the 109 mapped loci, only 17 loci (13% of the mapped loci) were mapped into all the mapping populations and 47 markers were common in at least two mapping populations. About 18%-34% of the SSRs depending on the population showed significant segregation distortion. We detected 7 different genomic regions where the segregation distortions were observed in at least two populations. Especially, the region covering contig 7 at chromosome 1R consistently revealed the deviated segregations in the mapped SSR loci (six SSR loci out of 8 SSR loci). The total genetic length and loci density of the three genetic maps are summarized in Table 5. The total genetic distances varied in the three line-cross populations; N2 cross, 547.1 cM with an average marker distance (amd) of 13.0 cM; N4 cross, 882.7 cM with an amd of 13.0 cM; N6 cross, 934.8 cM with an amd of 13.7 cM.

Utilizing the polymorphic SSR markers, 188 F1 haploid progeny derived from each mapping population were genotyped. We coalesced 109 SSR loci out of 140 SSR loci into the three genetic maps: N2 (50 out of 71 SSRs, 70.04 %), N4 (69 out of 94 SSRs, 77.2 %), and N6 (70 out of 91 SSRs, 76.9 %).

To evaluate the co-linearity of the mapped SSR loci among the three populations and the physical map based on the sequenced strain, FGSC# 2489, SSR marker orders

among the three mapping population were cross examined by using commonly mapped loci and a physical map. The positions of the mapped SSR loci from the three mapping populations were highly consistent with the physical map positions, with few exceptions (Figure 2-9). These exceptions are found in closely linked markers, especially when the markers are located in the same contig, for example, MN153 and MN061 on linkage group 5. No errors in genotyping or significant segregation distortion at the adjacent markers were detected in most case (Table 2-9).

### **Discussion; Distribution of SSRs in the sequenced *N. crassa* genome**

There is a discrepancy between the numbers of the estimated SSRs in the *N. crassa* genome in our study and a previous report [41]. The discrepancy could be attributed to the following facts: 1) we used different algorithms from the one used in the previous analysis; and 2) we used the most up-to-date genome sequence (release 7), whereas, the previous analysis used an earlier version (release 3) of the genome sequence. In addition, it should be noted that there is currently no consensus among researchers regarding how to define SSRs [1]. To achieve our goals, we applied more stringent conditions to define the SSRs than previously used. Since rates of SSR mutations are positively correlated with SSR lengths, we chose to have the number of nt within the SSR locus to be greater than 21 [1, 42, 43].

We found similarities and differences in SSR compositions and distributions between the sequenced *N. crassa* genome and other eukaryotic genomes. In the mono-nt SSRs, which accounted for 36% of the total SSRs, poly-A/T was far more abundant than poly-G/C. Indeed, A/T is most abundant across the *N. crassa* genome.

**Table 2-9. The marker quality of the mapped SSR loci.**

Population	Chromosome	SSR Locus	Contig	Number of progenies tested <sup>a</sup>	Proportion of progenies tested <sup>b</sup>	$\chi^2$ value	p value <sup>c</sup>
n4	3	MN003	1	175	93%	4.17	0.01< p<0.05
n4	1	MN010	2	139	74%	28.55	<0.001
n6	1	MN013	2	175	93%	10.57	<0.001
n6	1	MN015	3	166	88%	11.66	<0.001
n2	1	MN017	3	138	73%	10.46	<0.001
n6	1	MN018	3	154	82%	14.96	<0.001
n4	1	MN019	3	154	82%	10.39	<0.001
n2	2	MN027	5	135	72%	41.67	<0.001
n6	1	MN128	6	184	98%	33.07	<0.001
n6	1	MN035	7	160	85%	13.23	<0.001
n2	1	MN129	7	166	88%	16.29	<0.001
n6	1	MN129	7	179	95%	16.90	<0.001
n4	2	MN037	8	162	86%	5.56	0.01< p<0.05
n6	2	MN037	8	177	94%	10.45	<0.001
n6	2	MN038	8	167	89%	11.07	<0.001
n6	1	MN042	9	167	89%	31.91	<0.001
n6	1	MN136	9	155	82%	6.20	0.01< p<0.05
n6	5	MN051	11	175	93%	25.65	<0.001
n4	5	MN150	11	159	85%	6.85	0.001< p<0.01
n6	6	MN053	12	165	88%	7.42	0.001< p<0.01
n6	6	MN054	12	151	80%	10.07	0.001< p<0.01
n4	5	MN059	13	169	90%	6.44	0.01< p<0.05
n6	5	MN155	15	184	98%	7.85	0.001< p<0.01
n2	6	MN067	16	185	98%	50.86	<0.001
n6	6	MN157	16	161	86%	9.45	0.001< p<0.01
n4	4	MN074	20	166	88%	29.52	<0.001
n4	4	MN075	20	153	81%	6.28	0.01< p<0.05
n4	4	MN167	20	141	75%	15.67	<0.001
n6	7	MN078	21	183	97%	10.10	0.001< p<0.01
n2	7	MN168	21	130	69%	7.88	0.001< p<0.01
n6	6	MN080	22	169	90%	6.44	0.01< p<0.05
n2	7	MN082	23	133	71%	8.19	0.001< p<0.01
n6	5	MN083	24	167	89%	6.52	0.01< p<0.05
n4	4	MN086	26	180	96%	28.80	<0.001
n6	3	MN089	27	165	88%	11.21	<0.001
n4	4	MN090	28	141	75%	24.69	<0.001
n6	4	MN191	35	181	96%	14.37	<0.001
n2	1	MN199	39	182	97%	12.66	<0.001
n2	4	MN215	47	123	65%	61.54	<0.001
n2	4	MN220	51	137	73%	11.10	<0.001
n6	4	MN220	51	171	91%	16.43	<0.001
n6	4	MN223	53	170	90%	15.91	<0.001
n4	5	MN236	63	134	71%	28.69	<0.001

**Table 2-9 (continued)**

Population	Chromo-some	SSR Locus	Contig	Number of progenies tested <sup>a</sup>	Proportion of progenies tested <sup>b</sup>	$\chi^2$ value	p value <sup>c</sup>
n6	5	MN236	63	183	97%	23.09	<0.001
n6	7	MN247	76	164	87%	7.05	0.001<p<0.01
n6	5	MN250	80	166	88%	16.29	<0.001

<sup>a</sup> The number progenies genotyped in each SSR locus.

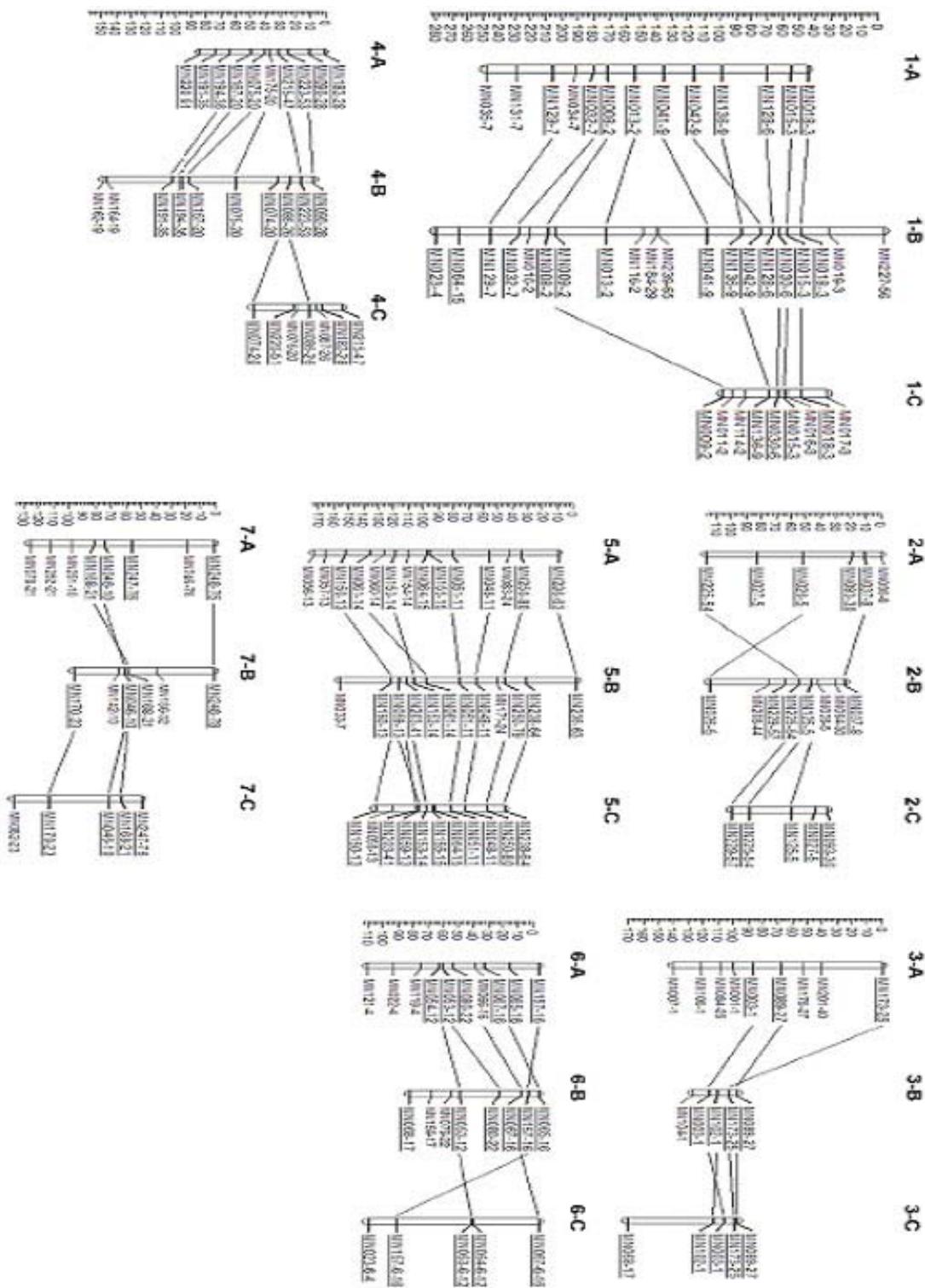
<sup>b</sup> # of genotypes progeny / # of total progeny (188) \* 100

<sup>c</sup> The p values were determined based on c2 distribution in df=1. In df=1, p =0.05 is equal to 3.84 in c2 value and p=0.01 is 6.65, p=0.001 is 10.8 respectively.

**Figure 2-9. Genetic linkage maps of three mapping populations.**

Each linkage group is named according to their cross number and the corresponding chromosome number. Cross number is expressed by a letter; A (N6, FGSC#4825 x FGSC#2223), B (N4, FGSC#4720 x FGSC4715), and C (N2, FGSC#3223 x FGSC#4724). For example, 1-A indicates the linkage group that corresponds to chromosome 1 in the cross N2. The corresponding linkage groups from different crosses are aligned based on the relative positions of anchor markers. The anchor markers are underlined and connected by thin lines among the corresponding linkage groups. The physical location of each marker is indicated by the super-contig number (Broad Institute,

<http://www.broad.mit.edu/annotation/genome/neurospora/Home.html>) followed by the marker name, e.g. MN018-3 and MN015-3 are two markers that are located in the super-contig 3. The scale on the left of each linkage group shows a relative map position denoted by centi-morgan (cM).



Most mono-nt SSRs were located in the intergenic and the intronic regions but rarely located in the exonic regions. This overrepresented A/T SSR tract in the *N. crassa* genome resembles the pattern found in the primate genome [2, 41, 44, 45]. In general, di-nt SSRs are the most common SSRs in many organisms [2, 45]. However, di-nt SSRs represent only 6.9% of the total SSRs in the *N. crassa* genome. Among the di-nt SSRs, the proportion of the AT/TA SSR type was smaller than those of the other di-nt SSRs: AG/GA, GT/TG, AC/CA, and CT/TC. This result may reflect the difference in SSR compositions between fungal and other organisms [1]. It is also possible that the AT/TA SSR type could have been underestimated because of our stringent SSR definition (discussed earlier), thus accounting for the difference between the studies.

Tri-nt SSR is the major class of SSRs in the *N. crassa* genome. In our analysis, tri-nt SSRs accounted for 39.4 % of the total. This is larger than the di-, tetra-, penta- and hexa-SSRs combined (24.5%). The predominance of tri-nt SSRs in *N. crassa* appeared to be unique feature compared to other sequenced fungal genomes [41]. In terms of relative abundance, there were twice as many tri-nt SSRs present in the exonic region than in the intronic and intragenic regions combined. The enrichment of the tri-nt in the exonic region has been observed in other eukaryotic organisms across taxa [2, 10, 44, 46]. This pattern was attributed to a tight negative selection on the other SSRs (other than tri-nt SSRs) that would perturb the reading frame in the coding regions [2, 10, 47-49]. Our analysis shows that most SSRs (74%) are predominantly distributed in the intergenic and intronic regions, with tri-nt and hexa-nt SSRs being exceptions (Table 1 and Fig. 1). Moreover, the presence of SSRs, such as ATG variants that could act as a start codon, or TTA variants that could act as a stop codon, are restricted in the exonic region (See Table 2-1, Figure 2-1 and Figure 2-2).

## **Discussion; Potential role of AAR encoded by tri-nt SSRs**

Our results suggest that AAR encoded by tri-nt SSRs have undergone positive and negative selections, depending on their sequence types: three AAR (Gln, Glu and Ser) were over-represented and three AAR (Leu, Cys and Val) were under-represented in the genome (Figure 2-3). This suggested that the observed size variation of tri-nt SSRs within a gene may be differential, possibly due to functional selection on the amino acid reiteration in encoded proteins [10, 11, 50]. Previous analyses of protein database and genomic sequence in different taxa found that AAR stretches of small hydrophilic amino acids were more tolerated in proteins [10, 50]. In agreement with previous reports, our data showed that the hydrophilic amino acids including Gln, Glu and Ser repeats are over-represented in the *N. crassa* genome (Fig. 2-3). However, the tolerance of AAR stretches in proteins has certain restrictions. Our results showed that the proportion of AARs exponentially decreased as the number of repeat units increased in all types of AARs, with 25 repeats being a critical threshold (Figure 2-4). This may be because longer AAR repeats have such detrimental effects on protein functions that they are apt to be selected out in the genome [10, 50, 51].

Numerous lines of evidence have been accumulated to support the potential roles of the AAR encoded by tri-nt SSRs in the functional divergence of proteins [10, 12, 52, 53]. Hydrophilic AAR stretches can be a major source of phenotypic variations [10, 18]. For instance, expansion of CAG repeats resulting in poly-Gln repeats in various neurological genes in humans can cause changes to their original gene functions and lead to various neuronal disorders including Huntington's disease, dentatorubro-pallidoluysian atrophy, spinbulbar muscular atrophy, and spinocerebellar ataxia [53].

It is suggested that gene duplication has a fundamental role in diversifying gene function [54]. However, diversifying gene function by gene duplication is

probably not a good option for *N. crassa* because it has a genome defense mechanisms, Repeat-Induced Point mutation or RIP [54]. Neurospora detects duplicated copies of sequences in the genome and mutates both sequences by repeated point mutations during the sexual cycle [54]. Thus, the questions of whether and how *N. crassa* could generate diversified functional genes has been raised [55]. We propose that the AAR encoded by tri-nt SSRs might have a crucial role in creating functional variability of gene regulation. Since RIP requires a minimal duplicated sequence length of about 400 base pairs (bp)[55], a tandem repeat of SSRs less than 400 bp within a gene may escape from the influence of RIP and hence may modify the original gene functions efficiently [10, 56]. The proteins including AAR are in diverse functional groups. Further, the size variations of tri-nt SSR in exonic regions are variable across the *N. crassa* genome (Figure 2-6). These raise the possibility that highly active contraction and expansion of the tri-nt SSRs in exonic regions may play roles in the evolution of gene functions that may facilitate adaptation in new environments [10, 56].

We explored the possibility of SSRs being a target of functional variation of circadian rhythms in nature. First, we surveyed the variation of repeat numbers of SSRs located in ORFs of known circadian clock genes, *white collar-1* (*wc-1*), *white collar-2* (*wc-2*), *vivid* (*vvd*), and *frequency* (*frq*), among 143 *N. crassa* natural accession collected from all over the world. We found significant size variations of SSRs in clock genes. Furthermore, these variations were associated with circadian rhythms [57]. WC-1 is a blue-light receptor for circadian clock and it functions as an activator in a complex with a partner, *wc-2*. We focused our study on the polyglutamine repeat domain in the amino-terminal of *wc-1*, NpolyQ, which has been proposed as an activation domain [58, 59]. Previous studies also suggested that NpolyQ plays a role in clock-specific activation [60, 61]. We found that NpolyQ is a

target for period variation. Furthermore, we found evidence that variation in the circadian clock was associated with latitude of collection, which suggested that the WC-1 genotype provided an adaptive advantage in natural populations. The quantitative role of variation in the amino-terminal polyglutamine (NpolyQ) domain of WC-1 in period variation among accessions has been confirmed in an independent experimental line cross population [57]. Further functional characterization will be directed toward determining the effects of the variable AAR encoded by tri-nt SSRs on the corresponding gene functions and their ecological implication.

### **Discussion; Evolutionary inference of SSR variations in *N. crassa***

We attempted to infer factors on size variation of SSRs in the *N. crassa* genome by statistical analysis of size variations of 33 markers in seven accessions. Our results suggest that there were at least three different levels of statistically significant factors (genome-wide, chromosome-specific, and local effects) involved in size variations of SSRs in the *N. crassa* genome. Our study does not address the actual mechanism of variation in SSR repeat numbers; however, it provides foundations for further experimental verification. One of the widely discussed theories on the genesis of the length variation of SSRs is the strand-slippage theory, that the variation of length in SSRs is caused by slipped-strand mis-pairing and subsequent errors during DNA replication, repair, and recombination [1, 11, 62, 63]. This could be a good explanation for the strand-slippage theory with an assumption that there is no bias in the rate of mis-pairing in genomic regions during the replication process. However, there are reports that the length variation does not follow in a step-wise manner because the efficiency of the length variation may differ due to numerous local circumstances in the genome [1, 10, 11]. Our data also suggest that the genome-wide mechanism cannot be the only source variation for SSR size variations. The existence

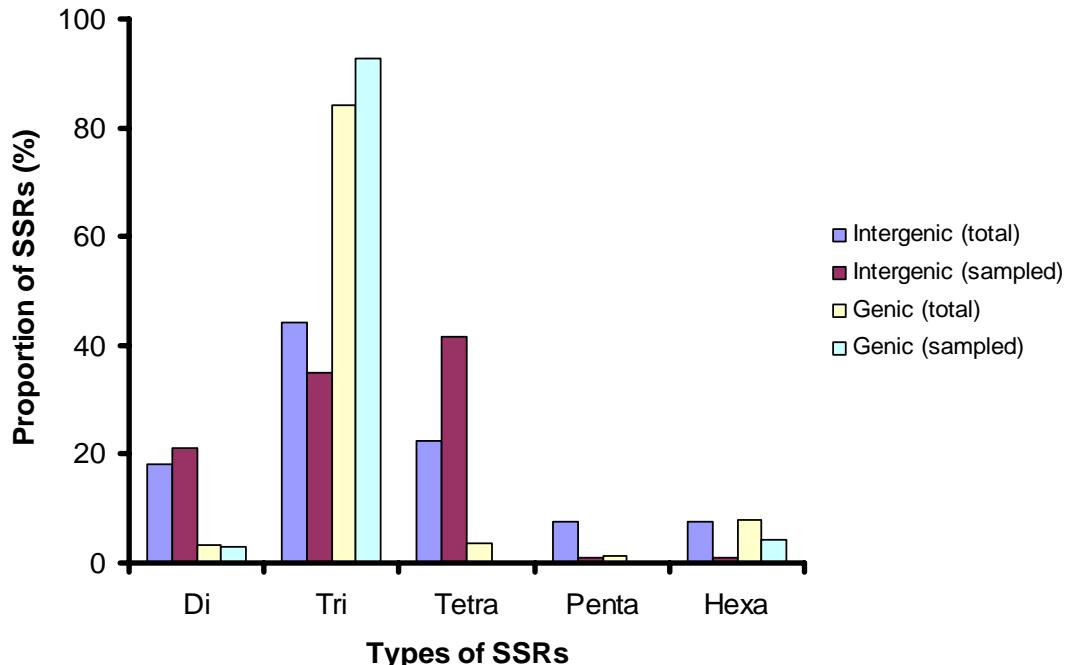
of chromosome-specific and local effects suggests that genomic context is an important factor for the variation of SSR repeat numbers. More research should be focused on the factors influencing the local variation.

RIP could be a potential mechanism for the observed species-specific bias in SSR distribution. RIP refers to a genetic phenomenon that mutates duplicated sequences in a genome during the sexual cycle [22, 64]. Both duplicated regions go through C:G to T:A mutations preferentially at CpA di-nt [65, 66]. For example, a segment of the *Tad* 1-1 sequence,...ACACA..., is mutated to,...ATATA..., after RIP in all progeny the authors analyzed (Fig. 4 in [21]). The systematic mutations of these types could accelerate the genesis of certain types of SSRs and interrupt others. We did not find CG/GC repeats in the *N. crassa* genome. This observation could be explained by RIP since the expanded CG/GC repeats could be a target for repeated RIP. Characterizing the roles of RIP in SSR evolution requires more careful study.

### **Discussion; Marker potential of *Neurospora* SSR**

The estimated PIC value was comparable and relatively high for SSR markers compared to other organisms [67-69] where the SSR marker system has been applied to many genetic analyses. The average PIC in rice is 0.637[70], in soybean 0.43 [71], and in wheat 0.40 [72]. The mean PIC score in the rust fungus, *Puccinia graminis*, is 0.49 [73], and the mean PIC score in *Diplodia pinea* [= *Sphaeropsis sapinea*], a well-known pathogen causing a shoot or tip blight of numerous pine species and some other conifers, is 0.43 [74]. Compared to these organisms, SSR size variability estimated by PIC scores in *N. crassa* seems to be relatively high. To estimate a mean PIC value objectively, the tested SSR loci should be randomly sampled. However, unbiased sampling, as done in our study (Figure 2-10), is not easy to achieve even in sequenced

organisms. Since the PIC calculations in many studies, including those mentioned above, are based on the available SSR marker set rather than a random sample, they could produce biased estimates of the PIC for the genome.



**Figure 2-10. The distribution of the randomly selected SSRs by the unit number**  
 The genome is scan by a 250 kb window to select a SSR within each window. Mononucleotide SSRs are skipped in the count. The distribution of selected 162 SSRs shows comparable to those of the complete genome-wide collection in the SSR repeat type.

The high PIC value in *N. crassa* implies that the SSR marker system has sufficient resolution/polymorphism to be used for genetic studies. Even though our current study of polymorphism of the selected SSR types uses a rather small sample of 7 strains, our PIC estimation of the SSR type AC/CA was consistent with the estimation using a larger sample of 32 strains (Figure 2-7). A larger scale study of

genome-wide SSR analysis with a bigger population would be required to obtain a more comprehensive understanding of the distribution of SSRs in fungal genomes.

We investigated whether the polymorphism of SSRs could be affected by any of the factors including different repeat units, SSR types, chromosomes, repeat numbers, and total SSR lengths (Figure 2-6). Our result showed that there were no significant differences in PIC scores among those criteria (Figure 2-6) in the *N. crassa* genome. Since the mutation rate seems to be random across the genome, it was difficult to estimate the mutation rates of SSRs in different categories, i.e. SSR types or functional regions. Thus, empirical characterization of size variability for each SSR is necessary to estimate the usefulness of a particular SSR as a molecular marker.

Currently, genomic sequences of many fungal organisms are accessible through public genome databases. Identification of SSRs can be easily done using several publicly available software packages. However, despite the many advantages of SSR markers in various biological studies, the lack of experimental data on polymorphic SSR markers is still a major limitation for utilizing SSR markers in biological studies in fungal systems. Thus, community based databases for SSRs will expedite the implementation of SSR markers in genetic and genomic studies in *N. crassa* as well as in other fungal organisms.

### **Discussion; SSR based genetic map construction for *N. crassa* genome**

Recently, molecular marker techniques for assisting efficient mapping/gene-cloning have been developed in *N. crassa* system [75-77]. All of these techniques utilize polymorphisms at the nucleotide level. The usefulness of polymorphisms found in SSRs for evolutionary studies was explored by Dettman and Taylor [8]; 13 SSRs in 147 strains from eight species of *Neurospora* have been analyzed. The authors sequenced 5 SSRs and about 500 nucleotides of the flanking sequences, and then

characterized the genealogical relationships between SSR alleles by mapping them onto a tree drawn by flanking sequence data. This study revealed that SSRs are not appropriate for studies on inter-phylogenetic relationship among species due to high mutation rates in SSRs (about 2500 times greater than those of flanking sequences) and allele length homoplasy [8]. The same report also suggested that SSRs could be used for population studies in inter-species populations. In the current study, we wanted to test if we could experimentally confirm this prediction by constructed three linkage maps using three independent F1 populations.

Based on our SSR polymorphism data, we were able to construct three different genetic maps from the three different pairs of *N. crassa* natural accessions (Table 2-6). A previous study estimated that the *Neurospora* genome is about 1000 cM [35]. The discrepancy in the estimated genome-wide map units, between our estimation and the previous study, is mostly due to the different coverage of either molecular or genetic markers for each strain. Our linkage maps roughly agree with the previous estimation, about 1000 cM [35].

The order of SSR markers along the chromosomes were conserved well among the three mapping populations in our analysis. Furthermore, the positions of mapped SSR loci from the three mapping population are highly consistent with the positions in the physical map, with a few exceptions, suggesting that the genetic architecture of the 6 natural accessions are highly similar to each others. One of the exceptions was the loci order between closely linked markers. Because of this, the inconsistency is mostly attributable to statistical complications cased by a lack of recombination information between two tightly linked markers, rather than chromosomal rearrangements due to missing values or segregation distortions. A previous simulation study also supports our interpretation[78] .

## REFERENCES

1. Ellegren H: **Microsatellites : simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5**(6):435-445.
2. Lawson MJ, Zhang L: **Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes.** *Genome Biol* 2006, **7**:R14.
3. Selkoe KA, Toonen RJ: **Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers.** *Ecol Lett* 2006, **9**:615-629.
4. Choi H-K, Kim D, Uhm T, Limpens E, Lim H, Mun J-H, Kalo P, Penmetsa RV, Seresd A, Kulikovac O *et al*: **A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*.** *Genetics* 2004, **166**(3):1463-1502.
5. Yu JK, Dake TM, Singh S, Benschoter D, Li WL, Gill B, Sorrells ME: **Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat.** *Genome* 2004, **47**(5):805-818.
6. Yu JK, La Rota M, Kantety RV, Sorrells ME: **EST derived SSR markers for comparative mapping in wheat and rice.** *Mol Gen Genomics* 2004, **271**(6):742-751.

7. Suwabe K, Tsukazaki H, Iketani H, Hatakeyama K, Kondo M, Fujimura M, Nunome T, Fukuoka H, Hirai M, Matsumoto S: **Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: the genetic origin of clubroot resistance.** *Genetics* 2006, **173**(1):309-319.
8. Dettman JR, Taylor JW: **Mutation and evolution of microsatellite loci in *Neurospora*.** *Genetics* 2004, **168**:1231-1248.
9. Buschiazzo E, Gemmell NJ: **The rise, fall and renaissance of microsatellites in eukaryotic genomes.** *Bioessays* 2006, **28**(10):1040-1050.
10. Li YC, Korol AB, Fahima T, Nevo E: **Microsatellites within genes: structure, function, and evolution.** *Mol Biol Evol* 2004, **21**(6):991-1007.
11. Li YC, Korol AB, Fahima T, Beiles A, Nevo E: **Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review.** *Mol Ecol* 2002, **11**(12):2453.
12. Jasinska A, Michlewski G, de Mezer M, Sobczak K, Kozlowski P, Napierala M, Krzyzosiak WJ: **Structures of trinucleotide repeats in human transcripts and their functional implications.** *Nucleic Acids Res* 2003, **31**(19):5463-5468.
13. Prasad MD, Muthulakshmi M, Madhu M, Archak S, Mita K, Nagaraju J: **Survey and analysis of microsatellites in the silkworm, *Bombyx mori*:**

**frequency, distribution, mutations, marker potential and their conservation in heterologous species.** *Genetics* 2005, **169**:197-214.

14. Kashi Y, King DG: **Simple sequence repeats as advantageous mutators in evolution.** *Trends Genet* 2006, **22**(5):253-259.
15. Sawyer LA, Hennessy JM, Peixoto AA, Rosato E, Parkinson H, Costa R, Kyriacou CP: **Natural variation in a Drosophila clock gene and temperature compensation.** *Science* 1997, **278**(5346):2117-2120.
16. Zamorzaeva I, Rashkovetsky E, Nevo E, Korol A: **Sequence polymorphism of candidate behavioural genes in Drosophila melanogaster flies from 'Evolution Canyon'.** *Mol Ecol* 2005, **14**(10):3235-3245.
17. Fahima T, Roder MS, Wendehake K, Kirzhner VM, Nevo E: **Microsatellite polymorphism in natural populations of wild emmer wheat, Triticum dicoccoides, in Israel.** *Theor Appl Genet* 2002, **104**(1):17-29.
18. Fondon JW, Garner HR: **Molecular origins of rapid and continuous morphological evolution.** 2004, **101**:18058-18064.
19. Verstrepen KJ, Jansen A, Lewitter F, Fink GR: **Intragenic tandem repeats generate functional variability.** *Nature Genetics* 2005, **37**(9):986-990.

20. Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, Wray GA: **Positive selection on MMP3 regulation has shaped heart disease risk.** *Curr Biol* 2004, **14**(17):1531-1539.
21. Kinsey JA, Garrett-Engele PW, Cambareri EB, Selker EU: **The Neurospora transposon Tad is sensitive to repeat-induced point mutation (RIP).** *Genetics* 1994, **138**(3):657-664.
22. Galagan JE, Selker EU: **RIP: the evolutionary cost of genome defense.** *Trends Genet* 2004, **20**(9):417-423.
23. Borkovich KA, Alex LA, Yarden O, Freitag M, Turner GE, Read ND, Seiler S, Bell-Pedersen D, Paietta J, Plesofsky N *et al*: **Lessons from the genome sequence of Neurospora crassa: tracing the path from genomic blueprint to multicellular organism** *Microbiol Mol Biol Rev* 2004, **68**:1-108.
24. Mannhaupt G, Montrone C, Haase D, Mewes HW, Aign V, Hoheisel JD, Fartmann B, Nyakatura G, Kempken F, Maier J *et al*: **What's in the genome of a filamentous fungus? Analysis of the Neurospora genome sequence.** *Nucleic Acids Res* 2003, **31**(7):1944-1954.
25. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**(2):573-580.

26. Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TFC, Aquadro CF: **The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster***. *Mol Biol Evol* 1998, **15**(12):1751-1760.
27. Bachtrog D, Weiss S, Zangerl B, Brem G, Schlotterer C: **Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome**. *Mol Biol Evol* 1999, **16**(5):602-610.
28. Agresti A: **An Introduction to Categorical Data Analysis**. . John Wiley & Sons, Inc New York, New York, USA 1996.
29. Kim TS, Booth J, Gauch HG, Sun Q, Park J, Lee Y, Lee K: **Simple Sequence Repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference**. *BMC genomics* 2008, **9**(1):31.
30. Schuelke M: **An economic method for the fluorescent labeling of PCR fragments**. *Nat Biotechnol* 2002, **18**(2):233-234.
31. Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S: **Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.)**. *Theor Appl Genet* 2000, **100**(5):713-722.
32. Botstein D, White RL, Skolnick M, Davis RW: **Construction of a genetic linkage map in man using restriction fragment length polymorphisms**. *Am J Hum Genet* 1980, **32**(3):314-331.

33. Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME: **Optimizing parental selection for genetic-linkage maps.** *Genome* 1993, **36**(1):181-186.
34. Sun X, Liu Y, Lutterbaugh J, Chen WD, Markowitz SD, Guo B: **Detection of mononucleotide repeat sequence alterations in a large background of normal DNA for screening high-frequency microsatellite instability cancers.** *Clin Cancer Res* 2006, **12**(2):454-459.
35. Perkins DD: **Neurospora crassa genetic maps and mapped Loci.** *Fungal Genet News* 2000, **47**:40-58.
36. Mardia KV, Kent JT, Bibby JM: **Multivariate Analysis.** Duluth, London: Academic Press; 1979.
37. Gauch H: **Statistical analysis of regional yield trials: AMMI analysis of factorial designs.** *Elsevier, Amsterdam* 1992.
38. Manly K, Olson J: **Overview of QTL mapping software and introduction to map manager qtx.** *Mamm Genome* 1999, **10**:327-334.
39. Holloway JL, Knapp SJ: **G-MENDEL 3.0 user guide.** *Oregon State University, Corvallis, OR* 1993:1-130.
40. Kosambi D: **The estimation of map distances from recombination values.** *Annals of Eugenics* 1944, **12**:172-175.

41. Karaoglu H, Lee CM, Meyer W: **Survey of simple sequence repeats in completed fungal genomes.** *Mol Biol Evol* 2005, **22**(3):639-649.
42. Brohede J, Primmer CR, Moller A, Ellegren H: **Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci.** *Nucleic Acids Res* 2002, **30**(9):1997-2003.
43. Harr B, Schlotterer C: **Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation.** 2000, **155**:1213-1220.
44. Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L: **Triplet repeats in human genome: distribution and their association with genes and other genomic regions.** *Bioinformatics* 2003, **19**(5):549-552.
45. Chistiakov DA, Hellemans B, Volckaert FAM: **Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics.** *Aquaculture* 2006, **255**(1-4):1-29.
46. Subramanian S, Mishra RK, Singh L: **Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions.** *Genome Biol* 2003, **4**(2):R13.
47. Field D, Wills C: **Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of**

- microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces.** *Proc Natl Acad Sci U S A* 1998, **95**(4):1647-1652.
48. Edwards YJ, Elgar G, Clark MS, Bishop MJ: **The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses.** *J Mol Biol* 1998, **278**(4):843-854.
49. Young ET, Sloan JS, Riper Kv: **Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*.** *Genetics* 2000, **154**:1053-1068.
50. Katti MV, Ranjekar PK, Gupta VS: **Differential distribution of simple sequence repeats in eukaryotic genome sequences.** *Mol Biol Evol* 2001, **18**(7):1161-1167.
51. Katti M, Sami-Subbu R, Ranjekar P, Gupta V: **Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications.** *Protein Sci* 2000, **9**(6):1203-1209.
52. Boeva V, Regnier M, Papatsenko D, Makeev V: **Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression.** *Bioinformatics* 2006, **22**(6):676-684.

53. Zoghbi HY, Orr HT: **Glutamine repeats and Neurodegeneration**. *Annu Rev Neurosci* 2000, **23**:217-247.
54. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S *et al*: **The genome sequence of the filamentous fungus Neurospora crassa**. 2003, **422**(6934):859-868.
55. Watters MK, Randall TA, Margolin BS, Selker EU, Stadler DR: **Action of Repeat-Induced Point Mutation on Both Strands of a Duplex and on Tandem Duplications of Various Sizes in Neurospora**. *Genetics* 1999, **153**(2):705-714.
56. Kashi Y, King D, Soller M: **Simple sequence repeats as a source of quantitative genetic variation**. *Trends Genet* 1997, **13**(2):74-78.
57. Michael TP, Park S, Kim T-S, Booth J, Byer A, Sun Q, Chory J, Lee K: **Simple sequence repeats provide a substrate for phenotypic variation in the Neurospora crassa circadian clock**. *PLoS ONE* 2007, <http://www.plosone.org/doi/pone.0000795>.
58. Ballario P, Vittorioso P, Magrelli A, Talora C, Cabibbo A, Macino G: **White collar-1, a central regulator of blue light responses in Neurospora, is a zinc finger protein**. *EMBO J* 1996, **15**(7):1650-1657.
59. Liu Y: **Molecular mechanisms of entrainment in the Neurospora circadian clock**. *J Biol Rhythms* 2003, **18**(3):195-205.

60. Lee K, Dunlap CJ, Loros JJ: **Roles for WHITE COLLAR-1 in circadian and general photoperception in *Neurospora crassa*.** *Genetics* 2003, **163**(1):103-114.
61. Toyota K, Onai K, Nakashima H: **A new *wc-1* mutant of *Neurospora crassa* shows unique light sensitivity in the circadian conidiation rhythm.** *Mol Gen Genomics* 2002, **268**:56-61.
62. Borstnik B, Pumpernik D: **Evidence on DNA slippage step-length distribution.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **71**(3 Pt 1):031913.
63. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution.** *Mol Biol Evol* 1987, **4**(3):203-221.
64. Selker EU: **Premeiotic instability of repeated sequences in *Neurospora crassa*.** *Annu Rev Genet* 1990, **24**:579-613.
65. Cambareri EB, Jensen BC, Schabtach E, Selker EU: **Repeat-induced G-C to A-T mutations in *Neurospora*.** *Science* 1989, **244**(4912):1571-1575.
66. Grayburn WS, Selker EU: **A natural case of RIP: degeneration of the DNA sequence in an ancestral tandem duplication.** *Mol Cell Biol* 1989, **9**(10):4416-4421.

67. Blair MW, Hedetale V, McCouch SR: **Fluorescent-labeled microsatellite panels useful for detecting allelic diversity in cultivated rice ( *Oryza sativa* L.).** *Theor Appl Genet* 2002, **105**(2-3):449-457.
68. Rungis D, Llewellyn D, Dennis ES, Lyon BR: **Simple sequence repeat (SSR) markers reveal low levels of polymorphism between cotton ( *Gossypium hirsutum* L.) cultivars.** *Aust J of Agri Res* 2005, **56**(3):301-307.
69. Heckenberger M, Bohn M, Ziegler JS, Joe LK, Hauser JD, Hutton M, Melchinger AE: **Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties. I. Genetic and technical sources of variation in SSR data.** *Mol Breeding* 2002, **10**(4):181-191.
70. Pessoa-Filho M, Belo A, Alcochete AA, Rangel PH, Ferreira ME: **A set of multiplex panels of microsatellite markers for rapid molecular characterization of rice accessions.** *BMC Plant Biol* 2007, **7**:23.
71. de Campos T, Benchimol LL, Carbonell SAM, Chioratto AF, Formighieri EF, de Souza AP: **Microsatellites for genetic studies and breeding programs in common bean.** *Pesq agropec bras* 2007, **42**(4):589-592.
72. Zhang LY, Bernard M, Ravel C, Balfourier F, Leroy P, Feuillet C, Sourdille P: **Wheat EST-SSRs for tracing chromosome segments from a wide range of grass species.** *Plant Breeding* 2007, **126**(3):251-258.

73. Szabo LJ: **Development of simple sequence repeat markers for the plant pathogenic rust fungus, Puccinia graminis.** *Mol Ecol Notes* 2007, **7**(1):92-94.
74. Burgess TI, Wingfield MJ, Wingfield BD: **Global distribution of Diplodia pinea genotypes revealed using simple sequence repeat (SSR) markers.** *Aust Plant Pathol* 2004, **33**(4):513-519.
75. Dunlap JC, Borkovich KA, Henn MR, Turner GE, Sachs MS, Glass NL, McCluskey K, Plamann M, Galagan JE, Birren BW *et al*: **Enabling a community to dissect an organism: overview of the Neurospora functional genomics project.** *Adv Genet* 2007, **57**:49-96.
76. Jin Y, Allan S, Baber L, Bhattacharai EK, Lamb TM, Versaw WK: **Rapid genetic mapping in Neurospora crassa.** *Fungal Genet Biol* 2007, **44**(6):455-465.
77. Lewis ZA, Shiver AL, Stiffler N, Miller MR, Johnson EA, Selker EU: **High density detection of restriction site associated DNA (RAD) markers for rapid mapping of mutated loci in Neurospora.** *Genetics* 2007, **177**(2):1163-1171.
78. Hackett CA, Broadfoot LB: **Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps.** *Heredity* 2003, **90**(1):33-38.

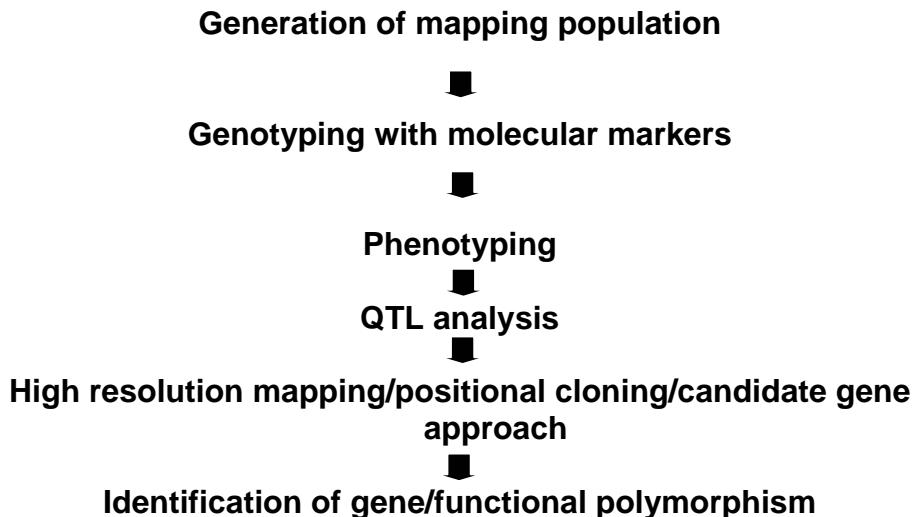
## CHAPTER 3

### TOOL DEVELOPMENTS TO FACILITATE QTL ANALYSIS AND SUBSEQUENT POSITIONAL CLONING PROCEDURES

#### **Introduction; The need for tool development for the efficient QTL analysis**

A QTL approach to the study of the *Neurospora* clock may be promising for the following reasons: First, *Neurospora crassa* is a very genetically tractable organism, having appropriate resources and also reliable phenotypes for the circadian clock regulation. Second, this is forward genetic approach so no prior information is required to find new genes underlying the clock phenotype. Third, we may find a novel genetic components underlying clock phenotype if we exploit natural variations between natural strains that have accumulated for a long period time as a consequence of the adaptive process to local environment. Fourth, we may be able to characterize a causal genetic variant that from an adaptive process since the genetic variation that affects clock function is relevant to the adaptive process. Thus, for these reasons, a QTL approach may increase our understanding of the mechanisms of circadian regulation and also give insight for the role of the circadian clock in the process of adapting to local environments, a topic that is somewhat overlooked in current research.

However, identifying a gene from the QTL approach is not easy; it requires a sequence of multiple procedures and each step may require a plenty of work (Figure 3-1). Thus, I have developed two efficient tools that facilitate using QTL approaches in genetic marker development/management and phenotyping procedures.



**Figure 3-1. Procedures to discover a gene/genes from QTL analysis**

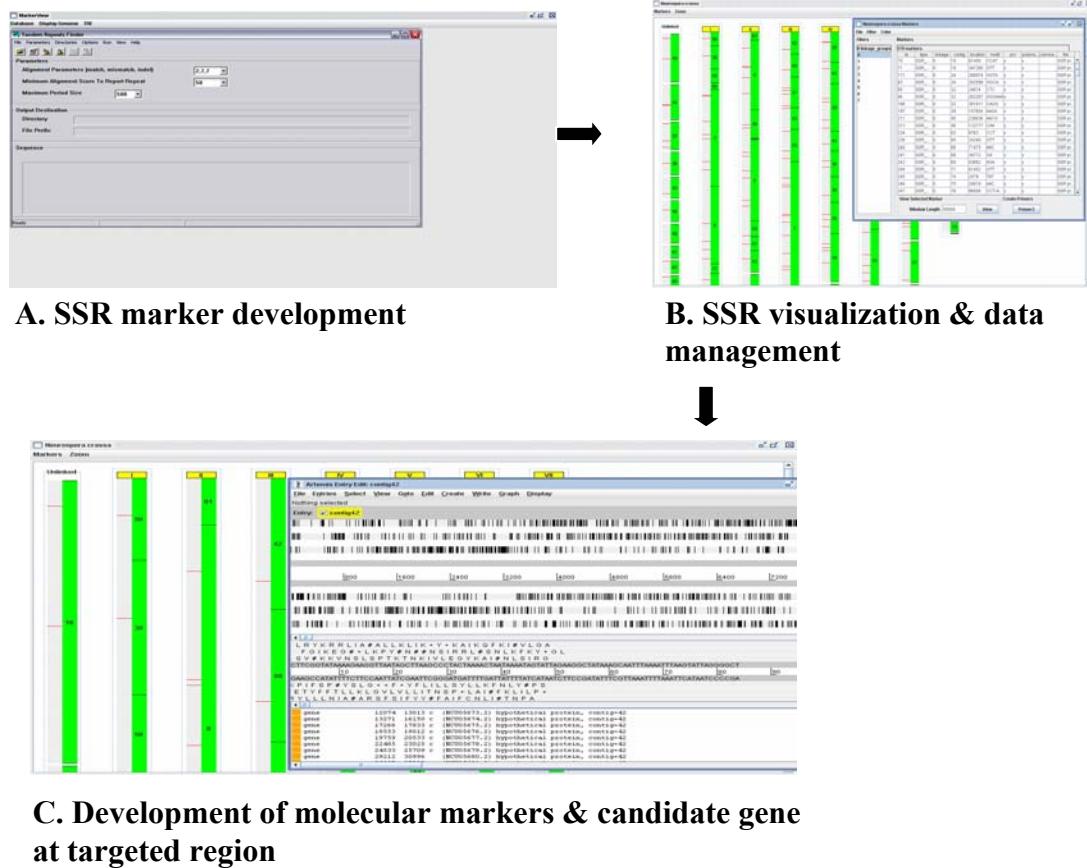
### **Introduction; MoMMS: A Molecular Marker Management System**

Whole genome sequence data facilitates the development and application of sequence based markers. Microsatellite or simple sequence repeat (SSR) markers are one of most frequently used sequence based markers in current genomics studies because they are ubiquitous, abundant and highly polymorphic across eukaryotic genomes [1-4]. Moreover, since SSRs are PCR (Polymerase Chain Reaction) based markers, they are easily assayable with relatively low cost and compatible with high throughput methods [5, 6].

Among many applications, SSR markers has been employed efficiently in quantitative trait loci (QTL) analysis to identify genetic polymorphisms in a segregating population that are associated with a phenotype of interest [5]. A significant advantage of using sequence based markers for QTL analyses in sequenced organisms is that a genetically defined QTL position can be converted directly to a physical position, which facilitates subsequent map-based or positional cloning [6].

However, identifying a gene that is responsible for a QTL from a coarse QTL map through positional cloning still requires much work; one has to perform high resolution mapping to narrow down the genomic region of interest [7, 8]. Efficient ways to develop molecular markers for a specific region of interest are thus indispensable [9].

To facilitate the process of developing and managing SSR marker data, we developed the MoMMS. This software provides a graphical user interface to display positions of SSR markers in a sequence segment or a genetically ordered physical map if the related information is available. Selected markers can be exported or viewed at the sequence level using the ARTEMIS package (<http://www.sanger.ac.uk/Software/Artemis/>). Further, this visualization function in the software can, in principle, be extended to any other sequence based marker systems such as SNPs (single nucleotide polymorphisms) if a proper input file is constructed. Tandem Repeats Finder (trf, <http://tandem.bu.edu/trf/trf.html>) and Primer3 are integrated in this software to maximize the efficiency of the application of SSR marker. The incorporated trf assists to find SSRs through an uploaded genomic sequence. MoMMS can automatically convert the output of trf to a new input file to visualize SSR positions in the genome. Primers flanking SSRs can be created by Primer3 program (<http://primer3.sourceforge.net/>) in a high throughput manner. Local database in this software assist to organize and filter SSR marker information regarding their genomic position, their physical characteristics (e.g. SSR motif) and their polymorphism. These functions greatly assist genetic marker developments both genome-wide and at specific target sites, which facilitate procedures from QTL analyses and positional cloning in any eukaryotic model organism. The operational process is summarized in Figure 3-2. In this report, we demonstrate an application of MoMMS in QTL analysis and positional cloning using the *Neurospora crassa* genome.



**Figure 3-2. Graphical summary in the operational processes of MoMMS**

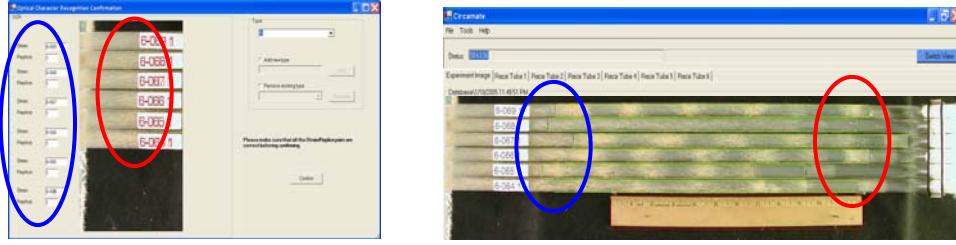
A. SSR marker development step. A proper input file is automatically constructed after the SSR discovery. Tandem Repeats Finder (trf) B. SSR visualization & data management step. The genetic information along with sequence of a genome of interest information allow a construction of a physical map (green bar). The defined SSR by trf can be visualized in the physical map (red line). Local database in this software assist to organize and filter SSR marker information regarding their genomic position, their physical characteristics (e.g. SSR motif) and their polymorphism. C. Development of molecular markers & candidate gene at targeted region. Selected markers can be viewed at the sequence level using the ARTEMIS package and the flanking primer of the selected SSRs can be designed in a high though-put manner.

## **Introduction; Circamate, the high through-put image processing system for the circadian clock analysis**

As mentioned earlier, since a clock phenotype may be easily detected through a race tube assay, the *Neurospora* circadian rhythm becomes an attractive system for genetic studies. However, phenotyping the number of race tubes manually that is required in QTL analysis is so laborious that one must think about overcoming this problem before applying QTL to *Neurospora* circadian clock study.

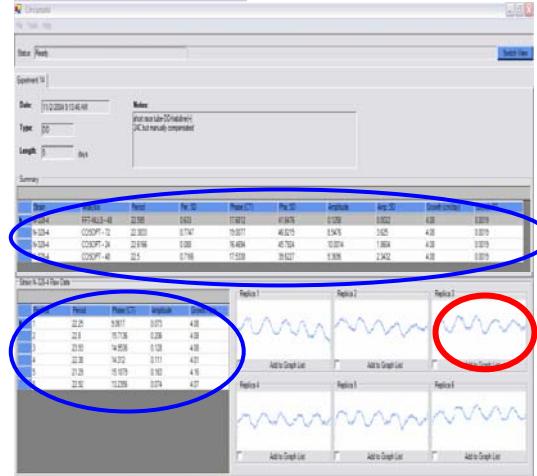
Circamate, a program developed in our lab, was designed to assist the phenotyping procedure in race tube assay. Specifically, it is designed to analyze the circadian rhythm of *N. crassa* or other fungi from digital images by performing the following three steps automatically (figure 3-3): 1) label recognition 2) band recognition and interpretation 3) the visualization and management of the result. This software was applied to our QTL analysis, and the result turned to be very satisfactory since it not only saved time and effort but also introduced a precision and objectiveness in the interpretation of the banding phenotype.

In the original version of the Circamate program, we focused on the function that analyzes period and phase under FRC (need to define FRC?). Using the clock phenotypes under the FRC, the circadian clock study may be not comprehensive; using those phenotypes alone, we may not go further to find genetic components underlying the function of circadian clock at natural environments using QTL analysis, since the output phenotypes of circadian rhythm only reflect endogenous regulation of circadian oscillator. However, the natural environment is cyclic in light and temperature, by which harmonious actions of input, oscillator and output pathways are orchestrated to control the physiological processes according to daily or seasonal basis change. Thus, we updated Circamate to accommodate a function to interpret *Neurospora* circadian clock phenotype under cycling conditions to understand biological functions of *Neurospora* circadian clock in natural environment through QTL.



**A. 1<sup>st</sup> step; label recognition.**

**B. 2<sup>nd</sup> step; banding pattern**



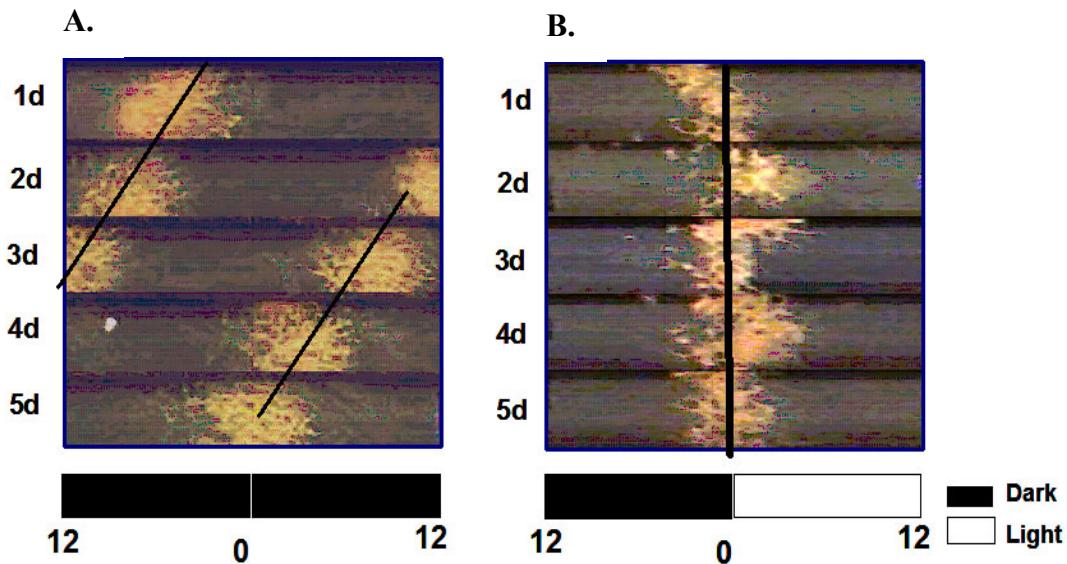
**C. 3<sup>rd</sup> step; statistical analysis and saving data in the SQL server**

### Figure 3-3. The three analysis steps by Cricamate.

A. 1<sup>st</sup> step; label recognition. Automatic label recognition (blue circle) of labels in the race tubes (red circle). B. 2<sup>nd</sup> step; banding pattern recognition. Automatic recognition of beginning and end of experiment (blue and red circle). C. 3<sup>rd</sup> step; statistical analysis and saving data in the SQL server. Data summary of statistical analysis (blue circle) and graphical presentation of the banding pattern (red circle). This figure is adopted and modified from the figure shown in Lee's lab home page (<http://genesis.plantpath.cornell.edu/Tools.aspx.>)

## **Experiment and Result; Establishment of a phase evaluation function which is compatible to cycling condition.**

Under the entraining condition, the conidiation rhythm is synchronized to the external cycling rhythm. For example, if the race tube is subject to a 24 hour light-dark cycle, the conidiation occurs at the same position everyday, in contrast to the rhythm under FRC since the endogenous rhythm is synchronized to 24 hours (Figure 3-4A and B).



**Figure 3-4. The comparison of the conidiation rhythm between free running (A) and entraining condition to 24 hour cycle (B).**

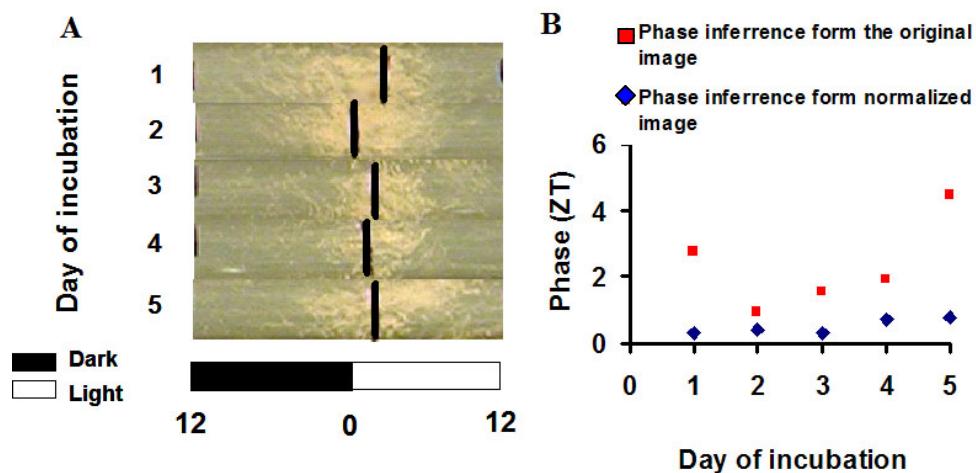
Race tube image is segmented by 24 hours' interval and is arrayed horizontally. x axes in (A) and (B) represent duration of light treatment each day. Open box in x axis represents light treatment and box with black color represent dark treatment. In y axis, d stands for day thus, for example, 1d represent day 1 of incubation. In this figure, endogenous rhythm (20.5hr) under FRC (A) is synchronized to 24 hour (B) of the external cycle.

Calculation of the phase in entraining condition is simply formulated as follows:

Phase (ZT) = (distance of band center from light-off/total growth for 24 hours) x 24 hours +12 (ZT); where, 12 (ZT) is a constant that represents ambient light condition (light-off) of a cycling environment.

The basic idea of this calculation is to represent when the band center occurs each day under cycling conditions. However, the calculation is based on the assumption that the growth rate is constant. It is found that growth rate is not constant in some cases dependent on different environmental stages of a cycle or any unknown microenvironment in a day; some natural *Neurospora* strains show differential growth rates under a different ambient environment of a cycle. Thus, one cannot generally use the above calculation to infer the entraining phase or phase of *Neurospora* natural stains. Furthermore, since the differential growth rate is not thought to be under circadian clock regulation, the estimation of the phase from the calculation could be incorrect. In order to bypass this problem, the updated Circamate has a function that processes a reconstructed band image that has been normalized by the growth rate at each different environmental stage of a cycle when the phase is calculated (data not shown). Our prediction was that the updated function could estimate the phase accurately regardless of the differential growth rate since the growth rate is not under circadian control. To confirm this prediction, we examined the phase of a strain that has inconsistent growth during race tube experiments using a normalized image processed by the undated Circamate and an original image as a control. The race tube experiment for the phase analysis was conducted over 5 days under a 12h:12h light-dark cycle at 25C° inoculated with FGSC2225, a strain that tends to have unstable growth rates during race tube experiments. As expected, the growth rate of FGSC 2225 was variable during the race tube experiment, with somewhat faster growth under light condition in the 12h:12h light-dark cycle (Figure 3-5A). We then compared the phase values of the normalized

image processed by the updated Circamate to values of the original image. The phase measurements from the normalized image among incubation days are much less variable than those measurements from the original image (the standard deviations are 0.2 vs 1.3 ZT hours respectively, Figure 3-5B). Thus, this experiment confirmed the prediction and also strongly suggested that the updated function of Circamate is accurate enough to estimate the phase of Neurospora natural strains.



**Figure 3-5. Comparison of phase inference from between the normalized and original image.**

A. Race tube image that is inoculated with FGSC 2225. The original image is segmented by 24 hours' interval and then is arrayed horizontally. Open box in x axis represents light treatment and box with black color represents dark treatment. y axis represents day of incubation. Black vertical bar is a marking drawn at ZT0. B. Comparison of phase inference between the normalized (diamond with blue color) and original image (rectangle with red color) in each incubation day through Circamate. x-axis represents the day of incubation and y axis represents phase value.

In addition to the main change ad described above, the updated Circamate accommodate many other efficient functions for Neurospora circadian phenotype. First, the image reconstruction solution to correct differential growth rate can be applied to period analysis since the daily growth could be variable during race tube experiments from unexplained environmental noise. Second, we can specify any cycling condition

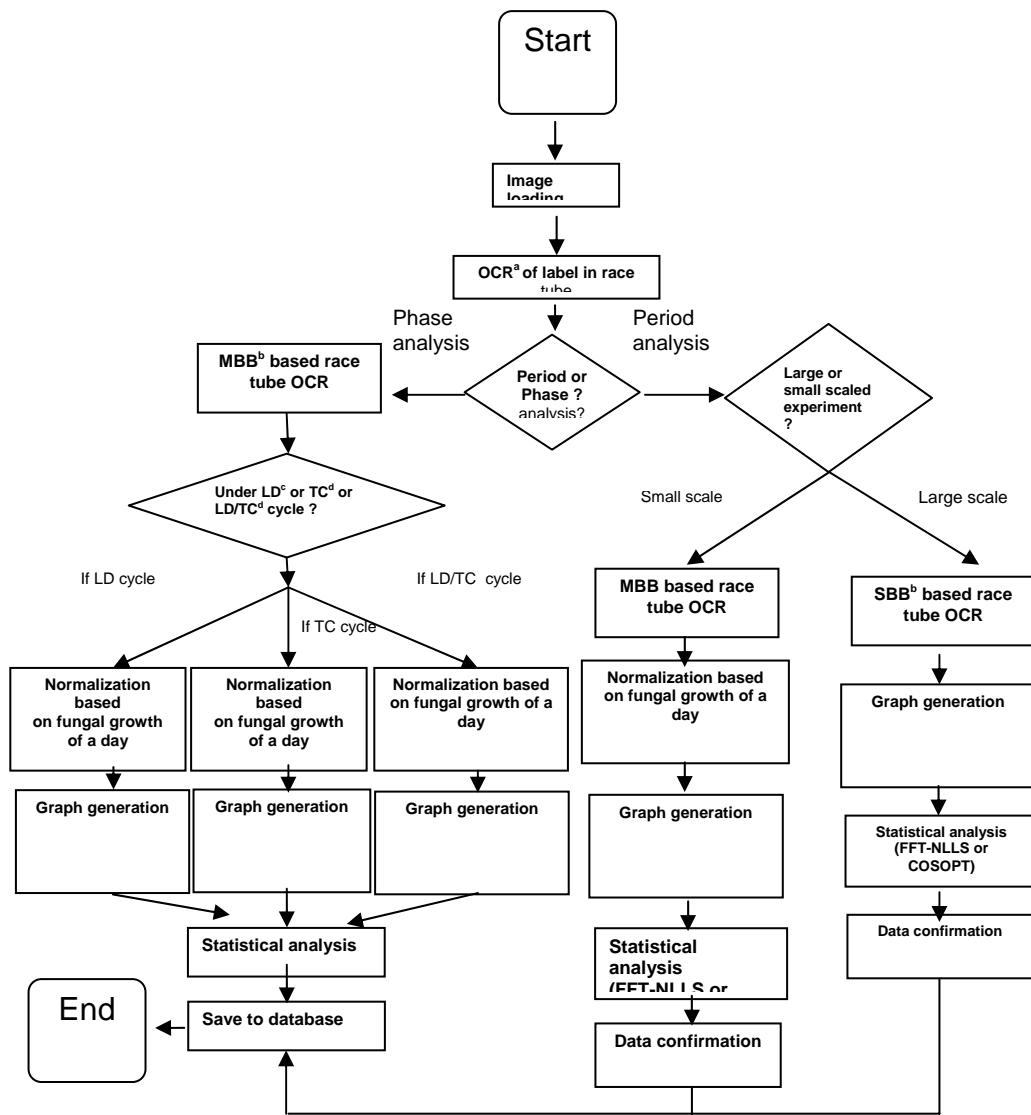
other than light/dark conditions (for example, temperature cycle) in the database to manage the information efficiently. Third, we can change period of a certain environmental condition in a given cycle in phase analysis. For example, in the light/dark cycle with 24 hour condition, we can choose any photoperiod to define the photoperiod-specific phase. The flow chart of the operational process is described in Figure 3-4.

### **Method; Implementation of MoMMS**

MoMMS is implemented in Java using the Swing graphics package and is compatible with J2SE 1.4.0 and later. It reads common bioinformatics sequence file formats such as FASTA and EMBL. Chromosomes that are split into contigs can be displayed with the aid of an additional linkage file. A local MySQL database is used to store chromosomal sequences and annotations. Large sequences are 'chunked' into subsequences for faster storage and retrieval. A Java .jar file for ARTEMIS release 8 is included in MoMMS and accessed as a Java package. trf v 3.3.0 and Primer3 are included as separate programs.

### **Experiment and Result; Application of MoMMS in QTL analysis and subsequent positional cloning.**

Through a QTL analysis, we aimed to find genetic regions associated with one circadian clock property, namely, period under a constant environment [12, 13]. First, we obtained 2745 candidate SSR markers using the trf program. The outcome of trf was automatically converted into a format compatible for MoMMS. After uploading the output of trf analysis into MoMMS, the corresponding location of each marker was visualized as a graph (Figure 3-7).



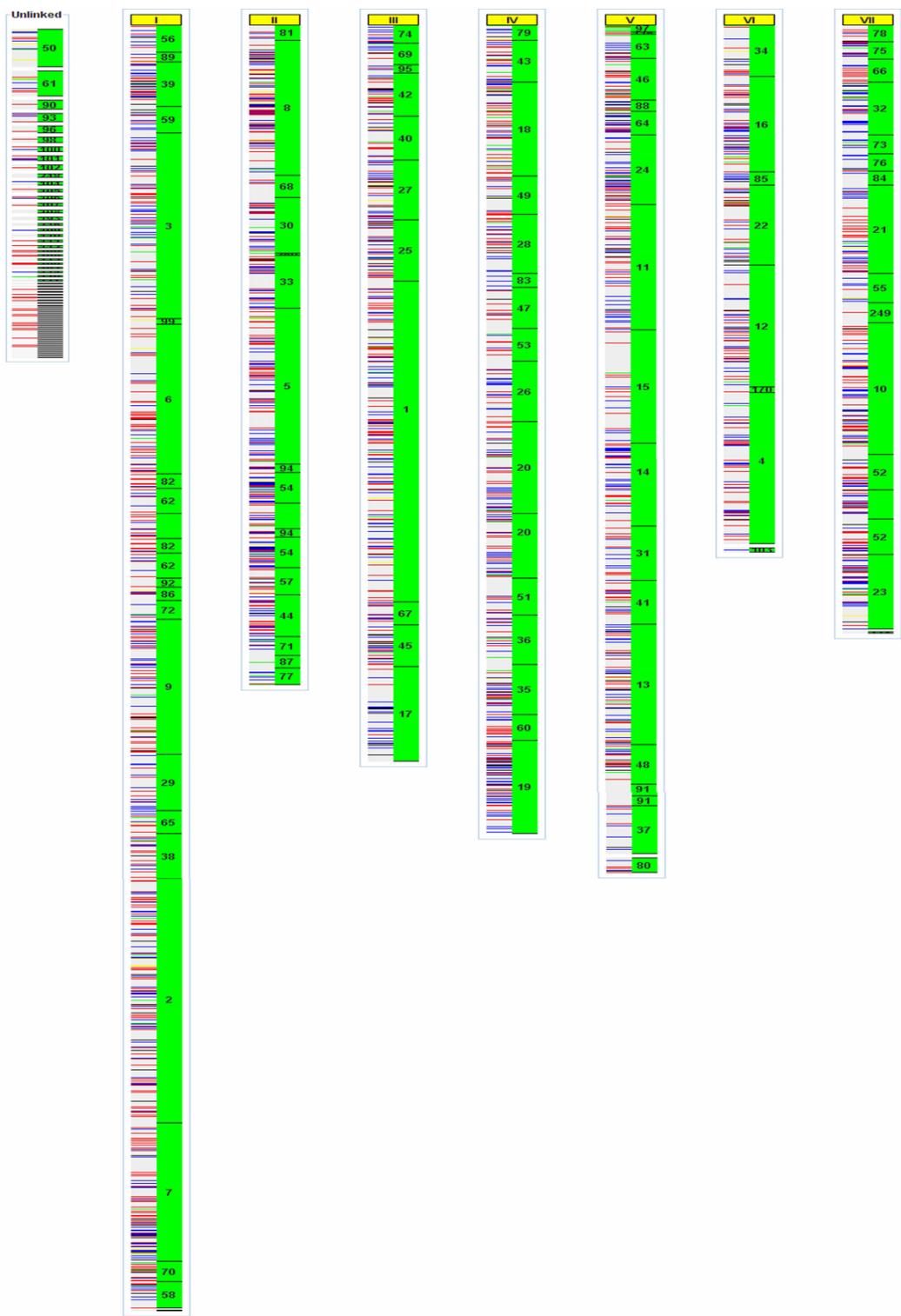
**Figure 3-6. Flow chart in the operational process of the updated Circamate**  
<sup>a</sup>, OCR stands for optical center recognition. <sup>b</sup>, Mbb stands for multi bound box. The MBB method represents a marking method that require multiple times of markings done at every transition of cycle (in cycling condition) or day (in constant environment). <sup>c</sup>, SBB stands for single bonding box. The SBB method represents a marking method that require two times of marking done at the start and end of a race tube experiment.

To identify polymorphic SSR markers evenly distributed across the genome, we randomly selected one candidate SSR marker within each 250 kb window, and experimentally tested for heterogeneity between two parents (Figure 3-8A) [10]. Primers flanking SSRs were created by the integrated Primer3 program (Figure 3-6B). MoMMS can simultaneously generate primer pairs for SSR for any number of selected loci. Once we had experimental data for each marker, we updated the marker information. The updated marker information, both computational and experimental data, was visualized as a table (Figure 3-8B). The information in the table can be reorganized according to the user's purpose with a filter function (Figure 3-8B).

In this fashion we constructed a linkage map covering the whole genome using polymorphic SSR markers[10], and then performed a QTL analysis to identify chromosomal regions associated with the period phenotype. Several QTLs affecting period were detected. One QTL mapped to a region around MN046 on chromosome 7. To find a potential gene corresponding to this QTL [11], we searched for candidate genes in the QTL region using genome browser, ARTEMIS, linked to MoMMS (Figure 3-4C). We could easily identify a candidate gene, *wc-1*(NCU02356.2, Broad Institute) that is located just next to MN046 (Figure 3-8C). WC-1 is a blue light receptor and a key circadian clock gene that has been extensively characterized[12].

**Figure 3-7. Simple sequence repeat (SSR) marker developments of *Neurospora crassa* genome using MoMMs.**

Virtual image of the 2745 SSR sequences defined by Tendam repeat finder (trft) across *Neurospora* genome through MoMMs. The Roman numerals at the head of each column (in yellow) are chromosome numbers. The vertical green bars under the yellow blocks represent chromosomes of *N. crassa* that consist of supercontig sequences that are oriented by genetic order. The numbers written by black clock in green bar denote supercontig IDs, which are obtained from *Neurospora* genome database on the Broad Institute website (<http://www.broad.mit.edu/annotation/genome/neurospora/Home.html>). The horizontally drawn red lines located in green bar in the left side on each *Neurospora* chromosome represent the SSRs defined by trf.





**Figure 3-8. Finding candidate genes for QTL in the *N. crassa* genome.**

An example chromosome view (A) and table view (B) for SSR markers on chromosome 7. (C). The image of genome browser, Artemis, which displays the selected contig sequence and annotation data (e.g. contig 10). The circled marker in the left panel is the marker linked to the circadian QTL, marker MN046. In the genome view of contig 10 (right), the marker MN046 is indicated by red arrows. A well-characterized clock gene *wc-1* (NCU02356.2, Broad Institute) is indicated by blue arrows. The top panel shows the physical location of the SSR marker and the annotated genes, and the bottom panel shows the order of annotated genes and SSR markers.

## REFERENCES

1. Subramanian S, Mishra RK, Singh L: **Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions.** *Genome Biol* 2003, **4**(2):R13.
2. Suwabe K, Tsukazaki H, Iketani H, Hatakeyama K, Kondo M, Fujimura M, Nunome T, Fukuoka H, Hirai M, Matsumoto S: **Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: the genetic origin of clubroot resistance.** *Genetics* 2006, **173**(1):309-319.
3. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S: **Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential.** *Genome Res* 2001, **11**(8):1441-1452.
4. Selkoe KA, Toonen RJ: **Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers.** *Ecology Letters* 2006, **9**:615-629.
5. Mackay TF: **The genetic architecture of quantitative traits.** *Annu Rev Genet* 2001, **35**:303-339.
6. Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL: **Arabidopsis map-based cloning in the post-genome era.** *Plant Physiol* 2002, **129**(2):440-450.

7. Glazier AM, Nadeau JH, Aitman TJ: **Finding genes that underlie complex traits.** *Science* 2002, **298**(5602):2345-2349.
8. Korstanje R, Paigen B: **From QTL to gene: the harvest begins.** *Nat Genet* 2002, **31**(3):235-236.
9. Borevitz JO, Chory J: **Genomics tools for QTL analysis and gene discovery.** *Curr Opin Plant Biol* 2004, **7**(2):132-136.
10. Kim TS, Booth J, Gauch HG, Sun Q, Park J, Lee Y, Lee K: **Simple Sequence Repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference.** *BMC genomics* 2007, **9**(1):31.
11. Kim TS, Logsdon BA, Park S, Mezey JG, Lee K: **Quantitative Trait Loci for the Circadian Clock in *Neurospora crassa*.** *Genetics* 2007, **177**(4):2335-2347.
12. Ballario P, Vittorioso P, Magrelli A, Talora C, Cabibbo A, Macino G: **White collar-1, a central regulator of blue light responses in *Neurospora*, is a zinc finger protein.** *Embo J* 1996, **15**(7):1650-1657.

## CHAPTER 4

### QUANTITATIVE TRAIT LOCI FOR THE CIRCADIAN CLOCK IN *NEUROSPORA CRASSA*

#### **Introduction: Quantitative Trait Loci (QTL) approach for Neurospora Clock study**

Biological rhythms with about a 24 hour period have been found in all forms of life from bacteria to humans [1-6]. The availability of powerful genetic analysis tools and the easily assayable clock phenotype in *Neurospora* has made the system one of the most successful model organisms to dissect the circadian clock by forward genetics approaches [6-8]. Mutant screens for clock genes have focused on mutants with altered period or arrhythmic phenotypes caused by a single mutation inherited through Mendelian segregation [6]. The most interesting gene discovered in these mutant screenings is *frequency (frq)*, which, when mutated, leads to strains with long period, short period or arrhythmic phenotype [6-8]. This finding led to the proposal that a single gene could function as a “state variable” for the circadian oscillator [9]. Cloning and characterizing the *frq* gene significantly advanced molecular understanding of eukaryotic circadian oscillators [10, 11]. Despite the intensive molecular characterizations of *frq* and other known clock genes in *N. crassa*, there is still not a comprehensive understanding of the *Neurospora* circadian clocks. With advances in our understanding of the molecular structure of *Neurospora* clocks, there has been a realization that the circadian clock is more tightly linked to other cellular machineries than speculated previously [12].

It has been suggested that the *frq*-based oscillator might not be the only oscillator. For example *frq*-less oscillators (FLO) coupled to other oscillators in a cell have been proposed [5]. Currently, we know very little about the genetic basis of these

loosely defined oscillators [10]. The conventional forward genetics approach is limited in the discovery of genetic loci with subtle clock phenotypes or with essential cellular functions [13, 14]. Furthermore, most of the genetic screening done previously was focused on identifying period determinants. Thus, just a handful of genetic loci have been characterized that are responsible for other clock properties such as entrained phase or temperature compensation. Thus, we explored an alternative strategy for detecting novel genetic loci affecting the *Neurospora* circadian clock.

Within natural populations there reside important clues to genetic variation, which are vital in unraveling the mysteries of gene function [15]. Identifying the molecular components and characterizing the molecular mechanisms of the natural variations will provide us novel insights into molecular mechanisms of complex circadian traits. For identifying regulatory elements for complex traits, quantitative genetics techniques have been utilized successfully over the past decade to describe how known mutant loci genetically interact with one another, as well as to isolate new loci in the same pathway [16]. Quantitative genetics is an extension of fundamental Mendelian principles to polygenic traits, phenotypes encoded by multiple loci. Much of the phenotypic variation seen in natural populations is due to multiple loci contribute to variation [15]. Each of these Quantitative Trait Loci (QTLs) has relatively small effects on the phenotype. For circadian clock phenotypes, the clock QTLs have been identified in multiple organisms and a few were characterized in *Arabidopsis* and in mice [17-22].

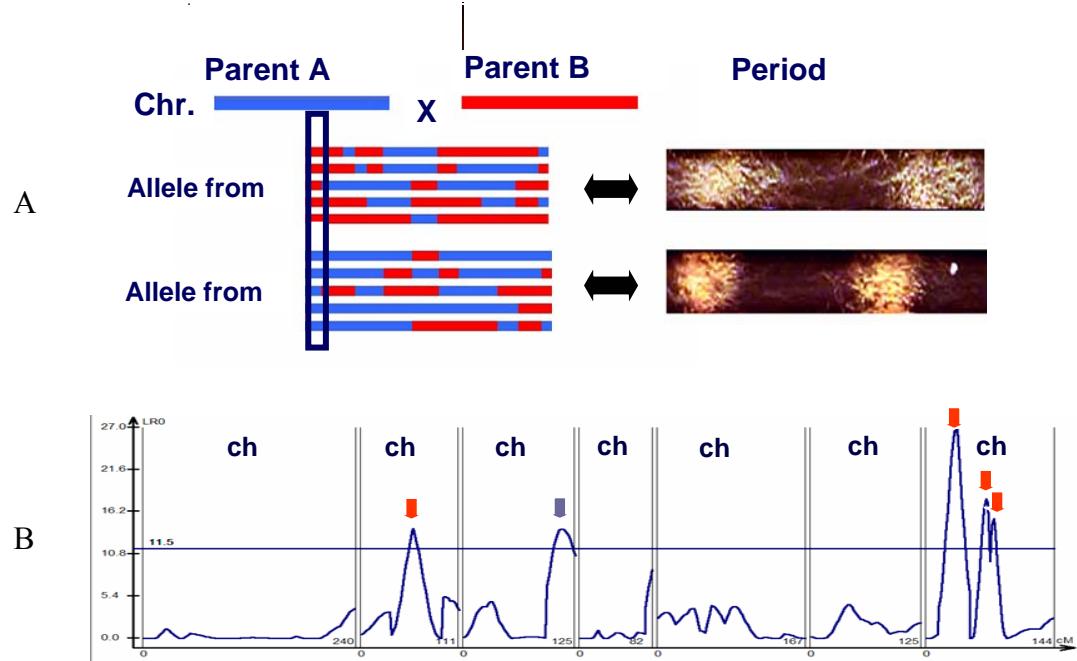
Although clock phenotypes are ideal subjects for QTL analysis, there have been no reports on QTL analysis for *Neurospora* circadian clock phenotypes. This may have been due to technical barriers. The most common assay to measure the *Neurospora* clock has been the race tube assay. As a result of circadian clock-controlled asexual development, *Neurospora* produces orange spores. These dense orange colored spores create a “banding” phenotype in a long glass tube or “race tube” [7]. This easily detected clock phenotype made the *Neurospora* circadian rhythm an attractive system

for genetic studies. All laboratory *Neurospora* strains used in clock studies contain a useful mutation *band*, *bd* [23]. This mutation allows a robust conidial banding pattern even in the high CO<sub>2</sub> environment of the race tube culture. In the wild type strains without the *bd* mutation, the rhythmic asexual development of *Neurospora* in a race tube is suppressed, which has been a major obstacle in clock study [24]. However, a modified race assay has recently been developed that allows study of *Neurospora* circadian clocks in natural accessions without the *bd* mutation in their genetic background [25].

Haploid organisms provide multiple advantages in quantitative studies; 1) they can be maintained clonally to allow large numbers of individuals to be assayed (thereby reducing error from environmental effects); 2) genetic dominance does not contribute to the genetic variation; 3) a permanent segregating population is available after the first generation; 4) the sexual cycle is short [26]. Thus, the QTL analysis in *Neurospora* system as described in Figure 4-1 could provide a unique opportunity to extend quantitative genetics to molecular genomics due to its sequenced haploid genome and many useful resources [27]. For these reasons, *Neurospora* will serve as a valuable model organism for elucidating fundamental questions of quantitative genetics for complex behaviors such as the circadian clock. The high quality of genomic sequence, genetic tractability and bioinformatics supports of *Neurospora* system along with the current breakthrough in the molecular marker technologies and sophisticated mapping algorithms in QTL analysis could make it possible to dissect the circadian clock phenotype down to a single sequence polymorphism strategically (Figure 4-2) [28].

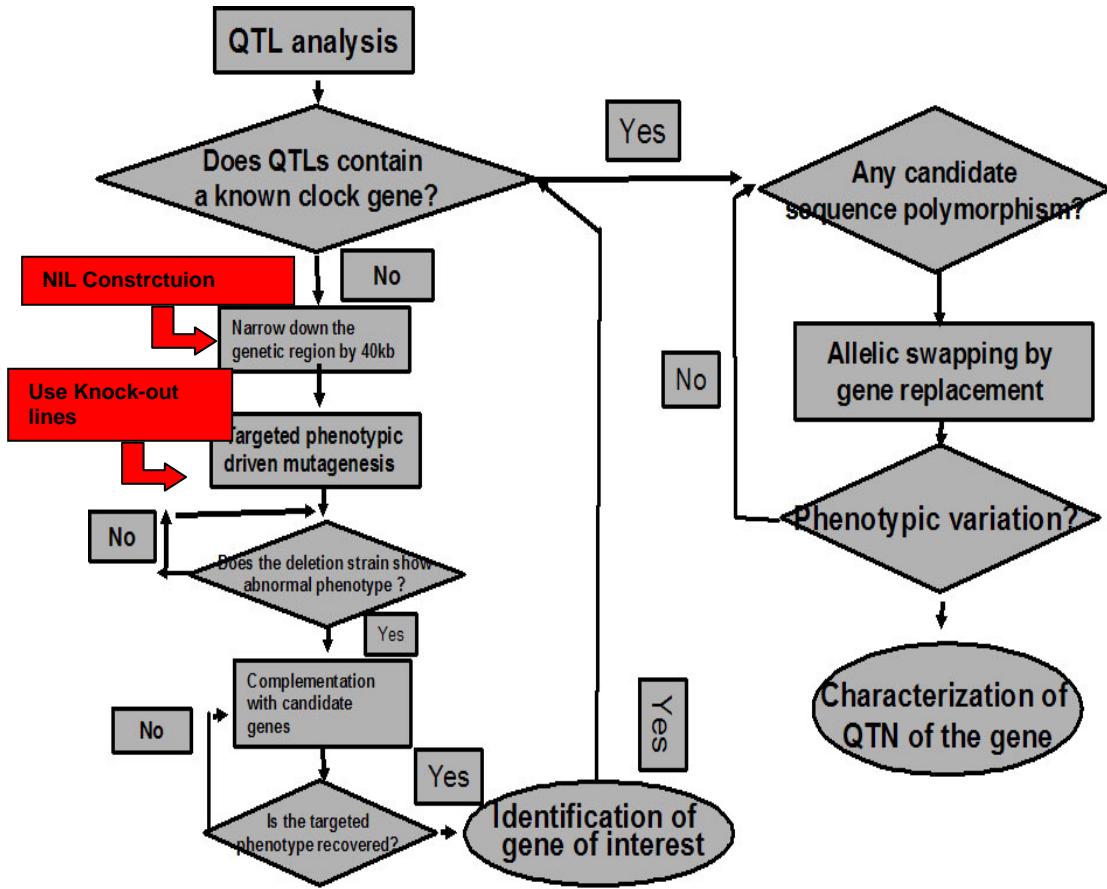
In this report, we describe the QTL analysis of the two clock phenotypes, period and entrained phase using natural populations. In an effort to efficiently find natural genetic variations affecting the clock phenotype, we employed two QTL analyses, composite interval mapping (CIM) and Bayesian multiple QTL (BMQ) analysis in three independent mapping populations derived from the natural accessions that were

collected from geographically isolated areas [29]. In the further characterization, we confirm the QTL effect at BC4 generation which predicted in F1 generation from the construction of the near isogenic line (NIL).



**Figure 4-1. Sechmatic diagram of *Neurospora* QTL analysis.**

A. Genotyping and phenotyping step. Parents with different genetic history (as described different colors) are crossed to make F1 mapping population where the genetic and phenotypic segregation can be occurred due to haploid nature. The genome wide genotyping with multiple genetic markers and phenotyping for the circadian phenotypes in all the F1 progeny are performed in this step. B. QTL analysis. The associations between genotype and phenotype in genome wide are inferred from the genotyping and phenotyping results by various statistical algorithms so as to find a significance genetic region (red arrow) relevant to circadian phenotypic changes where the molecular characterizations are targeted subsequently.



**Figure 4-2. The flow chart that describes the process for the characterizations QTLs defined in F1 mapping population.**

Each QTL is categorized in terms of availability of candidate gene. If obvious candidate genes (for example, previous characterized clock genes) are available at a QTL, candidate gene approach is performed to characterize a molecular mechanism underlying natural variations. In contrast, if any candidate genes are not available, the QTL is narrow down by forward genetic approach which is dependent on genetic purification (by NIL construction) and recombination and gene-knock out (which is publicly available) to verify a function of gene in the clock regulation.

## **Method; Strains and growth conditions**

Natural accessions, FGSC#3223, FGSC#4724, FGSC#4720, FGSC#4715, FGSC#4825 and FGSC#2223, were obtained from Fungal Genetics Stock Center, www.fgsc.net. F1 progenies from three line-crosses were obtained as previously described [30]. The detailed information of the parental strains in their estimated geographical origin and circadian properties is summarized in Table 1. *N. crassa* strains in this study were cultured as previously described [31]. The overt clock phenotypes including period and phase in those strains described above were measured utilizing the Inverted Race Tube Assay [25].

## **Method; Race tube experiment**

Race tubes were incubated in constant light (LL) for 12 hrs at room temperature. After confirming the normal mycelial growth in the race tube, the race tubes were transferred to the growth chamber, I-36LL Percival Scientific (Perry, IA) and incubated an additional 12 hr under LL. For all experiments, temperature was set at 25 °C. After the 24hr LL treatment, light was off for the rest of the experiment for the period measurement. The growing front was marked at light to dark transition and on the last day of experiment. In the race tube experiment for the phase phenotype, the light condition was light12 hr: dark12 hr (LD) cycle. The growing front of the culture in the race tube was marked every 24 hr at the time when light to dark transition occurred. The fluence rate was 250 µE/m<sup>2</sup>/s in LL. Light source was the white fluorescent bulbs and incandescent bulbs (Osram Sylvania Inc.). In both period and phase experiments, tubes were randomly positioned within the chamber to reduce the possibility of positional effects. In each experiment, three replicates of each progeny were assayed. In the event that we could not have at least three replicate data for each strain, we repeated the experiment to generate at least three biological replicates for each strain.

For the analysis of period phenotype, individual period estimates of F1 progenies of each population were produced after 4-5 day of consecutive conidial banding using the Fast Fourier Transform Nonlinear Least-Squares (FFTNLS) program [19, 32]. For the phase analysis, the reference time for phase of each individual genotype/progeny was the band center. Thus, the phase of each individual progeny was determined based on the time elapsed from one band center to the next after light to dark transition within a day. Band center was visually determined by the spore density. The time when cultures were transferred to dark cycle is, by definition, CT12 (dusk). Thus in these experiments, the time in band center of each individual was calculated by the following formula: ZT phase= (growth to band center/ overall growth) x 24+ 12. For example, if conidial band occurs 180 mm out of 280 mm total growth after light dark transition, ZT phase=  $24 \times (180/280) + 12 = 27.43$ . By convention, ZT is always expressed as 24 hr period. For example, the example ZT phase above will be expressed as ZT 3.43 (27.43 – 24).

### **Method; Genotyping and genetic map construction**

The genotyping method and linkage group analysis has been done as described previously [33-35]. The 564 F1 progenies (188 F1 haploid progeny from each line-cross, Table 2-6) were genotyped to determine the linkage maps for each cross. Genetic linkage maps of each population were constructed using two different algorithms, Map Manger QTX v. 0.3 [36] and GMENDEL v.3.0 [37] with the Kosambi mapping function [38]. Using Map manager, the initial linkage grouping was performed using the Double Haploid option with a threshold level of  $P= 0.001$ . Subsequently, Monte Carlo simulation with 500 iterations was used to test the marker locus order generated by GMENDEL. We also utilized the physical map information at the Broad Institute database, <http://www.broad.mit.edu/annotation/genome/neurospora/maps/Index.html>.

For semi-automated genotyping analysis, the 5' M13 sequence was attached to a forward primer in order to incorporate a fluorescent dye into the PCR product. Fluorescent dye labelled M13 forward primer and a marker specific reverse primer were used to generate fluorescent-labelled PCR product as previously described [39]. The composition of the PCR master mix was prepared as described in Cho et al [40], and the PCR profile was modified from Schuelke as follows [39]. The basic profile was: 5 min at 94°C, 30 cycles of 30 sec at 94°C, 45 sec at 55°C, 1 min at 72°C, and 25 cycles of 15 sec at 94 °C, 30 sec at 53°C, 1 min at 72°C, and 10 min at 72°C for final extension. Fluorescent-labelled PCR products for SSR loci were multiplexed with regard to each molecular weight and fluorescent dye. Each multiplexing set of primers was called a panel. One panel consisted of 12-15 SSR marker sets. The multiplexed PCR products were analyzed by an ABI 3730 (Applied Biosystems) according to the manufacturer's instructions.

### **Method; QTL analysis**

QTL analysis was carried out on the mean value of free running period and entrained phase in N2, N4 and N6 population (Table 2). The markers with significant segregation distortion ( $\chi^2$  test,  $p=0.05$ ) were disregarded from our QTL analyses. Composite Interval mapping (CIM) and Bayesian QTL mapping (BMQ) was used to identify putative clock QTL. CIM analysis using the QTLCartographer v.2.5 [41] for F1 mapping population was conducted with a walking speed of 0.5 and a window size of 3 cM under forward and backward regression model (probability into 0.01, probability out 0.1). To determine experimental type 1 error, 1,000 permutations test were performed in each phenotype of each population [42]. We defined the confidence interval as the physical genome region above the threshold defined by this 1,000 permutation test. This functional confidence interval region was in average 10 cM or about 200-300 kilobase pair (kb) around the genetic positions with the maximum LR

score. We searched the candidate clock QTL genes among these genome regions within confidence interval regions. LR critical values ranged from 11 to 12 ( $P = 0.05$ ) across phenotypes and populations. Additive effect estimates and percentages of variance explained by the QTL were generated with Eqtl, testing hypothesis 10 and using model 6 from Zmapqtl. Likelihood ratio (LR) profiles for two circadian properties including free running period and entrained phase in the three populations of our study

CIM analysis using the QTLCartographer v.2.5 [41] for backcross 4<sup>th</sup> generation (BC4F1) mapping population was conducted with a walking speed of 0.5 and a window size of 3 cM by standard model under forward regression mode. To determine experimental type 1 error, 1,000 permutations test were performed in each phenotype of each population [42]

The BMQ approach uses a hierarchical modeling scheme where at the “top” level, each marker has a probability of being categorized into one of three classes: linked to a QTL with a positive effect on (i.e. increases) the value of the phenotype ( $p_+$ ), linked to a QTL with a negative effect on the phenotype ( $p_-$ ), and not linked to a QTL ( $1 - p_+ - p_-$ )[43]. At the “bottom” level, the actual effects of QTLs are defined in the usual way using a linear model. The advantage of this hierarchical classification approach is that, with an appropriate choice of prior for marker class hyper-parameters [43], we can implement an efficient stochastic search in low-dimensional model subspaces. This avoids the tendency to over-shrink estimates of QTL effects observed with other multi-QTL Bayesian approaches [44, 45]. Following the previous report [43], we used a “spike and slab” prior [46] which incorporates the assumption that most markers will not be linked to a QTL. In our Bayesian classification framework, the probability that a marker is linked to a QTL is reflected by the posterior probability distribution associated with the marker classes  $p_+$  and  $p_-$ . We implemented the Gibbs sampler described in Zhang et al. 2005 to generate samples from these posterior distributions. Marginal posteriors for both the additive effect ( $\beta$ ) and probabilities of

categorization ( $p_+$ ,  $p_-$ ) were estimated by sampling 5000 iterations after an initial burn in of 5000.

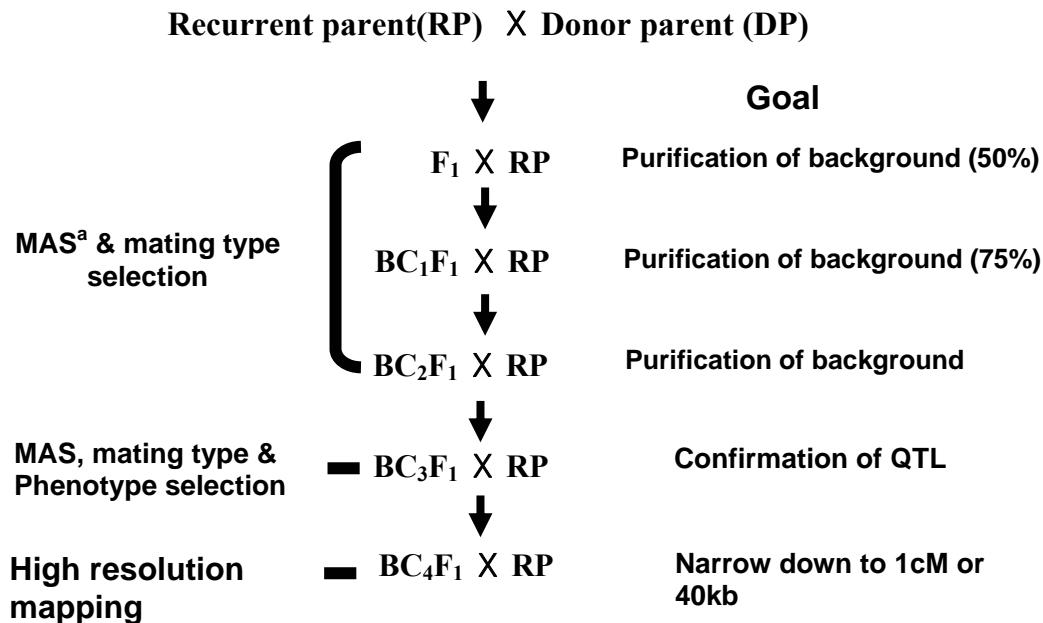
We considered the cumulative probability greater than 0.5 that the marker is in the  $p_+$  ( $p_-$ ) class to determine whether a marker is linked to a QTL (hereafter we refer to this as the Posterior Probability Threshold or PPT). This is a univariate version of the heat map summary provided in Zhang *et al.* (2005) and reflects the probability that a marker has a greater than 50% chance of being linked to a QTL.

QTL names were formulated in order of the name of the mapping population, a QTL method used ("C" for CIM specific QTLs and "B" for BMQ specific QTLs, BC for the QTLs detected by CIM and BMQ), the trait targeted (for example, "per" for period and "pha" for phase), chromosome (chr.) number, and numeric numbers to differentiate QTLs within a chr. For example, N6CBper7-1 refers to 1<sup>st</sup> QTL located chromosome 7 of period phenotype in N6 that were detected both by CIM and BMQ method.

### **Method; The Near isogenic line (NIL) construction**

The near isogenic lines (NIL) were constructed to confirm the QTL effect. Dependent on an origin of allele that generates a QTL effect, one of mapping parent is used as a recurrent parent (RP) and the other parent is used as a donor parent (DP). In F1 population, several candidate strains are selected as a starting genetic material where the genetic background except the targeted region is relatively similar to the RP genome. Those selected strains are backcrossed (BC) to RP in order to clean the genetic background of DP except the target site. In every BC generation, two hundred ascospores were selected to test the mating type and genotypes in the marker loci around the QTL region. Individuals (BC progeny) with opposite mating type to RP and favorable genotype and favorabl were selected. The selected strains were then bulked and crossed to RP to progress another BC generation. This step was repeated up to BC

<sup>2<sup>nd</sup> generation (BC<sub>2</sub>F<sub>1</sub>). To test the allelic effect of QTL, the phenotype of individuals with target allele was evaluated at BC 3<sup>rd</sup> generation (BC<sub>3</sub>F<sub>1</sub>) where about 85 % of genetic background is switched to of the RP genome. The strategy for the NIL construction is graphically described in Figure 4-3.</sup>



**Figure 4-3. Strategy for the QTL confirmation and the subsequent high resolution mapping step.**

, RP or DP can be determined dependent on an origin of allele of a QTL effect. <sup>a</sup>, MAS represent marker assisted selection.

## **Experiments and Results; Phenotypic analysis of *Neurospora* circadian rhythm (Period and phase)**

We generated three F1 populations derived from mapping parents described in Table 4-1 and Figure 4-4 to map QTLs for two circadian phenotypes, the free running period and the light entrained phase. Each mapping population was composed of 188 progeny derived from a cross between two *N. crassa* natural accessions (Table 4-2). The continuous patterns of the distribution of both of the circadian phenotypes in F1 progenies were observed in all three populations, indicating that inheritance of the circadian properties in *N. crassa* is polygenic (Figure 4-5 and Figure 4-6), which is consistent with results with previous studies in other systems [17-19, 21, 22, 47-54]. The mean period lengths of our mapping populations were 21.4 hrs, 21.7 hrs and 21.7 in N2, N4 and N6 populations, respectively (Table 4-2). The period of the parental lines of each population were approximately similar to mean value of the periods in the F1 progeny (Figure 4-5 and Table 4-2). The ranges of the period length in N2, N4 and N6 were 4.55, 5.79 and 4.12 hr, respectively. The broad sense of heritability ( $H^2$ ) of *N. crassa* clock phenotype was high in all populations, 0.62, 0.87 and 0.85, which suggests the phenotypic variation in the segregating populations was due to mostly genetic effects.

**Table 4-1.** *N. crassa* accessions used as parental stains in crosses.

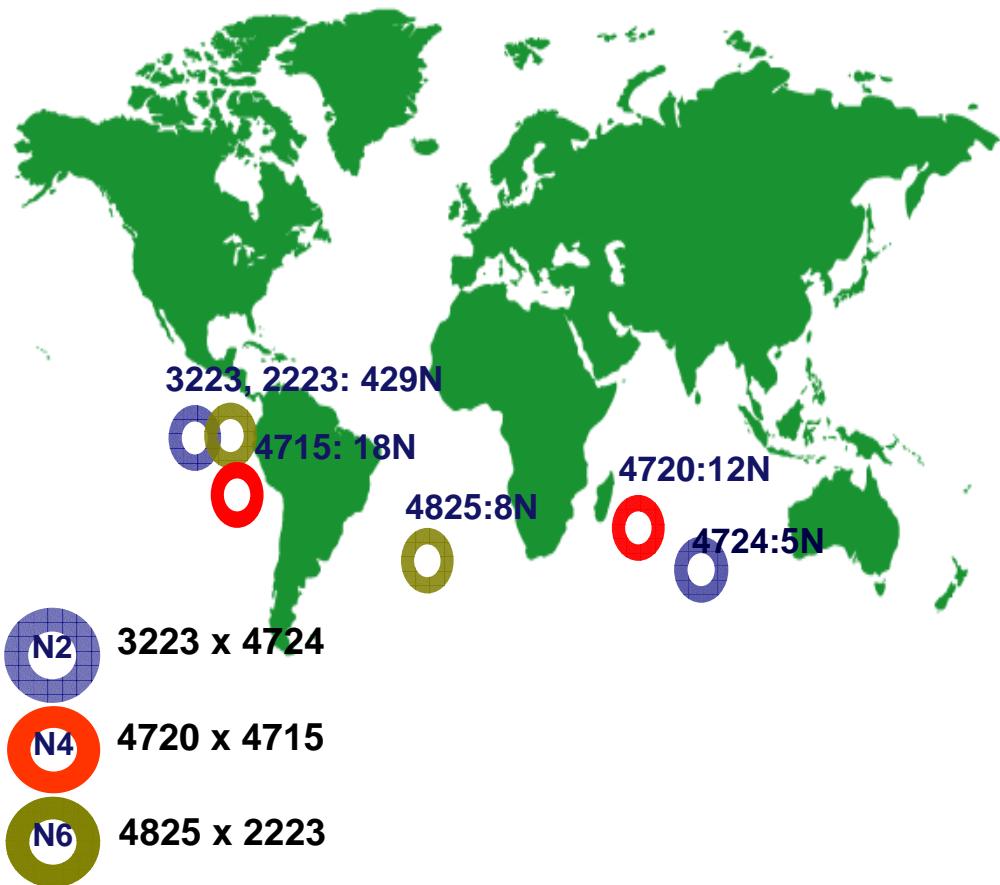
Cross	Strain <sup>a,b</sup>	Mating type	Origin of collections	Period <sup>c</sup>	Phase <sup>d</sup>
N2	3223 <sup>e</sup> (♀)	mat A	Louisiana,	21.3 ± 0.20	0.6 ± 0.18
N2	4724 (♂)	mat a	Penang,	21.0 ± 0.17	0.5 ± 0.24
N4	4720 (♀)	mat A	India	21.4 ± 0.42	23.2 ± 0.44
N4	4715 (♂)	mat a	Haiti	21.7 ± 0.14	23.5 ± 0.26
N6	4825 (♀)	mat A	Tiassale, Ivory	22.2 ± 0.10	2.5 ± 0.25
N6	2223 (♂)	mat a	Iowa, U.S.A.	21.3 ± 0.10	1.4 ± 0.19

<sup>a</sup> ♀, ♂ represent female or male parent, respectively, of each population

<sup>b</sup> Strain number (<http://www.fgsc.net/scripts/StrainSearchForm.asp>)

<sup>c</sup> period values refer to the period values under free running condition, unit = hr

<sup>d</sup> phase values refer to the phase values under 12 hr light :12 hr dark cycles, unit = ZT



**Figure 4-4. Graphical summary of the collection site of *N. crassa* accessions used as parental stains in crosses**

Traditionally, the phase phenotype has been expressed in subjective hours, or zeitgeber (ZT) hours. In contrast to period, the means of the phase values among progenies in different populations were different; the mean phase value in N2 and N4 was 0.5 ZT hr, whereas, that in N6 was 1.6 ZT hr (Figure 4-6 and Table 4-2). The phase of the parental lines of N2 and N6 populations was close to the mean of the phase of the progenies. However, in N4 population, 93% of N4 progeny were distributed toward the right side beyond the mean value of parental strains in the phase phenotype (Figure 4-6 and Table 4-2). The ranges of phase distribution in N2, N4 and N6 were 4.7, 6.1 and 4.1 ZT hr, respectively. As observed in period value, relatively high heritability in phase was also observed in each population; the heritabilities of N2, N4 and N6 were 0.84, 0.87 and 0.74, respectively. There was no correlation between period and phase under entrained environment within a population in all three populations (Figure 4-4).

### **Experiments and Results; Comparison of two QTL methods (CIM vs BMQ)**

In an effort to pinpoint the clock QTLs and identify genetic elements responsible for subtle phenotypic variation in the *N. crassa* clock efficiently, two independent QTL analyses methods were used, CIM and BMQ. In BMQ approach, we considered the cumulative probability greater than 0.5 (Posterior Probability Threshold or PPT) to determine whether a marker is linked to a QTL (Methods). To assess the appropriate PPT cutoff to use when determining whether a marker is linked to a QTL, we simulated QTL data using the marker data for population N6 and determined the PPT for all markers not linked to QTL, neutral markers [43].

**Table 4-2. Phenotypic variation in period length and phase in N2, N4 and N6 population.**

Pheno-type	Cross	Mean <sup>a</sup>	STDEV	Range <sup>b</sup>	Heritability <sup>c</sup>	Percentage of progenies that show the transgressive segregation		
						Total (%)	(-) side <sup>d</sup> (%)	(+) side <sup>e</sup> (%)
period	N2	21.4	0.62	4.60	0.62	47	35	75
	N4	21.7	0.98	5.80	0.87	43	46	54
	N6	21.7	0.57	4.10	0.85	20	54	46
phase	N2	0.5	0.75	4.67	0.84	54	49	51
	N4	0.5	0.92	6.11	0.87	80	7	93
	N6	1.9	1.05	4.14	0.74	44	49	51

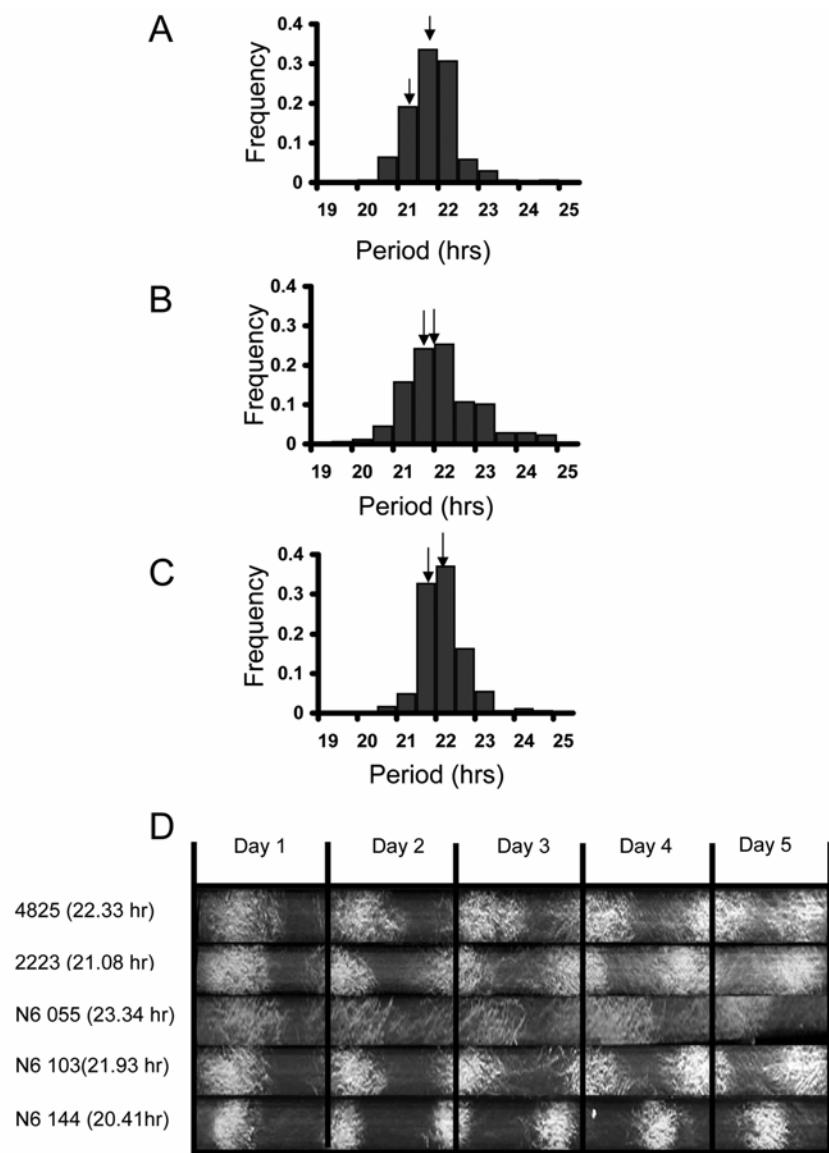
<sup>a</sup> The unit for period is hr and the unit for phase is ZT hr.

<sup>b</sup> Range for period = most highest phenotype - most lowest phenotype

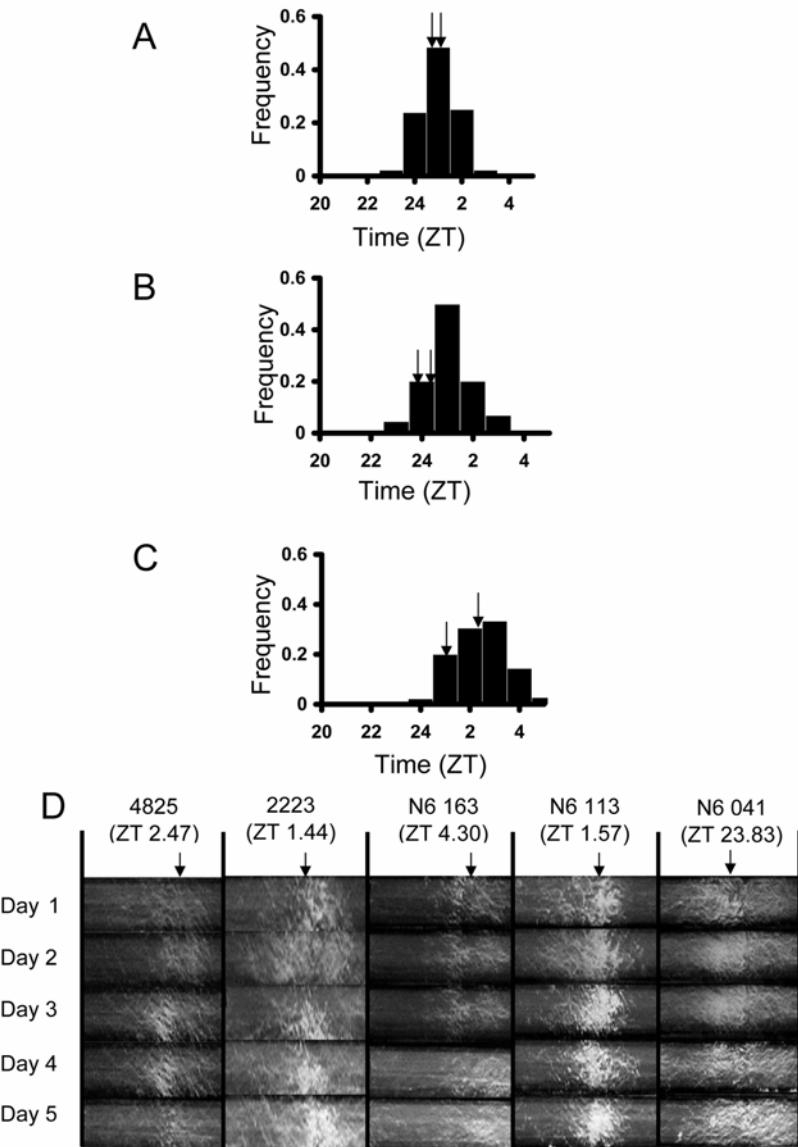
<sup>c</sup> Variance associated with the genotype effect by 2-way ANOVA and its significance, \*p <0.05; \*\*\*p <0.001.

<sup>d</sup> (-) side , percentage of progeny that show transgressive phenotypic segregation in lower phenotypic value than a mean value of parental phenotypes in each mapping population.

<sup>e</sup> (+) side , percentage of progeny that show transgressive phenotypic segregation in higher phenotypic value than a mean value of parental phenotypes in each mapping population



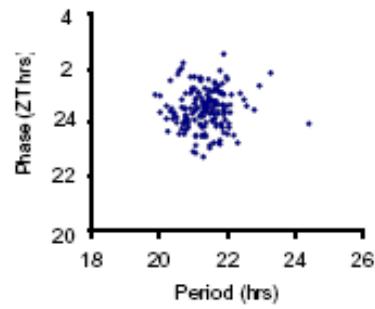
**Figure 4-5. The circadian period variation in F1 populations. The phenotypic distributions of F1 progenies of N2 (A), N4 (B), and N6 (C) populations.**  
 x-axis represents circadian period and y-axis represents frequency of period in the corresponding F1 progenies. Arrows indicate the periods of the parents for each cross.  
 (D) Race tube images of conidial banding patterns under constant darkness (free running condition). The panel shows two N6 parents (FGSC 4825 and FGSC 2223) and three representative progeny (N6 055, N6 103 and N6 144) with different free running periods. The vertical black lines represent growing front in 24 hr period. The number in the parenthesis is the average period of the strain.



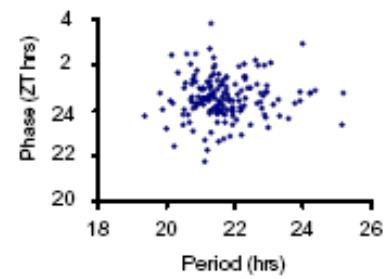
**Figure 4-6. The entrained phase variation under 12:12 light/dark condition in F1 populations. The phenotypic distribution of F1 progeny of N2 (A), N4 (B) and N6 (C) populations.**

x-axis represents the entrained phase in ZT (zeitgeber) time (see Materials and Methods) and y-axis represent frequency of the periods in the corresponding F1 progenies. ZT 24 is the same as ZT 0. ZT 0 is when light is on (dawn) and ZT 12 is when light is off (dusk). Arrows indicate the phases of the parents for each cross. (D). Race tube images of conidial banding pattern under 12:12 LD cycles for 5 days. The panel shows N6 parents (FGSC 4825 and FGSC 2223) and three representative progeny (N6 163, N6 113 and N6 041) with different phases. The number in the parenthesis is the average phase of the strain for 5 days. The arrow indicates the average phase value of each strain.

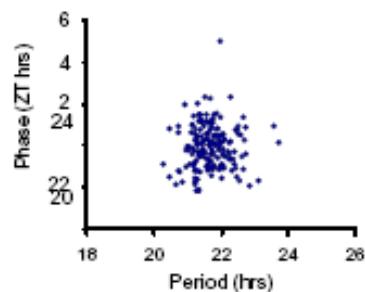
N2



N4



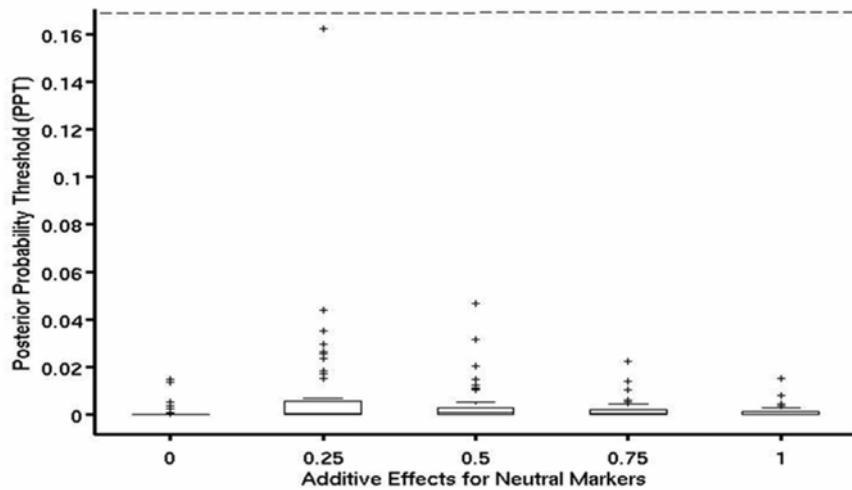
N6



**Figure 4-7. Scatter plot analysis between period and phase in N2, N4 and N6 population**

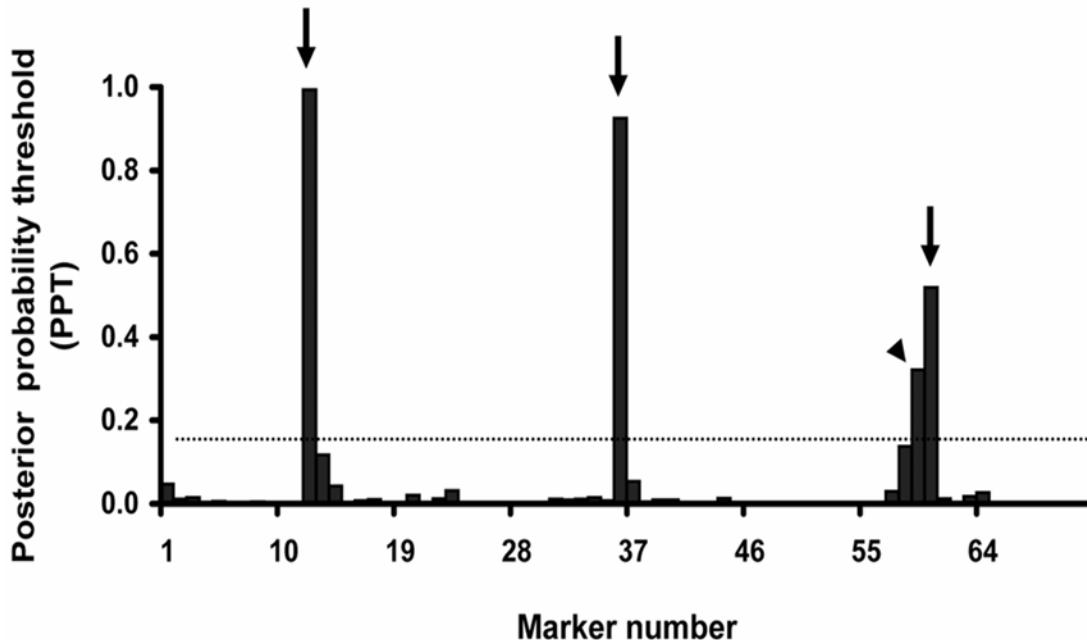
The results of the simulations are summarized in Figure 4-8. When there is no QTL, i.e. additive effects = 0, no neutral markers had PPT>0.01 (or < -0.01). When three QTL of equal effects spaced throughout the genome are simulated, the PPT can be larger but the bulk of the markers still have a PPT<0.05. In fact, even as the effects of these QTLs are decreased to an additive effect of 0.25 (heritability of 0.13), only 1 neutral marker had PPT > 0.05, showing PPT= 0.16 (Figure 4-8). Missing genotype data can increase the type I error rates for neutral loci when there are QTL present as we see in Figure 4-8, where the distribution of PPT across neutral markers has greater outliers with smaller additive effects. We therefore used PPT=0.17 as a cutoff for deciding whether markers were linked to QTL to minimize a false positive result. We also performed a simulation experiment with three pre-defined QTLs (Fig. 4-9) Note that neutral markers surrounding the marker in strongest linkage disequilibrium with a QTL also have reasonably high PPT but that the highest PPT occurs at the marker where the true QTL is positioned (Figure. 4-9). For a set of consecutive markers with  $PPT \geq 0.17$ , we therefore determined the maker with the greatest PPT to be linked to a QTL.

We detected twice the number of QTLs from BMQ compared to that from CIM (Figure 4-10A). BMQ identified all QTLs that were found in CIM in both phenotypes of our study (Table 4-3, Figure 4-10B) except two QTLs (*N6CPer7-2* and *N6Cper7-3*, Table 4-3). The peak positions on the marker loci that linked to significant QTL were highly consistent in the two methods (Figure 4-10B).



**Figure 4-8. Distributions of PPT for neutral markers.**

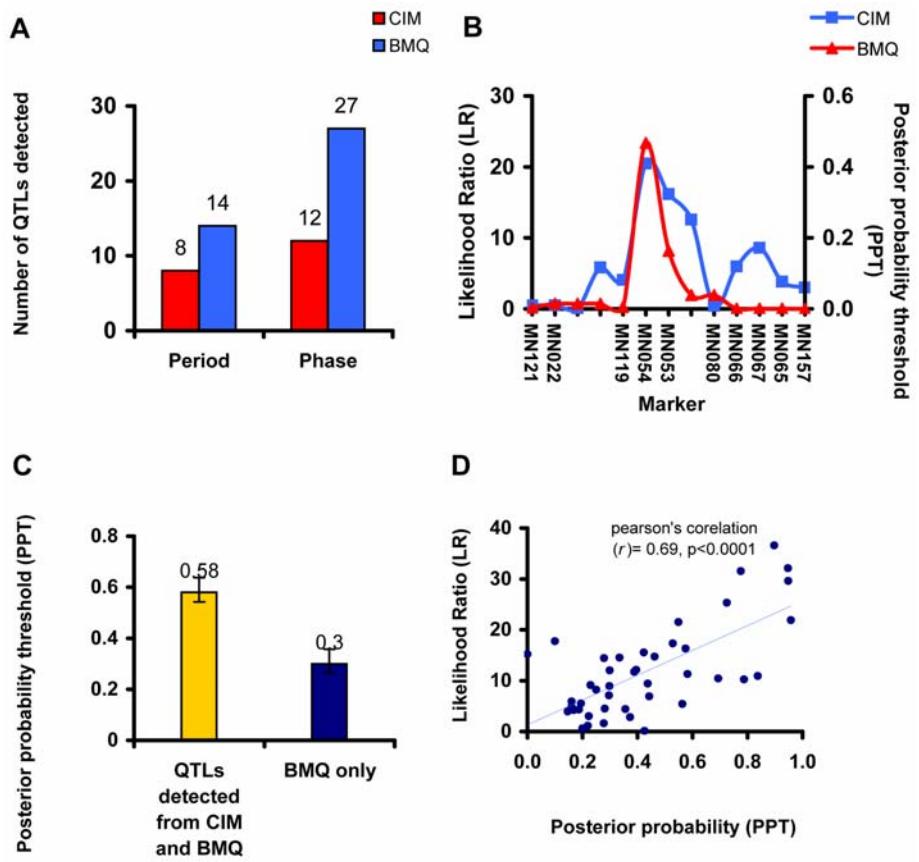
A simulation with no QTLs (additive effects of 0) and four simulations with three QTLs distributed uniformly throughout the genome with additive effects 0.25, 0.5, 0.75, and 1 (heritabilities 0.13, 0.38, 0.58, and 0.72 respectively) are shown. The distributions of PPT values for markers unlinked to QTLs across the genome (neutral markers) are represented as box plots. The middle line in each box plot is the median, the boxes span the interquartile range, and the whiskers span the maximum and minimum observations, unless there are outliers, which are defined as observations greater than 1.5 times the interquartile range above or below the box. Outliers are represented as pluses. The PPT value 0.17 (dotted line) was used as the threshold to eliminate false positives.



**Figure 4-9. PPT for three simulated QTLs.**

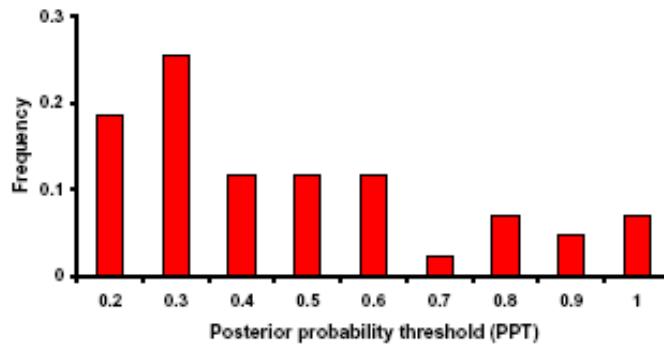
Three QTLs were simulated with additive effect 0.5 (heritability 0.38). The markers, 12, 36 and 60, were the true QTLs (black arrows). The simulation showed that the same markers have the highest PPT values. Although the peak PPT occurs at the marker linked to the QTL, there was one neutral marker has a PPT value higher than 0.17 (arrow head). The dotted line represents the threshold PPT, 0.17.

The ranges of the PPT were variable, spanning from 0.17 to 0.96 (Figure 4-11), in which the median value is 0.43. Average PPT for QTLs detected by both methods is significantly higher than the average value PPT for QTLs that were detected only by BMQ (0.58 versus 0.30, Figure 4-10C). The LR score in CIM showed a highly significant positive correlation with PPT in BMQ (Pearson's correlation coefficient = 0.69 p<0.0001, Figure 4-10D). Thus, hereafter, we will use PPT value when we describe QTLs except those two QTLs (*N6CPer7-2* and *N6Cper7-3*) that were not detected by BMQ.



**Figure 4-10. Summary of Bayesian QTL analysis (BMQ, CIM).**

(A). Comparison of number of QTLs using CIM and BMQ methods (red bar, CIM and blue bar, BMQ approach). (B). The graphical description of BMQ and CIM analysis (red line, CIM and blue line, BMQ analysis). The x-axis represents the marker position in the linkage map. The primary y-axis on the left is LR score for CIM analysis and secondary y-axis is PPT score for the BMQ approach. (C). The average PPT (y-axis) in between QTLs mapped by BMQ and CIM simultaneously and QTLs mapped by BMQ specifically (x-axis). Error bar represents for mean standard error. (D). The scatter plot analysis between LR score by CIM and by BMQ in each QTL locus. In the panel (D), the variable plotted on the x-axis represents PPT of a QTL detected by BMQ analysis and the y-axis is the LR score of the corresponding locus measured by CIM. The diamond shaped dots with pink color are scatter plots representing QTL loci that are commonly detected by BMQ and CIM. The QTLs that are detected by BMQ specifically are denoted by the rectangular shaped dot with blue color.



**Figure 4-11. Distribution of PPTs of QTLs in BMQ analysis.**  
x-axis is range of PPT and y-axis is frequency in given PPT.

### Experiments and Results; Comparisons with the previous clock study

From two statistical methods, we detected 43 QTLs from three populations that affect the two circadian clock properties, period and phase (Table 4-3). We detected similar number of QTLs in two clock phenotypes per population; 8 QTLs in period and 9 QTLs in phase per population (Table 4-3) except the period phenotype in N2 where we did not detect any significant QTLs with either CIM or BMQ analyses.

We searched candidate QTL genes around the detected clock QTL regions to see whether previously characterized clock genes were co-localized with the clock QTLs. We defined the confidence interval region for the clock QTL by performing a permutation test. We also developed *wc-1* and *vvd* specific SSR markers as positive controls. This strategy was based on the idea that one of QTLs maybe co-localized with these two key genes (*wc-1* in period, *vvd* in phase) in *N. crassa* clock regulation.

Although, these genes were known to influence phase of the *Neurospora* clock, the specific roles of theses candidate genes for the phase-determination have not been clearly studied except *vvd* [55, 56]. These results suggest that our QTL studies were concordant with previous clock studies and give insight in the mechanism of *N. crassa* regulation, especially in phase regulation.

**Table 4-3. Summary of the additive QTLs in circadian properties that are segregated in three different population formed by *N. crassa* natural accessions using Bayesian QTL analysis.**

Cross	Trait	QTL ID	Marker	Chr <sup>a</sup>	PPT <sup>b</sup>	LR <sup>c</sup>	Additive genetic variance <sup>d</sup>	Origin of allelic effect <sup>e</sup>	Candidate Gene <sup>f</sup> (ncu <sup>g</sup> )
N2	Phase	<i>N2BPha</i> 2-1	MN125	2	0.30	7.10	0.31	3223	na
N2	Phase	<i>N2CBPh</i> a2	MN229	2	0.72	25.30	0.46	4724	na
N2	Phase	<i>N2BPha</i> 3	MN173	3	0.19	5.50	0.25	3223	na
N2	Phase	<i>N2BPha</i> 4	MN182	4	0.17	5.90	0.26	3223	<i>pp2a</i> (ncu06630.2)
N2	Phase	<i>N2CBPh</i> a5-1	MN051	5	0.78	31.50	0.48	4724	na
N2	Phase	<i>N2CBPh</i> a6	MN054	6	0.46	14.70	0.36	3223	<i>vvd</i> (ncu03967.2)
N2	Phase	<i>N2BPha</i> 7-2	MN247	7	0.17	4.20	0.22	4724	na
N4	Period	<i>N4CBPe</i> r1-1	MN008	1	0.16	4.8	0.28	4715	na
N4	Period	<i>N4CBPe</i> r1-3	MN129	1	0.19	4.3	0.29	4715	<i>prd-4</i> (ncu02814.2)
N4	Period	<i>N4CBPe</i> r1-2	MN042	1	0.30	12	0.41	4720	<i>ckII catalytic subunit</i> (ncu03124.2))
N4	Period	<i>N4BPer2</i>	MN094	2	0.44	9.4	0.38	4715	na
N4	Period	<i>N4BPer3</i>	MN003	3	0.25	8.2	0.31	4715	na
N4	Period	<i>N4BPer4</i>	MN162	4	0.44	6.9	0.34	4720	na
N4	Period	<i>N4CBPe</i> r5	MN153	5	0.58	11.3	0.45	4720	na
N4	Period	<i>N4CBPe</i> r7	MN046	7	0.95	32.1	0.66	4720	<i>wc-1</i> (ncu02356.2)
N4	Phase	<i>N4CBPh</i> a1-1	MN008	1	0.42	15.5	0.35	4715	na
N4	Phase	<i>N4BPha</i> 1-3	MN019	1	0.28	4.5	0.25	4715	na
N4	Phase	<i>N4BPha</i> 1-2	MN129	1	0.22	3	0.26	4715	<i>prd-4</i> (ncu02814.2)
N4	Phase	<i>N4BPha</i> 4-1	MN074	4	0.22	1.1	0.32	4715	na
N4	Phase	<i>N4CBPh</i> a4-2	MN090	4	0.57	16.3	0.39	4715	<i>pp2a</i> (ncu06630.2)
N4	Phase	<i>N4CBPh</i> a5	MN061	5	0.90	36.58	0.59	4715	na

**Table 4-3 (continued)**

Cross	Trait	QTL ID	Marker	Chr <sup>a</sup>	PPT <sup>b</sup>	LR <sup>c</sup>	Additive genetic variance <sup>d</sup>	Origin of allelic effect <sup>e</sup>	Candidate Gene <sup>f</sup> (ncu <sup>g</sup> )
N4	Phase	<i>N4CBPh a6</i>	MN157	6	0.40	12.1	0.34	4715	<i>na</i>
N6	Period	<i>N6CBPe r2</i>	MN026	2	0.34	14.5	0.30	2223	<i>na</i>
N6	Period	<i>N6CBPe r3</i>	MN108	3	0.28	14.4	0.22	4825	<i>na</i>
N6	Period	<i>N6BPer4 -1</i>	MN075	4	0.20	0.6	0.22	4825	<i>na</i>
N6	Period	<i>N6BPer4 -2</i>	MN220	4	0.23	9.1	0.23	4825	<i>na</i>
N6	Period	<i>N6BPer5 -2</i>	MN061	5	0.17	3.9	0.21	4825	<i>na</i>
N6	Period	<i>N6CPer-7-3</i>	MN046	7	0.00	15.2	na	2223	<i>wc-1</i> (ncu02356.2)
N6	Period	<i>N6CPer-7-2</i>	MN046	7	0.10	17.75	na	2223	<i>frq</i> (ncu02265.2)
N6	Period	<i>N6CBPe r-7-1</i>	MN168	7	0.95	29.6	0.35	4825	<i>fwd-1</i> (ncu045450.2)
N6	Phase	<i>N6BPha 1-1</i>	MN018	1	0.56	5.4	0.39	2223	<i>na</i>
N6	Phase	<i>N6CBPh a1-2</i>	MN131	1	0.79	10.2	0.37	4825	<i>prd-4</i> (ncu02814.2)
N6	Phase	<i>N6CBPh a1-3</i>	MN041	1	0.55	21.5	0.38	2223	<i>ckII catalytic subunit</i> (ncu03124.2)
N6	Phase	<i>N6BPha 2-2</i>	MN027	2	0.28	1.6	0.25	2223	<i>na</i>
N6	Phase	<i>N6BPha 2-1</i>	MN038	2	0.84	10.9	0.51	4825	<i>na</i>
N6	Phase	<i>N6BPha 3-1</i>	MN084	3	0.30	8.9	0.37	2223	<i>na</i>
N6	Phase	<i>N6BPha 3-2</i>	MN089	3	0.39	11.7	0.34	2223	<i>na</i>
N6	Phase	<i>N6CBPh a4</i>	MN215	4	0.53	17.3	0.47	4825	<i>na</i>
N6	Phase	<i>N6BPha 5-2</i>	MN153	5	0.43	0.1	0.40	2223	<i>na</i>
N6	Phase	<i>N6BPha 5</i>	MN155	5	0.36	4.4	0.49	4825	<i>na</i>
N6	Phase	<i>N6BPha 5-1</i>	MN083	5	0.37	2.8	0.35	2223	<i>na</i>
N6	Phase	<i>N6CBPh a6</i>	MN054	6	0.96	21.9	0.79	4825	<i>vvd</i> (ncu03967.2)
N6	Phase	<i>N6CBPh a6-1</i>	MN067	6	0.69	10.4	0.41	2223	<i>na</i>

<sup>a</sup> chr., chromosome number<sup>b</sup> PPT, Posterior Probability<sup>c</sup> LR, Likelihood ratio<sup>d</sup>, In this column, the estimation of additive genetic variance value originates from Bayesian multiple QTL analysis.<sup>e</sup>, Each number in this column represents the accession number used in Fungal Genetics Stock Center ([www.fgsc.net](http://www.fgsc.net))

Table 4-3 (continued)

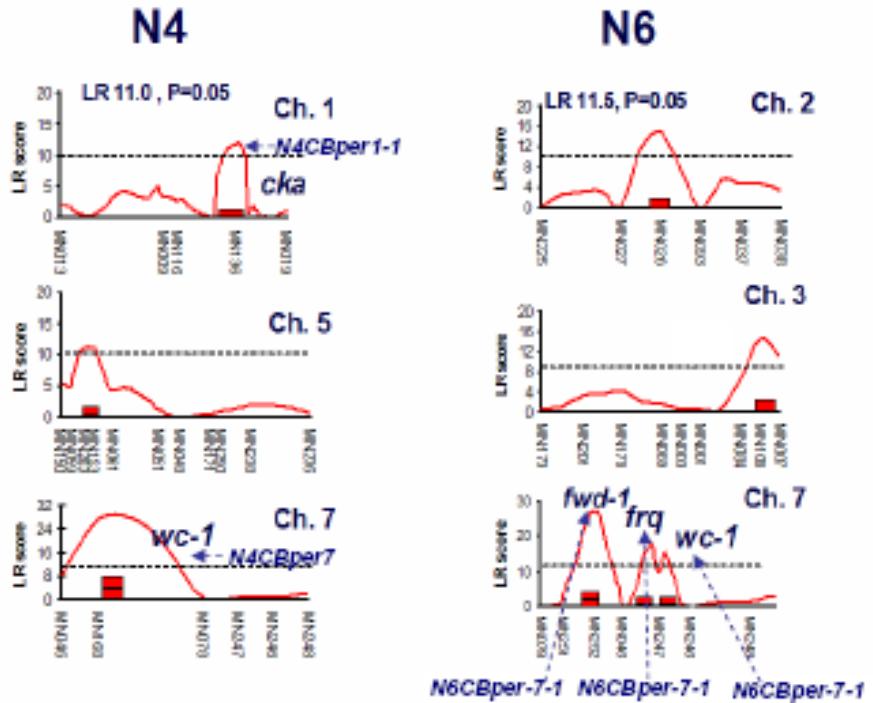
<sup>f</sup> The range of candidate gene are plus and minus 200-300 kilobase pair (kbp) at the genetic locus where LR score or PPT is maximized. In this column, *na* is abbreviation of not available, which means no previously characterized clock gene is available.

<sup>g</sup> In case a candidate gene is available, the corresponding ncu number of the candidate gene which is obtained from Neurospora genome database in Broad institute website is recorded in a parenthesis.

Nine (out of 16 QTLs in period) and 16 QTLs (out of 26 QTLs in phase) were characterized as unknown clock loci, which suggest there is a lot more to understand about *N. crassa* circadian clock (Figure 4-13). Several QTLs, especially in phase phenotype, with high significance level are still uncharacterized, including *N2Bpha5-1* (PPT=0.78, LR=35.5), *N2Bpha2* (PPT=0.72, LR=25.3) and *N4Bpha5* (PPT=0.90, LR=35.8). The co-localized candidate genes with QTLs are also summarized in Table 4-3 and Figure 4-13.

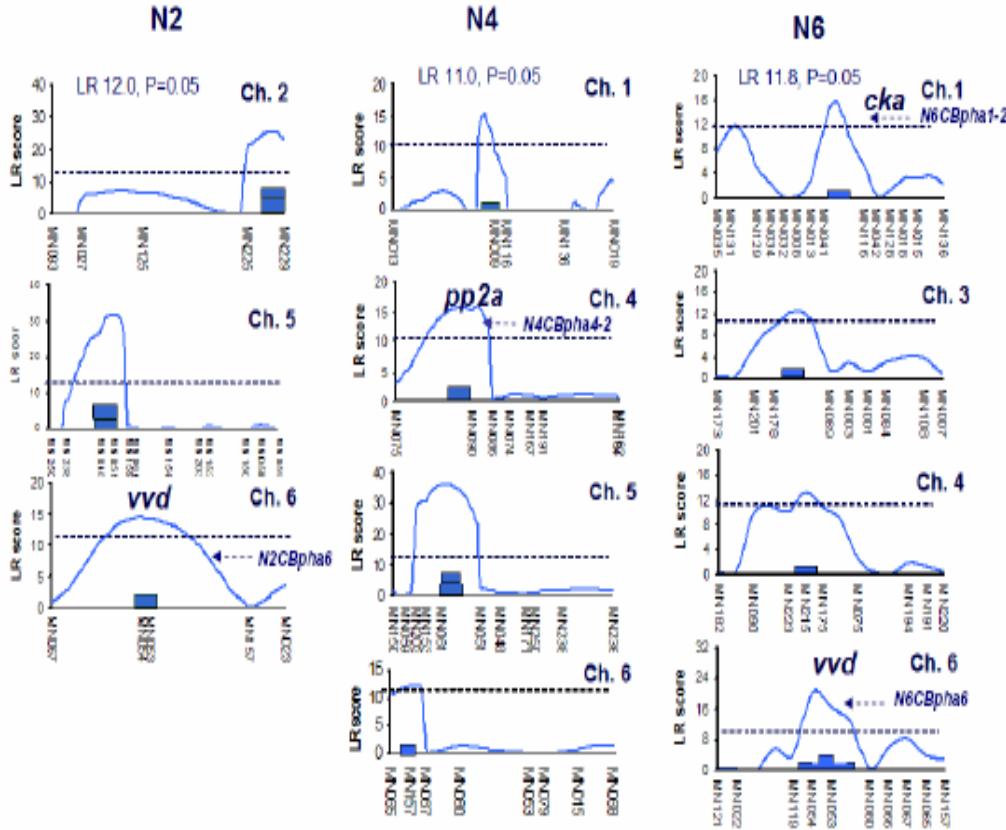
We wanted to estimate how many clock QTLs are identified more than one time in different populations. Obviously, we could increase the chance of identifying all potential clock QTLs by increasing the number of mapping populations. However, for practical reasons, we chose to characterize three independent line-cross populations. To avoid the over-estimation of the number of clock QTLs, we excluded the common QTLs identified in different populations. We found that there were no significant chromosome re-arrangements among *N. crassa* natural isolates that we studied (Figure 4-14). Thus, we defined the common QTL as a QTL linked to the same SSR marker in more than one population for the same phenotype regardless of their relative genetic positions.

Three QTLs out of 16 QTLs for the period phenotype and 8 QTLs out of 27 QTLs for the phase phenotype are common QTLs (Figure 4-15). Thus, our data suggest that at least 13 different QTLs contribute to the period phenotype, and 19 different QTLs contribute to the phase phenotype respectively.



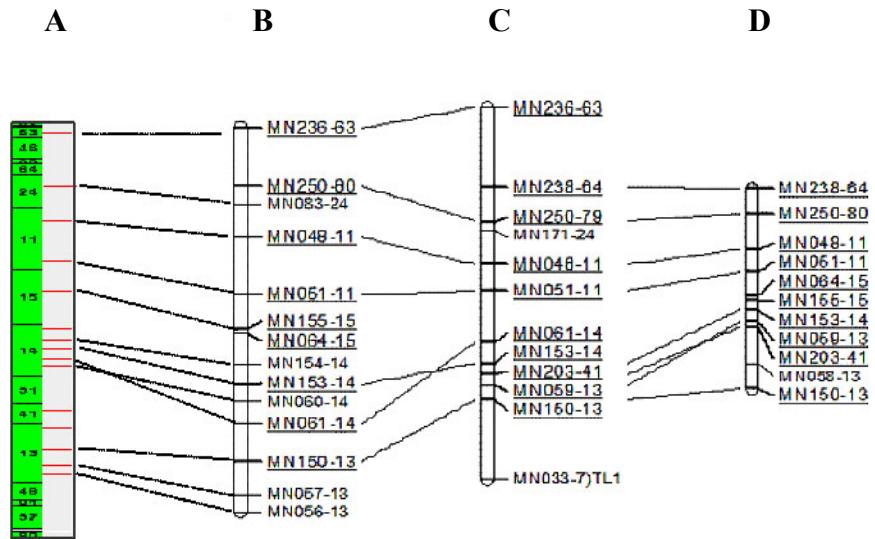
**Figure 4-12. The graphical description of composite interval mapping (CIM) analysis in period length under free running condition in N4 (A) and N6 (B) populations.**

x-axis of each panel represents marker position within the linkage map. y-axis of each panel represents likelihood ratio (LR) score of each genetic position denoted by cM. Dotted line in each panel stands for threshold level determined by 1000 permutation test. QTL names (indicated by arrows with dotted line) and candidate genes in the corresponding QTL are shown around the peak position of the QTL (refer to Table 3).



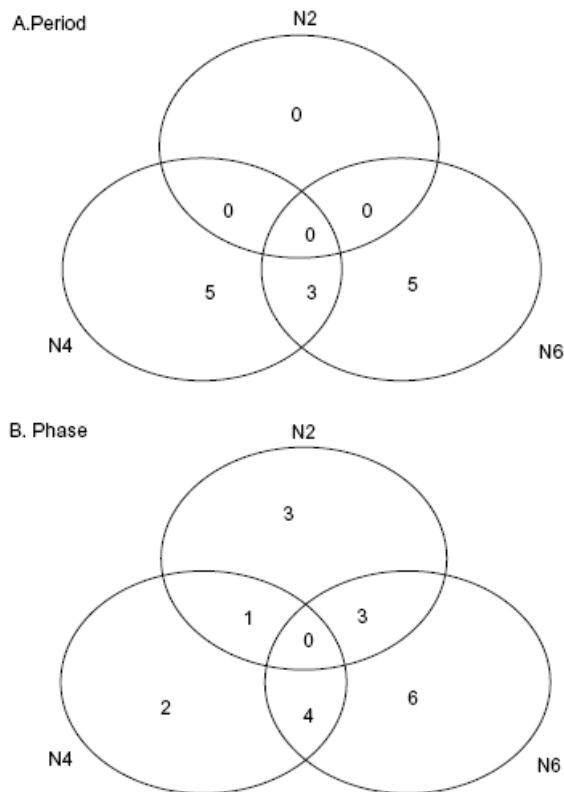
**Figure 4-13. The graphical description of composite interval mapping (CIM) analysis in the phase under the 12:12 LD cycle in N2(A), N4 (B) and N6 (C) population.**

x-axis of each panel represents marker position within the linkage map. y-axis of each panel represents likelihood ratio (LR) score of each genetic position. Dotted line in each panel stands for threshold level determined by 1000 permutation test. QTL names (indicated by arrows with dotted line) and candidate genes in the corresponding QTL are shown around the peak position of the QTL (refer to Table 3).



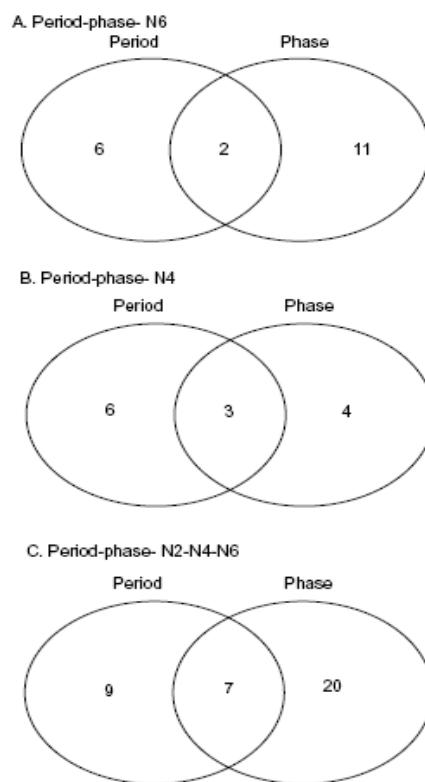
**Figure 4-14. Comparison of mapped loci between Neurospora physical map (A) and linkage maps derived from N6 (B), N4 (C), N2 (D) cross respectively.**

The image of physical map is obtained from MoMMs. The chromosome 5 is shown here as a example. Anchored markers in which the marker name is underlined are connected among each other.



**Figure 4-15. Venn diagram analysis of period (A) and phase QTL (B) among populations.**

Lastly, we wanted to know how closely the period and entrained phase phenotypes are genetically interrelated. To answer that question, we investigated on how many QTLs were contributing both to period and phase phenotypes. Since we could not detect any period QTL in N2, we excluded the comparison between the phase and period QTLs in N2 population. Three QTLs in N4 and two QTLs in N6 contribute to both in period and phase variations respectively (Table 4-3 and Figure 4-16). We also found seven QTLs that contribute to both period and phase variations when we consider all three populations (Table 4-3). This suggests that there are common genetic elements contributing for both period and phase phenotypes.



**Figure 4-16. Venn diagram analysis of between period and phase in population specific (A and B) and all three populations (C).**

## **Experiment and Results; Further characterizations of the QTL effect**

The reliable QTLs which include N6CBpha6, N4CBpha5, N6CBper7-1 and N6CBper5 were chosen for further characterizations. In F1 population, several candidate strains are selected as a starting genetic materials based on their genotypes (method). Those individuals were bulked and crossed to a recurrent parent (RP) to generate BC1F1 population. In the generation, individuals which showed the desirable genotypes at targeted marker loci along with an appropriate mating type were screened. About 15 % of progeny were selected for those two criteria and those selected were combined and backcrossed to the RP. This step is repeated up to the following generation (BC2F1). The steps required at each backcross generation and the information regarding genetic marker loci tested and recurrent or donor parent (RP or DP) of the targeted QTL are summarized at Table 4-4 and 4-5 respectively.

**Table 4-4. The information of backcross parent and tested marker.**

Targeted QTL	Backcross parent information		Tested markers
	Recurrent parent	Donor parent	
N4CBper5	4715 <sup>a</sup> (a) <sup>b</sup>	4720 (A)	MN153 <sup>c</sup>
			MN061
			MN051
			MN153
N4CBpha5	4720 (A)	4715 (A)	MN061
			MN051
			vii109
N6CBper7-1	2223 (a)	4825 (A)	MN046
			MN168
			MN053
N6CBpha6	2223 (a)	4825 (A)	MN054
			MN119

<sup>a</sup>, The strain ID in backcross parent section belongs is the number used in Fungal Genetics Stock Center (FGSC, Kansan Univ.). <sup>b</sup>, The A or a in the parenthesis stands for the mating type of a strain. <sup>c</sup>, The detailed genetic marker information can be found in Table 2-6.

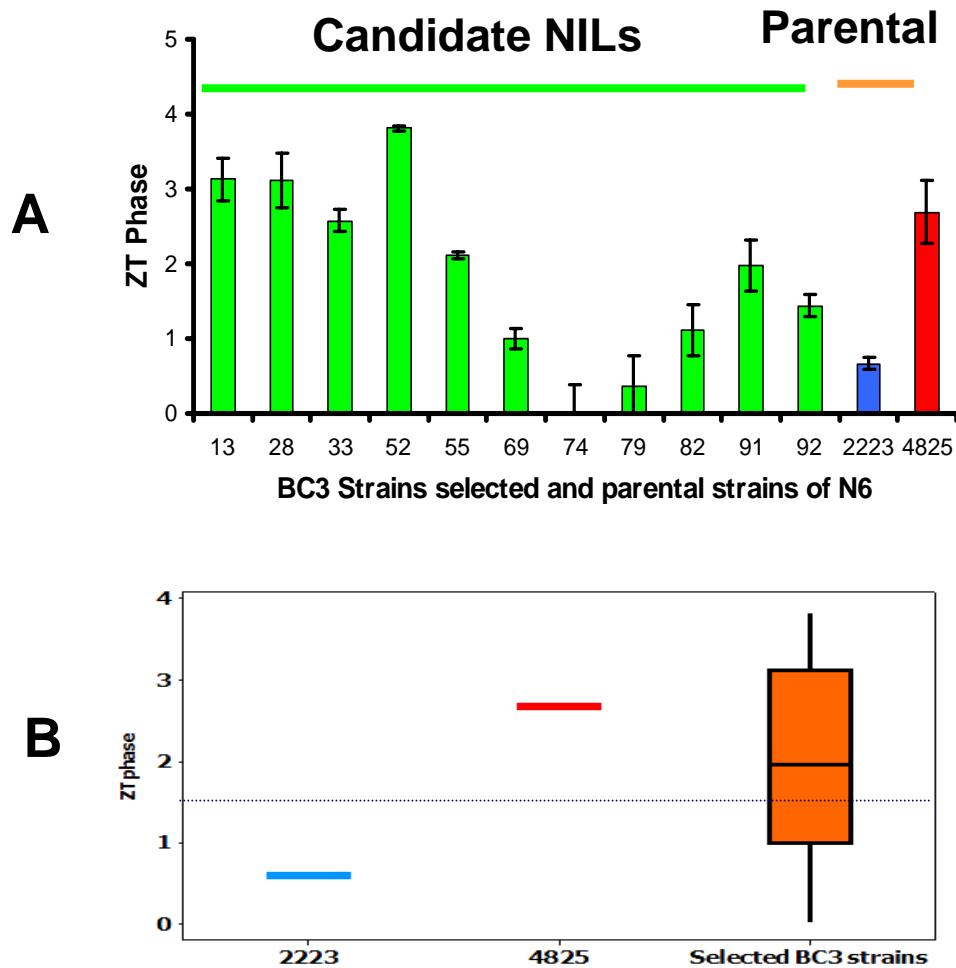
**Table 4-5. The steps required at each backcross (BC) generation in the NIL construction.**

Steps required at each backcross	Number of individuals that is expected to be selected as a result of each step in each BC <sup>a</sup>	Result of strain selections at each step in BC1 for NIL constructions in targeted QTLs			
		N6CBper7-1	N6Cbper6	N4Cbpha5	N4Cbper5
Spore picking	200	168	158	156	156
Mating type determination	100	94	93	82	69
Genotyping	30	29	31	32	16
The overall proportion selected for the next BC	15%	17.3%	19.6%	20.5%	10.2

<sup>a</sup>, The values of the column are estimated from the mean values derived from observed values among the targeted QTLs

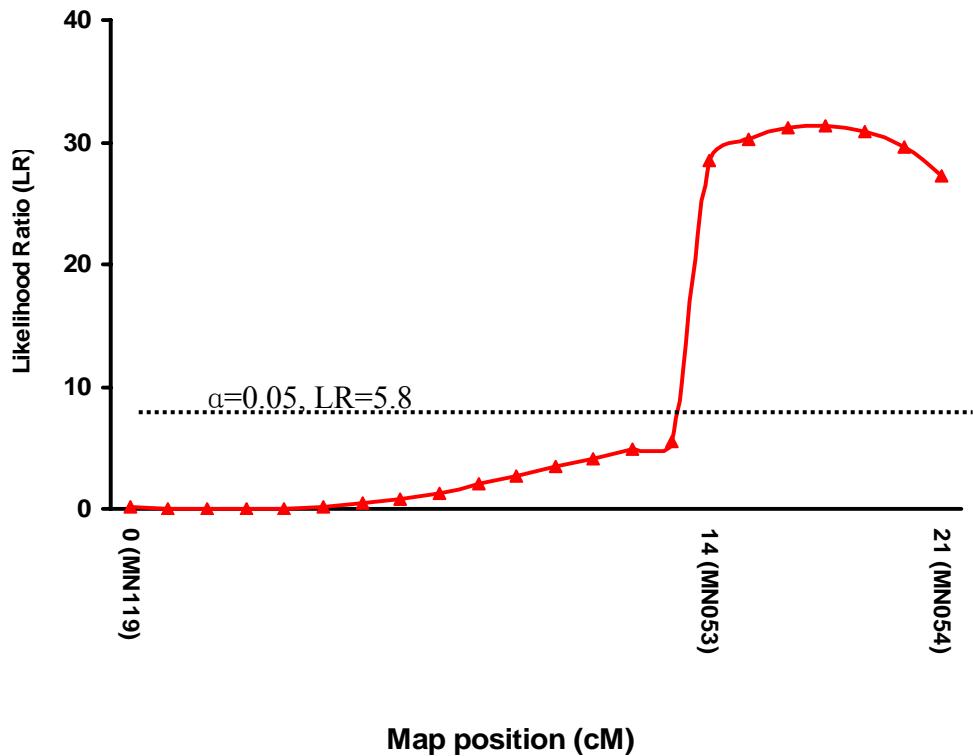
We did not perform the phenotyping until BC 1<sup>st</sup> and 2<sup>nd</sup> generations since we wanted to evaluate the QTL effect at more cleaned genetic background, which is more objective way to test QTL effect. Thus, the phenotype analysis for the QTL effect was evaluated at BC 3<sup>rd</sup> generation in which about 85% of the genome is purified.

Figure 4-17 shows the phenotypic analysis to evaluate the QTL effect underlying N6CBpha6. The QTL effect originates from the donor parent (FGSC 4825) was a delayed phase. Thus, if the QTL effect predicted at F1 were real, we should observe the expected phenotype at most of BC3F1 individuals with the targeted allele. To test the hypothesis, we selected 13 individuals of BC3F1 which have a desirable allele from FGSC 4825 and subsequently compared the phenotype with recurrent (FGSC 2223) and donor (FGSC 4825) parents (Figure 4-17). The result showed that most of the individuals selected showed more delayed phases than the recurrent parent, FGSC2223 (Figure 4-17A) as expected. The median value of the phase of the selected individuals was ZT 2.3 which is delayed by 1.5ZT comparing to the phase of the recurrent parent (FGSC 2223) (Figure 4-17B). This result strongly suggested that the QTL effect of N6Cbpha6 predicted at F1 was real. We further characterized the QTL effect of N6Cbpha6 at backcross 4<sup>th</sup> generation (BC4F1); CIM analysis was performed in BC4F1 population to confirm the phenotypic effect which is observed in the selected BC3F1 individuals. We detected highly significant QTL effect (LR 30 with R-Square=0.23) at the expected genetic position (MN053) for the phase phenotype, which confirmed the QTL effect of N6Cbpha6.



**Figure 4-17. Comparison of phase phenotype between 13 individuals of BC4F1 with a targeted allele and those of recurrent and donor parents.**

A. Comparison of phase phenotype between 13 individuals of BC3F1 with a targeted allele and those of recurrent and donor parents. x-axis represents the strain ID tested and y axis stands for the phase value of a strain tested. Green bars are the measurements of phase in the 13 BC3F1 individuals selected which is called candidate NILs in this graph. Red and Blue bars represent the phase measurement of the recurrent (FGSC 2223) and donor (FGSC 4825) parents. B. Box plot analysis that shows range of phase value from the selected strains. The median value is denoted by the black line in the box with light brown color.



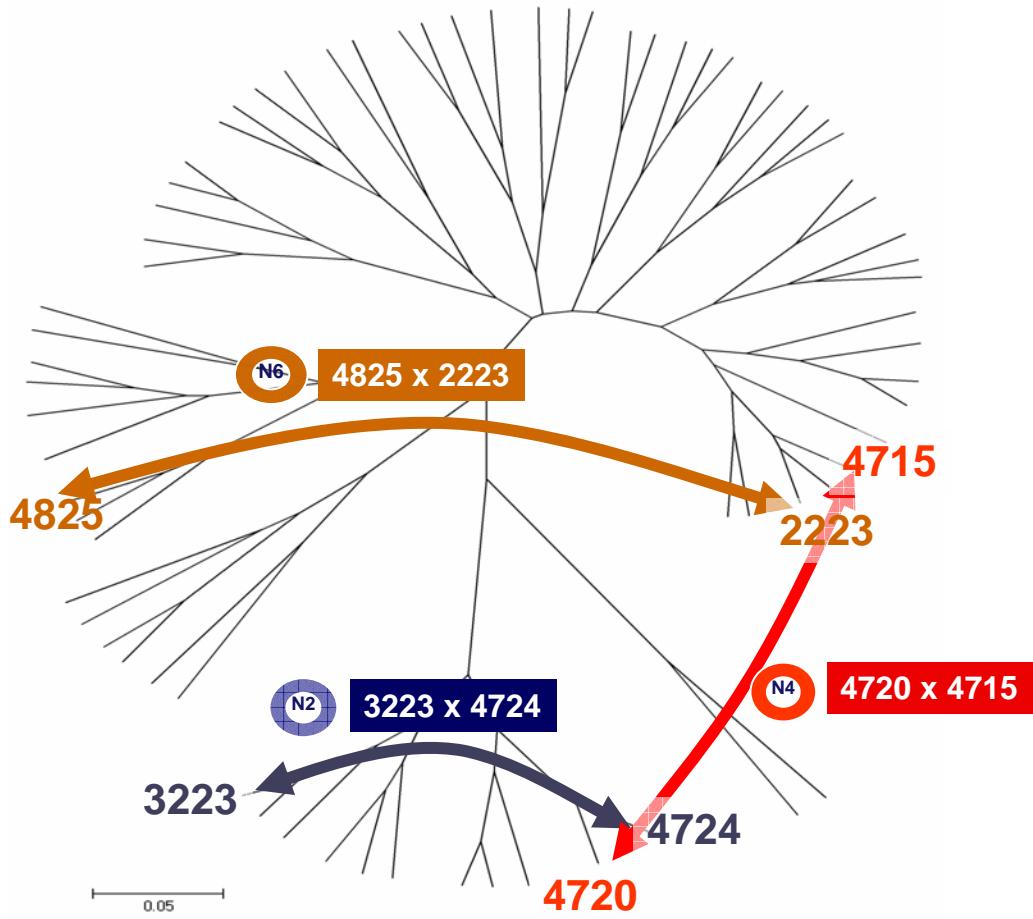
**Figure 4-18. CIM analysis to confirm the QTL effect of N6Cbpha6 at BC4F1**  
The highly significant QTL at consistent position (MN053) as F1 were detected., which confirm the QTL effect. The threshold level was calculated from 1000 permutation test

### **Discussion; Phenotypic variations of circadian rhythm in F1 mapping population.**

Our study showed that the fungal F1 population can be employed as a mapping population for the QTL study in circadian clock phenotypes by confirming the QTL effect (Figure 4-7 and Figure 4-18). From the three independent F1 populations in our study (N2, N4 and N6), a wide range phenotypic transgressive segregation were observed in both free running period and light entrained phase clock phenotypes. Since those phenotypes shows high heritability consistently in the line-cross populations (average  $H^2 = 0.79$ , standard deviation=0.12) and the genome structure of two parental strains are so divergent (Figure 4-19), the phenotypic transgressive segregations of the phenotypes that are observed in those populations were presumably attributable to segregation of the accumulated genetic variations to a specific environment (Figure 4-1) between the parental strains.

### **Discussion; Advantage of haploid organism in QTL analysis**

Haploid organism have an important advantage in constructing mapping population; due to the haploid nature of genome organization, one generation (F1) is enough to make an immortal and true breeding population similar to that of recombinant inbred line (RIL), where it takes at least 8-9 generations of selfing in plant species or about 20 generation of full-sib mating for out- breeding animals. The important consideration of the employment of F1 population is the size of population. Due to the relatively less number of meiotic events compared to RIL, a small number of progeny can cause errors in estimating genetic distance and order. HACKETT and BROADFOOT (2002) performed simulation studies to give a reasonable guess for the mapping population size[57].



**Figure 4-19. Phylogenetic analysis of *Neurospora* natural strains collected from diverse geographic region.**

This phylogenetic inference originates from unrooted tree with the neighbor joining method based on the polymorphism of 17 randomly chosen SNP loci. The parental stains are positioned in the phylogenetic three. The combination of mapping parents are denoted by same color and also connected with arrows with same color.

They investigated locus ordering performance in genetic linkage map construction of double haploid (DH) population under the conditions where effects of missing values, typing errors and distorted segregation are allowed [57]. They concluded that with 150 DH progenies, locus order spacing 10cM is relatively robust to missing values and typing errors, even in the case of the combination of 3% typing errors and 20% missing values. In accordance with their result, the order of mapped loci in our study is quite consistent with the physical map (Figure 4-14). Furthermore, significant QTLs associated with period and phase variation were detected in the given resolution (Table 4-3).

### **Discussion; Comparisons of two different QTL methods (CIM vs BMQ)**

The statistical power to detect meaningful QTLs can be determined by many factors including the number of individuals in genome organization of the target organism, the experimental designs for the mapping population, the types of molecular marker, the qualities of phenotyping and genotyping analysis and method of statistical analysis [43, 58, 59]. Thus, it is important to find an efficient statistical method so as to detect meaningful QTLs more sensitively in a given experimental conditions. We compared the result of the QTLs analysis with the two different statistical methods, CIM [58] and BMQ [43]. We detected twice the number of QTL using BMQ analysis compared to CIM (20 QTLs from CIM vs. 41 QTLs from BMQ) (Table 3, Figure 4-10A). And also, there is highly significant positive correlation between significance levels of the two methods (Figure 4-10D), which suggests BMQ could detect a QTL much sensitively for circadian clock phenotype. CIM has been used in quantitative genetics studies in other research areas [58, 60, 61]. While CIM incorporates additional markers into the regression analysis that can, in theory, account for the effects of other QTL, the method does not necessarily remove the confounding effects

of these other QTL. The Bayesian approach utilized in this study directly fits a multi-QTL model using a hierarchical variable selection approach that avoids many of the difficulties associated with model selection in likelihood based approaches [43, 62, 63]. By directly modeling how multiple QTL contribute to phenotype variation, BMQ is expected to be able to identify true QTLs, particularly those with more subtle effects. In our analysis, 22 additional clock QTLs were detected by BMQ.

### **Discussion; Important considerations in the application of QTL approach to *Neurospora* circadian clock study**

Numerous study have demonstrated that QTL analysis is one of the most powerful ways to find naturally conserved intact genetic variation affecting traits [64] but the method has a major limitation; QTLs detected in one cross are limited to the different alleles fixed in parental strain [65]. Thus, regardless of the amount of divergence between parents, those QTLs detected in the cross may be one snapshot of the total variation [14]. Therefore to overcome this problem, we increased the number of populations and derived each population by crossing different accessions adapted in different regions so as to widen our scope in searching various genetic loci that potentially contribute to the circadian clock traits [65, 66]. Our study found 43 QTLs affecting the two *N. crassa* circadian clock phenotypes of the period length and the entrain phase. As expected, QTLs of both phenotypes in our study showed population specific patterns, suggesting that those divergent mapping parents have accumulated genetic variation at independent loci. Thus, similar trait values in circadian properties among mapping parents observed in Table 4-1 and originate from different genetic variation of different loci that were built as a result of distinct evolutionary history. Besides the population specific QTLs, common QTLs affecting period (3 QTLs) or phase (8 QTLs) variation were detected from our mapping populations. Thus, at least

thirteen and nineteen different QTLs could be involved in the determination of period and phase respectively in *N. crassa* natural population (Table 4-3). Cloning and characterizing those common QTLs may reveal the molecular nature of clock variation in nature.

### **Discussion; Neurospora clock QTLs**

From QTLs affecting period length, some QTLs co-localized with previously characterized clock genes, which includes the catalytic subunit (*cka*) (*N4CBper1-2*) of casein kinase II (*ckII*), frequency (*frq*) (*N6Cper7-2*) and *fwd-1* (*N6CBper7-1*) (Table 4-3). This result suggests that our QTL study agrees with a previous clock study in period, where progressive phosphorylation of FRQ (by *cka*), FRQ degradation (by *fwd-1*) are suggested as major determinants of period length of *N. crassa* circadian clock [11].

A total of 27 QTLs affecting the phase variation were detected from the all three population (N2, N4 and N6). As observed in period QTLs, QTLs affecting phase determinant underlie numerous known clock genes including *ckba* (*N6CBpha1-3*), *prd-4* (*N4Bpha1-2*, *N6Bpha1-2*), *pp2a* (*N2Bpha4*, *N4Bpha4-2*) and *vvd* (*N2Bpha6* and *N6Bpha6*) (Table 3). However, the roles of these candidates gene are undefined currently except *vvd* [56]. Thus, characterizations of the role of these candidate genes in the phase determination will provide valuable insights into the regulation of this phenotype.

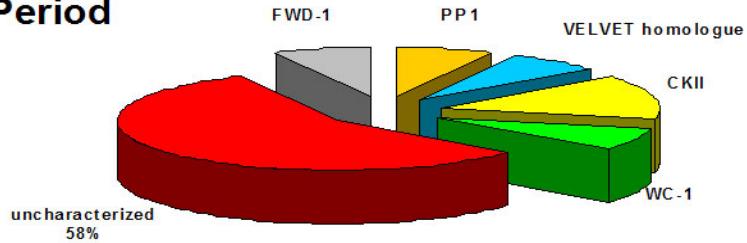
One of the interesting findings in our study was the phase QTL linked to the marker MN129, which was detected both in N4 and N6 populations. Those QTLs were closely linked with *prd-4* as a candidate gene. One of the known roles in *prd-4* in circadian clock oscillation is enhancing FRQ phosphorylation in response to DNA-damaging agents, resulting in resetting the *N. crassa* clock [67]. Interestingly, the

mutants of *prd-4* failed to show an appropriate circadian phase shift in response to a light pulse [67, 68]. It is tempting to propose that *prd-4* plays a role in phase determination in light/dark cycling environment. In general, light information is one of the important environmental signals for fungi. However, light also could be a DNA-damaging agent. *prd-4* might play a role in determining the phase in such a way to avoid adversary photooxidative damage/stress in light phase, which may function as a DNA-damaging agent [68].

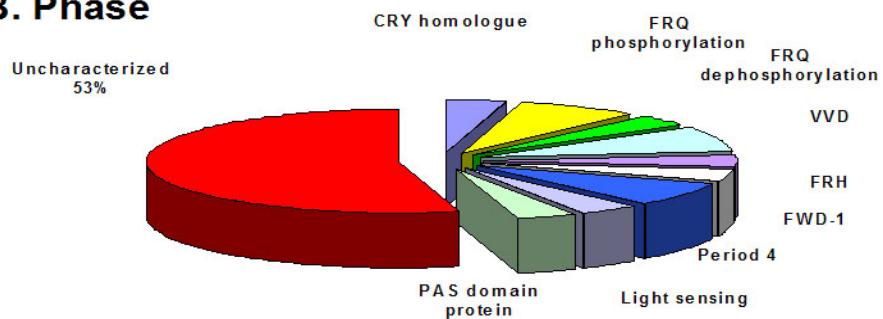
### **Discussion; Functional relationship between clock phenotypes (period vs phase)**

From the correlation analysis between period and phase, we found no evidence that there is significant correlation between period and phase in the three populations of our study (Pearson's correlation, p value in N2 =0.61, p value in N4=0.64, p value in N6=0.68, Figure 4-4). Consistent with this result, we found few common QTLs between the two phenotypes within a population (Figure 4-15). However, when we consider three populations, 7 QTLs were overlapped between period and phase phenotype (Figure 4-16). This suggests at least some pleiotropic effects for the regulation of phase and period. More in-depth study of those common QTLs may provide an important clue of how phase and period are functionally associated. Since 30 QTLs out of 43 (70%) are not linked to any previous characterized clock genes (Figure 4-20). This result strongly suggests that our current understanding of *N. crassa* circadian clock regulation is not complete. Further characterization of these novel 30 genomic regions will aid our understanding of *N. crassa* circadian clock regulation.

### A. Period



### B. Phase



**Figure 4-20. Graphical summary of our QTL study for *Neurospora* circadian clock phenotypes.**

The QTL regions which include known clock loci are partitioned and denoted with various colors. The proportion with novel QTL is denoted with red color.

## REFERENCES

1. Schibler U: Circadian time keeping: the daily ups and downs of genes, cells, and organisms. *Prog Brain Res* 2006, 153:271-282.
2. Stanewsky R: Genetic analysis of the circadian system in *Drosophila melanogaster* and mammals. *J Neurobiol* 2003, 54(1):111-147.
3. Young MW, Kay SA: Time zones: a comparative genetics of circadian clocks. *Nat Rev Genet* 2001, 2(9):702-715.
4. Bell-Pedersen D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, Zoran MJ: Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nat Rev Genet* 2005, 6(7):544-556.
5. Lakin-Thomas PL: Transcriptional feedback oscillators: maybe, maybe not. *J Biol Rhythms* 2006, 21(2):83-92.
6. Feldman JF, Hoyle MN: Isolation of circadian clock mutants of *Neurospora crassa*. *Genetics* 1973, 75(4):605-613.
7. Loros JJ, Dunlap JC: Genetic and molecular analysis of circadian rhythms in *Neurospora*. *Annu Rev Physiol* 2001, 63:757-794.
8. Dunlap JC, Loros JJ: The *Neurospora* circadian system. *J Biol Rhythms* 2004, 19(5):414-424.

9. Aronson BD, Johnson KA, Dunlap JC: Circadian clock locus frequency: protein encoded by a single open reading frame defines period length and temperature compensation. *Proc Natl Acad Sci U S A* 1994, 91(16):7683-7687.
10. Dunlap JC, Loros JJ: How fungi keep time: circadian system in *Neurospora* and other fungi. *Curr Opin Microbiol* 2006, 9(6):579-587.
11. Liu Y, Bell-Pedersen D: Circadian rhythms in *Neurospora crassa* and other filamentous fungi. *Eukaryot Cell* 2006, 5(8):1184-1193.
12. Bell-Pedersen D, Crosthwaite SK, Lakin-Thomas PL, Merrow M, Okland M: The *Neurospora* circadian clock: simple or complex? *Philos Trans R Soc Lond B Biol Sci* 2001, 356(1415):1697-1709.
13. He Q, Shu H, Cheng P, Chen S, Wang L, Liu Y: Light-independent phosphorylation of WHITE COLLAR-1 regulates its function in the *Neurospora* circadian negative feedback loop. *J Biol Chem* 2005, 280(17):17526-17532.
14. Mackay TF: The genetic architecture of quantitative traits. *Annu Rev Genet* 2001, 35:303-339.
15. Alonso-Blanco C, Koornneef M: Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci* 2000, 5(1):22-29.

16. Shimomura K, Low-Zeddies SS, King DP, Steeves TD, Whiteley A, Kushla J, Zemenides PD, Lin A, Vitaterna MH, Churchill GA *et al*: Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res* 2001, 11(6):959-980.
17. Darrah C, Taylor BL, Edwards KD, Brown PE, Hall A, McWatters HG: Analysis of phase of LUCIFERASE expression reveals novel circadian quantitative trait loci in *Arabidopsis*. *Plant Physiol* 2006, 140(4):1464-1474.
18. Edwards KD, Anderson PE, Hall A, Salathia NS, Locke JC, Lynn JR, Straume M, Smith JQ, Millar AJ: FLOWERING LOCUS C mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock. *Plant Cell* 2006, 18(3):639-650.
19. Edwards KD, Lynn JR, Gyula P, Nagy F, Millar AJ: Natural allelic variation in the temperature-compensation mechanisms of the *Arabidopsis thaliana* circadian clock. *Genetics* 2005, 170(1):387-400.
20. Michael TP, Salome PA, Yu HJ, Spencer TR, Sharp EL, McPeek MA, Alonso JM, Ecker JR, McClung CR: Enhanced fitness conferred by naturally occurring variation in the circadian clock. *Science* 2003, 302(5647):1049-1053.

21. Suzuki T, Ishikawa A, Yoshimura T, Namikawa T, Abe H, Honma S, Honma K, Ebihara S: Quantitative trait locus analysis of abnormal circadian period in CS mice. *Mamm Genome* 2001, 12(4):272-277.
22. Swarup K, Alonso-Blanco C, Lynn JR, Michaels SD, Amasino RM, Koornneef M, Millar AJ: Natural allelic variation identifies new genes in the *Arabidopsis* circadian system. *Plant J* 1999, 20(1):67-77.
23. Sargent ML, Woodward DO: Genetic determinants of circadian rhythmicity in *Neurospora*. *J Bacteriol* 1969, 97(2):861-866.
24. Sargent ML, Kaltenborn SH: Effects of Medium Composition and Carbon Dioxide on Circadian Conidiation in *Neurospora*. *Plant Physiol* 1972, 50(1):171-175.
25. Park S, Lee K: Inverted race tube assay for circadian clock studies of the *Neurospora* accessions. *Fungal Genet Newsl* 2004, 51:12-14.
26. Davis RH, Perkins DD: *Neurospora*: a model of model microbes. *Nat Rev Genet* 2002, 3(5):397-403.
27. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S *et al*: The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 2003, 422(6934):859-868.

28. Doerge RW: Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 2002, 3(1):43-52.
29. Turner BC, Perkins DD, Fairfield A: Neurospora from natural populations: a global study. *Fungal Genet Biol* 2001, 32(2):67-92.
30. Davis RH: **Neurospora, contributions of a model organism.** New York: Oxford University Press, Inc.; 2000.
31. Lee K, Dunlap JC, Loros JJ: Roles for WHITE COLLAR-1 in circadian and general photoperception in *Neurospora crassa*. *Genetics* 2003, 163(1):103-114.
32. Plautz JD, Straume M, Stanewsky R, Jamison CF, Brandes C, Dowse HB, Hall JC, Kay SA: Quantitative analysis of *Drosophila* period gene transcription in living animals. *J Biol Rhythms* 1997, 12(3):204-217.
33. Schuelke M: An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 2000, 18(2):233-234.
34. Yu J-K, Kantety RV, Graznak E, Benschoter D, Tefera H, Sorrells ME: A genetic linkage map for tef [*Eragrostis tef* (Zucc.) Trotter]. *Theor Appl Genet* 2006, 113(6):1093-1102.
35. Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S: Diversity of microsatellites derived from

- genomic libraries and GenBank sequences in rice (*Oryza sativa L.*).  
*Theoretical and Applied Genetics* 2000, 100(5):713-722.
36. Manly K, Olson J: Overview of QTL mapping software and introduction to map manager qtx. *Mamm Genome* 1999, 10:327-334.
37. Holloway JL, Knapp SJ: G-MENDEL 3.0 user guide. *Oregon State University, Corvallis, OR* 1993:1-130.
38. Kosambi D: The estimation of map distances from recombination values. *Annals of Eugenics* 1944, 12:172-175.
39. Schuelke M: An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 2002, 18(2):233-234.
40. Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S: Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa L.*). *Theor Appl Genet* 2000, 100(5):713-722.
41. Basten CJ, Weir BS, Zeng Z-B: Windows QTL Cartographer 2.5. *Department of Statistics, North Carolina State University, Raleigh, NC* (<http://statgenncsuedu/qtlcart/WQTLCarthtm>) 2006.
42. Churchill GA, Doerge RW: Empirical threshold values for quantitative trait mapping. *Genetics* 1994, 138(3):963-971.

43. Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D: Mapping multiple Quantitative Trait Loci by Bayesian classification. *Genetics* 2005, 169(4):2305-2318.
44. Yi N, Xu S, Allison DB: Bayesian model choice and search strategies for mapping interacting quantitative trait Loci. *Genetics* 2003, 165(2):867-883.
45. ter Braak CJ, Boer MP, Bink MC: Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* 2005, 170(3):1435-1438.
46. George EI, McCulloch RE: Variable selection via Gibbs sampling. *Amer Statist Assoc* 1993, 88:881-889.
47. Hofstetter JR, Mayeda AR, Possidente B, Nurnberger JI, Jr.: Quantitative trait loci (QTL) for circadian rhythms of locomotor activity in mice. *Behav Genet* 1995, 25(6):545-556.
48. Hofstetter JR, Possidente B, Mayeda AR: Provisional QTL for circadian period of wheel running in laboratory mice: quantitative genetics of period in RI mice. *Chronobiol Int* 1999, 16(3):269-279.
49. Kernek KL, Trofatter JA, Mayeda AR, Hofstetter JR: A locus for circadian period of locomotor activity on mouse proximal chromosome 3. *Chronobiol Int* 2004, 21(3):343-352.

50. Kopp C: Locomotor activity rhythm in inbred strains of mice: implications for behavioural studies. *Behav Brain Res* 2001, 125(1-2):93-96.
51. Mayeda AR, Hofstetter JR: A QTL for the genetic variance in free-running period and level of locomotor activity between inbred strains of mice. *Behav Genet* 1999, 29(3):171-176.
52. Salathia N, Edwards K, Millar AJ: QTL for timing: a natural diversity of clock genes. *Trends Genet* 2002, 18(3):115-118.
53. Toth LA, Williams RW: A quantitative genetic analysis of locomotor activity in CXB recombinant inbred mice. *Behav Genet* 1999, 29(5):319-328.
54. Welch SM, Dong ZS, Roe JL, Das S: Flowering time control: gene network modelling and the link to quantitative genetics. *Australian Journal of Agricultural Research* 2005, 56(9):919-936.
55. Elvin M, Loros JJ, Dunlap JC, Heintzen C: The PAS/LOV protein VIVID supports a rapidly dampened daytime oscillator that facilitates entrainment of the *Neurospora* circadian clock. *Genes Dev* 2005, 19(21):2593-2605.

56. Heintzen C, Loros JJ, Dunlap JC: The PAS protein VIVID defines a clock-associated feedback loop that represses light input, modulates gating, and regulates clock resetting. *Cell* 2001, 104(3):453-464.
57. Hackett CA, Broadfoot LB: Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* 2003, 90(1):33-38.
58. Zeng ZB: Precision mapping of quantitative trait loci. *Genetics* 1994, 136(4):1457-1468.
59. Zeng ZB, Kao CH, Basten CJ: Estimating the genetic architecture of quantitative traits. *Genet Res* 1999, 74(3):279-289.
60. Leips J, Gilligan P, Mackay TF: Quantitative trait loci with age-specific effects on fecundity in *Drosophila melanogaster*. *Genetics* 2006, 172(3):1595-1605.
61. Jordan KW, Morgan TJ, Mackay TF: Quantitative trait loci for locomotor behavior in *Drosophila melanogaster*. *Genetics* 2006, 174(1):271-284.
62. Kao CH, Zeng ZB, Teasdale RD: Multiple interval mapping for quantitative trait loci. *Genetics* 1999, 152(3):1203-1216.

63. Liao JG: A hierarchical Bayesian model for combining multiple 2 x 2 tables using conditional likelihoods. *Biometrics* 1999, 55(1):268-272.
64. Abiola O, Angel JM, Avner P, Bachmanov AA, Belknap JK, Bennett B, Blankenhorn EP, Blizzard DA, Bolivar V, Brockmann GA *et al*: The nature and identification of quantitative trait loci: a community's view. *Nat Rev Genet* 2003, 4(11):911-916.
65. Mackay TF: Quantitative trait loci in *Drosophila*. *Nat Rev Genet* 2001, 2(1):11-20.
66. Xie CQ, Gessler DDG, Xu SZ: Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics* 1998, 149(2):1139-1146.
67. Pregueiro AM, Liu Q, Baker CL, Dunlap JC, Loros JJ: The *Neurospora* checkpoint kinase 2: a regulatory link between the circadian and cell cycles. *Science* 2006, 313(5787):644-649.
68. Okamura H: Clock genes in cell clocks: roles, actions, and mysteries. *J Biol Rhythms* 2004, 19(5):388-399.