1-1-2012

# A Study of Selection on Microsatellites in the Helianthus Annuus Transcriptome

Sreepriya Pramod

# A STUDY OF SELECTION ON MICROSATELLITES IN THE *HELIANTHUS ANNUUS* TRANSCRIPTOME

By

Sreepriya Pramod

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Biological Sciences
in the Department of Biological Sciences

Mississippi State, Mississippi

May 2012

A STUDY OF SELECTION ON MICROSATELLITES IN THE *HELIANTHUS*

*ANNUUS* TRANSCRIPTOME

By

Sreepriya Pramod

Approved:

| | |
|---|---|
| Mark E. Welch<br>Assistant Professor of Biological Sciences<br>Committee Chair | Lisa E. Wallace<br>Assistant Professor of Biological Sciences<br>Committee Member |
| Andy D. Perkins<br>Assistant Professor of Computer Science And Engineering<br>Committee Member | Brian A. Counterman<br>Assistant Professor of Biological Sciences<br>Committee Member |
| David J. Chevalier<br>Assistant Professor of Biological Sciences<br>Committee Member | |
| Gary N. Ervin<br>Professor of Biological Sciences<br>Graduate Coordinator | Gary L. Myers<br>Professor and Dean<br>College of Arts & Sciences |

Name: Sreepriya Pramod

Date of Degree: May 12, 2012

Institution: Mississippi State University

Major Field: Biological Sciences

Major Professor: Mark E. Welch

Title of Study: A STUDY OF SELECTION ON MICROSATELLITES IN THE *HELIANTHUS ANNUUS* TRANSCRIPTOME

Pages in Study: 129

Candidate for Degree of Doctor of Philosophy

The ability of populations to continually respond to directional selection even after many generations instead of reaching response plateaus suggests the presence of mechanisms for rapidly generating novel adaptive variation within organismal genomes. The contributions of cis regulation are now being widely studied. This study details the contributions of one such mechanism capable of generating adaptive genetic variation through transcribed microsatellite mutation. Microsatellites are abundant in eukaryotic genomes, exhibit one of the highest known mutation rates; and mutations involve indels that are reversible. These features make them excellent candidates for generating variation in populations. This study explores the functional roles of transcribed microsatellites in *Helianthus annuus* (common sunflower). More specifically, I explored the role of microsatellites as agents of rapid change that act as "tuning knobs" of phenotypic variation by influencing gene expression in a stepwise manner by expansions and contractions of the microsatellite tract.

A bioinformatic study suggests that selection has favored expansion and maintenance of transcriptomic microsatellites. This inference is based on the non-random

distribution of microsatellites, prevalence of motifs associated with gene regulation in untranslated regions, and the enrichment of microsatellites in Gene Ontologies representing plant response to stress and stimulus. A population genetics study provides support for selection on these transcribed microsatellites when compared to anonymous microsatellites that were assumed to evolve neutrally. The natural populations utilized in this study show greater similarity in allele frequencies, mean length, and variance in lengths at the transcribed microsatellites relative to that observed at anonymous microsatellite loci. This finding is indicative of balancing selection, and provides evidence that allele lengths are under selection. This finding provides support for the tuning knob hypothesis. The findings of a functional genomic study with regard to the tuning knob hypothesis are ambiguous. No correlation between allele lengths and gene expression was detected at any of three loci investigated. However, the loci utilized exhibited narrow ranges in length. The tuning knob hypothesis implies that similar allele lengths are likely to exhibit similar gene expression levels. Hence, variation in the populations studied may be tracking the optimal gene expression levels.

# DEDICATION

This dissertation is dedicated to my husband Punith Naik, without whom it would not been possible. Your love and support in these years made all the difference!

ACKNOWLEDGEMENTS

I would like start by expressing my sincerest gratitude to my major advisor Dr. Welch, who has not only trained and guided me towards the successful completion of my doctoral dissertation, but also has helped develop me into the person I am now; a person much more confident in her abilities and strengths. Much appreciated are his patience in training me with population genetics and his willingness to help even if it meant sacrificing personal time. The encouragement and support I received from him, while I explored ideas and projects that were outside his comfort zone make him the best possible mentor anyone could hope for.

My dissertation committee members Dr. Wallace, Dr. Perkins, Dr. Chevalier and, Dr. Counterman, were always helpful and forthcoming when needed. I have learnt a lot from each one of you. I am also thankful to Dr. Bridges, whose classes inspired me to delve further into bioinformatics and programming. These classes helped me decide the course of my doctoral dissertation work as well as my career. I am also thankful to all the wonderful undergraduate researchers who helped with data collection; Armed Rasberry, Tiffini Butler, Jessica Martin, and Katie Downs.

I would like to thank Chris Varghese for his friendship and help during these years. My friends in the department over the years, Leah Chinchilla, Kristen Sauby, Jamen Berk, Ben Shepard, Lavanya Challagundla, Chris Doffitt, Giuliano Colosimo, Hanna Dorman, and Steve Shaak, have been of great support to me. I can't thank you all enough for you love and kindness. I am also indebted to my best friends, Aditi Bhan,

Ragini Raj Chitranshi, and Garima Shivani Talwar, for all good times during my undergraduate years in Birla Institute of Technology and for being there when I needed them.

I am also indebted to my parents Mr. Pramod Kannadath and Mrs. Jaya Pramod, and my younger brother, Dr. Bijun Kannadath, for their unconditional love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

Rates of adaptive evolution can be rapid. The classic explanation for this observation is that populations maintain a large amount of genetic variation mediating rapid change. However, it has become increasingly clear through studies of experimental evolution, and even short-term ecological studies that presumably inbred lineages with little genetic variation can still respond quickly, and even dramatically, to selective pressures (Carroll et al. 2007, Strauss et al. 2008). The role of molecular mechanisms that serve as sources of novel adaptive genetic and phenotypic variation are not well understood. In fact, some of the major questions that the scientific community strives to seek answers for include, the relative importance of cis-regulatory changes and novel structural changes in proteins (Wittkopp et al. 2007, Stern and Orgogazo 2008, Fay and Wittkopp 2008), the functional role of repetitive sequences as well as unique DNA polymorphisms in generating heritable genetic variation (Britten and Davidson 1969, Britten and Davidson 1971, Gemayel et al. 2010), etc. This dissertation aims to improve our understanding of the contributions of one source of variation, microsatellite polymorphisms, in generating novel phenotypes that could potentially be adaptive.

Microsatellites are sequences found in abundance in eukaryotic genomes (Tautz and Renz 1984). A microsatellite sequence is composed of numerous front to back repetitions of a simple motif usually 1-6 base pair (bp) in length. Alleles at a microsatellite locus typically tend to differ in length. These length differences are due to

differences in copy number, resulting from expansion or contraction of a tract (Hancock 1999, Ellegren 2000). Microsatellite mutations are thought to result from slipped strand mispairing during DNA replication. As per this mechanism, DNA polymerase slips on a microsatellite tract during replication, resulting in loop formation on the microsatellite tract followed by mispaired alignment of the stands that either increases or decreases the length of the strand (Levinson and Gutmann 1987). The mutation rates observed in these sequences are much higher than that for any other source for novel polymorphism (Gemayel et al. 2010) and range between 10-5 to 10-2 mutations per base pair per generation (Weber and Wong 1993, Wang et al. 1994). The high mutation rates and the role of trinucleotide microsatellites located within genes associated with hereditary neurodegenerative disorders in humans resulted in the widespread belief that microsatellites are not likely to be prevalent with genes (Ellegren 2004). However, as genome and transcriptome sequence databases became available, it became clear that microsatellites are much more abundant in the transcriptomes than had been assumed. Microsatellites are present in 10% of the protein coding regions or open reading frames (ORFs) of primates, and occur at similar frequencies in the genomes of rabbits (15%), pigs (9.1%), chickens (10.6%), cereals (1.5-7.5%) such as maize, wheat, barley, sorghum and rice (Li et al. 2002; Li et al. 2004) and in 13.6% of transcripts in the model dicot *Arabidopsis thaliana* (Gemayel et al. 2010). These high prevalence rates in eukaryotic genes along with the mutational dynamics of microsatellites that include the reversible nature of these mutations make them potential source of adaptive variation. Several instances where microsatellite variation was shown to have a marked effect on phenotype have been identified. This is achieved either by altering amino acid composition of proteins or by modulating gene expression variation. Comprehensive lists of these

anecdotes have been compiled in multiple reviews (Li et al. 2002, Kashi and King 2006, Gemayel et al. 2010). The growing body of scientific literature, as well as the unique features of microsatellites, has led many researchers to explore the idea of microsatellites playing a role as "tuning knobs" of phenotypic variation by either modulating gene expression or protein function in a stepwise manner (Trifanov 2004, Kashi and King 2006, Gemayel et al. 2010). In this model, increases and decreases in repeat number on the microsatellite tract that serve to adjust the phenotype are likened to a tuning knob. Adjusting the length of microsatellites is anticipated to move organismal phenotypes toward or away optimal phenotypic values.

The goal of this dissertation is to test the hypothesis that transcribed microsatellites play a functional role in adaptive evolutionary processes, by using *Helianthus annuus* (common sunflower) as a model system. Bioinformatic, population genetic and functional genomic approaches are utilized to address this hypothesis. The bioinformatics study detailed in Chapter II looks at the distribution of microsatellites in the transcriptome of *H. annuus*, and identifies if the distribution is consistent with a neutral model of microsatellite evolution or if it is suggestive of selection. The key findings from these analyses include a non-random distribution of microsatellites across domains of transcripts, prevalence of motifs associated with gene regulation in untranslated regions (UTRs), and the enrichment of microsatellites in gene ontologies (GO) associated with plant response to stress and environmental stimuli. The selective enrichment of microsatellites within these GOs supports the role of microsatellites as per predictions of the "tuning knob" model since these genes may require frequent modulatory changes to track plant responses to the environment. The nonrandom distribution of microsatellites, and their enrichment within genes in plant response

pathways, suggests that transcribed microsatellites could play an important role in aiding adaptive evolution

The bioinformatics study also identified transcribed microsatellites that were outliers for length and percentage of impurities or proportion of interruptions harbored by the tract. These outlier microsatellites were chosen to develop long transcribed molecular markers for a population genetics study, since the pattern of length and impurities observed in these microsatellites suggested either the role of selective forces in maintaining these elongated tracts or relaxed constraints on these tracts. The marker development process and the markers are described in Chapter III. These microsatellites were the subject of a population genetics study detailed in Chapter IV. The population structure at the transcribed microsatellites was compared to that at anonymous microsatellites. This study provides evidence that these transcribed microsatellites are under different selection pressures compared to the anonymous microsatellites. The study reveals that transcribed microsatellites exhibit similar allele frequencies as well as a tighter regulation of variance in lengths within the populations. This indicates the role of balancing selection and also provides a potential signal that allele lengths are under greater selection. The maintenance of similar allele lengths at the transcribed microsatellites is consistent with expectations of the tuning knob model.

The final study was designed to determine if microsatellite allele lengths influence gene expression in natural populations. A functional genomics approach was utilized. Gene expression was quantified at three well characterized loci for which conditions for genotyping were previously established (Chapman et al. 2008). The development and characterization of the gene expression assays, including two assays for constitutively expressed genes designed for standardization purposes, is described in

Chapter V. These three targeted genes harbor the microsatellites in the UTRs. These tools were utilized to study the effect of microsatellite length on gene expression in natural populations of *H. annuus* from Kansas. This study is detailed in Chapter VI. A significant correlation between gene expression levels and microsatellite allele was not detected at any of the three loci, nor was a significant effect observed when allele length was treated as a continuous variable. Hence, evidence that transcribed microsatellites modulate gene expression is lacking. However, these results were not entirely inconsistent with these microsatellites playing a functional role either. This is especially true given the findings of the population genetics study that indicated that balancing selection is maintaining similar allele lengths across populations. Hence, both results are consistent with selection favoring optimal allele lengths.

## References


Britten RJ and Davidson EH. 1969. Gene regulation for higher cells: A theory. *Science* 165:349-358.

Britten RJ and Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origin of evolutionary novelty. *The Quarterly Review of Biology* 46(2): 111-138.

Carroll SP, Hendry AP, Reznick DN, and Fox CW. 2007. Evolution on ecological time scales. *Functional Ecology* 21:387-393.

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ and Burke JM. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *The Plant Cell* 20: 2931–2945.

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* 16(12):551-558.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5:435-445.

Fay JC and Wittkopp PJ. 2008. Evaluating the role of natural selection in evolution of gene regulation. *Heredity* 100:191-199.

Gatchel JR and Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics* 6:743–55.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* 44:445-477.

Hancock JM. 1999. Microsatellites and other simple sequences: genomic context and mutational mechanisms. In "*Microsatellites: Evolution and Applications*" (D.B. Goldstein & C. Schlötterer, eds) (Oxford University Press) pp. 1-9.

Kashi Y and King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* 22(5): 253-259.

Levinson G and Gutman GA.1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4 (3): 203–221.

Li YC, Korol AB, Fahima T, Beiles A, and Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* 11: 2453-2465.

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: structure, function and evolution. *Molecular Biology and Evolution* 21(6):991-1007.

Stern DL and Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62(9): 2155–2177.

Strauss SY, Lau JA, Schoener TW, and Tiffin P. 2008. Evolution in ecological field experiments: implications for effect size. *Ecology Letters* 11:199-207.

Tautz D and Renz M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* 12: 4127-4138.

Trifanov EN. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation; In "*Evolutionary theory and processes: Modern Horizons, papers in honor of Eviatar Nevo"* (Ed. Wasser SP); Kluwer Academic Publishers. pp 115-138.

Wang Z, Weber JL, Zhong G, Tanksley SD. 1994. Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* 88:1-6.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Human Molecular Genetics* 2:1123-1128.

Wittkopp PJ. 2007. Variable gene expression in eukaryotes: a network perspective. *The Journal of Experimental Biology* 210:1567-1575.

# CHAPTER II

# PATTERNS OF MICROSATELLITE EVOLUTION INFERRED FROM THE *HELIANTHUS ANNUUS* TRANSCRIPTOME


## Abstract

The transcriptomic database of Helianthus annuus (common sunflower) was analyzed to determine if patterns of microsatellite distribution, length and impurities are consistent with proposed mechanisms by which microsatellites might influence phenotypic evolution. Although the coding regions harbor numerically more microsatellites compared to the untranslated regions (UTRs); the density of microsatellites in UTRs is greater than that in coding regions, with the greatest density found in the 5'UTRs. This reflects either relaxed constraints or directional selection on microsatellites in 5'UTRs; suggesting microsatellites more readily evolve to regulate gene expression rather than bringing about structural changes in proteins. Motifs overrepresented in the 5'and 3'UTRs included AG, AAG, AT and AAT. These classes of microsatellites are associated with gene expression changes via translational and transcriptional modification in 5'UTRs, as well as by influencing mRNA stability in 3'UTRs. The total absence of CG dinucleotide repeats, and other CG containing motifs, suggests that modulating levels of CpG island methylation is not a common mechanism for microsatellite mediated transcription regulation in *H. annuus*. Gene Ontologies associated with response to various abiotic and biotic stimuli including water deprivation and heavy metal tolerance were significantly enriched for microsatellites. This finding is

consistent with predictions of the "tuning knob" model since these are the most likely candidates for regulating plant responses to the environment. The nonrandom distribution of microsatellites, and their enrichment within genes in plant response pathways, suggests that transcribed microsatellites are sequences that are under strong selection and hence could play an important role in aiding adaptive evolution.

Keywords: *Helianthus annuus*, microsatellite, evolution, transcriptome, selection, untranslated region

## Introduction

Microsatellites are highly polymorphic regions in the genome, composed of tandem repeats of 1-6 base pair (bp) long sequences. Different alleles at a locus are defined by length or repeat number (Hancock 1999). It is generally believed that the repetitive nature of microsatellites results in higher incidences of polymerase slippage during replication making microsatellites particularly susceptible to mutation (Levinson and Gutman 1987). The results of microsatellite mutations are typically alleles with incremental increases or decreases in the number of tandem repeats (Ellegren 2000; Hancock 1999). The mutation rates of microsatellites, while variable, are unusually high, ranging from 10-5 to 10-2 per locus per generation (Weber and Wong 1993; Ellegren 2000). For comparison, estimations of substitution rates for the nuclear genomes of multicellular organisms vary from 10-8 to 10-10 per base per generation (Nachman and Crowell 2000; Baer et al. 2007), and recently estimates from mutation accumulation experiments in *Arabidopsis thaliana* suggests that plant nuclear genomes mutate at a rate between 10-8 and 10-9 per base per generation (Ossowski et al. 2010).

Until the late 1990's it was widely believed that microsatellites were largely restricted to non-coding regions of the genome. The high mutation rate was seen as a potentially disruptive force within genes since; the few microsatellites known to be located within exons of the genome were associated with hereditary neurodegenerative diseases such as Huntington's disease and Fragile-X syndrome in human populations (Gatchel and Zoghbi 2005). Availability of Expressed Sequence Tag (EST) and whole genome sequence databases of organisms revealed that microsatellites are present in as much as 10-20% of all examined Open reading frames (Li et al. 2002). This prevalence of microsatellites in organismal transcriptomes suggests that they may play a far more active role in generating functional genetic variation than was indicated by the isolated cases of microsatellite instability diseases. Moreover, microsatellites in plants are preferentially associated with unique DNA sequences and not with repetitive DNA (Morgante et al. 2002). The higher than anticipated prevalence rates along with the benefits conferred by high mutation rates and the reversible nature of microsatellite mutations, makes these genetic elements a source of genetic variation that can be rapidly generated. Although, traditionally viewed as neutrally evolving sequences of no value to organismal genomes, a body of literature has explored the possibility of functional beneficial roles of microsatellites (Trifanov 2003; Li et al. 2004; King et al. 2006; Gemayel et al. 2010). Reviews by Li et al. (2004), Kashi and King (2006) and Gemayel et al. (2010), detail studies that suggest a possible beneficial role for exonic microsatellites.

In this study, hypotheses regarding the role of microsatellites in adaptive evolution are considered through a database survey of their frequency and distribution within the transcriptome represented by the UniGene database of the common sunflower,

*Helianthus annuus*. Observed patterns of microsatellite distribution are utilized as proximate estimates either indicating a possible beneficial role for transcribed microsatellites in *H. annuus* or indicating the role of neutral processes in shaping evolution of microsatellites. Further, we utilized appropriate measures to circumvent ascertainment bias resulting from sequencing differences associated with EST sequences that compose the UniGene database of *H. annuus*. Our mechanistic hypotheses designed to explain potential roles that transcribed microsatellites might play, fall into two general categories; modulation of gene regulation, and alteration of protein structure. The assumption being made is that the relative abundance of different microsatellite classes in the transcriptome reflects the probability that specific mechanisms for altering gene regulation, or protein function via microsatellite expansion and contraction are more likely to evolve, or offset constraints limiting their genesis.

Our mechanistic hypotheses regarding the role of microsatellites in gene regulation are founded on the understanding that the composition of UTRs can alter transcription, translation, mRNA transport, mobility and stability. The 5'UTR may play a role in modulating transcription as well as mRNA translation and 3'UTRs may regulate mRNA stability (Mignone et al. 2002; Yin and Blanchard 2000). Here we identify as well as quantify classes of microsatellites in the 5'UTR and 3'UTR of the *H. annuus* transcriptome to make predictions about the role of these untranslated microsatellites based on well documented empirical studies conducted by other researchers, such as the association of CUG repeats in the 5'UTR in modulating translation efficiency (Timchenko et al. 1999; Raca et al. 2000), CG microsatellites in 5'UTR with transcription silencing via CpG methylation (Yin and Blanchard 2000), and AU rich microsatellites in 3'UTR with modulation of mRNA stability (Mignone et al. 2002).

Hence, similar microsatellites within certain regions of the gene in the *H. annuus* transcriptome are identified as potential candidates for fine scale regulation.

The second mechanistic hypothesis regarding translated microsatellites is based on studies by some researchers who have suggested that translated microsatellites or homopolypeptide regions may serve as flexible chains between structural domains that could enable modification of protein-protein interaction or binding (Faux et al. 2005; Karlin and Burge 1996). Marcotte et al. (1999) observed an increased presence of homopolypeptide stretches encoded by microsatellites in proteins of eukaryotes as compared to prokaryotes, leading them to suggest that microsatellites could have played a role in the evolution of eukaryotes. Several other studies have also suggested similar evolutionarily advantageous roles for homopolypeptides (Gerber et al. 1994; Mularoni et al. 2007; Shimohata et al. 2000). Further, we report Gene Ontology (GO) classes within which enrichment of coding microsatellites is observed in the *H. annuus* transcriptome. This approach would help make predictions regarding gene classes which could potentially be evolving faster than gene classes not harboring microsatellites.

To identify the relative importance of microsatellites in gene regulation to that of structural changes in the protein, we have identified differences in prevalence of microsatellites between the UTRs and coding regions. Further, correlations between region and microsatellite length as well as percentage impurities harbored by the tract are quantified. This approach also tests the hypothesis that constraints against evolution of microsatellites are stronger in the coding region.

Further, we report GO classes where significant clustering of microsatellites is observed. Differences as well as similarities observed in gene classes enriched with microsatellites between the three exonic regions are utilized to make inference regarding

pattern of microsatellite evolution in *H. annuus*. For example, similarities in enrichment across exonic regions are inferred to suggest that only particular GOs or genes are likely to evolve and maintain microsatellites. At the same time, differences between exonic regions in gene classes that are enriched with microsatellites are inferred to suggest that different mechanisms are important for different GOs. For example, the role of 5'UTR microsatellites in translation regulation is well known; hence we would expect clustering of 5'UTR microsatellites in genes associated with regulation or in trans acting genes that are involved in 5'UTR binding (Legendre et al. 2007). Only a handful of empirical studies have reported functional roles for 3'UTR and coding region microsatellites (Fondon and Garner 2004; Rattenbacher et al. 2010; Lee et al. 2010; Li et al. 2004; Wheeler et al. 2009). Hence we have not yet seen adequate empirical data to make predictions regarding enrichment of microsatellite containing genes within any particular GO. This study provides insight regarding probable targets of evolutionary change from the observed clustering of microsatellites within specific domains of transcripts. This bioinformatics study provides useful insights on candidate genes that are likely to be influenced by microsatellite expansion and contraction.

## Materials and methods

### Data collection

The *H. annuus* UniGene database (Assembly ID: HELI_ANNU.CSA1) available from the Compositae Genome Project (CGP) website (http://compgenomics.ucdavis.edu/) contains 31,605 unique sequences including 10,796 contigs and 20,809 singletons. An exhaustive search for microsatellites in this database was conducted using the software package SciRoKo v. 3.4 (Kofler et al. 2007). Parameter settings used to detect

microsatellites include a minimum copy number of 4, minimum score of 8, mismatch variable penalty mode to score and detect imperfect or impure microsatellites, mismatch penalty of 4, maximum number of continuous mismatches allowed being 3. Default parameter setting scores were used for all parameters except the minimum score, where it was set to the minimum allowable score in SciRoKo, eight to allow detection of shorter microsatellites as well as shorter microsatellites with impurities. Hence, only microsatellites meeting these parameter values were utilized for our analysis. SciRoKo v. 3.4 also identifies the position of the microsatellite in the sequence, motif size, motif type and level of impurity in the microsatellite. A standardized motif type is also reported, which makes it easier to perform further statistical analyses.

To infer the relative position of microsatellites with regard to start and stop codons, UniGene sequences (nucleotide) containing microsatellites were compared to potentially homologous sequences in the completely annotated dicot genome of *Arabidopsis thaliana*, using the Linux based software package ESTScan v. 2.0 (Iseli et al. 1999; Lottaz et al. 2003). The matrix parameter, 'm' was set to specify the *A. thaliana* scoring matrix as the comparison matrix. ESTScan v. 2.0 uses a fifth order Markov model to detect and extract coding regions from EST sequences. The program makes predictions based on taxon dependent codon usage biases as well as G+C isochore set data. The program also corrects frameshift errors inherent in single pass EST sequences prior to predicting the position of translated sequences. ESTScan was run a second time to extract the translated amino acid sequence encoded by each coding region from the input nucleotide sequence. All parameters were kept the same as previously specified except that an additional parameter 'p' was added to retrieve translated amino acid

sequences. The translated amino acid sequences were further run on a custom Perl script to extract all homopolypeptide regions.

**Data Analysis**

Unigene sequence data were used to perform a series of statistical analyses designed to test our hypotheses. To test the prediction that less constrained regions were more likely to harbor more microsatellites as well as microsatellites with greater mutability, we determined the relative abundance of microsatellites and the abundance of each repeat unit type in the different exonic regions (5'UTR, coding and 3'UTR). EST based databases are likely to be biased as far as the sequencing differences between different regions are concerned, and the direction of bias is unknown. To overcome ascertainment bias associated with sequencing differences, we utilized mean densities of microsatellites ($N'_{Region}$) to identify prevalence, as opposed to mere counts ($N_{Region}$). Density is measured as number of microsatellites sampled per 2000 bp of a region (Equation 2.1).

$N'_{Region\ 'x'} = (N_{Region\ 'x'} / \text{Total number of bases sampled in Region 'x'}) * 2000$ (Equation 2.1)

Since mutability should also be reflected in the mean lengths of microsatellite tracts found in the database, we used ANOVA to test if microsatellite length significantly differs (i) between the three exonic regions, (ii) between different motif size classes or (iii) by an interaction effect between region and motif size. Effect of motif size class on lengths was included to determine if length of a tract was an artifact of greater mutability observed in smaller repeat size classes. Once an effect of region, motif size or interaction effect was found significant, we performed Tukey's Honestly Significant Difference

(HSD) test at $\alpha$=0.05 level of significance, to determine the exact region, motif size or the specific interaction effect that significantly contributed to the observed differences.

To test for constraints on mutability of microsatellite tracts, we performed a regression analysis to determine whether there is a relationship between the first predictor of constraint, i.e., number of interruptions per base pair in the microsatellite and the length of the microsatellite. Further, to test the prediction that constraints on microsatellites are more likely greater on microsatellites in the coding regions than the UTRs, we performed ANOVA to see if the impurity harbored by microsatellites significantly differs i) between the three exonic regions, (ii) between different motif size classes, and (iii) by an interaction effect between region and motif size. Further, an ANCOVA was performed with impurities as a continuous outcome variable and length and region as the continuous and nominal predictor variables, respectively. Hence, effects on impurities resulting from variance in lengths were removed while determining the effect of region. Underlying cause for significant effects established by ANOVA were identified using Tukey's HSD test at $\alpha$ =0.05 level of significance. Further, to determine if certain motif types (e.g. AT rich over GC rich) were more prone to evolve in the transcriptome, we performed ANOVA to quantify statistically the differences in abundance, lengths, and mean degree of impurity between the three regions for all the major repeat types. All statistical analyses were performed using JMP® version 9 for Windows (SAS Institute Inc.) and R statistical package v. 2.11.1 for Windows (R Core Development Team 2010).

A BLAST (Basic Local Alignment Search Tool) analysis was performed on all sequences containing microsatellites. BLASTX (Altschul et al. 1997) was performed against the "Refseq" protein database of *A. thaliana*. To identify general patterns of

microsatellite enrichment within the *H. annuus* database, hypergeometric tests to determine the enrichment of microsatellites within Gene Ontology (GO) classes were done using the GO Term enrichment tool within the AmiGO module (Carbon et al. 2009) with the BLASTX hit identifiers (accession numbers) used as the input query data, TAIR as the database filter, a cutoff value of $p < 0.001$ and minimum number of genes as 20. To test the prediction that gene classes that are under greater selection will be enriched for microsatellites in all regions, a region wise analysis of enrichment was then performed using AmiGO. Further, to test for associations between specific microsatellite motif types and gene classes, motif types with the highest density within each region were analyzed using AmiGO.

## Results

### Microsatellite distribution

We identified 6,057 microsatellites from 31,605 Unigene sequences. For 4,555 of these microsatellites, we were able to infer exonic regions namely 5'UTR, coding and 3'UTR. The remaining 1,502 were classified as non-coding transcripts and were excluded from further analyses because our focus is on the role of microsatellites in gene regulation and protein structure. Microsatellites in the unigene database of *H. annuus* were numerically more common in coding regions than in UTRs (Figure 2.1). However, density of microsatellites in UTRs is more than double that seen in the coding region. Observed density of microsatellites in the coding region is 1 per 1188 bp, whereas the density in the UTRs is 1 per 416 bp. Within the UTRs, 5'UTR has the highest density of microsatellites, followed by 3'UTR (Table 2.1). These results are comparable to several other microsatellite surveys in eukaryotes where greater density of microsatellites is

observed in the UTRs compared to the coding region (Wren et al. 2000; Lawson and Zhang 2006). Figure 2.1 provides the detailed distribution of microsatellites by repeat size in the three exonic regions (5'UTR, coding and 3'UTR). Mononucleotides, dinucleotides and trinucleotides were significantly enriched in the dataset as compared to tetranucleotide (n=369), pentanucleotides and hexanucleotide repeats (Chi square test, $P<0.001$). This observation suggests that smaller motif size classes are likely to be more abundant in the transcriptome due to their higher mutability.

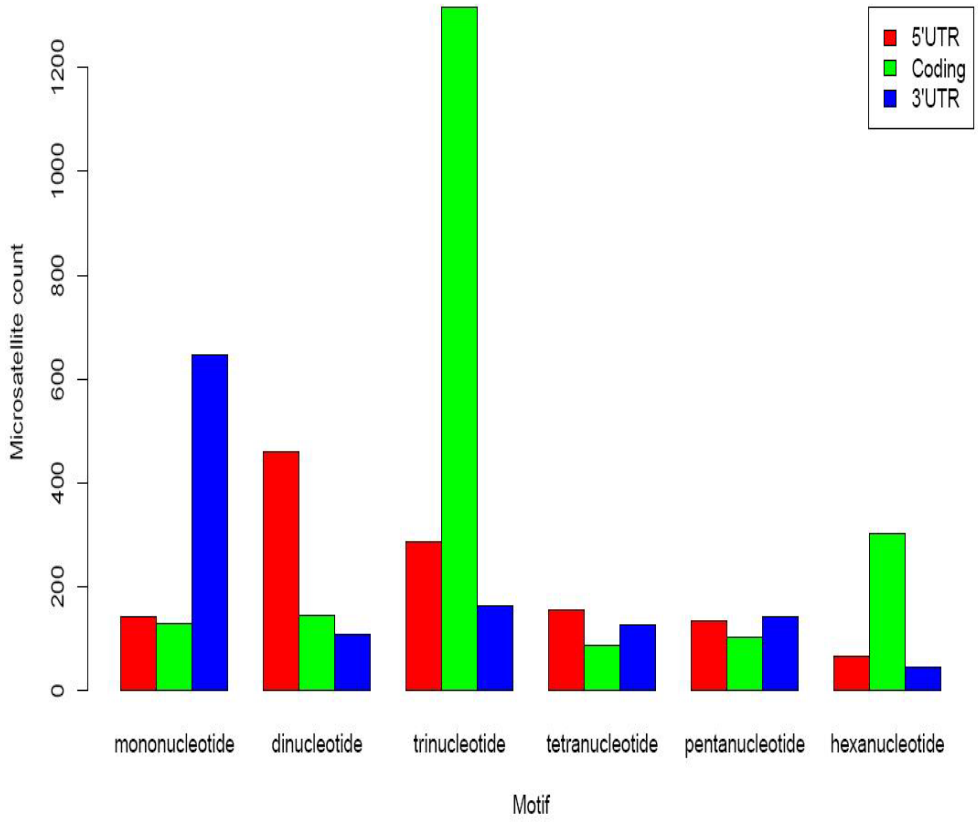


Figure 2.1 Observed counts of different microsatellite motif size classes in the three exonic regions of *Helianthus annuus*

Table 2.1 Density of microsatellites found in different regions of the *Helianthus annuus* transcriptome

| Region | Total number of base pairs | Number of Microsatellites | Density (per 2kbp) |
|---|---|---|---|
| Coding | 2,472,710 | 2082 | 1.7 |
| UTR | 1,029,636 | 2473 | 4.8 |
| 5'UTR | 133,163 | 1243 | 18.7 |
| 3'UTR | 896,473 | 1230 | 2.7 |

**Length**

The minimum observed tract length of microsatellites in the database is 15 bp. The mean length of microsatellites is 24.9 bp (±10.4 standard deviation). Results of ANOVA on microsatellite lengths show a general reduction in length with increasing motif length (Figure 2.2). A large number of statistical outliers for length were found to be tri- and hexanucleotide repeats in the coding region as well as mononucleotides in 3'UTR (Figure 2.2). There is a significant effect of motif size as well as exonic region on microsatellite length (ANOVA, $P<0.0001$). Mean length of microsatellites in the coding regions are significantly shorter than those in 3'UTR and 5'UTR (Tukey's HSD test, $P<0.05$). The interaction effect of region and motif size on length of microsatellites is found to be significant (ANOVA, $P<0.0001$). Microsatellite length is also observed to be significantly dependent on the interaction effect of repeat type and region (ANOVA, $P<0.0001$).

Further, an analysis was performed in the absence of motif size classes 1, 4, 5 and 6. It shows that there is a significant difference in lengths between di- and trinucleotides (ANOVA, $P<0.0001$). The 5'UTR microsatellites are significantly longer than the 3'UTR and coding region microsatellites (Tukey's HSD, $P<0.05$). At the same time,

trinucleotides and dinucleotides in the coding region are significantly longer than trinucleotides and dinucleotides in 5'UTR (Tukey's HSD test, P<0.05).

**Effect of Motif type**

*Mononucleotides*

A/T repeat frequencies are significantly different from C/G repeat frequencies across the three exonic regions (Chi-square test, $P<0.001$). The regionwise distribution of A/T repeats and C/G repeats are shown in Figure 2.3a. There is no significant difference in average lengths between A/T and C/G repeat tracts across the three regions. Significant differences in percentage of impurities harbored by the A/T repeats and C/G repeats are shown in Figure 2.4a.

*Dinucleotides*

The frequencies of the motif types were standardized as (i) AT, grouping AT and TA , (ii) AC, grouping AC, CA, TG and GT, and (iii) AG, grouping AG, GA, TC and CT; are shown in Figure 2.3b. The dinucleotides have the greatest density within the 5'UTR. CG dinucleotides were absent from the *H. annuus* dataset. AG dinucleotides are the most abundant, with AGs in 5'UTR accounting for nearly 57% of all dinucleotides in protein coding sequences. AGs and ACs have their greatest density within 5'UTR. However, ATs are most abundant within the 3'UTR.

There is no significant difference in lengths between standardized motif types AG, AC and AT across regions (Tukey's HSD, $P>0.05$). Significant differences (Tukey's HSD, $P<0.05$) in percentage of impurity harbored by the standardized motif types are shown in Figure 2.4b.

*Trinucleotides*

Ten standardized motif types were reported. There is a significant difference in frequencies of the motif types in different regions (Chi square test, $p < 0.0001$). The regionwise distribution of standardized motif types is shown in Figure 2.3c. All motif types have a higher abundance in the coding region. AAGs are most abundant in 5'UTR and AATs in 3'UTR. Motif types with a "CG" within it were significantly reduced within the *H. annuus* dataset. There is a significant difference in lengths of the microsatellites between the standardized motif types (ANOVA, $P<0.05$). Significant differences in lengths of standardized motif types are shown in Figure 2.5 (Tukey's HSD, $P<0.05$). There is no significant effect of the exonic regions in percentage impurities harbored by the motif types (ANOVA, $P>0.05$). There is a significant effect of motif type on the percentage impurity harbored by the microsatellite (ANOVA, $P<0.005$). There is also a significant effect of motif type by region interaction on the percentage impurity harbored by the microsatellite (ANOVA, $P<0.05$). The differences in percentage impurity harbored by the motif types are shown in Figure 2.4c.

The analysis could not be performed on tetranucleotide, pentanucleotide and hexanucleotide repeats due to small sample sizes.

**Abundant homopolypeptide tracts**

The distribution of amino acid repeats observed in the *H. annuus* transcriptome is detailed in Figure 2.7. There is a significant difference in frequencies of amino acids found as homopolypeptide tracts (Chi square test, $P<0.001$). Polar amino acid repeats were significantly more abundant in the transcriptome compared to non-polar amino acid repeats (Chi square test, $P<0.001$). There is a significant difference in abundance of various amino acid repeats within the two polarity classes, polar and non-polar (Chi

square test, $P<0.001$). Alanine, cysteine and tryptophan homopolypeptide tracts were not present in the *H. annuus* transcriptome. The amino acid type is found to have a significant effect on the length of the tract (ANOVA, $P<0.001$). The mean tract lengths were found to be significantly different between , (i) valine and glycine, (ii) valine and asparagine, (iii) valine and glutamine, and (iv) glutamine and arginine (Tukey's HSD, $P<0.05$).

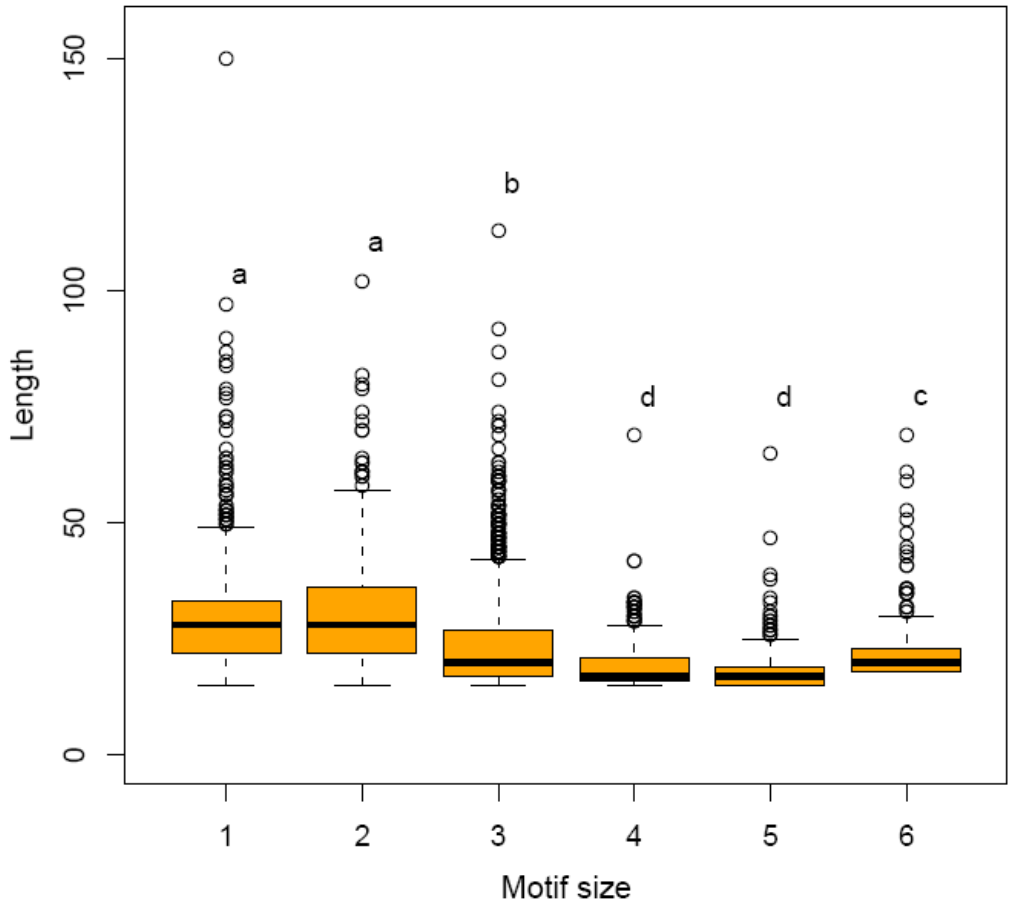


Figure 2.2 Relationship between microsatellite length and microsatellite motif size in the transcriptome of *Helianthus annuus* is shown.

Motif size classes that do not share a letter are significantly different from each other (Tukey's HSD, P<0.05).

**Major Gene Ontology (GO) classes**

The microsatellite containing genes were broadly classified as belonging to GOs associated with biological processes (P), cellular components (C) or molecular functions (F). There is significant enrichment of microsatellites in a total of 200 GO classes (Table A.1). Further, an analysis of enrichment of microsatellites within each exonic region separately shows significant enrichment within GO terms associated with biological processes, representative of various plant responses to stresses, biotic and abiotic stimuli. The most specific GO terms with significant microsatellite enrichment in each region are reported in Table 2.2. GO terms associated with response to chemical and organic stimuli were significantly prominent in transcripts containing AG microsatellites, which are the most abundant class of dinucleotides within the 5'UTR. Transcripts harboring the most abundant 3'UTR dinucleotide repeat, AT, in their 3'UTRs were significantly associated with GO term "Response to water deprivation". Transcripts with AAGs, the most abundant trinucleotides in the 5'UTR, were enriched in Biological processes GO term "Response to Cadmium ion". Similarly, transcripts with AAT repeats, the most abundant trinucleotides in the 3'UTR, were enriched in GO term "Response to water deprivation".

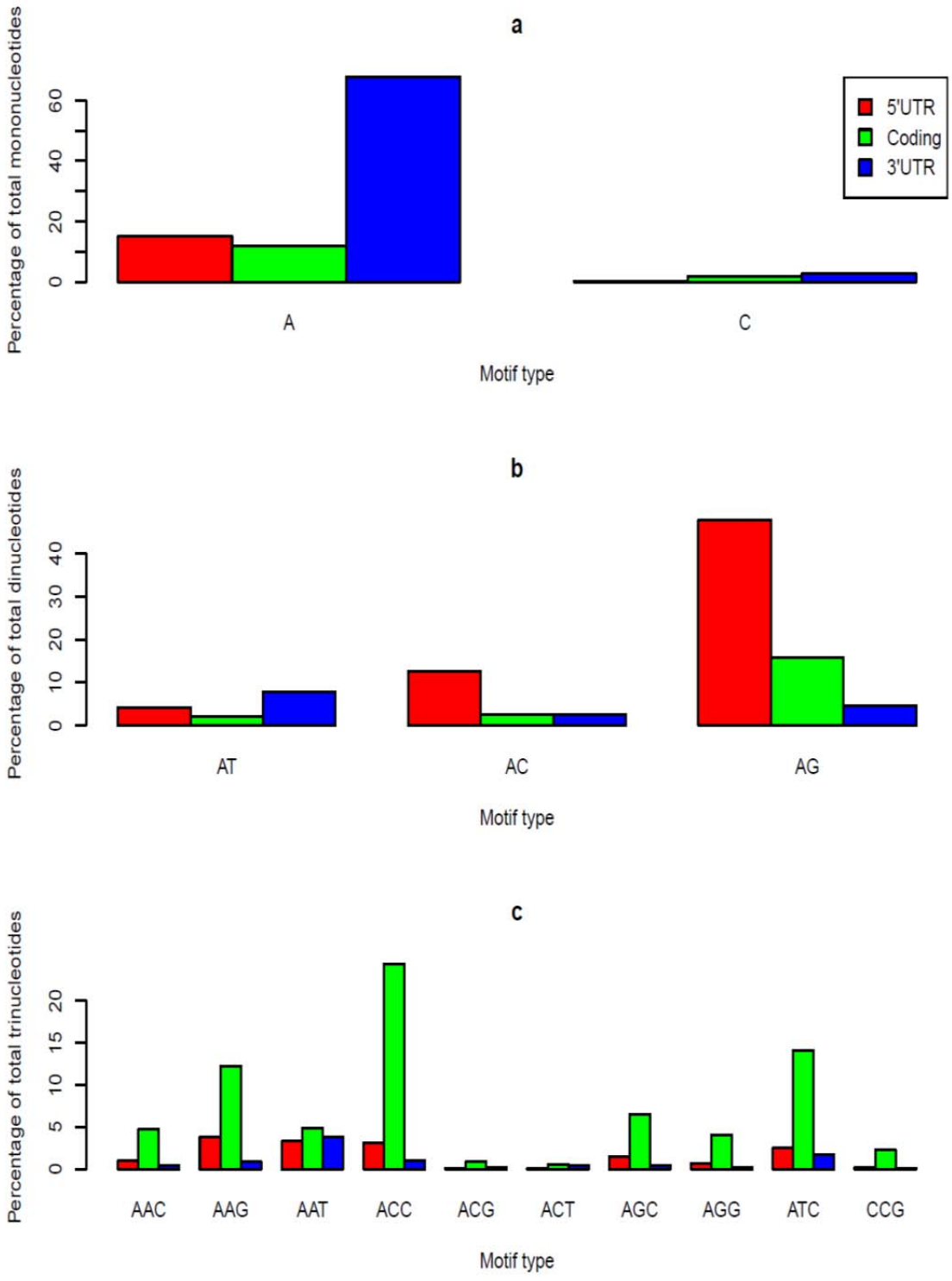


Figure 2.3 Distribution of standardized (a) mononucleotide, (b) dinucleotide, and (c) trinucleotide repeat types in the *Helianthus annuus* transcriptome

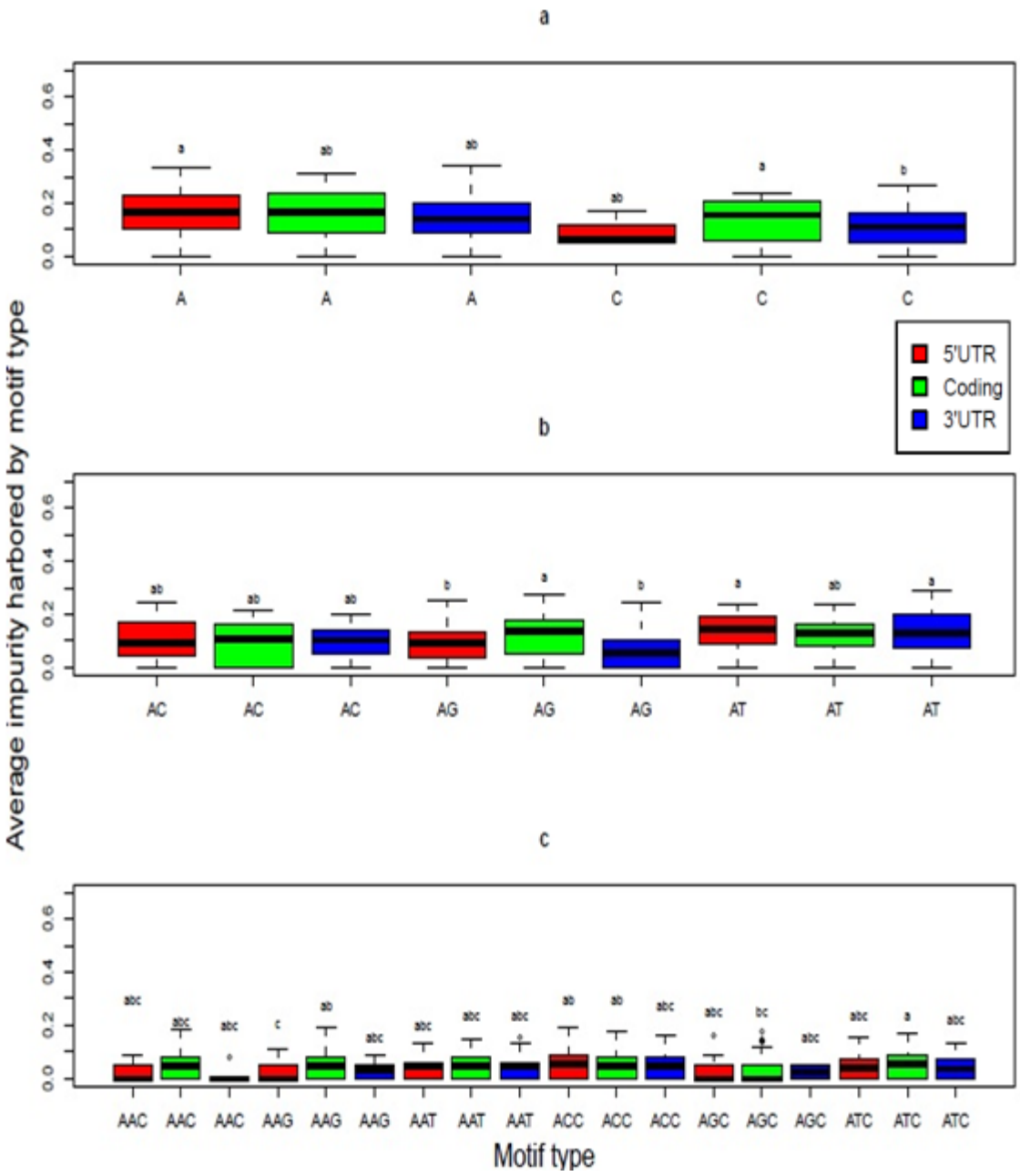


Figure 2.4 Relationships between impurities harbored by a) mononucleotide, b) dinucleotide, and c) trinucleotide microsatellite repeats within the three exonic regions of *Helianthus annuus* are shown.

Average impurities are calculated as the number of impurities divided by the length of the microsatellite tract. The means, standard deviations, and 95% confidence intervals are depicted for each motif type in each region. Motif types that do not share a letter are significantly different from each other (Tukey's HSD, $P<0.05$).

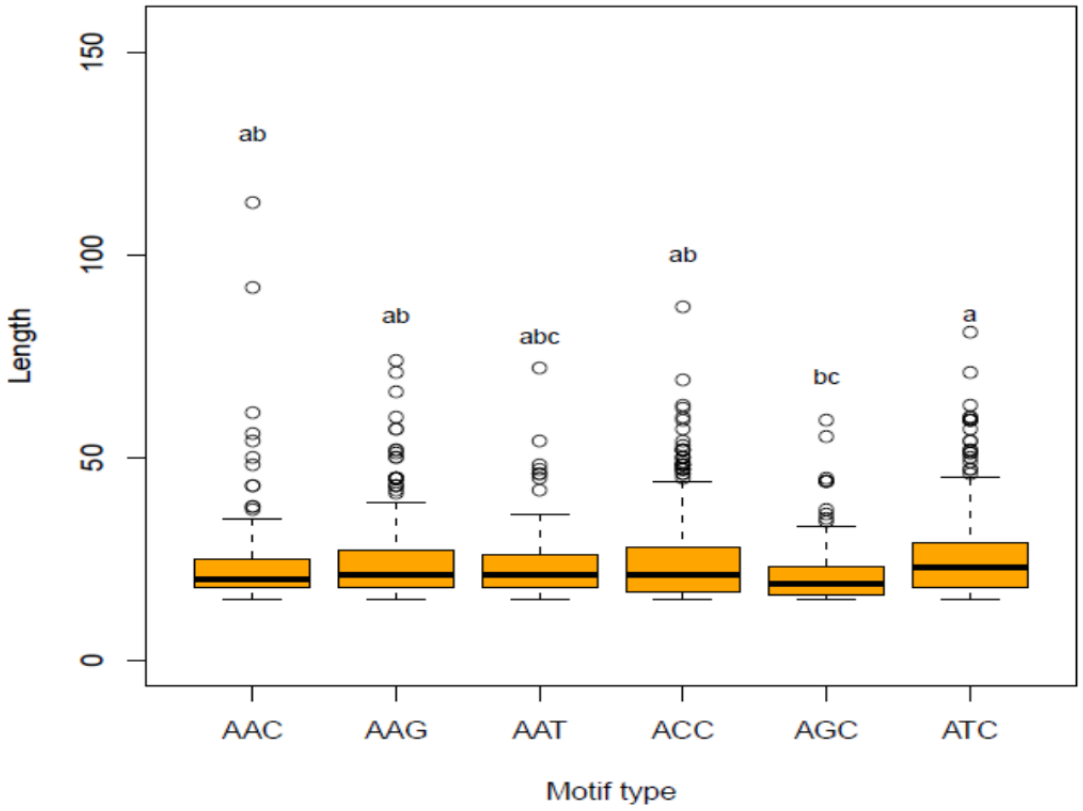


Figure 2.5 Relationship between length and major trinucleotide motif types found in *Helianthus annuus* transcriptome is shown.

Motif types that do not share a letter are significantly different from each other (Tukey's HSD, $P<0.05$).

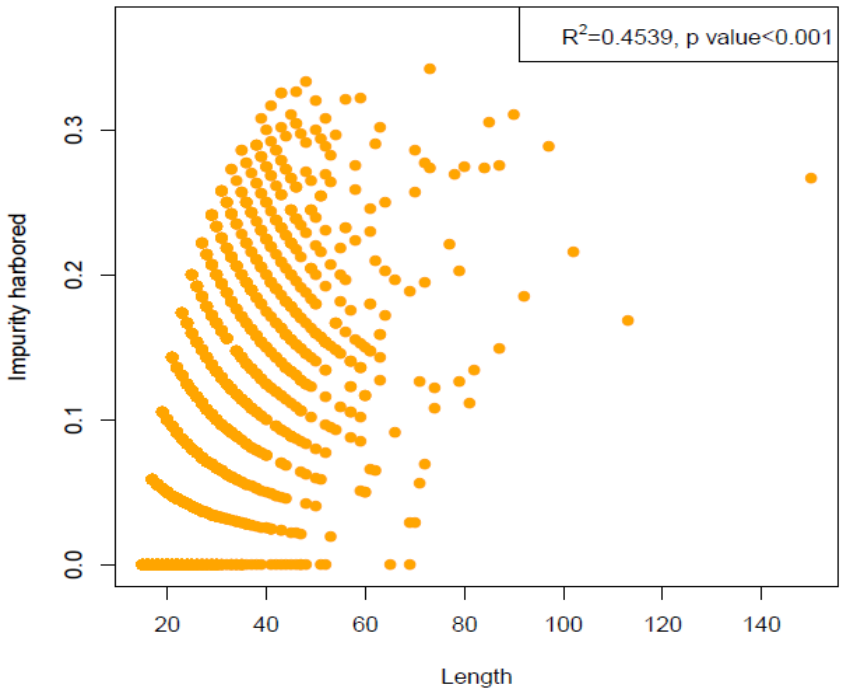


Figure 2.6 Relationship between impurities harbored by microsatellites in the *Helianthus annuus* transcriptome and the length of the microsatellites

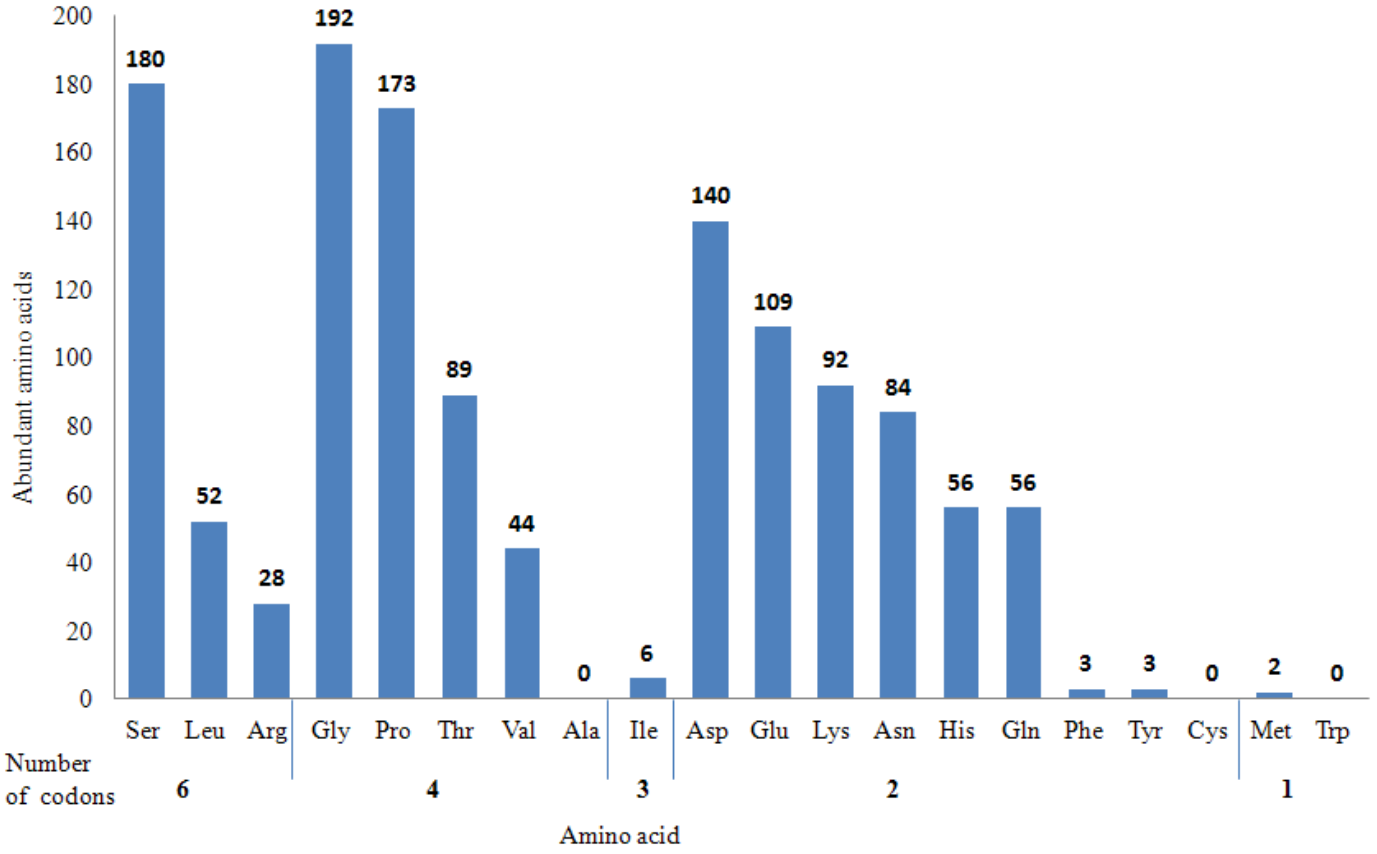


Figure 2.7 Number of amino acids found as homopolypeptide repeats in the *Helianthus annuus* transcriptome

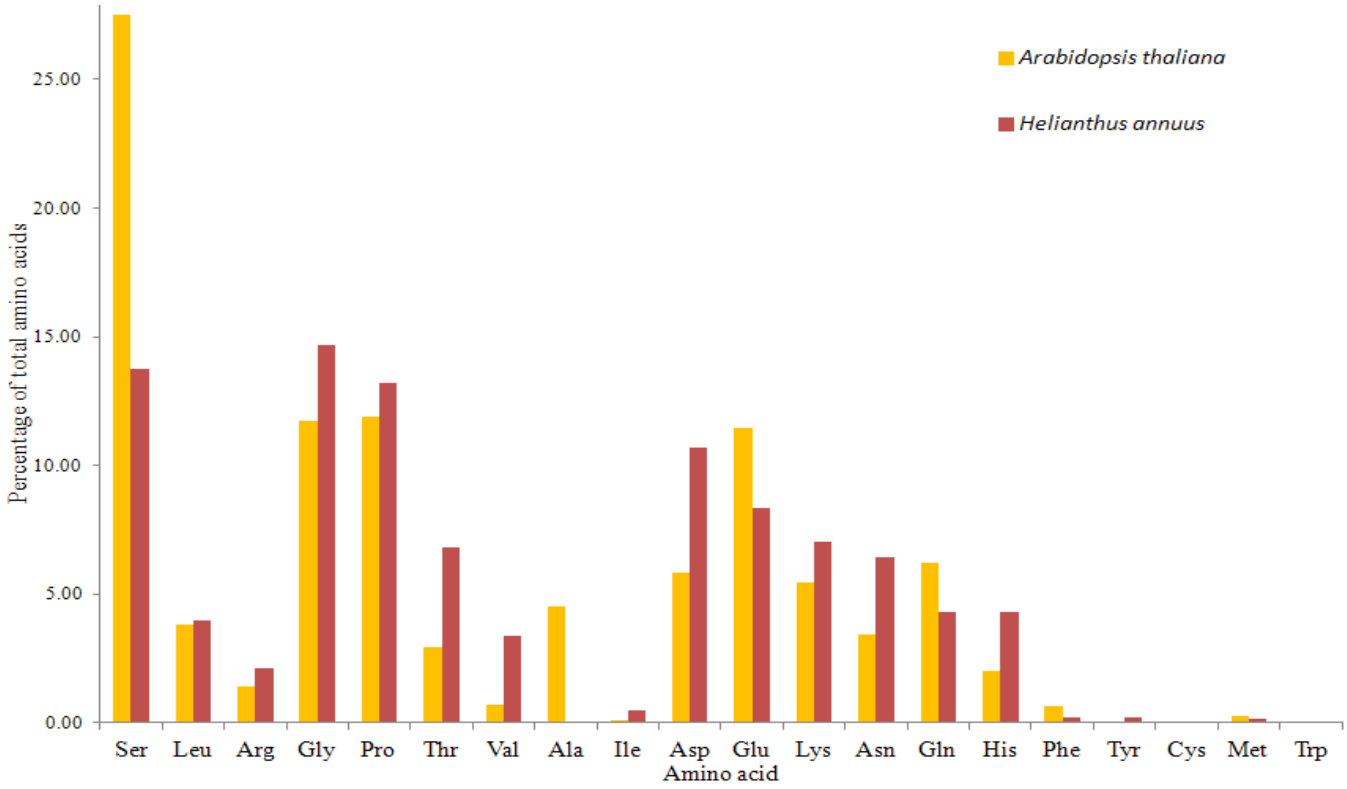


Figure 2.8 A comparison between percentages of amino acid repeats in the *Helianthus annuus* and *Arabidopsis thaliana* transcriptomes. The *A. thaliana* data was obtained from Lawson and Zhang (2006)

Table 2.2 The most specific Gene Ontology terms with significant enrichment of microsatellites in the *Helianthus annuus* transcriptome within each region is shown. The analyses were performed using the AmiGO module.

| Region | GO Term | Aspect | *P* value | Number of Genes | Percentage |
|---|---|---|---|---|---|
| 5'UTR | GO:0010033 response to organic substance | P | 1.12E-08 | 82/805 | 10.20% |
| | GO:0006810 transport | P | 1.31E-06 | 95/805 | 11.80% |
| | GO:0044281 small molecule metabolic process | P | 3.28E-05 | 75/805 | 9.30% |
| | GO:0071310 cellular response to organic substance | P | 7.87E-06 | 38/805 | 4.70% |
| | GO:0009725 response to hormone stimulus | P | 4.00E-05 | 59/805 | 7.30% |
| | GO:0009651 response to salt stress | P | 1.47E-04 | 33/805 | 4.10% |
| | GO:0005515 protein binding | F | 1.88E-12 | 132/805 | 16.40% |
| Coding | GO:0009737 response to abscisic acid stimulus | P | 4.02E-06 | 41/1091 | 3.80% |
| | GO:0046686 response to cadmium ion | P | 1.87E-04 | 34/1091 | 3.10% |
| | GO:0034641 cellular nitrogen compound metabolic process | P | 7.59E-04 | 186/1091 | 17.00% |
| | GO:0005515 protein binding | F | 9.34E-11 | 159/1091 | 14.60% |
| | GO:0005198 structural molecule activity | F | 4.15E-06 | 50/1091 | 4.60% |
| 3'UTR | GO:0009414 response to water deprivation | P | 2.49E-05 | 24/789 | 3.00% |
| | GO:0009651 response to salt stress | P | 2.62E-05 | 34/789 | 4.30% |
| | GO:0009737 response to abscisic acid stimulus | P | 4.35E-05 | 32/789 | 4.10% |
| | GO:0046686 response to cadmium ion | P | 5.19E-05 | 29/789 | 3.70% |
| | GO:0005515 protein binding | F | 1.44E-13 | 133/789 | 16.90% |
| | GO:0003735 structural constituent of ribosome | F | 9.07E-11 | 42/789 | 5.30% |

## Discussion

The survey of *H. annuus* unigene database establishes many known features of the microsatellite prevalence as well as provides some unique insights into the gene families that are more likely to harbor microsatellites. We also approached the database survey to provide insights into the relative strength of the functional role of microsatellites in gene regulation and in modulating protein structural changes, compared

to presumed neutrality of these sequences. We have compared the relative prevalence of microsatellites as well as characteristics such as length, percentage impurity, and motif type, of microsatellites in exonic regions that are likely to be involved in gene regulation, i.e., 5'UTR and 3'UTR, to microsatellites in coding regions that are likely to be associated with structural changes in proteins. As the coding region is more conserved, we expected to find reduced microsatellite prevalence in the coding region compared to the UTRs.

This study reveals that microsatellites are present in 19% of all transcripts in the sunflower unigene database. The largest proportions of transcribed microsatellites are located within the coding regions (Figure 2.1). This finding runs contrary to the expectation of increased constraint on coding regions. However, from the perspective of density, microsatellites per base pair, which is our preferred method to circumvent the ascertainment bias associated with sequencing differences, they are far more common within the UTRs, a finding that is consistent with that of other similar studies (Lawson and Zhang 2006; Hong et al. 2007; Mun et al. 2006; Wren et al. 2000; Subramanian et al. 2003). It is well known that microsatellites are highly prone to mutation by means of slippage (Levinson and Gutman 1987); hence the observed higher density of microsatellites within the UTRs suggests that these regions are more labile or are under fewer evolutionary constraints than the coding regions. Further, 5'UTRs have six times greater density of microsatellites than 3'UTRs in the sunflower UniGene database. This suggests that 5'UTRs either have much reduced constraints limiting microsatellites in these regions relative to 3'UTRs, or that directional selection has favored their expansion within 5'UTRs. In the absence of any general observation that 3'UTRs are under greater constraints, this finding is consistent with the hypothesis that microsatellites evolve

preferentially to regulate transcription and translation relative to the frequency at which they evolve to regulate mRNA stability. However, other explanations for this finding are clearly possible. One piece of evidence that runs counter to this interpretation, concerns the actual counts of microsatellites in 5' and 3' UTRs. These are roughly equal. Hence, greater densities in 5'UTRs may simply reflect the fact that these sequences are shorter rather than being more prone to evolving regulatory microsatellites.

The distribution of microsatellites within the coding regions of the *H. annuus* transcriptome indicates a higher prevalence of trinucleotides and hexanucleotides when compared to the remaining motif size classes, supporting predictions from studies suggesting that elongation or contraction of these microsatellites should be less likely to cause frameshift mutations (Metzgar et al. 2000; Wren et al. 2000). These results are also consistent with studies in other plants such as *A. thaliana* and rice (Lawson and Zhang 2006), *Brassica rapa* (Hong et al. 2007), *Medicago tunculata* (Mun et al. 2006) as well as animals such as mice and humans (Wren et al. 2000, Subramanian et al. 2003). Evolutionary constraints on microsatellites could also be inferred from the mean length of the tracts, with shorter than average lengths indicating higher evolutionary constraint. When the major microsatellite motif size classes (dinucleotides and trinucleotides) were compared, microsatellites in 5'UTRs were on average longer than those in 3'UTRs and coding regions. This could either suggest reduced constraint on evolution of microsatellites in 5'UTR compared to coding region and 3'UTR or directional selection favoring the expansion of microsatellites in 5'UTRs.

There was a length constraint on a majority of the microsatellites identified; however a few outliers were also identified. According to Garza et al. (1995), shorter microsatellite tracts have low mutation rates and longer ones have higher mutation rates.

This likely reflects the increased opportunity for slipped strand mispairing presented by longer microsatellite tracts (Levinson and Gutman 1987; Ellegren 2004; Hancock 1999). More recently, Kelkar et al. (2010) have estimated by both computational and experimental analyses, that slippage rates at mono- and dinucleotide tracts of lengths of 10 bases or above, is significantly greater than the background slippage rates, when compared to shorter tracts. The outliers for length were all the shorter motif size classes, which also harbored a greater number of interruptions per bp when compared to other microsatellites in the *H. annuus* database. Dokholyan et al. (2000) suggested that UTRs are likely to harbor greater numbers of pure microsatellite tracts, since these non-coding regions exhibit higher tolerance for mutation. Our results show that dinucleotides as well as trinucleotides in coding regions had significantly more interruptions per base pair than dinucleotides and trinucleotides in 5'UTR respectively. The purity of the 5'UTR microsatellites could also indicate a higher selection for mutability of these microsatellites, which also supports the hypothesis regarding the "tuning ability" of microsatellites in UTRs, suggested by King et al. (1992) and Trifanov (2004). That is, purifying selection may in essence be limiting the accumulation of impurities in some exonic domains while directional selection favors them in others.

Purity of a tract may also indicate a recent origin compared to impure tracts encountered in the database since with time microsatellites are likely to accumulate point mutations by neutral processes. Hence, purer tracts in 5'UTR could be relatively "young" compared to the coding region microsatellite tracts, suggestive of recent "seeding" of these UTR microsatellites. If functionally beneficial, the recently seeded microsatellites may play a role in modulation of genes via transcription or translation regulation. Presence of a greater number of pure tracts as well as a greater density of microsatellites

in the 5'UTR as compared to the coding regions could also indicate that a large proportion of microsatellites in 5'UTRs are relatively younger than coding region microsatellites. This may also indicate that a large proportion of evolutionary novelty in genomes is derived from sequence variation in the 5'UTRs of genes. Although the 5'UTRs form a relatively small proportion of the *cis* regulatory elements of a gene, the results point towards a more important role for 5'UTRs as compared to changes in the amino acid sequence. The relative importance of gene regulation over changes in the gene product has been explored as early as the 1960s (Britten and Davidson 1969). Several studies have since considered more closely the evolutionary changes brought about by gene regulation (Bradley et al. 2010; Fay and Wittkopp 2008; Wittkopp 2007; Wittkopp 2010). The prevalence of microsatellites in 5'UTRs of *H. annuus* as well as several other such surveys (Lawson and Zhang 2006; Hong et al. 2007; Mun et al. 2006; Subramanian et al. 2004; Wren et al. 2000) could possibly be pointing towards the relative importance of microsatellites in gene regulation. Further, the most abundant microsatellites in the 5'UTRs of *H. annuus* were AGs (among dinucleotides) and AAGs (among trinucleotides), which are also observed to be most abundant in *A. thaliana* (Lawson and Zhang 2006). Anecdotal evidence from other plant species has suggested an involvement of AGs and AAGs in 5'UTRs in transcription and translation regulation (Newcomb et al. 2006; Zhang et al. 2004). Similarly, AREs or AT rich repeats in 3'UTRs may modulate gene expression changes by influencing the stability of the transcript. Our study has found significant enrichment of AT and AAT repeats in the 3'UTR, suggesting a possible prevalence of gene regulation by modulating transcript stability. AT microsatellites are the most abundant dinucleotides in 3'UTR of *A. thaliana*, however, AAG repeats and not AAT repeats are the most abundant trinucleotide repeats in *A.*

*thaliana* (Lawson and Zhang 2006). This difference in the prevalence of different microsatellite motifs between the two dicots is not surprising given the differences observed between more closely related species such as humans and chimpanzees (Galindo et al. 2009). The absence of CG repeats as well as significantly low densities of trinucleotide motifs containing "CG" in the transcriptome, suggest that CpG island methylation is not a common mechanism by which exonic microsatellites regulate transcription rates in *H. annuus*.

Previous studies have shown that longer stretches of microsatellites with point mutations are relatively more stable than those without interruptions (Yu et al. 2005). If a microsatellite tract attains a threshold length, it may expand rapidly via slipped strand mispairing (Hancock 1999; Eisen 1999). However, point mutations that interrupt otherwise homogeneous tracts appear to improve fidelity during replication. In the case of microsatellites located in coding regions, purifying selection is hypothesized to prevent accumulation of nonsynonymous mutations as they can potentially disrupt homopolypeptide repeats. Consistent with this hypothesis is the relative prevalence of synonymous mutations within coding region microsatellites (Wren et al. 2000). Empirical studies such as those of Yu et al. (2005) may provide evidence that point mutations accumulate once the microsatellite tract has attained a length that confers the optimally fit phenotype (Kashi and King 2006). Microsatellite mutation rates are reduced by the introduction of point mutations. Hence, selection on microsatellites may act to allow for tract expansion till it achieves the optimum gene regulatory activity possible, which is then followed by introduction of point mutations to stabilize microsatellite alleles at optimum lengths. An example to demonstrate this phenomenon is a study by Fondon and Garner (2004) where they suggested that morphological differences in 92 breeds of

recently domesticated dogs could result from highly variable microsatellites present in coding regions. These tracts were also highly homogenous, pointing to their recent origin. Elongated microsatellites are known to be involved in congenital neurodegenerative diseases in humans, which are caused by a 10 or 100 fold increase in the number of DNA triplet repeats that code for homopolymeric amino acid stretches. These stretches are usually involved in binding nuclear proteins involved in transcription regulation (Wren et al. 2000). Although we would expect selection to act against such deleterious mutations we find that these mutation prone alleles are widespread in several functional sites. This was found by Yu et al. (2005) in their study of prevalence of pre-expansion spinocerebellar ataxia 2 (SCA 2) disease allele, and also by Rockman et al. (2004) with their study on relative prevalence of an allele associated with heart disease risk. Kashi and King (2006) suggested evolutionary trade-offs as a possible answer to the selective maintenance of disease alleles which might possibly confer some evolutionary advantage that might not be readily apparent. That is, natural selection may be favoring individuals that can produce diverse offspring over those producing similar offspring. This suggests that the outliers could be regions that expanded until they reached their optimum functional activity and were later stabilized by the introduction of point mutations that prevented it from mutating or elongating further. The presence of exceptionally long repeats in coding regions that also harbor a greater number of interruptions may suggest the role of selection.

The amino acids stretches encoded by microsatellites in the *H. annuus* unigene database are comparable to distributions in *A. thaliana* as reported by Lawson and Zhang (2006) (Figure 8). Amino acids encoded by more codons are likely to have enhanced ability to form homopolypeptide tracts, since these tracts are encoded by nonhomogenous

and hence stable microsatellite tracts. Glycine repeats form the most abundant homopolypeptide stretch in this database. Although Glycine is encoded only by four different codons, it has the simplest structure among amino acids, increasing its likelihood of forming homopolypeptide stretches. Homopolypeptides of serine are also abundant in the database and this could be explained in part by the fact that serine can be encoded by six different codons. Proline homopolypeptides exhibit similar levels of abundance but are only encoded by four different codons. Two out of the three amino acids that were absent from the *H. annuus* database, cysteine and tryptophan match expectations from other such surveys (Lawson and Zhang 2006; Wren et al. 2000). Homopolypeptides of cysteine are thought to result in abnormal cross links due to their ability to form sulphide bridges. Similarly, tryptophan and tyrosine result in bulky structures that are highly unstable (Wren et al. 2000). Alanine tracts were found to be abundant in the *A. thaliana* database. Absence of alanine tracts from the *H. annuus* database could most possibly be due to biases introduced by the nature of sequences sampled.

We further wanted to test if microsatellites are enriched within genes that would be predicted to require frequent "tuning" to aid in plant adaptation. Some studies have suggested that certain genes, such as transcription factors which can potentially regulate many other genes, are more likely targets of selection (Bradley et al. 2010; Fay and Wittkopp 2008). Lawson and Zhang (2006) did not observe a statistically significant enrichment of microsatellites in any GO terms in *A. thaliana*, but significant enrichment was observed in genes associated with biological processes GO terms; "response to stress" and "response to biotic stimulus" in rice. Significant enrichment of microsatellites in *H. annuus* is observed in GO terms associated with Biological processes that involve

plant response to stress, biotic and abiotic stimulus. These GO terms consistently were shown to be enriched when each of the exonic regions were analyzed separately. These results suggest that the microsatellites are enriched in GO classes that are most likely targets of selection in a plant system. The highest enrichment of 5'UTR and coding microsatellites has been found in GO categories associated with genes that are involved in response to various abiotic stresses including salt stress, osmotic stress, and heavy metal stress. Moreover, motifs most prevalent in 5'UTR and 3'UTR are known to influence transcription, translation as well as transcript stability. In other words, UTRs of genes involved in stress response are enriched with highly mutable microsatellites. This potentially suggests that there is selection for enriching microsatellites within UTRs of genes that are required more frequently in adaptive responses than other classes of genes. Hence, our results suggest that microsatellites in gene clusters that could potentially affect the phenotype (gene expression) are being targeted by selection. This study has identified several microsatellites that could potentially be of importance to adaptive responses observed in *H. annuus*. Experiments to estimate the effects of all 5'UTR and 3'UTR microsatellites in the unigene database detected to be involved in gene regulation could provide a better understanding of the "tuning knob hypothesis" discussed by Trifanov (2004). Our study also provides abundant resources that could be used to test for relationship between gene expression and microsatellite allele length in wild as well as experimental populations of *H. annuus*.

**Acknowledgements**

The authors would like to acknowledge; Kristen Sauby, Leah Chinchilla and Christopher Brooks for help during initial stages of this work, Susan M. Bridges for help

with preliminary data analyses, and David Chevalier and Donna Gordon for helpful suggestions on the Gene Ontology analyses. This work was supported in part by the National Science Foundation under grant NSF EPS-0903787 to AD Perkins. The Office of Research and Economic Development, College of Arts and Sciences, and the Department of Biological Sciences at Mississippi State University funded this research

## References


Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics* 8:619–631. doi:10.1038/nrg2158

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biology* 8(3): e1000343. doi:10.1371/jornal.pbio.1000343

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: A theory. *Science* 165:349-358

Carbon S, Ireland A, Mungall CJ, Shu SQ, Marshall B, Lewis S, the AmiGO Hub and the Web Presence Working Group. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* 25(2): 288–289

Dokholyan NV, Buldyrev SV, Havlin S, Stanley HE. 2000. Distributions of dimeric tandem repeats in noncoding and coding DNA sequences. *Journal of Theoretical Biology* 202:273-282

Eisen JA. 1999. Mechanistic basis for microsatellite instability. In "*Microsatellites: Evolution and Applications*" (D.B. Goldstein & C. Schlötterer, eds) (Oxford University Press) pp. 34-48

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* 16(12):551-558

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5:435-445

Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Bandaand MG, Whisstock JC. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Research* 15: 537-551

Fay JC and Wittkopp PJ. 2008. Evaluating the role of natural selection in evolution of gene regulation. *Heredity* 100:191-199

Fondon JW III and Garner HR. 2004. Molecular origins of rapid and continous morphological evolution. *Proceedings of the National Academy of Sciences* 99:1991-2004

Galindo CL, McIver LJ, McCormick JF, Skinner MA, Xie Y, Gelhausen RA, Ng K, Kumar NM, Garner HR. 2009. Global Microsatellite Content Distinguishes Humans, Primates, Animals, and Plants. *Molecular Biology and Evolution* 26(12):2809–2819

Garza JC, Slatkin M, Freimer NB. 1995. Microsatellite allele frequencies in Humans and Chimpanzees with implications for constraints on allele size. *Molecular Biology and Evolution* 12:594-603

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics* 6:743–55

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* 44:445-477

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, SchaffnerW. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* 263: 808–811.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW. 1995. An Evaluation of Genetic Distances for Use with Microsatellite Loci. *Genetics* 139: 463–471

Hammock EAD, Young LJ. 2004. Functional microsatellite polymorphism associated with divergent social structure in vole species. *Molecular Biology and Evolution* 21:1057–1063.

Hammock EAD, Young LJ. 2005. Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science* 308(5728):1630 – 1634.

Hancock JM. 1999. Microsatellites and other simple sequences: genomic context and mutational mechanisms. In "*Microsatellites: Evolution and Applications*" (D.B. Goldstein & C. Schlötterer, eds) (Oxford University Press) pp. 1-9

Hong CP, Piao ZY, Kang TW, Batley J, Yang TJ, Hur YK, Bhak J, Park BS, Edwards D, Lim YP. 2007. Genomic Distribution of Simple Sequence Repeats in *Brassica rapa*. *Molecules and Cells* 23(3):349-35

Iseli C, Jongeneel CV, Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol.*138-48. url: http://www.ch.embnet.org/software/ESTScan.html

Jarne P, Lagoda PJL. 1996. Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution* 11: 424 – 429

Jongeneel CV. 2000.Searching the Expressed Sequence Tag (EST) databases: planning for genes. *Briefings in Bioinformatics* 1(1):76-92

Karlin S, Burge C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci.* 93: 1560-1565.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* 22(5): 253-259.

Kashi Y, King DG, Soller M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics* 13:74-78

Kashi Y, King DG, Soller M. 1999. Functional roles of microsatellites and minisatellites. In "*Microsatellites: Evolution and Applications*" (D.B. Goldstein & C. Schlötterer, eds) (Oxford University Press). pp. 10-22

Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. 2010. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution* 2: 620–635

King DG, Soller M, Kashi Y. 1992. Evolutionary tuning knobs. *Endevor* 21: 36-40

King DG, Trifonov EN, Kashi Y. 2006. Tuning knobs in the genome: evolution of simple sequence repeats by indirect selection. In "*The Implicit Genome*" (Ed. Caporale LH), Oxford University Press, New York, pp 77-90

Kim TS, Booth JG, Gauch HG Jr, Sun Q, Park J, Lee YH, Lee K. 2008. Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. *BMC Genomics* 9: 31. doi:10.1186/1471-2164-9-31

Kofler R, Schlötterer C, Lelley T. 2007. SciRoKo: A new tool for whole genome microsatellite search and investigation. *Bioinformatics* 23(13): p. 1683-1685.

Lawson MJ, Zhang L. 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biology* 7(2):R14

Lee JE, Lee JY, Wilusz J, Tian B, Wilusz CJ. 2010. Systematic analysis of cis-elements in unstable mRNAs demonstates that CUGBP1 is a key regulator of mRNA decay in muscle cells. *PloS ONE* 5(6) e11201

Levinson G, Gutman GA.1987. Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution. *Molecular Biology and Evolution* 4 (3): 203–221.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* 11: 2453-2465.

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: Structure, Function and evolution. *Molecular Biology and Evolution* 21(6):991-1007

Lottaz C, Iseli C, Jongeneel CV, Bucher P. 2003. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* 19:103-112.

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. 1999. A census of protein repeats. *Journal of Molecular Biology* 293:151-160.

Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. 2005. Microsatellite instability regulates transcription factor binding and gene expression. *Proceedings of the National Academy of Sciences. U.S.A*. 102, 3800

Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research* 10: 72-80

Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. *Genome Biology* 3(3):reviews0004.1-0004.10

Morgante M, Hanafey M, Powell W. 2002. Microsatellites are preferentially associated with non repetitive DNA in plant genomes. *Nature Genetics* 30:194-200.

Mularoni L, Veitia RA, Mar-Alba M. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* 89:316-325.

Mun JH, Kim DJ, Choi HK, Gish J, Debelle´ F, Mudge J, Denny R, Endre´ G, Saurat O, Dudez AM, Kiss GB, Roe B, Young ND, Cook DR. 2006. Distribution of Microsatellites in the Genome of *Medicago truncatula*: A Resource of Genetic Markers That Integrate Genetic and Physical Maps. *Genetics* 172: 2541–2555

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156: 297–304

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EHA, et al. (14 more authors). 2006. Analyses of Expressed Sequence Tags from Apple. *Plant Physiology*141:147–166

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327: 92–94

Persico AM, Levitt P, Pimenta AF. 2006. Polymorphic GGC repeat differentially regulates human reelin gene expression levels. *Journal of Neural Transmission* 113:1373–1382

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. url [http://www.R-project.org]

Raca G, Siyanova EY, McMurray CT, Mirkin SM. 2000. Expansion of the $(CTG)_n$ repeat in the 5′-UTR of a reporter gene impedes translation. *Nucleic Acids Research* 28(20): 3943–3949

Ranum L P W and Day JW. 2002. Dominantly inherited, non-coding microsatellite expansion disorders. *Current Opinion in Genetics and Development* 12:266–271

Rattenbacher B, Beisang D, Wiesner DL, Jesche JC, Von Honhenberg M, St. Louis Vlasova IA and Bohjanen PR. 2010. Analysis of CUGBP1 targets identifies GU-repeat sequences that mediate rapid mRNA decay. *Molecular and Cellular Biology* 30(16): 3970-3980

Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, Wray GA. 2004. Positive selection on MMP3 regulation has shaped heart disease risk. *Current Biology* 14:1531-1539.

Rubinsztein DC. 1999. Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations. In "*Microsatellites: Evolution and Applications*" (D.B. Goldstein & C. Schlötterer, eds) (Oxford University Press) pp. 80-96

Schlötterer C, Wiehe T. 1998. Microsatellites, a neutral marker to infer selective sweeps; In "*Microsatellites: Evolution and Applications*" (D.B. Goldstein & C. Schlötterer, eds) (Oxford University Press) pp.238-247

Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y. 1999. Shortened microsatellite d (CA) 21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS letters* 455(2): 70-74

Shimohata T, Nakajima T, Yamada M, Uchida C, Onodera O, Naruse S, Kimura T, Koide R, Nozaki K, Sano Y, et al. 2000. Expanded polyglutamine stretches interact with TAFII130, interfering with CREB dependent transcription. *Nature Genetics* 26: 29–36.

Subramanian S, Mishra RK and Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology* 4:R13

Tautz D, Renz M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* 12: 4127-4138

Timchenko NA, Welm AL, Xiaohui L, Timchenko LT. 1999. CUG repeat binding protein (CUGBP1) interacts with the 5′ region of C/EBPβ mRNA and regulates translation of C/EBPβ isoforms. *Nucleic Acids Research* 27 (22): 4517-4525

Trifanov EN. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation; In "*Evolutionary theory and processes: Modern Horizons, papers in honor of Eviatar Nevo"* (Ed. Wasser SP); Kluwer Academic Publishers. pp 115-138

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science* 324:1213-1216

Wang Z, Weber JL, Zhong G, Tanksley SD. 1994. Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* 88:1-6

Wittkopp PJ. 2007. Variable gene expression in eukaryotes: a network perspective. *The Journal of Experimental Biology* 210:1567-1575

Wittkopp PJ. 2010. Variable transcription factor binding: A mechanism of evolutionary change. *PLoS Biology* 8(3): e1000342. doi:10.1371/jornal.pbio.1000342

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Human Molecular Genetics* 2:1123-1128

Wheeler TM, Sobczak K, Lueck JD, Osborne RJ, Lin X, Dirksen RT, Thornton CA. 2009. Reversal of RNA dominance by displacement of protein sequestered on triplet repeat RNA. *Science* 325: 336-339

Wren JD, Forgacs E, Fondon 3rd JW, Minna JD, Garner HR.2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *American Journal of Human Genetics* 67:345-356

Yamada M, Tsuji S, Takahashi H. 2002. Involvement of lysosomes in the pathogenesis of CAG repeat diseases. *Annals of Neurology* 52:498-503

Yin H and Blanchard KL. 2000. DNA methylation represses the expression of human erythropoietin gene by two different mechanisms. *Blood* 95(1):111-119

Yu F, Sabeti PC, Hardenbol P, Fu Q, Fry B, et al. 2005 Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLoS Genetics* 1(3): e41

Zhang L, Yuan D, Yu S, Li Z, Cao Y, Miao Z, Qian H, Tang K. 2004. Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. *Bioinformatics* 20(7):1081-1086

# CHAPTER III

# CHARACTERIZATION OF LONG TRANSCRIBED MICROSATELLITES IN *HELIANTHUS ANNUUS* (ASTERACEAE)

This chapter is a modification of a published work: S Pramod, AB Rasberry, TG Butler and ME Welch. 2011. Characterization of long transcribed microsatellites in *Helianthus annuus* (Asteraceae). American Journal of Botany e388–e390


## Abstract

Premise of the study: Research on the evolutionary role of exonic microsatellites currently lacks from an understanding of the evolutionary pressures that promote or limit their expansion. Contrasting levels of variability and genetic structures at anonymous and transcribed microsatellite loci of varying lengths is likely to provide useful insights regarding the relative strength of selection acting on different classes of microsatellites. We have developed primers for long, transcribed microsatellites in Helianthus annuus to make these comparisons.

Methods and Results: Eight relatively long microsatellites from sequences in the Expressed Sequence Tag database of *H. annuus* were characterized. A total of 63 individuals from three populations in Kansas were genotyped. The number of alleles per locus ranged from four to 11 with an average observed heterozygosity of 0.723. Conclusions: Our study has generated suitable tools for studying the population genetics of long transcribed microsatellites that are potentially influenced by selection.

Keywords: EST; evolution; *Helianthus annuus*; microsatellites; natural populations; transcriptome

## Introduction

The goal of this study is to develop long microsatellite loci from sequences available in the Expressed Sequence Tag (EST) database of *Helianthus annuus* (common sunflower). Microsatellites are typically thought to be associated with non-functional regions of the genome, with few instances of their presence in or near coding regions where they are likely to be disruptive. Examples include diseases such as Huntington's and Spino Cerebellar Ataxias in humans (Ellegren, 2000). In recent years, several studies of EST databases have revealed a large number of microsatellites in transcribed regions. These findings led some to propose that microsatellites might act as "tuning knobs" of genes with microsatellites in 5'UTR and 3'UTR influencing gene expression and microsatellites in coding regions influencing enzyme function by stepwise increases or decreases in length (Li et al., 2004). Further, several empirical studies implicate beneficial roles for microsatellites within genes that are consistent with the tuning knob model (Kashi and King 2006). To explore these hypotheses, we have characterized eight microsatellite loci from sequences in the *H. annuus* EST database that will be useful as molecular makers for a population genetics studies on this system and may allow for gene expression levels (mRNA transcript abundance) to be linked to genotype (microsatellite allele lengths). We targeted long microsatellites located within the UTRs or coding region of sequences in the available *H. annuus* EST database. This database has been successfully used for the characterization of hundreds of shorter microsatellites for various genetic diversity studies and as tools for quantitative trait locus mapping

(Chapman et al., 2008). Because microsatellite length appears to be constrained in most species, we are also interested in studying the dynamics and patterns of selection on longer transcribed microsatellites (>=24 bp) in natural populations of *H. annuus* because this evidence of expansion implies either weaker constraints or directional selection at these loci.

## Methods and results

The EST database of *H. annuus* was downloaded from the NCBI website. The software package Tandem Repeat Finder (TRF) V.4.0.3 (Benson 1999) was used for identifying microsatellites from the database. The position of the start and stop codons in sequences were determined using the software ESTScan V.2 (Iseli et al., 1999; Lottaz et al., 2003). The relative positions of microsatellites as identified in TRF were then used to ascribe regions for each microsatellite (Table 3.1). A BLAST analysis (Altschul et al., 1997) of these sequences was performed against the *Arabidopsis thaliana* database to determine putative function (Table 3.1). Primers for polymerase chain reaction (PCR) amplification were then designed using Primer 3 development software v. 0.4.0 (Rozen and Skaletsky 2000). For performing three-primer PCR (Schuelke, 2000), we added the M13 sequence (5'CAC GAC GTT GTA AAA CGA C 3') at the 5’ end of each forward primer.

Sunflower leaf tissue was collected from three populations in Kansas in September 2010 (Table A.2). Approximately 5-10 mg of leaf tissue was macerated using a Retsch MM200 ball mill (Retsch Incorporated, Newtown, PA); followed by extraction of whole genomic DNA on an ABI prism 6100 Nucleic Acid Prep Station (ABI, Foster City, CA). Three primer PCR was performed in 10 µl volumes with ~10 ng DNA, 2mM

MgCl2, 30mM Tricine (pH 8.4- KOH), 50mM KCl, 100µM of each dNTP, 200 nM of reverse primer and M13 primer labeled with fluorescent dye (HEX, NED or FAM), between 40 and 150 nM forward primer (Table 3.1), and 0.4 U of Taq DNA polymerase. PCR involved an initial denaturation at 95º C for 4 mins, followed by 10 touchdown cycles, with a temperature profile of 94º C for 30 s, 30 s at the annealing temperature (the annealing temperature drops 1 ºC in each subsequent touchdown cycle), and 72º C for 45 s. The remaining 25 cycles had a thermal profile of 94º C for 30 s, 30 s at the optimal annealing temperature (Table 3.1), and 45 s at 72º C. This was followed by a 7 min final elongation period at 72ºC. The fluorescently labeled PCR products were pooled without dilution. Fragment analysis of microsatellites was performed on ABI 3730 capillary sequencers (ABI, Foster City, CA) in the Arizona State University DNA Lab with MapMarker® 1000 as a size standard (Bioventures, Murfreesboro, TN). Alleles were scored using Peak Scanner™ Software v1.0 (ABI, Foster City, CA). A total of 63 individuals from three populations were genotyped for the eight loci. Significance of deviations from Hardy Weinberg equilibrium, expected and observed heterozygosities were calculated (Table 3.2) using GenAlEx v. 6.2 (Peakall and Smouse, 2006). The number of alleles at a locus ranged from 4 to 11 in each population. Expected heterozygosity averaged 0.723. Deviations from Hardy Weinberg equilibrium (HWE) were observed at locus L21 in all populations and locus L1 in population KS3 ($P<0.01$), which may reflect the presence of null alleles. However, other factors including population substructure or locus specific effects may also account for these deviations from HWE.

## Conclusions

Previously, Chapman et al. (2008) developed nearly 500 primers for short microsatellites from the *H. annuus* EST database. We have characterized eight relatively long transcribed microsatellite loci from sequences in the same database. Even though these microsatellites are longer than typical and harbored within the coding region or UTRs; these loci are polymorphic, and heterozygosity is high. Hence, we have developed loci that will be an added resource for genetic analyses in *H. annuus*. Further, contrasting the genetic structure observed at these long transcribed loci to the shorter transcribed loci as well as anonymous microsatellite loci is likely to provide useful insights about the relative strength of selection acting on these three classes of microsatellites. These loci will also be useful for genotyping individuals for studying the dynamics of microsatellite allele length on gene expression levels.

## Acknowledgements

The authors would like to acknowledge; the Office of Research and Economic Development, College of Arts and Sciences, and the Department of Biological Sciences at Mississippi State University for support, North East Mississippi Daily Journal for an undergraduate research grant to TGB, SG Shaak for help with sampling, C Doffit, G Wheeler, and LE Wallace for help with specimen voucher preparation, and AD Perkins for help with the bioinformatic analyses.

Table 3.1 Forward (F) and Reverse (R) Primer sequences, microsatellite repeat motif and copy number, exonic region or location of microsatellite within the gene, expected product size, Genbank accession number of EST sequence, putative function as per BLAST analysis, F primer concentration and annealing temperature ($T_a$) of 8 microsatellite primers developed for *Helianthus annuus* are provided here

| Primer | Primer Sequences | Repeat motif & Copy Number | Region | Product size | Genbank Accession # | Putative function | Forward Primer (μM) | Ta (°C) |
|---|---|---|---|---|---|---|---|---|
| L1 | F: 5'cac gac gtt gta aaa cga ttg aag aac gac acc aga gaa a3'<br>R: 5'tca gct gct gct tca aca tt3' | $(CAC)_{45}$ | Coding | 311 | DY929358.1 | protein kinase family protein | 0.1 | 52 |
| L2 | F: 5'cac gac gtt gta aaa cga cgg ctc aca tgt tgt ctt ca3'<br>R: 5'aag acg aaa aga aat tca aac ca3' | $(GGT)_{39.5}$ | Coding | 380 | DY926758.1 | no significant hits | 0.15 | 52 |
| L3 | F: 5'cac gac gtt gta aaa cga agc acc gac aaa aag aat gg3'<br>R: 5'ttc atc atc gta gtg gtc ttc g3' | $(GAT)_{54.5}$ | Coding | 420 | DY926122.1 | C2 domain-containing protein | 0.08 | 52 |
| L4 | F: 5'cac gac gtt gta aaa cga ccg aca aaa aga atg gca ct3'<br>R: 5'agc acc aaa caa atc cct ga3' | $(GAT)_{51.5}$ | Coding | 476 | DY924930.1 | Heavy metal associated domain containing protein | 0.04 | 52 |
| L28 | F: 5'cac gac gtt gta aaa cga cta aac gtg tcg tgc gat tgt3'<br>R: 5'gct cgt cag agc cca gta tc3' | $(AG)_{14}$ | 5'UTR | 376 | DY907230.1 | unknown or hypothetical protein | 0.15 | 48 |
| L25 | F: 5'cac gac gtt gta aaa cga cga tca acc aaa acc cac cat3'<br>R: 5'gca gaa aga cca gca aga cc3' | $(TC)_{15}$ | 5'UTR | 410 | DY916696.1 | proton-transporting ATPase | 0.1 | 52 |
| L21 | F: 5'cac gac gtt gta aaa cga cga gga ttg gtg gag aaa cga3'<br>R: 5'ccc cac aca ttc tgt ttt ca3' | $(TA)_{14}$ | Coding | 586 | BQ914526.1 | CESA1 (cellulose synthase 1) | 0.1 | 50 |
| L15 | F: 5'cac gac gtt gta aaa cga cgc cat gtt gga gga gac tgt3'<br>R: 5'ccg ctc cct ctt ctt ctt ct3' | $(GA)_{18}$ | Coding | 242 | BU021006.1 | no significant hits | 0.15 | 50 |

Table 3.2 The number of alleles sampled (A) in three populations of *Helianthus annuus* (KS1, KS2 and KS3), the sample sizes of each population (N) as well as observed heterozygosity ($H_o$) and expected heterozygosity ($H_e$) for the 8 primers are provided here

| Primer | Population | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | KS1 (N=13) | | | KS2 (N=18) | | | KS3 (N=32) | | |
| | A | $H_o$ | $H_e$ | A | $H_o$ | $H_e$ | A | $H_o$ | $H_e$ |
| **L1** | 6 | 0.692 | 0.716 | 10 | 0.667 | 0.840 | 10 | 0.700 | 0.836 |
| **L2** | 6 | 0.571 | 0.786 | 10 | 0.533 | 0.793 | 9 | 0.667 | 0.607 |
| **L3** | 8 | 1.000 | 0.780 | 9 | 0.933 | 0.844 | 9 | 0.828 | 0.801 |
| **L4** | 11 | 0.923 | 0.828 | 9 | 1.000 | 0.840 | 8 | 0.897 | 0.771 |
| **L15** | 8 | 0.909 | 0.843 | 11 | 0.944 | 0.867 | 11 | 0.844 | 0.819 |
| **L28** | 7 | 0.818 | 0.781 | 8 | 0.706 | 0.766 | 8 | 0.538 | 0.770 |
| **L25** | 7 | 1.000 | 0.795 | 12 | 0.889 | 0.878 | 9 | 0.935 | 0.832 |
| **L21** | 4 | 0.100 | 0.645 | 7 | 0.071 | 0.737 | 6 | 0.179 | 0.749 |

## References


Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402

Benson G.1999. Tandem repeat finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27(2):573-580

Chapman MA, Pashley CH, Wenzler J, Hvala JJ, Tang S, Knapp SJ, and Burke JM. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *The Plant Cell* 20: 2931–2945

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* 16(12):551-558

Iseli C, Jongeneel CV, and Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*.138-48

Kashi Y and King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* 22(5): 253-259

Li YC, Korol AB, Fahima T, and Nevo E. 2004. Microsatellites within genes: Structure, Function and evolution. *Molecular Biology and Evolution* 21(6):991-1007

Lottaz C, Iseli C, Jongeneel CV and Bucher P. 2003. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* 19:103-112

Peakall R and Smouse PE. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288-295

Rozen S and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Humana Press, Totowa, NJ, pp 365-386

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18:233-234

# CHAPTER IV

# MICROSATELLITES ARE LABILE, MOBILE AND CONSTRAINED


## Abstract

In this study we look for differences between functional and neutral rates of gene flow between populations of *Helianthus annuus*, by estimating allele frequency variation at both transcribed and anonymous microsatellites across populations. Previous work in *Helianthus* implies that neutral rates of gene flow are much greater between species than functional rates of gene flow contributing to the maintenance of species boundaries. By contrasting relative variance in terms of allelic diversity and length of microsatellites found within and among populations we inferred differences in rates of gene flow between regions of the *H. annuus* genome. We also considered a secondary hypothesis that implicates transcribed microsatellites in playing a role in aiding adaptive evolution. A bioinformatics survey of *H. annuus* has shown that the distribution of microsatellites in sequenced transcripts is non-random, and that genes associated with response to stress and stimuli are significantly enriched for microsatellites. We found that heterozygosity at the transcribed microsatellites and the anonymous microsatellites are similar. Further analysis based on allele frequency differences reveals that transcribed microsatellites show patterns consistent with that of balancing selection relative to that observed for anonymous microsatellites. We also found that a similar pattern, yet more extreme, is present when variance in allele lengths is considered in lieu of allelic diversity. This pattern is consistent with allele lengths being under selection. Further, analysis of

population structure suggests that anonymous microsatellites are evolving in a neutral fashion, hence providing accurate estimates of population structure, whereas with transcribed microsatellites we were unable to detect any significant population structure, further supporting role of balancing selection in shaping transcribed microsatellite distributions. Evolution at transcribed microsatellites implies that there are fewer functional differences than neutral differences among populations of *H. annuus* suggesting that selection favors the maintenance of the species across the sampled geographic range.

Keywords: gene flow, microsatellites, *Helianthus annuus*, evolution, selection

## Introduction

Barton and Bengtsson (1986) in their landmark simulation study demonstrated that it is theoretically possible for locus specific selective pressures to result in different realized rates of gene flow across the genome. Their major finding was that neutral alleles would cross readily between hybridizing taxa unless hybrids had effectively no fitness. There is now empirical evidence that functional variation or markers physically linked to functional loci are much less likely to pass between species through hybrid intermediaries. For example, Yatabe et al (2007) considered allele frequency differences among three hybridizing annual sunflowers at anonymous and transcribed microsatellite loci. They found greater similarity among species at anonymous loci when compared to transcribed microsatellite loci. This finding suggests that neutral genetic variation is more likely to move across species boundaries than are functional genes that may define differences between species.

These views on variable realized locus specific rates of gene flow between species contrast starkly with some of the literature on functional rates of gene flow among populations within species (Rieseberg and Burke 2001). These authors have suggested that gene flow estimated at functional loci may be higher than at neutral loci. Hence, populations that share very few migrants may be able to evolve collectively. However, it also seems likely that functional differences among populations would reveal the opposite trend. That is, locally adaptive alleles may serve to enhance allele frequency differences across populations while a more or less free exchange of alleles occurs throughout the majority of the genome.

In this study, we examine similarities and dissimilarities in inferred patterns of gene flow in natural populations of *Helianthus annuus*, using anonymous microsatellite markers that are typically assumed to evolve neutrally, and transcribed microsatellite markers, which are located within genes. If inferred rates of gene flow were to differ significantly between the two classes of microsatellites, it would suggest that populations either have fewer functional differences than neutral differences distinguishing them, or more. If functional similarities exceed that at ostensibly neutral sites, it would suggest that selection serves to maintain species as cohesive units. Conversely, if functional similarities were muted relative to neutrality, the inference would be that selection favors locally adaptive alleles at functional sites even with stable gene flow between populations. Our study aims to address this question by comparing and contrasting the strength of selection acting on transcribed and anonymous microsatellites in *H. annuus.* Although there are several methods available to test if markers are influenced by selection, these methods are not designed to test if the marker is actually under selection.

Through this study we also hope to gain a better understanding of the role of transcribed microsatellites in natural populations, especially the role they may play in generating locally adaptive phenotypic variation. If transcribed microsatellites were to play a role in aiding natural populations adapt to local conditions, we could expect them to play a greater role in abetting species divergence. Microsatellites are typically assumed to evolve neutrally, thus making them a marker of choice for population genetics analyses (Jarne and Lagoda 1996). However, several studies have provided empirical evidence of functional roles for both intronic and exonic microsatellites. These include the association of coding microsatellites with morphological variation in dogs (Fondon and Garner 2004); a 5'UTR microsatellite with the expression of vasopressin and its role in socio-behavioral changes in voles (Hammock and Young 2005); and coding microsatellites with circadian rhythm regulation in *Drosophila melanogaster* (Sawyer et al. 1997). This list is in no way exhaustive. Kashi et al (1997) and Trifanov (2004) have suggested a "tuning knob model" for microsatellites wherein expansion and contraction of the repeat motif functions to tune or modulate gene expression, or enzyme function potentially hastening the rate at which natural populations can evolve more optimal gene function. More recently, Vinces et al (2009) have observed a genome wide association between the presence of microsatellites in upstream regions of genes and greater gene expression variance between these genes in several *Saccharomyces cerevisiae* strains.

Specific examples from the plant kingdom include studies from emmer wheat populations in Israel and Turkey (Fahima et al 2002), and barley populations in "Evolutionary canyon" in Israel (Nevo et al 2005), that have observed intronic as well as exonic microsatellite allele frequency distributions that correlate with environmental conditions. A triplet repeat microsatellite expansion in the first intron of the IIL locus in

*Arabidopsis thaliana* is responsible for the abnormal phenotype, "irregularly impaired leaves", which becomes apparent at higher temperatures (>23°C) and higher light intensities (Sureshkumar et al 2009). Apart from these studies very little is known in terms of functional roles for microsatellites in the plant kingdom.

Early bioinformatic analyses of economically important plant transcriptomes such as cereals (maize, wheat, barley, sorghum and rice) showed that at least 1.5-7.5% of transcripts contain a microsatellite (Li et al. 2002; Li et al. 2004). More recent studies suggest this proportion is much higher with 13.6% of transcripts in the model dicot *A. thaliana* known to contain a microsatellite (Gemayel et al. 2010). A recent bioinformatic survey has shown that nearly 20% of all *H. annuus* transcripts contain a microsatellite (Chapter II). Further, 5'UTRs of genes in *H. annuus* have a greater density of microsatellites, and the microsatellites were enriched within genes belonging to Gene Ontology (GO) terms associated with plant responses to various stresses and stimuli (Chapter II). These results are consistent with observed densities of microsatellites in 5'UTRs of *A. thaliana* and rice (Lawson and Zhang 2006), *Brassica rapa* (Hong et al. 2007), *Medicago truncatula* (Mun et al. 2006). Similarly, GO terms significantly enriched with genes containing microsatellites in rice included response to abiotic stresses (Lawson and Zhang 2006). These results from the *H. annuus* transcriptome suggest that the presence of microsatellites is non random, and may reflect selection favoring their expansion in genes where "tuning" is beneficial.

Although the results from the bioinformatic survey could lend support for the role of transcribed microsatellites in adaptive evolution, and population genetic analyses could be performed to detect signals of selection, it is a difficult task to uncouple the effects of the transcribed microsatellites from linkage effects associated with potential

functional variation elsewhere within genes. Hence, our study mainly focuses on identifying population level differences in selection pressures on transcribed and anonymous microsatellites.

The study system, *H. annuus* is native to North America and is found in diverse habitats (Heiser et al 1969, Kane and Rieseberg 2007). It is a summer annual, with discrete non-overlapping populations. It is also an obligate outcrosser unlike the model dicot, *A. thaliana*, which is a selfing plant. Further, the plant is not known to reproduce clonally. Wild populations are known to harbor high allelelic diversity (Tang and Knapp 2003), with populations in the Midwestern United States (US) harboring the highest genetic diversity (Cronn et al 1997). Hence, its life history traits make *H. annuus* an excellent system to perform population genetic analyses. Natural populations of *H. annuus* can be used to test hypotheses regarding the similarities and dissimilarities in patterns of gene flow between anonymous and transcribed microsatellites as well as the potential role of microsatellites in aiding in adaptation to the wide range of ecological conditions implied by local environmental variation. The clinal differences in microsatellite allele frequencies, observed by Nevo et al (2005) between wheat populations have been inferred as evidence for functional variation at microsatellite loci. This system would also be a good system to compare and contrast to what has been observed by Nevo et al (2005) and Yatabe et al (2007).

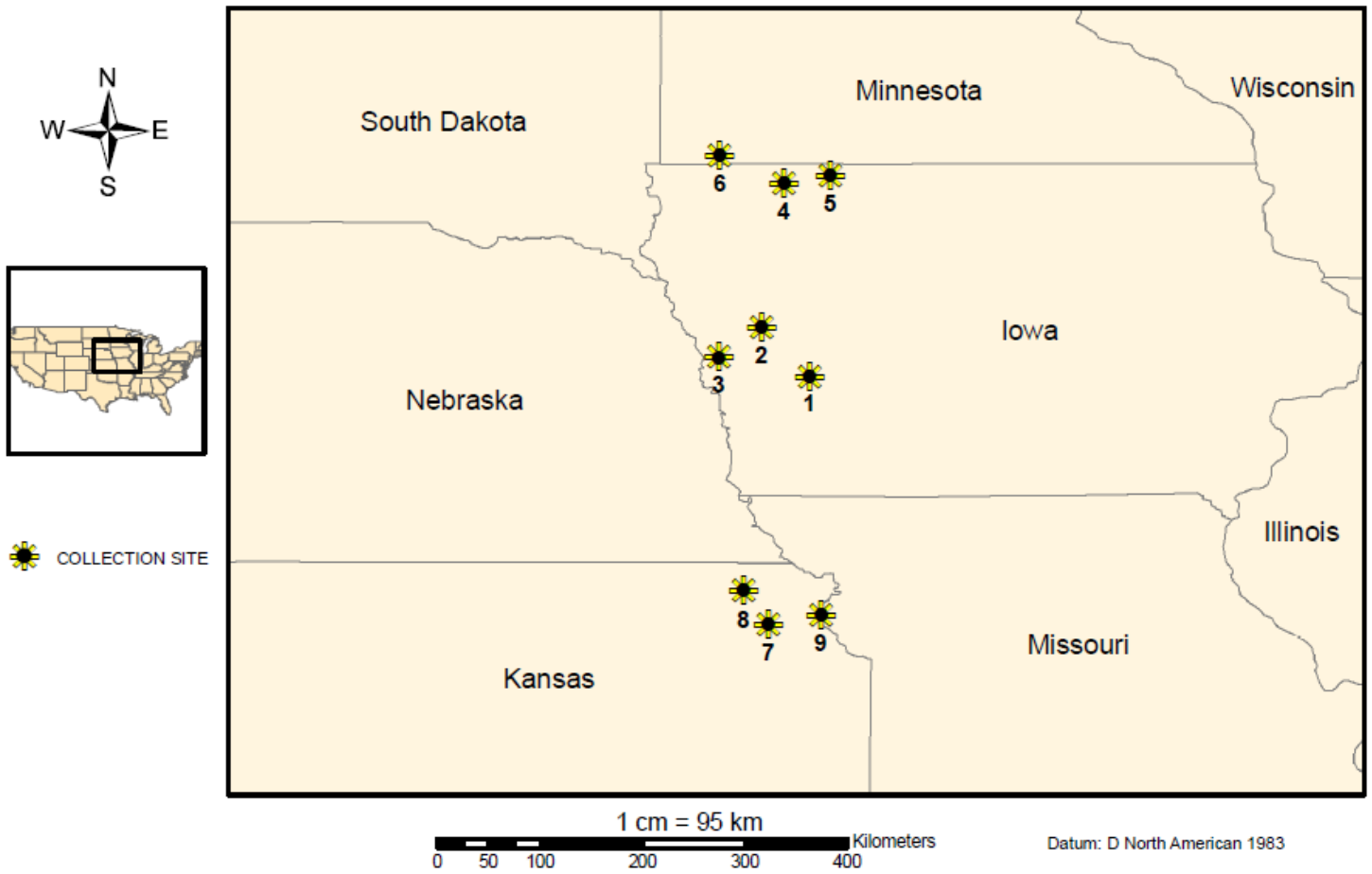


Figure 4.1 Sampling location of the nine *Helianthus annuus* populations are shown in the map. The populations were sampled along a latitudinal cline in seed oil content detailed in Linder (2000) and flowering time (Blackman et al 2011).

## Materials and methods

### Sampling

A total of 450 *H. annuus* individuals were collected from 9 different locations (50 individuals from each location), belonging to three different latitudinal zones in the Midwestern US (Figure 4.1). Since the question we are interested in addressing involves the differences in neutral and functional rates of gene flow, we selected populations that are likely to be experiencing local adaption. These populations were known to have clinal variation in seed oil content (Linder 2000), and flowering time (Blackman et al 2011), which served as proxies for local adaptation.

### Genotyping

Total DNA was extracted from leaf tissue using DNeasy plant preparations (QIAGEN Inc., Valencia, CA). DNA was PCR amplified using 17 anonymous primers using conditions specified in Tang et al (2002) and Yu et al (2002) and ten long transcribed microsatellites developed for this study. Eight of the primers are described in (Pramod et al 2011) and details of the remaining two primers are provided in Table A.3. For the transcribed microsatellites, three primer PCR (Schuelke 2000) was performed in 10 µl volumes with ~10 ng DNA, 2mM $MgCl_2$, 30mM Tricine (pH 8.4- KOH), 50mM KCl, 100µM of each dNTP, 300 nM of reverse and forward primer, in case of the anonymous primers; 200 nM of reverse primer and M13 primer labeled with fluorescent dye (HEX, NED or FAM), between 40 and 150 nM forward primer, in case of the exonic primers (Pramod et al 2011, Chapman et al 2008), and 0.4 U of *Taq* DNA polymerase. PCR involved an initial denaturation at 95º C for 4 mins, followed by 10 touchdown cycles, with a temperature profile of 94º C for 30 s, 30 s at the annealing temperature (the

annealing temperature drops 1ºC in each subsequent touchdown cycle), and 72º C for 45 s. The remaining 25 cycles had a thermal profile of 94º C for 30 s, 30 s at the optimal annealing temperature (Pramod et al 2011), and 45 s at 72º C. This was followed by a 7 min final elongation period at 72ºC. Around five to 6 non overlapping fluorescently labeled PCR products were pooled without dilution in a set. Fragment analysis of these microsatellites was performed on ABI 3730 capillary sequencers (ABI, Foster City, CA) in the Arizona State University DNA Lab with MapMarker® 1000 as a size standard (Bioventures, Murfreesboro, TN). Alleles were scored using Peak Scanner™ Software v1.0 (ABI, Foster City, CA).

**Data analysis**

Allelic diversity, as measured by heterozygosity ($H_e$) of populations, the mean heterozygosity of the subpopulations ($H_S$), and the mean total heterozygosity ($H_T$) were calculated at both transcribed and anonymous loci, since differences in allelic diversity between the transcribed and anonymous loci indicate differential selection pressures acting on them. If drift is shaping population differentiation, then the population differentiation measures should be similar at both transcribed and anonymous loci because neutral processes are likely to have a similar effect on different regions of the genome. Conversely, if the population differentiation measures differ significantly between the sets of loci at these populations, we would expect the role of natural selection in maintaining differential realized rates of gene flow. Both $F_{ST}$, a measure of variance in allele frequencies among populations, as well as $R_{ST}$ which is analogous to $F_{ST}$, but also takes into account the contribution of mutational processes in microsatellites to allele frequency differences, are being utilized as measures of population

differentiation in this study. The genotypes of individuals obtained were used to calculate expected and observed heterozygosity ($H_e$ and $H_o$ respectively) (Nei 1973, Nei 1978) as well as $F_{ST}$ (Weir and Cockerham 1984) and $R_{ST}$ values (Slatkin 1995, Rousset 1996, Goodman 1997) at each locus per population using FSTAT v.2.9.3.2 (Goudet et al 2001).

To test for population substructure among the three latitudinal gradients and among populations, an Analysis of Molecular Variance (AMOVA) was performed using Arlequin v3.5.1.2 (Excoffier and Lischer 2010). If population differentiation measured at the anonymous and transcribed loci, also shows a pattern of isolation by distance, it would suggest that the differences in seed oil content and flowering time at the latitudes are evolving neutrally. The latitude and longitude of each population was employed to create the geographic distance matrix in GenAlEx (Peakall and Smouse 2006). A Mantel's test using 1000 permutations was also performed using Arlequin v3.11 to test for correlation between the geographic distance matrix and genetic distance matrix. Two genetic distance matrices based on both pairwise $F_{ST}$ and $R_{ST}$ were employed for performing the Mantel's test. These two measures were utilized because genetic distance based on $F_{ST}$ and $R_{ST}$ could be different, which could potentially have an influence on the outcome of Mantel's test.

The genotypic data were analyzed using BayesFST (Beaumont and Balding 2004) to identify loci under selection. This approach was taken to test the hypothesis that allele frequencies at transcribed microsatellite loci are shaped by greater selection pressures as compared to anonymous microsatellite loci that are assumed to evolve neutrally. A large fraction of variation in $F_{ST}$-values among loci in populations, is expected to follow patterns indicating neutral evolution. The software, BayesFST identifies a confidence interval for Fst values based on the allele frequency data at the loci. Loci falling within

this confidence interval indicate neutrally evolving loci. Loci that fall outside the confidence interval are statistical outliers. These exhibit extremely high or low values of $F_{ST}$.

To address the question of whether the microsatellites themselves play a functional role, we contrasted variance in allele lengths, or repeat number, at microsatellite loci with allelic diversity. The number of alleles and the variance in repeat number (V) at each locus were calculated using Microsatellite analyzer (MSA) v.4.05 (Dieringer and Schlötterer 2003). The estimates of $H_e$ were used to calculate Schlötterer and Dieringer's (2005) ln *RH* as shown in Equation 4.1.

$$\ln RH=\ln [(((1/(1-H_{\text{Population x}}))^2 -1)/(((1/(1-H_{\text{global Population}}))^2-1)\ ] \quad \text{(Equation 4.1)}$$

The estimates of V were used to calculate Schlötterer's (2002) ln *RV* as shown in Equation 4.2.

$$\ln RV= \ln[(V_{\text{Population x}})/(V_{\text{global population}})] \quad \text{(Equation 4.2)}$$

In both these estimates, the equations simplify down to a ratio of the product of effective population size ($N_E$) and the mutation rate ($\mu$) at that locus. This is because the levels of variation at a locus are a function of the $N_E$ and $\mu$. Since mutation rates at a locus are assumed constant across populations, these tests essentially indicate the differences in $N_E$. For neutrally evolving loci the ratio of $N_E$ should be similar across populations. Typically, the test utilizes a regression between lnRV and lnRH to detect outlier loci. Neutrally evolving loci are expected to have values between -1.96 and 1.96.

Structure v.2.3.3 (Pritchard et al 2000; Hubisz et al 2009) was used to compare population structure observed at anonymous and transcribed microsatellite loci. The most likely number of clusters based on linkage disequilibrium across microstatellite loci was inferred from the data. This is achieved by calculating the likelihood that linkage

relationships between loci can be explained by a set number of populations or clusters (K). A Bayesian approach is used to determine the most likely allele frequencies in the K clusters. Since samples were collected from nine natural populations, a maximum K value greater than nine, the actual number of populations sampled, was chosen. Hence, Structure was set to run 10 iterations at each value of K that ranged from 1 to 15. At each run, the burn in period consisted of 25,000 chains, which was followed by 100,000 Markov Chain Monte Carlo (MCMC) iterations. The results were analyzed using Structure Harvester v0.6.7 (Dent and vonHoldt 2011) to determine the most appropriate number of populations to recognize given linkage relationships among loci according to the method established by Evanno et al (2005).

All aforementioned analyses were performed on each set of microsatellite loci; anonymous, transcribed dinucleotide, and transcribed trinucleotide loci. The transcribed loci were separated out based on the differences in their mean tract lengths. Length of a microsatellite tract could be an important factor influencing its mutation dynamics (Garza et al 1995). The five trinucleotide loci were statistical outliers for length in the *H. annuus* unigene database. Similarly, all dinucleotide loci had comparable tract lengths. Hence, the mean tract lengths of the dinucleotide and trinucleotide microsatellite loci were different enough to warrant their separate analysis. Further, to detect significant differences between anonymous and transcribed microsatellite loci a non-parametric test, , Mann-Whitney U test, that takes into account the differences in sample sizes was performed with a significance threshold or $\alpha$ value of 0.05 using the R statistical package v. 2.11.1 for Windows (R Core Development Team 2010).

## Results

The mean expected heterozygosity ($H_E$) across populations, ranged between 0.49 and 0.94 for the anonymous loci, between 0.78 and 0.90 for the dinucleotide transcribed loci, and between 0.82 and 0.94 for the trinucleotide transcribed loci (Table 4.1). The mean observed heterozygosity at the anonymous loci is 0.82 (± 0.11 s.d.), at transcribed dinucleotide loci heterozygosity is 0.87 (± 0.05 s.d), and at transcribed trinucleotide loci it is 0.90 (± 0.05 s.d.). The mean subpopulation heterozygosity ($H_S$) at the anonymous loci is 0.751 (± 0.104 s.d.), and at transcribed loci is 0.836 (± 0.055 s.d.). The mean total heterozygosity ($H_T$) at the anonymous loci is 0.822 (± 0.111 s.d.), and at the transcribed loci is 0.887(± 0.049 s.d.). The average heterozygosities observed at the two classes of microsatellite loci, transcribed and anonymous, do not differ significantly (Mann-Whitney U test) (Table 4.2). The measure of allele frequency variance as inferred using $F_{ST}$ differs between the anonymous and transcribed microsatellite loci (Mann-Whitney U test, $P$=0.002), with transcribed microsatellite loci exhibiting lower mean $F_{ST}$. However, $R_{ST}$ does not significantly differ between anonymous and transcribed microsatellites (Mann-Whitney U test, $P$=0.06) (Table 4.2).

An Analysis of Molecular Variance (AMOVA) reveals that for the anonymous loci, only 0.32% of the variation in terms of allelic diversity can be attributed to differences among the three latitudinal groups, and 9.2% of the variation is explained by differences among populations within each latitudinal group (Table 4.3a). For the transcribed dinucleotide loci, 1.5% of the variation is attributable to differences among groups and 5.0% to differences among populations within groups (Table 4.3c). Similarly, for the transcribed trinucleotide loci, 1.0% of the variation is attributable to differences among groups and 4.3% to differences between populations within groups (Table 4.3b).

Similarly, results from Mantel's test suggest that isolation by distance does not sufficiently explain the observed genetic variation observed within and among these populations at anonymous or transcribed loci (Table 4.4).

Analysis of allele frequency distributions of the anonymous and transcribed loci in these populations using BayesFST show that none of the anonymous microsatellites display allele frequency differences among populations that deviate significantly from the majority of the loci, but three of the transcribed microsatellite loci (L1, L3, and L21) exhibit significantly low $F_{ST}$-values suggesting that balancing selection may be influencing these loci (Figure 4.2). Although, the results from BayesFST indicate greater selection pressures shaping evolution at transcribed loci, it may not necessarily be indicating direct selection on the microsatellites. The likelihood of the transcribed microsatellites being in linkage disequilibrium with functional variation within genes in which they are located is much greater than that of anonymous microsatellite loci.

The second test for selection, utilized both allele frequencies and variance in repeat numbers and it is implemented as a regression between ln *RV* and ln *RH*. The results from this analysis did not indicate the presence of any outlier loci (Figure 4.3). *lnRV* and *lnRH* are poorly correlated at transcribed loci ($R^2$=0.03) when compared to anonymous loci ($R^2$=0.20; p=0.01). Mean *lnRV* values at transcribed microsatellite loci are significantly different from those at anonymous microsatellite loci (Mann-Whitney U test, *P*=0.0246). Further, the variance at transcribed loci is lower than the variance at anonymous loci at *lnRV* values (Figure 4.4a). At the same time, the mean *lnRH* values (Figure 4.4b) did not differ significantly between the anonymous and transcribed loci (Mann Whitney U test, *P*=0.2701). The variance in distributions of *lnRH* values vary widely in the nine populations at anonymous (Figure 4.5 A) and transcribed (Figure 4.5

B), whereas the variance in *lnRV* values is more uniform at transcribed loci (Figure 4.5 D) compared to anonymous loci (Figure 4.5 C).

Further, an analysis of population structure using STRUCTURE v.2.3.3 indicates that anonymous and transcribed loci identify different number of demes. The anonymous loci indicating strong statistical support for K=11 (Figure 4.9a). Both dinucleotide and trinucleotide transcribed loci providing statistical support for K=2 (Figure 4.9b, 4.9c). The results at transcribed loci essentially indicates the absence of any population structure, which in terms of the K value would indicate a K value of 1. The result indicating K=2 is essentially an artifact of the likelihood method (Evanno et al 2005), as per which probability of a K value is determined by subtracting the likelihood at this K value from the one at the previous K value. The structure results at the anonymous loci show a signal of over-differentiation. For example, individuals that compose population 2 were sampled from two separate clusters on opposite sides of the road at the same location. The individuals at this location show up as two distinct clusters, suggesting that anonymous loci can detect even the remotest signs of geographically based population substructure. At the same time, the allele frequency distributions at the transcribed loci indicate that gene flow is maintaining similar allele frequencies across populations, or that selection within populations has resulted in convergence on alleles of similar lengths. These results provide support for our BayesFST analysis that indicate that neutral forces are shaping evolution of anonymous microsatellite loci, and that balancing selection shapes evolution at functional transcribed loci in these *H. annuus* populations.

Table 4.1 The heterozygosities ($H_E$) observed in each latitudinal group at the transcribed and anonymous microsatellite loci are shown. Mean and standard deviation (s.d.) for each class of microsatellite is also provided.

| Locus Type | Locus | Group 1* | Group 2† | Group 3‡ | Mean | s.d. | Total $H_E$ |
|---|---|---|---|---|---|---|---|
| **Transcribed Trinucleotide** | **Mean** | 0.8303333 | 0.845 | 0.874 | 0.85 | 0.043 | 0.896 |
| | **s.d.** | 0.0876667 | 0.0733333 | 0.067 | 0.076 | 0.026 | 0.048 |
| **Transcribed Dinucleotide** | **Mean** | 0.805 | 0.8263333 | 0.83 | 0.82 | 0.034 | 0.871 |
| | **s.d.** | 0.0303333 | 0.0573333 | 0.0683333 | 0.052 | 0.028 | 0.053 |
| **Anonymous** | **Mean** | 0.7136667 | 0.737 | 0.798 | 0.75 | 0.054 | 0.822 |
| | **s.d.** | 0.1543333 | 0.1256667 | 0.107 | 0.129 | 0.03 | 0.112 |

* includes populations 1, 2 and 3, † includes populations 4, 5 and 6, ‡ includes populations 7, 8 and 9

Table 4.2 Estimates of allelic diversity and genetic variance are provided for the two classes of microsatellite loci; anonymous and transcribed. $H_O$, $H_S$, and $H_T$ represents the observed heterozygosity, average expected heterozygosity of subpopulations, and the total population respectively

| Estimate | Anonymous | | Transcribed | | Mann Whitney U test |
|---|---|---|---|---|---|
| | **Mean** | **Std dev** | **Mean** | **Std dev** | ***P* value** |
| $H_O$ | 0.539824 | 0.143568 | 0.5728 | 0.172043 | 0.5636 |
| $H_S$ | 0.751706 | 0.10396 | 0.8364 | 0.054872 | 0.0223 |
| $H_T$ | 0.822059 | 0.111328 | 0.8867 | 0.049279 | 0.1452 |
| $R_{ST}$ | 0.094059 | 0.055543 | 0.0546 | 0.039643 | 0.0596 |
| $F_{ST}$ | 0.087956 | 0.014591 | 0.06253 | 0.020532 | 0.001757 |

Table 4.3 Results from Analysis of Molecular Variance (AMOVA) detailing the partitioning of genetic variation among the three latitudinal groups, among the three populations within each group, among individuals within populations, and within individuals; at the different classes of microsatellite loci are provided here.

| | Locus Type | Source of Variation | Sum of squares | Variance component | Percentage variation |
|---|---|---|---|---|---|
| a | Anonymous | Among groups | 144.809 | 0.02293 | 0.32423 |
| | | Among populations within groups | 397.682 | 0.65241 | 9.2248 |
| | | Among individuals within populations | 3198.964 | 1.76202 | 24.91403 |
| | | Within individuals | 1898.5 | 4.63503 | 65.53695 |
| | | Total | 5639.955 | 7.07239 | |
| b | Transcribed trinucleotide | Among groups | 25.452 | 0.02296 | 1.01509 |
| | | Among populations within groups | 50.002 | 0.09805 | 4.33573 |
| | | Among individuals within populations | 743.21 | 0.70476 | 31.16489 |
| | | Within individuals | 404 | 1.43564 | 63.48429 |
| | | Total | 1222.663 | 2.26141 | |
| c | Transcribed dinucleotide | Among groups | 30.869 | 0.03274 | 1.48645 |
| | | Among populations within groups | 56.345 | 0.10988 | 4.98918 |
| | | Among individuals within populations | 715.637 | 0.6361 | 28.88143 |
| | | Within individuals | 401.5 | 1.42373 | 64.64294 |
| | | Total | 1204.351 | 2.20246 | |

Table 4.4 Results of the Mantel test performed on each class of microsatellite loci are provided in this table. The table shows the estimate of genetic variation ($F_{ST}$ and $R_{ST}$) used to create the genetic distance matrix, the percentage of genetic variation explained by the geographic distance matrix (%variation explained), the correlation between the genetic and geographic distance matrix ($R$) and the significance value ($P$)

| Locus | Genetic variation estimate | Mantel test results | | |
|---|---|---|---|---|
| | | % explained by geographic matrix | *R* | *P* |
| Anonymous | $R_{ST}$ | 0.006976 | -0.0835 | 0.615 |
| | $F_{ST}$ | 0.047 | -0.217 | 0.77 |
| Dinucleotide Transcribed | $R_{ST}$ | 0.013072 | -0.1143 | 0.661 |
| | $F_{ST}$ | 0.009789 | 0.099 | 0.376 |
| Trinucleotide Transcribed | $R_{ST}$ | 0.18396 | 0.4289 | 0.099 |
| | $F_{ST}$ | 0.003917 | -0.0626 | 0.499 |

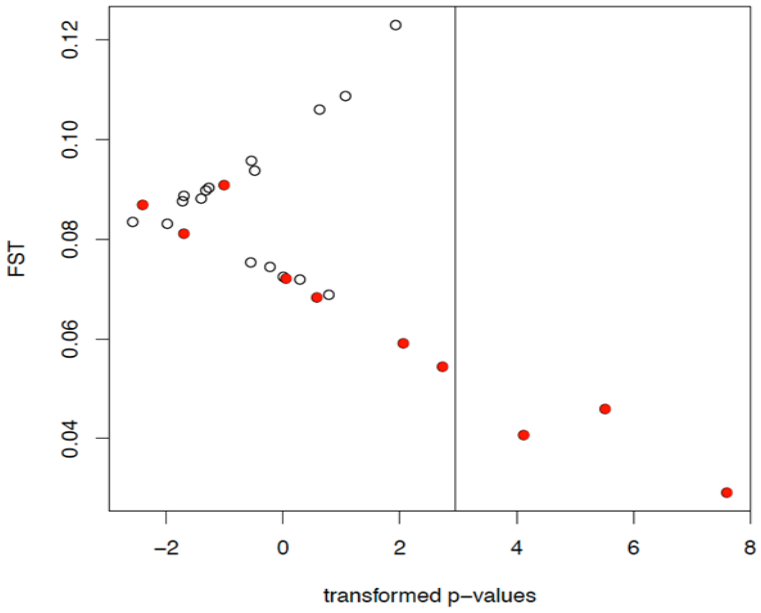


Figure 4.2 The results from BayesFST showing the relationship between Bayesian estimates of $F_{ST}$ and its significance value (transformed $P$ value is calculated as ln($P$/*1*-$P$).

Transcribed microsatellites are indicated by red circles and anonymous microsatellites are indicated by open circles. The vertical line indicates the significance threshold for $F_{ST}$ values. The figure indicates that three of the transcribed loci, which include two trinucleotide loci, L1, L5 and one dinucleotide locus, L21 are outliers for $F_{ST}$ values

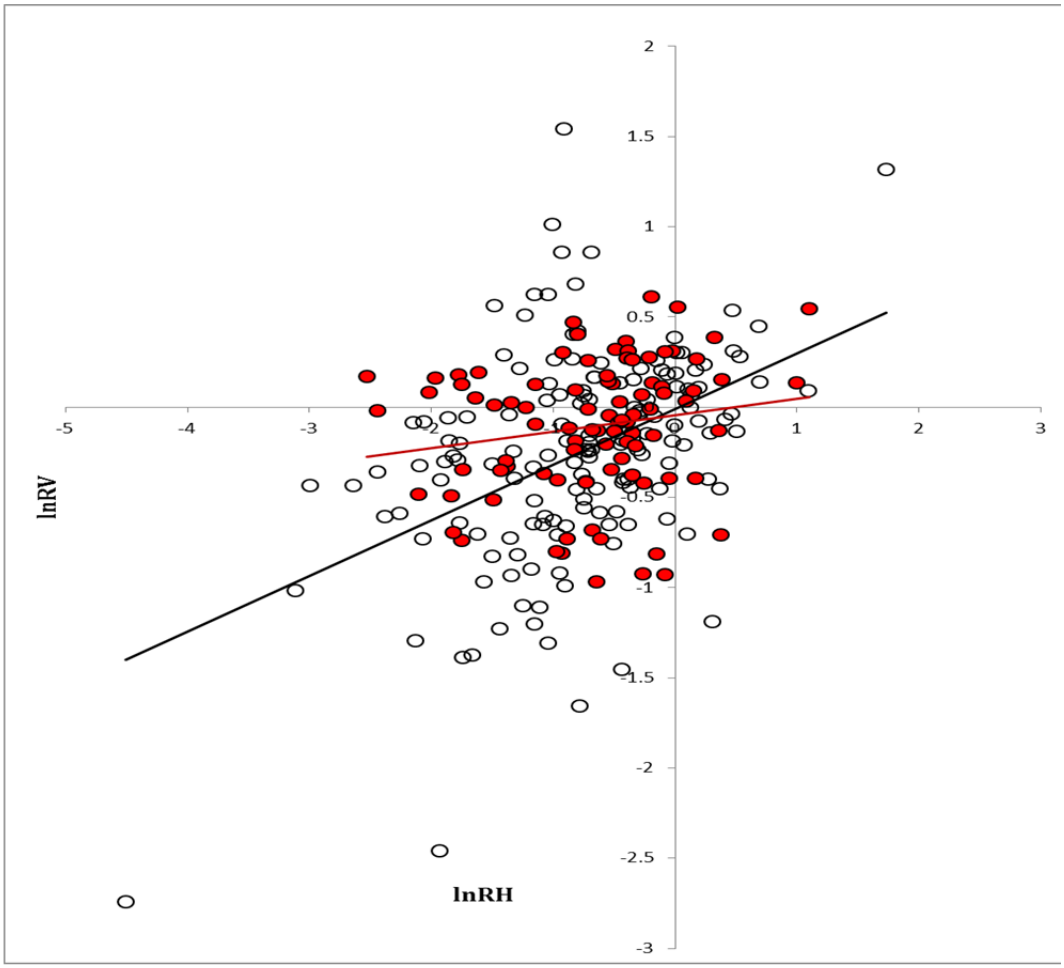


Figure 4.3 The results from a regression between *lnRV* and *lnRH* are shown here. Each data point indicates the regression of *lnRV* to *lnRH* at each population to the global population at a locus.

Anonymous microsatellites are indicated by open circles and transcribed microsatellites are indicated by red filled in circle

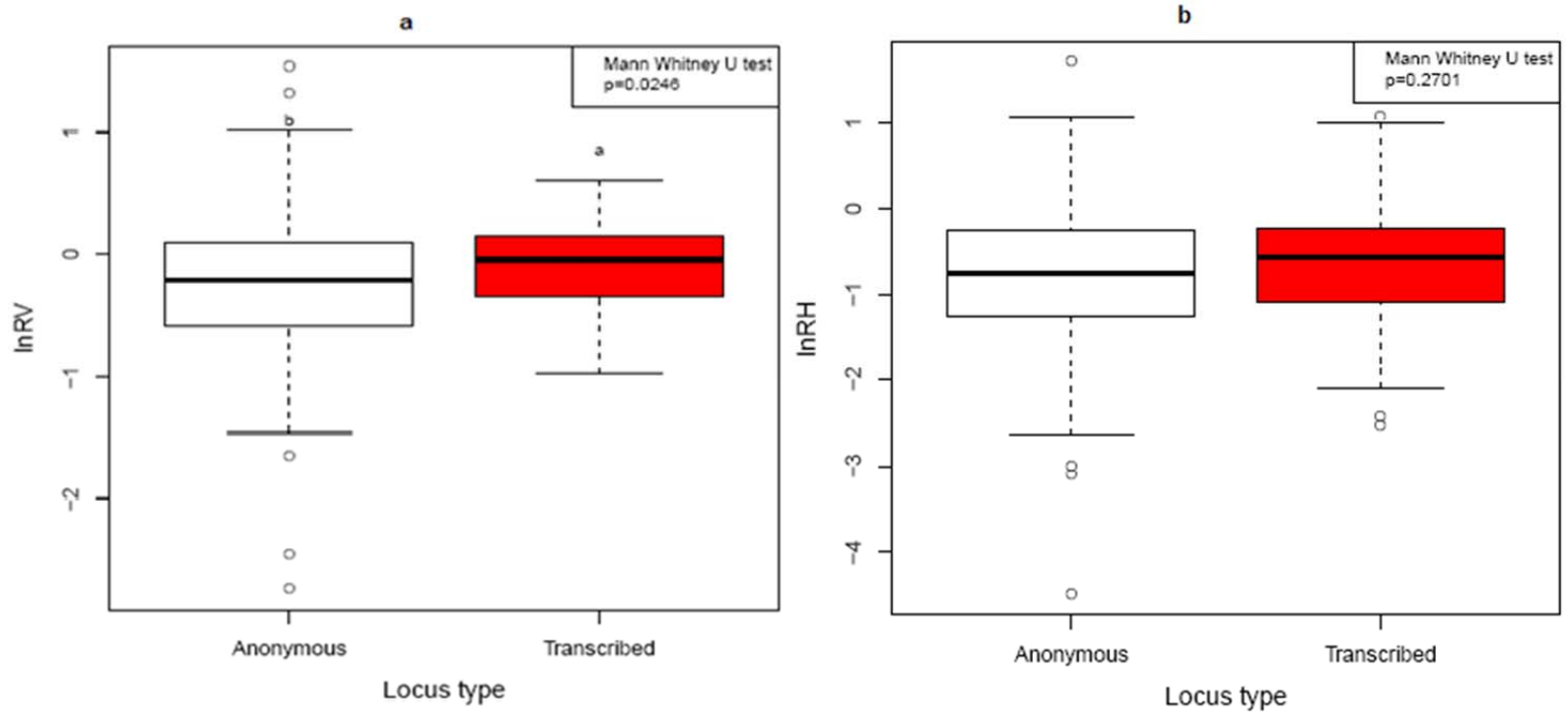


Figure 4.4 The comparisons between anonymous and transcribed microsatellite loci at the *lnRV* values (a) and at the *lnRH* values (b) are provided here. Mann Whitney U tests were performed to assess significant differences between the two classes of microsatellites.

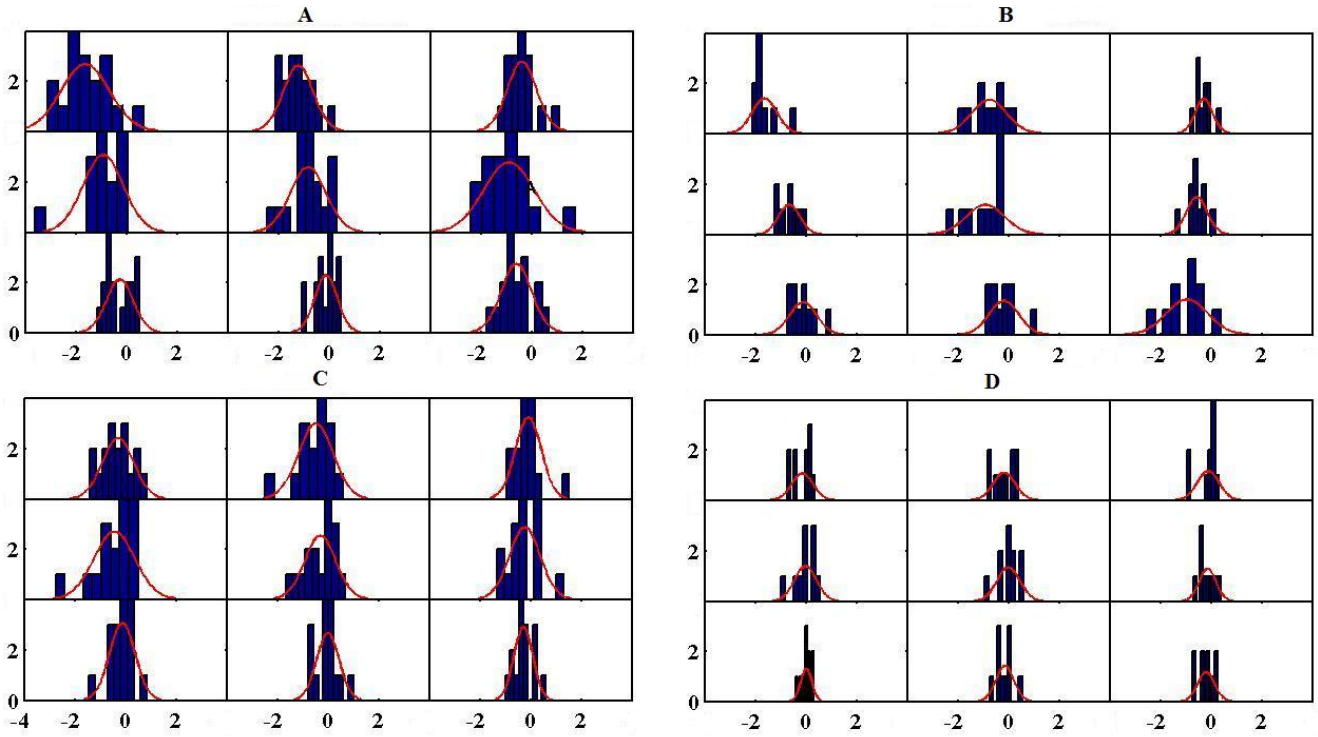


Figure 4.5 The normal distribution curve overlaying the histogram of, A-*lnRH* values at the anonymous loci, B-*lnRH* values at transcribed loci, C-*lnRV* values at anonymous loci, and D-*lnRV* values at transcribed loci in the nine populations is provided here.

The populations are arranged in serial order starting with population 1 in the top leftmost grid and ending with population 9 in the bottom rightmost grid. The x axis in A and B represents *lnRH* values, and in C and D represent the *lnRV* values. The Y axis represents the frequency of the observed *lnRH* or *lnRV* values, Differences between anonymous and transcribed loci at *lnRV* can be noted by contrasting C and D, where D shows a narrow length variance compared to C.

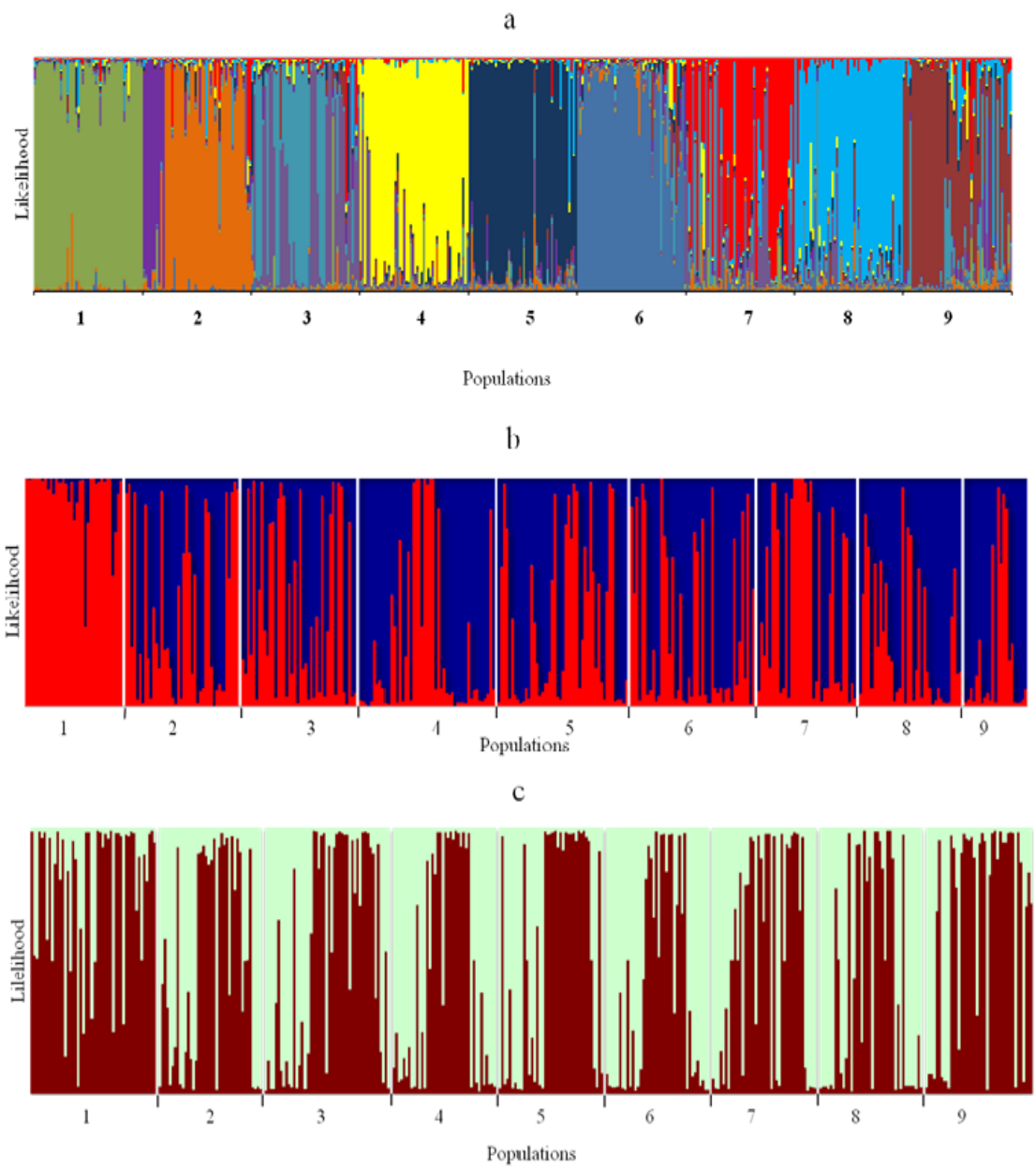


Figure 4.6 The most significant K values, indicating the most likely number of demes based on the type of microsatellite locus, using STRUCTUREv.2.3.3 is shown here.

Each bar on the x axis represents on individual within a specified population. The y axis represents the probability of an individual belonging to one deme or another. Structure from (a) anonymous, (b) transcribed trinucleotide, and (c) transcribed dinucleotide microsatellites are shown.

## Discussion

This study provides support for differential selection pressures acting on anonymous and transcribed regions of the genome. By contrasting patterns of microsatellite allele frequency variation at the anonymous and transcribed loci, we addressed two distinct but related hypotheses. The first hypothesis is that selection will either facilitate or impede the spread of alleles underlying functional variation, if only a few universally fit alleles are favored or if different alleles are favored in different populations, respectively. This hypothesis was assessed by contrasting realized rates of gene flow between anonymous and transcribed microsatellite markers. Data show that the allelic diversity at transcribed microsatellite loci was comparable to that observed at anonymous microsatellite loci. This is a marked difference from what has been observed in previous studies that have compared various species in the genus *Helianthus*. For example, Yatabe et al (2007) and Pashley et al (2006) have observed a significantly greater diversity for anonymous microsatellite loci, compared to transcribed loci among various *Helianthus* species, which lends support to the role of functional differences between species in helping maintain species differences. A study by Kane et al (2009) suggests a significant amount of functional gene flow between geographically proximal populations of sympatric species. The allopatric species pair showed a limited amount of gene flow at the same loci. Our results also indicate that allele frequency variance among populations is significantly lower at transcribed loci compared to the anonymous microsatellite loci. Hence, our study in conjunction with studies by Yatabe et al (2007), Pashley et al (2006), and Kane et al (2009) could be indicating that differences in functional and neutral rates of gene flow exist.

The second hypothesis addressed involves a functional role for transcribed microsatellites in generating adaptive evolutionary potential. Although quite a few empirical studies have indicated that transcribed microsatellites could have a functional role associated with them (Gemayel et al 2010), only a few handful of these look at more than one gene in the organism (Fondon and Garner 2004; Vinces et al 2009). Such studies are also currently lacking in plants, which are potentially constantly required to evolve adaptive responses as they lack mobility unlike animals. Hence, our study is also a novel attempt to identify contributions of transcribed microsatellites as tuning knobs in *H. annuus.* The results from our study make sense in light of both our hypotheses. For example, the $F_{ST}$ values were significantly lower at transcribed microsatellite loci relative to anonymous microsatellite loci, suggesting greater differentiation at anonymous microsatellite loci in these populations. We had expected to see greater differentiation at the transcribed loci, given our choice of populations that spanned the cline known to have variation in Oleic and Linolenic acid content in seed oil (Linder 2000), and the greater likelihood of these loci to be associated with functional genetic variation. However, the lower $F_{ST}$ values at transcribed microsatellite loci indicate that allele frequencies are similar among populations, suggesting maintenance of globally advantageous alleles across the sampled range. Further, BayesFST also detected three of the transcribed loci with extremely low $F_{ST}$ values to be statistical outliers. The pattern of $F_{ST}$ values depicted in Figure 4.2 indicates differences in evolution of transcribed and anonymous loci, with anonymous loci exhibiting higher $F_{ST}$ values relative to transcribed loci. The extremely low $F_{ST}$ values observed at these transcribed loci could possibly be indicating greater gene flow and hence, the role of balancing selection in shaping allele frequency distributions at these loci. This could be suggesting that functional rates of gene flow are

much greater within a species at functional loci relative to neutral loci increasing the likelihood that geographically isolated populations can evolve as a cohesive unit. This could particularly help explain the cohesiveness of *H. annuus*, a species with a very broad geographic range. The analysis to identify the statistically significant number of population clusters supports our BayesFST results, with anonymous loci almost perfectly inferring the number of clusters (Figure 4.9a), as opposed to the transcribed loci that are more consistent with the absence of genetic structure (figure 4.9b, Figure 4.9c). It was particularly interesting that genetic structure assessed using anonymous loci revealed even the minutest signatures of population structure, with identification of two distinct units within two of the sampled populations that was originally sampled from two distinct clusters in those general locations. This is consistent with anonymous microsatellite loci evolving nearly neutrally. The absence of population structure at the transcribed loci, also indicate the role of balancing selection in shaping allele frequencies in these populations. Both the BayesFST and STRUCTURE analyses utilize allele frequency data. Outliers detected by BayesFST indicate disparate selection pressures on these loci, however the microsatellites in question may not be the target of selection and it could be argued that it is merely an artifact of the microsatellites proximity to genes, which could be the actual unit under selection.

We were also interested in testing if the lower $F_{ST}$ values at transcribed microsatellites compared to the anonymous microsatellites could be due to the role of these transcribed microsatellites as tuning knobs where microsatellite allele length differences aid in adaptive processes, helping natural populations adjust or "tune" their phenotypes, via changes in allele lengths, in response to environmental change. The correlation analysis between *lnRV* and *lnRH* at anonymous and transcribed loci that

utilizes variance in allele lengths along with variance in allelic diversity was primarily used to infer whether selection was acting on a microsatellite. Although, the analysis did not detect outlier loci, a significant difference between the clustering of the anonymous and transcribed loci was noted (Figure 4.3). The results show a tight relationship between *lnRV* and *lnRH* for the anonymous loci and an absence of correlation for the transcribed loci. *lnRH* values, measures of allelic diversity were comparable between anonymous and transcribed loci. However, *lnRV* values were significantly different between the transcribed loci and anonymous loci, indicating that much of the difference between transcribed and anonymous loci is more extreme when differences in allele length variance is partitioned among populations than when allelic diversity is considered (Figure 4.4a, Figure 4.4b). The variance in *lnR*V values is lower at the transcribed loci compared to the anonymous loci. This further suggests that allele length diversity at transcribed loci is comparable across populations from the entire range regardless of differences in connectivity and local population size. These results also support our BayesFST findings, which indicate reduced average $F_{ST}$ values at transcribed loci. The presence of significantly low *lnRV* values at transcribed loci and its poor correlation with *lnRH* values indicates that an even greater proportion of variation in allele lengths relative to allelic diversity is housed within local populations, which could be indicating that allele length is under selection. The strong indication of difference between transcribed and anonymous microsatellites at *lnRV* values provides support for the prevalence of similar allele lengths across populations; thereby providing support for the tuning knob model, wherein microsatellite allele lengths are functional. If linkage disequilibrium were responsible for the more moderate genetic structure observed at transcribed microsatellites we anticipate that the length of microsatellite alleles that are in

linkage with adaptive variation elsewhere in genes to be a more random relationship. Hence, our finding from this analysis could be lending support to our hypothesis regarding selection acting directly on microsatellites possibly due to beneficial roles associated with microsatellite allele lengths in disparate environmental conditions, rather than selection acting on genes to which these microsatellites are physically linked to.

At the same time, the differences observed in our data could be due to our choice of microsatellites. Most population genetic studies employ short perfect microsatellites as markers of neutral rates of gene flow. However, we selected long microsatellite tracts in transcribed loci. The five trinucleotide microsatellites selected were statistical outliers for length in the *H. annuus* unigene database, with mean copy number ranging between 31-55 repeats. Trinucleotides form the single largest motif size class within the coding region in genes of *H. annuus* (Chapter II), which also supports the prediction by Metzgar et al (2000) and Wren et al (2000) as per which expansion or contraction on a trinucleotide microsatellite are more tolerated because these are less likely to cause frameshift mutation errors. These tracts also harbored a greater than average proportion of impurities. The remaining five loci were all dinucleotides with mean copy number ranging from 14 to 24 repeats. The greater than usual length range could have contributed to elevated mutation rates, which is observed in the form of greater than average genetic diversity compared to other studies that have observed a reduced allelic diversity in transcribed microsatellites (Pashley et al 2006, Yatabe et al 2007). At the same time, these trinucleotide microsatellite tracts harboring a greater number of impurities are expected to harbor lower mutation rates than pure tracts of comparable lengths (Yu et al 2005). However, the mean length of the pure stretch on the otherwise impure tract could be over the threshold length required for attaining the mutation rates

observed in our dataset. The fact that seven out of these 10 transcribed microsatellites are located in coding regions is interesting since these regions are expected to be highly conserved, especially within species. The remaining transcribed loci are located within the 5'UTR. Hence, our results indicate differential selection pressures on transcribed and anonymous markers chosen in this study. Population genetic studies utilizing shorter transcribed microsatellites that are comparable in size to anonymous microsatellites are needed to ascertain if all transcribed microsatellite markers may be under selection for tract lengths. Further, the availability of *next-generation sequencing* (NGS) data may help in providing conclusive support for the tuning knob hypothesis by looking at whole transcriptome sequences and linking transcript abundance to the presence or absence of microsatellite tracts as well as differences in microsatellite tract length.

**Acknowledgements**

The authors would like to thank members of the Wallace and Welch lab and Dr. BA Counterman at Mississippi State University, for helpful suggestions on improving this manuscript.

## References


Barton N and Bengtsson BO. 1986. The barrier to genetic exchange between hybridising populations. *Heredity* 56:357-376

Beaumont MA and Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* 13:969–980

Blackman BK, Michaels SD, and Rieseberg LH. 2011. Connecting the sun to flowering in sunflower adaptation. *Molecular Ecology* 20: 3503-3512

Cronn R, Brothers M, Klier K, Bretting PK, Wendel JF. 1997. Allozyme variation in domesticated annual sunflower and its wild relatives. *Theoretical Applied Genetics* 95:532-545

Dent EA and vonHoldt BM. 2011. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* DOI: 10.1007/s12686-011-9548-7 Version: v0.6.8 Oct 2011

Dieringer D and Schlötterer C. 2003. Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes* 3 (1):167-169

Excoffier L and Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564-567

Evanno G, Regnaut S, and Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14:2611 – 2620

Fahima T, Röder MS, Wendehake VM, Nevo E. 2002 Microsatellite polymorphism in natural populations of wild emmer wheat, *Triticum dicoccoides*, in Israel. *Theoretical Applied Genetics* 104:17–29

Fondon JW III and Garner HR. 2004. Molecular origins of rapid and continous morphological evolution. *Proceedings of the National Academy of Sciences* 99:1991-2004

Garza JC, Slatkin M, Freimer NB. 1995. Microsatellite allele frequencies in Humans and Chimpanzees with implications for constraints on allele size. *Molecular Biology and Evolution* 12:594-603

Gemayel R, Vinces MD, Legendre M and Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* 44:445-477

Goodman SJ. 1997. $R_{ST}$-calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology* **6**: 881-887

Goudet J. 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html. Updated from Goudet 1995

Heiser CB, Smith DM, Clevenger S, Martin WC. 1969. The North American Sunflowers (*Helianthus*). *Memoirs of the Torrey Botanical Club* 22:1-218

Hong CP, Piao ZY, Kang TW, Batley J, Yang TJ, Hur YK, Bhak J, Park BS, Edwards D, Lim YP. 2007. Genomic Distribution of Simple Sequence Repeats in *Brassica rapa*. *Molecules and Cells* 23(3):349-35

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* 9:1322-1332

Jarne P and Lagoda PJL. 1996. Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution* 11: 424 – 429

Kane NC, Rieseberg LH. 2007. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* 175: 1823-1834

Kane NC, King MG, Barker MS, Raduski A, Karrenberg S, Yatabe Y, Knapp SJ, Rieseberg LH. 2009. Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* 63(8):2061-2075

Kashi Y, King DG, Soller M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics* 13:74-78

Lawson MJ, Zhang L. 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biology* 7(2):R14

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* 11: 2453-2465

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: Structure, Function and evolution. *Molecular Biology and Evolution* 21(6):991-1007

Linder CR. 2000. Adaptive evolution of seed oils in plants: accounting for biogeographic distribution of saturated and unsaturated fatty acids in seed oils. *American Naturalist* 156: 442-458

Nei M.1973. Analysis of gene diversity in subdivided populations. *Proceedings of the Natural Academy of Sciences USA* 70:3321–3323

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89(3):583-590

Nevo E, Beharev A, Meyer RC, Hackett CA, Forster BP, Russel JR, Powell W. 2005 Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel. *Biological Journal of the Linnean Society of London*. 84: 205–224

Pashley CH, Ellis JR, Mccauley DE, and Burke JM. 2006. EST Databases as a source for molecular markers: Lessons from *Helianthus*. *Journal of Heredity* 97(4):381–388

Peakall R and Smouse PE. 2006. GENALEX 6:genetic analysis in excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288-295

Pramod S, Rasberry AR, Butler TG, Welch ME. 2011. Characterization of long transcribed microsatellites in *Helianthus annuus* (Asteraceae). *American Journal of Botany* e388–e390

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959

Rieseberg LH, Burke JM. 2001. The biological reality of species: gene flow, selection, and collective evolution. *Taxon* 50: 47-66

Rousset F. 1996. Equilibrium values of measure of population subdivision for stepwise mutation processes. *Genetics* 142: 1357–1362

Schlötterer C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160: 753–763

Schlötterer C and Dieringer D. 2005. A novel test statistic for the identification of local selective sweeps based on microsatellite gene diversity. In *Selective Sweep*, D. Nurminsky, ed (Boston: Kluwer Academic Publishers), pp. 55–64

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18: 233–234

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457-462

Sureshkumar S, Todesco M, Schneeberger K, Harilal R, Balasubramanian S, and Weigel D. 2009. A Genetic Defect Caused by a Triplet Repeat Expansion in Arabidopsis thaliana. *Science* 323 (5917):1060-1063

Tang S, Yu JK, Slabaugh MB, Shintani DK, and Knapp SJ. 2002. Simple sequence repeat map of the sunflower genome. *Theoretical Applied Genetics* 105: 1124–1136

Tang S and Knapp SJ. 2003. Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflower. *Theoretical Applied Genetics* 106:990–1003

Trifanov EN. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation; In "*Evolutionary theory and processes: Modern Horizons, papers in honor of Eviatar Nevo"* (Ed. Wasser SP); Kluwer Academic Publishers. pp 115-138

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science* 324:1213-1216

Weir BS and Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38:1358-1370

Yu JK, Mangor J, Thompson L, Edwards KJ, Slabaugh MB and Knapp SJ. 2002. Allelic diversity of simple sequence repeat markers among elite inbred lines in cultivated sunflower. *Genome* 45: 652–660

Yu F, Sabeti PC, Hardenbol P, Fu Q, Fry B, et al. 2005 Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLoS Genetics* 1(3): e41

# CHAPTER V

# GENE EXPRESSION ASSAYS FOR ACTIN, UBIQUITIN AND THREE MICROSATELLITE ENCODING GENES IN *HELIANTHUS ANNUUS* (ASTERACEAE)


## Abstract

*Premise of the study:* The "tuning knob" model postulates that microsatellite mutations can alter phenotypes in a stepwise fashion. Some proposed mechanisms involve regulation of gene expression. To study the effect of microsatellites harbored in untranslated regions on gene expression in *Helianthus annuus,* we have developed Taqman assays for three microsatellite encoding genes, and two constitutively expressed genes*, actin* and *ubiquitin* to serve as standards.

*Methods and Results:* To provide reliable gene expression quantification, Taqman assays were designed from expressed sequence tag based unigene sequences that contained multiple redundant sequences and were part of contigs. All five Taqman assays yielded strong log-linear relationships between $C_T$-values and cDNA concentrations ($R^2$ = 0.98-0.99). Standard curves were based on five concentrations for each of five individuals. Efficiencies ranged from 0.83-1.03.

*Conclusions:* The developed tools will allow for relative quantification of gene expression across individuals. Genotyping these loci will allow for testing the "tuning knob" hypothesis. Further, the *actin* and *ubiquitin* assays should be generally applicable to gene expression studies in *H. annuus*.

Keywords: *Helianthus annuus*; microsatellites; Taqman assays; untranslated regions

## Introduction

The goal of this study is to develop Taqman assays from the Expressed Sequence Tag (EST) database of *Helianthus annuus* (common sunflower), for further research on the role of microsatellites from the untranslated regions (UTRs) in gene regulation. Although microsatellites were classically viewed as neutral molecular markers, a recent comprehensive review by Gemayel et al. (2010) lists several empirical studies that have implicated microsatellites in phenotype modulation, by means of gene expression regulation or structural changes in proteins. The initial discovery that microsatellites were common in transcribed regions of genomes led to the development of the "tuning knob" model that suggests allele length can have a stepwise influence on phenotypes (Kashi et al., 1997; Trifanov, 2004). Some molecular mechanisms proposed include incremental changes in gene expression levels. Microsatellites in 5' and 3'UTRs of transcripts are deemed more likely to serve this function while coding microsatellites are more likely to influence enzyme function by stepwise expansion or contraction of the homopolypeptide tracts (Gemayel et al., 2010). To test the gene regulatory role of UTR microsatellites, we have developed three Taqman gene expression assays and two control assays to quantify baseline expression levels in *H. annuus* individuals. Actin and ubiquitin homolog sequences in *H. annuus* were utilized for designing the standard or control assays. The ultimate objective of the study is to design assays for microsatellite containing genes whose gene expression levels can be linked to microsatellite allele lengths. Primers and protocols for genotyping numerous transcribed microsatellites were previously developed

from EST databases available for *H. annuus* (Chapman et al., 2008; Pramod et al., 2011). The markers detailed in these studies typically exhibit high levels of polymorphism in natural populations. Since we are interested in ultimately developing a model to predict the relationship between microsatellite allele length and gene expression at a locus, we chose markers with fewer alleles. Ultimately, this strategy should facilitate the acquisition of larger sample sizes for each allele in a population.

## Methods and Results

Microsatellite markers developed by Chapman et al. (2008), with relatively lower variation in three natural populations of *H. annuus* (Table A.2) were identified. Selected sequences were retrieved from the unigene build used by Chapman et al. (2008) and the position of the start and stop codons in these sequences were determined using the software ESTScan V.2 (Iseli et al., 1999; Lottaz et al., 2003). Since, we were interested in quantifying effects of microsatellites on gene expression, only sequences with microsatellites in 5' or 3' UTRs were considered. Three microsatellite containing unigene sequences based on redundant contigs were identified for designing Taqman assays. Contigs were used preferentially because unigenes known only from singletons tend to express higher proportions of sequencing errors. *Helianthus annuus* unigene sequences of constitutively expressed genes, *Actin* and *Ubiquitin* were selected as standards or control assays. Identification of control genes, and the putative function of the microsatellite encoding genes, was determined by sequence homology with the annotated *Arabidopsis thaliana* genome database using BLAST (Altschul et al., 1997). Taqman® gene expression assays, presented in Table 1, were designed for the selected five unigene sequences using Primer Express v.3.0 (Applied Biosystems, Foster City, CA). The assay

probes are ZEN double quenched probes, that include an internal ZEN quencher, a 3'Iowa Black forward quencher (IABkFQ) and a 5'6-FAM reporter (Integrated DNA Technologies, Coralville, IA).

Young sunflower leaf tissue was collected from the three populations in Kansas detailed in Pramod et al. (2011). The voucher information for representative individuals from these populations is provided in Table A.2. Immediately after harvesting the tissue was stored in RNALater® (Ambion, Grand Island, NY). Approximately 50-65 mg of leaf tissue was used for RNA extraction using the Ambion® RNAqueous®-4PCR kit and protocol (Life Technologies, Grand Island, NY). Extracted mRNA was converted to cDNA using the High Capacity cDNA Reverse Transcription kit with RNase inhibitor (Applied Biosystems, Foster City, CA). For generating standard curves, five different concentrations were generated by means of 1:1 serial dilutions of cDNA samples with $ddH_2O$ for a total of five individuals collected from three natural populations. Real Time-PCR was performed in 10 µl reactions on an ABI StepOne$^{TM}$ Real-time PCR System (ABI, Foster City, CA). Reactions mixes contained 5 µl (1x concentration) of 2x iTaq$^{TM}$ supermix with ROX (Bio-Rad, Hercules, CA), 1 µl (1x concentration) of the Taqman assay, 3 µl of $ddH_2O$, and 1 µl of cDNA sample. The amplification profile included a 2 minute hold at 50°C and an initial denaturation at 95°C for 3 minutes, followed by 40 cycles of denaturation at 95°C for 30 seconds and annealing and elongation at 72°C for 40 seconds. The resultant $C_T$ values were used to calculate standard curves by means of ANCOVA with individual as the discrete variable and concentration as the continuous variable. All five assays revealed very strong log-linear relationships between $C_T$-values and concentration. The coefficient of determination ($R^2$) for each assay ranged from 0.98 to 0.99, and efficiencies (b) ranged from 0.83 to 0.99. See Table 2 for details on each

individual assay. Statistical analyses were conducted using JMP® 9 Pro (SAS Institute Inc., Cary, NC).

## Conclusions

We have developed Taqman gene expression assays from sequences derived from multiple accessions of the *H. annuus* EST sequence based unigene database. These assays proved reliable for relative quantification of gene expression across biological replicates from multiple wild populations, indicating that the assays detailed in this study should provide a reliable means of quantifying relative gene expression levels across both domesticated and wild *H. annuus*. Two of the assays, *Actin* and *Ubiquitin*, were designed to serve as constitutively expressed standards or controls. These two assays should be generally applicable for any study requiring relative gene expression quantification across multiple individuals of *H. annuus*. Further, the three Taqman assays designed for genes harboring microsatellites in the UTRs provide us with a means for studying the potential gene regulatory role of microsatellites in natural populations. Studying the relationship between microsatellite allele length and gene expression at these loci is now feasible because primers and protocols required for genotyping the microsatellites encoded by these genes are already available (Chapman et al., 2008).

## Acknowledgements

The work was supported and funded by the Department of Biological Sciences, College of Arts and Sciences, and the Office of Research and Economic Development at Mississippi State University

Table 5.1 Forward (F), Reverse (R) and Probe† (P) assay sequences, a description identifying if the assay is for a constitutively expressed sequence or a gene under regulatory control of microsatellite, amplicon length, sequence accession number in Chapman et al (2008) *Helianthus annuus* unigene build* or the current unigene database, and putative function as per BLAST analysis, of 5 Taqman assays developed for *H. annuus* are provided here.

| Assay | Taqman Assay Sequences | Assay Description | Amlicon Length (bases) | Sequence Accession # | Putative Function |
|---|---|---|---|---|---|
| c1181 | P: /5'6-FAM/ACA CAT CTC /ZEN/AAC CAC TCC CTC ATT CCC C/3'IABkFQ/<br>R: 5'GCT GTC GGC TGG GCT TAA G3'<br>F: 5'TTT CAG CAA ACC ACA CAA GCA3' | $(GA)_5$ in 5'UTR | 70 | QH_CA_Contig1181* | no significant hits |
| c3115 | P: /5'6-FAM/CCG GCC GAT /ZEN/GAC TCA TGT CGC /3'IABkFQ/<br>R: 5'CAG AAT GGA CGT GAA ACC TCA A3'<br>F: 5'GTG GGA CCG GCG ATT GT3' | $(TG)_8$ in 3'UTR | 69 | QH_CA_Contig3115* | nicotinate phosphoribosyltransferase |
| c5774 | P: /5'6-FAM/AGA TTG AAG /ZEN/GCA ACC ACC TCA CTG CTG /3'IABkFQ/<br>R: 5'GCC GTC ATG ACC GAT AAT AGC3'<br>F: 5'CGT CGA TGA CCA CTT GAT GTG3' | $(TC)_{11}$ in 5'UTR | 73 | QH_CA_Contig5774* | putative profilin protein |
| *Actin* | P: /5'6-FAM/AAG GTT ATG /ZEN/CAC TCC CCC ATG CCA TC/3'IABkFQ/<br>R: 5'GAC CAG CGA GAT CAA GAC GAA3'<br>F: 5'GGT GTG TCA CAT ACA GTT CCA ATT TAT3' | Control assay | 76 | Han#S18764473 | Sunflower actin, structural constituent of cytoskeleton |
| *Ubiquitin* | P: /5'6-FAM/TGC GTT CAT /ZEN/GGA CCG GCA CC/3'IABkFQ/<br>R: 5'GTA CAG AGT TGA GAG GAC CGA TGA3'<br>F: 5'GAT GGA TGA TGG AGA CGA CAT TT3' | Control assay | 71 | Han#S32095209 CHAY9826.b1_C10.ab1 | Ubiquitin, protein binding |

†ZEN double quenched probe with ZEN as an internal quencher, a 3' Iowa Black forward quencher (IABkFQ), and 5' 6-FAM reporter

Table 5.2 The correlation coefficient ($R^2$), efficiency (b), the mean $C_T$ values, and the intercept obtained in *Helianthus annuus* individuals at each assay are provided here.

| Assay | $R^2$ | b | Mean $C_T$ | Intercept |
|---|---|---|---|---|
| C1181 | 0.98 | 1.03 | 31.76 | 33.90 |
| C3115 | 0.98 | 0.89 | 32.72 | 34.63 |
| C5774 | 0.99 | 0.91 | 28.97 | 30.79 |
| *Actin* | 0.99 | 0.83 | 31.30 | 32.97 |
| *Ubiquitin* | 0.99 | 0.85 | 31.16 | 32.86 |

## References


Altschul SF., Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, and Burke JM. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *The Plant Cell* 20: 2931–2945

Gemayel R, Vinces MD, Legendre M, and Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* 44:445-477

Iseli C, Jongeneel CV, and Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*.138-48

Kashi Y., King DG, and Soller M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics* 13:74-78

Lottaz C, Iseli C, Jongeneel CV, and Bucher P. 2003. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* 19:103-112

Pramod S, Rasberry AB, Butler TG and Welch ME. 2011. Charecterization of long transcribed microsatellites in *Helianthus annuus* (Asteraceae). *American Journal of Botany* 98(12): e388-e390

Trifanov EN. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation; In "*Evolutionary theory and processes: Modern Horizons, papers in honor of Eviatar Nevo"* (Ed. Wasser SP); Kluwer Academic Publishers. pp 115-138

# CHAPTER VI

# LINKING GENE EXPRESSION VARIATION IN NATURAL POPULATIONS OF *HELIANTHUS ANNUUS* TO LENGTH VARIATION IN MICROSATELLITES


## Abstract

The role of microsatellites as agents of phenotypic change has been relatively unexplored in the plant kingdom. In general, this research area lacks population level information required to demonstrate its importance in nature. Previous work shows that *Helianthus annuus* (common sunflower) Gene Ontology terms associated with plant responses to various stresses and stimuli are enriched with transcribed microsatellites. Population genetic analyses of *H. annuus* comparing variation in allele frequencies and length at anonymous and transcribed microsatellite loci suggest that transcribed microsatellites are frequently influenced by selection. To detect if microsatellites from untranslated regions (UTRs) modulate gene expression in natural populations, population level analyses were carried out to quantify gene expression variation at three microsatellite containing genes. Although differences in allele frequencies among populations did not differ significantly, population explained a significant amount of gene expression variation, possibly indicating environmental effects or epistatic interactions. Allele lengths at the three genes utilized in the study did not predict gene expression variation within populations. Further, gene expression levels did not significantly differ among the different alleles at a gene. This finding does not support a role for these microsatellites in tuning gene expression. However, variation in allele length at the three

loci was narrow, 6, 14 and 20 bases. Hence, it is conceivable that these microsatellites can alter gene expression, but that purifying selection is maintaining alleles with near optimal expression levels.

Keywords: gene regulation, microsatellites, *Helianthus annuus*, natural populations.

## Introduction

One of the early mentions of the relative role of gene regulation and protein evolution is in the theory proposed by Britten and Davidson (1969, 1971) on the evolutionary role of repetitive elements and the importance of gene regulation in adaptive evolution. In recent years various studies have analyzed the relative contribution of cis-regulatory mechanisms and protein coding changes in species divergence (Townsend et al. 2003, Wittkopp et al. 2008, Bedford and Hartl 2009). Stern and Orgogozo (2008) analyzed a large body of empirical data and observed that the relative frequencies of cis versus protein coding changes differ from case to case. The same study suggests that ontogenetic change is frequently cued by cis regulatory elements.

Research on the role of gene expression's influence over the evolution of the phenotype is still relatively young. This is especially true of work contrasting the relative importance of cis and trans regulatory mechanisms. The most frequently considered mode of generating polymorphisms in cis-regulatory regions involves subsititutions frequently referred to as single nucleotide polymorphisms (SNPs), and clearly SNPs are important. However, mutations resulting in SNPs are not the only source of novelty enjoyed by regulatory elements. Microsatellites also exhibit properties that may make them suitable candidates for generating functional genetic variation. These properties

include high mutation rates and the reversible nature of these mutations (Gemayel et al. 2010). For example, in humans, increased cis-regulatory sequence variation compared to protein coding changes has been shown to result from microsatellite polymorphisms in cis regulatory elements (Rockman and Wray 2004).

Over the years, several empirical studies have reported functionally beneficial roles for microsatellites. A comprehensive list of such studies has been recently detailed in a review by Gemayel et al. (2010). The fact that several studies have provided empirical evidence of functional roles for both intronic and exonic microsatellites, lead many researchers to propose the "tuning knob" hypothesis, wherein expansions or contractions of the microsatellite tract affect the fitness of organisms, by modulating gene expression levels or protein function (King et al. 1999, Trifanov 2004). The tuning knob hypothesis explores the functionality of microsatellites as a useful source of generating variability, based on the high mutation rate as well as the reversible nature of microsatellite mutations. In fact, mutation rates of microsatellites are orders of magnitude greater than that of any other source of generating polymorphisms (Gemayel et al 2010). While estimates of mutations rate in microsatellites include values as high as that ranging between $10^{-5}$ to $10^{-2}$ base pair per mutation per generation (Wang et al. 1994); estimates from substitution rates range from $10^{-10}$ to $10^{-8}$ per base per generation (Baer et al. 2007; Ossowaski et al. 2010). Moreover, microsatellite mutations via slipped stand mispairing (Levinson and Gutman 1987) results in expansion or contraction mutations on a microsatellite tract are also more likely to be reversible. The reversible nature of mutations along with the faster mutation rate would provide the organisms an efficient mechanism to maintain optimum fitness from generation to generation. More recently, the contributions of microsatellites have been viewed as an important source of

generating variation (Trifanov 2004, Li et al. 2004, Kashi and King 2006, Gemayel et al. 2010). In fact, various studies on microsatellite functionality reveals that specific mechanisms by which microsatellites present in the UTRs modulate genes may vary according to a number of factors including its position in the UTR and motif composition. For example, lengths of AU rich microsatellites in 3'UTR are known to influence gene expression levels by modulating mRNA stability (Mignone et al. 2002), while lengths of AG dinucleotide repeats are known to modulate transcription rates (Newcomb et al. 2006).

In this study, we attempt to find support for the tuning knob hypothesis by investigating the role of these cis-regulatory elements in modulating or tuning gene expression in natural populations of *Helianthus annuus* (common sunflower). We specifically test the role of microsatellites from untranslated regions (UTR) in gene regulation, by quantifying gene expression levels associated with different microsatellite allele lengths. Prior studies of microsatellites in *H. annuus* including a bioinformatics survey and a population level study, suggest a pattern of selection. The bioinformatics study identified nearly 20% of all *H. annuus* transcripts as containing a microsatellite. Further, microsatellites were observed to be preferentially associated with Gene Ontology (GO) terms associated with plant response to stress and stimulus (Chapter II). The results from the *H. annuus* bioinformatics survey suggests that the presence of microsatellites in genes is non-random, and could reflect preferential association with genes whose "tuning" is beneficial for the organism. A population genetics study looking at differences in selection pressures acting on transcribed versus anonymous microsatellites, identified a clear difference between the two classes, with transcribed microsatellites exhibiting signals consistent with balancing selection and anonymous microsatellites

showing patterns of neutral evolution (Chapter IV). Although we could detect signals of selection on transcribed microsatellites, we cannot conclusively identify effects of the microsatellite from effects resulting from mere linkage of the microsatellite to genes, which could more likely be the target of selection. We investigate to see if differences or similarities in allele frequencies of transcribed microsatellites observed in natural populations of *H. annuus* are functional, by looking at correspondence between microsatellite allele length and gene expression levels.

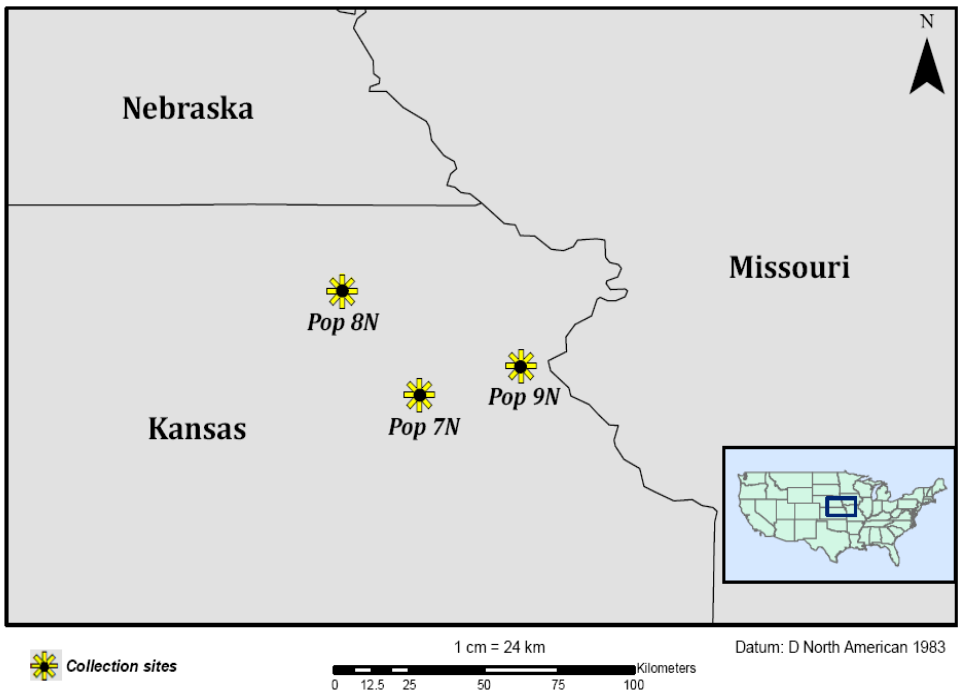


Figure 6.1 Sampling location of the three *Helianthus annuus* populations from Kansas are shown in the map.

## Materials and methods

### Sample collection and preparation

To get a good a estimate of allele frequencies in populations as well as to increase the probability of sampling rare alleles whose frequency is <5% of the sample, a total of 50 *H. annuus* individuals were collected from each of the three locations (150 individuals in total), in Kansas in September 2010 (Figure 6.1). From each individual sunflower, tissue for DNA and RNA extraction was collected. The tissue for DNA extraction was dried in silica gel and tissue for RNA extraction was stored in RNALater® (Ambion, Grand Island, NY). For DNA extraction, approximately 5-10 mg of dried leaf tissue was macerated using a Retsch MM200 ball mill (Retsch Incorporated, Newtown, Pennsylvania, USA). Whole genomic DNA was extracted from macerated tissue using an ABI PRISM 6100 Nucleic Acid Prep Station and proprietary chemistry (Applied Biosystems, Foster City, California, USA).

Approximately 50-65 mg of leaf tissue stored in RNALater® was used for extracting RNA. RNA extractions were performed using Ambion® RNAqueous®-4PCR Kit and its specified protocol (Life Technologies, Grand Island, NY). mRNA was converted to cDNA using High Capacity cDNA Reverse Transcription kit with RNase inhibitor (Applied Biosystems, Foster City, CA).

### Microsatellite genotyping

Three primer PCR (Schuelke 2000) was performed for all individuals at fifteen transcribed microsatellite loci listed in Table 6.2 following the general protocol of Chapman et al (2008). Specifically, each 10µl reaction consisted of ~10 ng whole genomic DNA, 2mM $MgCl_2$, 30mM Tricine (pH 8.4- KOH), 50mM KCl, 100µM of each

dNTP, 200 nM of reverse primer, 200 nM of M13 primer (DNA sequence) labeled with fluorescent dye (HEX, NED or FAM), between 40 and 150 nM forward primer, and 0.4 U of *Taq* DNA polymerase. Thermal cycling profiles consisted of an initial denaturation at 95º C for 4 mins, followed by 10 touchdown cycles involving temperature conditions of 94º C for 30 s, 30 s at the annealing temperature (the annealing temperature drops 1ºC in each subsequent touchdown cycle), and 72º C for 45 s. The remaining 25 cycles had a thermal profile of 94º C for 30 s, 30 s at the optimal annealing temperature, and 45 s at 72º C. The last stage consisted of a 7 min final elongation period at 72ºC. The fluorescently labeled PCR products were pooled without dilution for fragment analysis on ABI 3730 capillary sequencers (ABI, Foster City, CA) in the Arizona State University DNA Lab with MapMarker® 1000 as a size standard (Bioventures, Murfreesboro, TN). Alleles were scored using Peak Scanner™ Software v1.0 (ABI, Foster City, CA).

**Gene expression quantification**

The cDNA was quantified for three genes containing microsatellites in the UTRs and two constitutively expressed genes, namely actin and ubiquitin. These three microsatellite containing genes were selected from the list of fifteen microsatellite loci listed in Table 6.2. The assay sequences and the standard curves established at the five assays are provided in Chapter V. The thereshold cycle ($C_T$) values were obtained for each individual at each assay. These $C_T$ values were converted to a log concentration value utilizing the means and efficiencies obtained from the standard curves (Table 6.1). The concentrations at actin and ubiquitin assays for each individual were averaged to yield the standard or baseline concentration for each sample. The standard concentration

was subtracted from the concentration at each assay for an individual to yield the standardized concentrations at each gene.

**Population genetic analyses**

All of the transcribed microsatellite loci used in this study were characterized in a previous study (Chapman et al. 2008). To estimate the similarities and difference between the populations at the microsatellite loci, standard population genetic statistical measures were employed. To measure the allelic diversity of the microsatellite loci in the populations, heterozygosity was calculated (Nei 1978). To estimate the variance in allele frequencies among populations, $F_{ST}$ was calculated (Weir and Cockerham 1984). Further, $R_{ST}$, which is analogous to $F_{ST}$ but also takes into account the contribution of mutational processes in microsatellites to allele frequency differences, was calculated (Slatkin 1995, Rousset 1996, Goodman 1997). The estimates of Expected and observed heterozygosity ($H_E$ and $H_O$ respectively) as well as $F_{ST}$ and $R_{ST}$ values at each locus per population were obtained using FSTAT v.2.9.3.2 (Goudet et al. 2001). To test for population substructure among the three populations, an Analysis of Molecular Variance (AMOVA) ( Excoffier et al 1992) was performed using GenAlEx (Peakall and Smouse 2006).

Table 6.1 The mean $C_T$ value indicating the mean threshold cycle number, efficiency of the assay and the formula used for converting the $C_T$ values into a log scale concentration ([conc]) at each assay is provided in this table

| Assay | $R^2$ | b | Mean $C_T$ | Intercept | [Conc] |
|---|---|---|---|---|---|
| C1181 | 0.98 | 1.03 | 31.76 | 33.90 | $[C_T - 33.9]/-1.03$ |
| C3115 | 0.98 | 0.89 | 32.72 | 34.63 | $[C_T - 34.63]/-0.89$ |
| C5774 | 0.99 | 0.91 | 28.97 | 30.79 | $[C_T - 30.79]/-0.91$ |
| *Actin* | 0.99 | 0.83 | 31.30 | 32.97 | $[C_T - 32.97]/-0.83$ |
| *Ubiquitin* | 0.99 | 0.85 | 31.16 | 32.86 | $[C_T - 32.86]/-0.85$ |

**Linking gene expression to microsatellite allele length**

If microsatellites were to play a role in generating adaptive genetic variation, then there should be additive effects of microsatellite alleles; else variation is not heritable. The additive effects on gene expression levels of microsatellite alleles were quantified in two ways. First, the average effects of individual microsatellite alleles were estimated using multiple regressions (Falconer and Mackay 1996). Alleles at a locus were treated as discrete predictor variables for mRNA concentrations estimated. For each allele, the genotypic value was coded as zero, one or two. The presence of significant average effects for individual alleles does not necessarily provide support for the tuning knob model. Microsatellite alleles at marker loci could be in linkage disequilibrium with functional variation elsewhere in the gene. While the tuning knob model predicts stepwise relationships between allele lengths and gene expression levels, a random association between allele lengths and gene expression might be more consistent with linkage disequilibrium.

To help distinguish between competing explanations for significant average effects of alleles on gene a statistical test treating allele lengths as continuous variables was conducted. Evidence from other empirical studies suggests that the relationship between allele length and gene expression is likely to be quadratic (Vinces et al. 2009). Hence, second order polynomial regression was used to determine if the relationship between allele lengths and gene expression fit a quadratic function (Equation 6.1)

$$Y_{ijk} = B_0 + B_1X_1 + B_1X_2 + B_2X_1^2 + B_2X_2^2 \quad \text{(Equation 6.1)}$$

Where, $Y_{ijk}$ is gene expression in the $i^{th}$ individual in the $j^{th}$ population at loci $k$, $B_0$ is the mean response, $X_1$ and $X_2$ are the lengths or number of repeats for each allele at a locus. A linear parameter ($X'_1$) and second order parameter ($X'_2$), each combining both

alleles in an individual as a single estimate were also calculated as shown in equations (6.2) and (6.3), respectively.

$$X'_1 = (Z_l - Z_0) + (Z_m - Z_0) \quad \text{(Equation 6.2)}$$

$$X'_2 = (Z_l - Z_0)^2 + (Z_m - Z_0)^2 \quad \text{(Equation 6.3)}$$

Where, $Z_l$ and $Z_m$ represent the two observed allele lengths in the individual, and $Z_0$ represents the minimum observed fragment length at the locus in the individual. Statistical analysis was performed using JMP 9 Pro (Statistical Institute Inc., Cary, NC) and R v.2.11.1 for Windows (R Core Development Team 2010).

## Results

### Heterozygosity and population structure

The populations were polymorphic for all loci. The number of alleles sampled at the 15 loci ranged from 5 to 18 (Table 6.2). The mean expected heterozygosity at the loci, ranged between 0.358 and 0.906 (Table 6.2). Hence, all loci were polymorphic and exhibited a high degree of heterozygosity, making them suitable for the study. The measure of population differentiation between populations, as measured by estimates of $F_{ST}$ ranged from 0.01 to 0.156. The mean $F_{ST}$ at these loci was estimated at 0.0574. $R_{ST}$ values ranged from -0.009 to 0.541, with a mean value of 0.091 (Table 6.2). The heterozygosity, $F_{ST}$ and $R_{ST}$ values at the three loci (C5774, C1181 and C3115) chosen for the gene expression analysis were comparable to the general pool of loci utilized for population genetic analyses (Chapter IV).

An Analysis of Molecular Variance (AMOVA) indicates that 87% of genetic variation is found within individuals, 6% among populations, and 7% among individuals

within populations (Table 6.3). Hence, little genetic variation distinguishes the populations at these loci.

Table 6.2 The number of alleles sampled (A), estimates of allelic diversity ($H_O$) and genetic variance measured as $F_{ST}$ and $R_{ST}$ are provided for the 15 microsatellite loci. Loci for which Taqman assays were designed for are indicated by a *.

| **Locus** | **Alleles sampled** | **Ho** | $F_{ST}$ | $R_{ST}$ |
|---|---|---|---|---|
| C1181* | 8 | 0.598 | 0.032 | -0.012 |
| C3115* | 10 | 0.906 | 0.045 | 0.25 |
| C5774* | 8 | 0.742 | 0.045 | 0.049 |
| L19A05 | 12 | 0.811 | 0.04 | 0.055 |
| B12L21 | 8 | 0.581 | 0.01 | 0.047 |
| K151316 | 8 | 0.637 | 0.021 | -0.009 |
| E11B18 | 8 | 0.503 | 0.031 | 0.001 |
| B6A04 | 8 | 0.82 | 0.048 | 0.059 |
| C4529 | 7 | 0.476 | 0.069 | 0.054 |
| C2293 | 5 | 0.442 | 0.056 | 0.083 |
| M8NO3 | 10 | 0.678 | 0.081 | -0.005 |
| G18L13 | 8 | 0.554 | 0.156 | 0.541 |
| A16A09 | 18 | 0.687 | 0.031 | 0.031 |
| K12HO5 | 18 | 0.476 | 0.07 | 0.21 |
| B10NO5 | 11 | 0.358 | 0.126 | 0.01 |
| **Mean** | **9.8** | **0.617933** | **0.0574** | **0.090933** |

Table 6.3 Results from Analysis of Molecular Variance (AMOVA) detailing the partitioning of genetic variation among the three populations, among individuals within populations, and within individuals; at the 15 microsatellite loci are provided here

| Source | df | SS | MS | Est. Var. | % |
|---|---|---|---|---|---|
| **Among Populations** | 2 | 78.193 | 39.097 | 0.335 | 6% |
| **Among Individuals** | 147 | 823.715 | 5.604 | 0.394 | 7% |
| **Within Individuals** | 150 | 722.195 | 4.815 | 4.815 | 87% |
| **Total** | 299 | 1624.103 | | 5.544 | 100% |

Table 6.4 Length variation observed at the three loci in the three populations is shown in the table. The smallest allele observed at a locus serves as $Z_0$ in calculations of the linear and second degree parameter as predictors of gene expression variation

| Parameter | List of alleles at Loci | | |
|---|---|---|---|
| | C1181 | C3115 | C5774 |
| Alleles observed | 323 | 313 | 179 |
| | 324 | 315 | 181 |
| | 325 | 321 | 183 |
| | 326 | 323 | 185 |
| | 327 | 325 | 187 |
| | 328 | 327 | 189 |
| | | 329 | 191 |
| | | 331 | 193 |
| | | 333 | |
| | | 335 | |
| $Z_O$ | 323 | 313 | 179 |
| Length range | 5 | 22 | 14 |

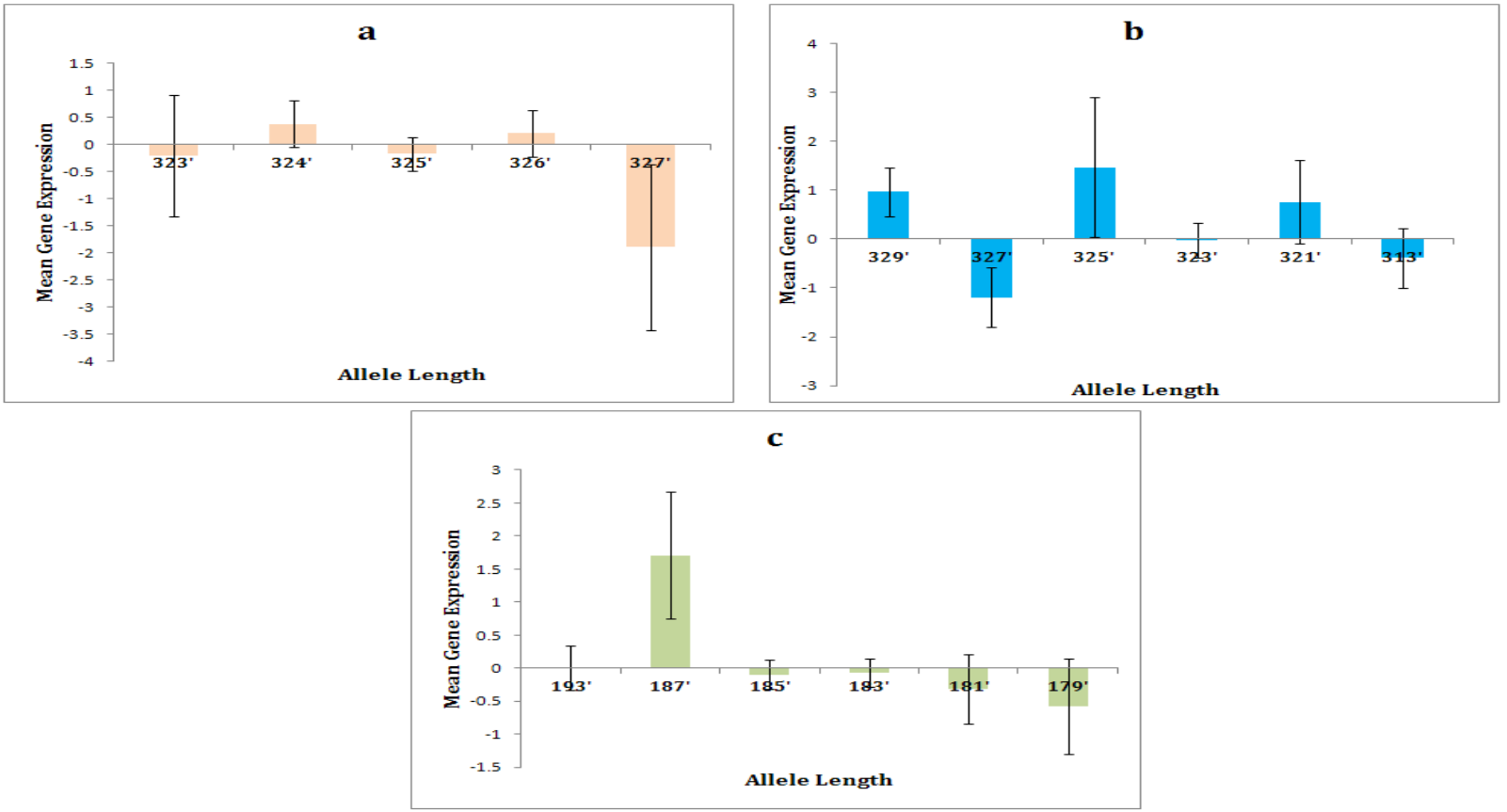


Figure 6.2 Mean and standard deviation of gene expression levels for the alleles found in population KS3 at (a) C1181, (b) C3115, and (c) C5774 are provided here.

**Microsatellite allele length and gene expression levels**

mRNA was extracted from a total of 62 individuals. Out of the 62 individuals, 44 belonged to population KS3, 9 belonged to population KS1, and the remaining 9 belonged to KS2. Amplicon lengths of all alleles observed at each locus are provided in Table 6.4. Mean gene expression levels at all the three genes were significantly different among the three populations ($P$<0.001), with population differences explaining a significant proportion of the variation in gene expression in C5774 ($R^2$=0.804, $P$<0.001), C1181 ($R^2$=0.421, $P$<0.001), and C3115 ($R^2$=0.415, $P$<0.001).

The gene expression levels at a locus were not significantly explained by allele lengths at a locus, while treating alleles as discrete variables. The $R^2$ at the three genes was estimated to be 0.19 at C5774 ($P$=0.2268), 0.154 at C1181 ($P$=0.2022), and 0.26 at C3115 ($P$=0.1396). The average effects of alleles present in population KS3 at the three loci are shown in Figure 6.2.

A significant association between gene expression levels and allele lengths was not detected, while treating allele lengths as a continous variable in form of the linear ($Y_1$) and second degree ($Y_2$) parameters. The correlation coefficient ($R^2$) for the relationship was determined to be 0.08 at C3115 ($P$=0.4041), 0.06 at C1181 ($P$=0.5256), and 0.015 at C5774 ($P$=0.6793).

## Discussion

In a previous study detailed in Chapter IV support was provided for differential selection pressures acting on transcribed microsatellites relative to that at anonymous microsatellites. The transcribed microsatellites showed signals consistent with balancing selection whereas the anonymous microsatellites are more neutrally evolving. Further,

when measures of variance in allele lengths were compared to variance estimates based on allelic diversity, a greater proportion of variance in length was found within populations. If linkage disequilibrium were to explain patterns of balancing selection, then the relationship between allele lengths and functional sites would likely be random, however, our results suggests that selection is acting to maintain length variation within populations. These results are also consistent with predictions of the tuning knob model, as per which functionality is determined by allele lengths. Hence, this study was designed to garner further support for the functional role of microsatellites. The loci used in the previous study were selected based on extreme lengths and impurities harbored within the microsatellite tract (Pramod et al. 2011). To remove biases introduced due to the nature of the microsatellites, shorter and pure microsatellites developed by Chapman et al. (2008) were utilized for this study, because they are more likely to exhibit mutational dynamics and mutation rates associated with more typical microsatellites found in organismal transcriptomes. Population genetic analyses on these loci show results similar to that observed in our previous study (Chapter IV). Transcribed microsatellites loci revealed low mean $F_{ST}$ values among the populations sampled for this study. The presence of similar allele frequencies across populations either indicates high gene flow, a possibility given that the populations were sampled from the same latitudinal gradient, or selection maintaining all the alleles across populations due to fitness effects associated with each. Interestingly, locus C3115 exhibited a $R_{ST}$ value of 0.25, which is approximately 2.5 times greater than the rest of the transcribed loci. Whereas, the $F_{ST}$ value at C3115 is comparable to rest of the transcribed loci. Since $R_{ST}$ estimates population divergence taking into account the allele length variance at the loci, locus

C3115 showed patterns predictive of microsatellites under the tuning knob model. Only, locus G18L13 exhibited a $R_{ST}$ value of 0.541 as well as elevated $F_{ST}$ of 0.15.

In order to test the "tuning knob" hypothesis, natural populations of *H. annuus* were utilized to detect the relationship between gene expression variation and microsatellite allele lengths. Gene expression levels were quantified for three of the fifteen genes that harbor microsatellites in UTRs selected for this study. The three genes were selected for assay development due to a number of reasons; the main reason being to minimize assay failure resulting from utilizing genes whose sequences are singletons in the error prone EST database. Other factors taken into consideration include presence of fewer alleles at the locus, availability of multiple redundant sequences in the contigs, etc. These reasons are discussed in further detail in Chapter V.

Our study failed to detect a relationship between allele length and gene expression levels at the three genes. The average effects of alleles found in the population with the largest sample size, KS3 were too low to provide us with a detectable signal (Figure 6.1). The average effects of alleles did not show patterns consistent with the tuning knob model. However, this does not necessarily indicate lack of support for the tuning knob model for microsatellite function. One of the main reasons for the lack of correlation between allele lengths and gene expression could be due to the small range of allele lengths at the microsatellites in the genes used for the study, since given our hypothesis, similar alleles are likely to exhibit similar expression levels. Allele length variation in C1181 spans 6 bases, C3115 spans 20 bases, and C5774 spans 14 bases. The microsatellites harbored are dinucleotide repeat tracts, making the range of copy numbers low as well. One of the reasons for choosing these loci, include the presence of fewer number of alleles. Yet, the sample sizes required to detect a measurable effect was

lacking at many of the alleles, even within the most extensively sampled population KS3. However, inclusion of other populations is not likely to provide any different answers. This is because the same alleles were identified across these populations in similar frequencies. The fact that allele frequencies at these loci were comparable across populations and a very narrow length range is observed in these populations, along with the fact that gene expression levels did not significantly differ between the differ alleles, could be suggestive of the role of purifying selection in selective maintenance of the alleles with optimum fitness.

Significant differences in basal gene expression levels at all three genes among the three populations were observed. This could suggest the role of environmental differences in influencing gene expression phenotypes of these genes. Further, we could be dealing with a very complex phenomenon that could not be explained by just looking at one gene in isolation. Wittkopp (2007) discusses the importance of studying gene regulatory mechanisms utilizing a network perspective, since a large number of interactions may determine expression variation. Studies of this scale need detailed knowledge of the gene regulatory networks in *H. annuus*, than is currently available. Hence, quantifying gene expression levels from plants grown under controlled environmental settings would reduce the effects of environmental factors as well as genotype by environment interaction effects. This may help enhance the proportion of gene expression variation attributable to additive effects resulting from microsatellite tract length changes.

Gemayel et al. (2010) have compiled a comprehensive list of empirical studies that provide evidence for a functional role of microsatellites. Almost all of these studies include a candidate gene approach, wherein the researchers have stumbled on

microsatellites as the causative mechanism while trying to identify why a gene was being differentially expressed. Fondon and Garner (2004) also followed a candidate gene approach in their study of genes influencing dog skeletal morphological features and found a correlation between microsatellite lengths and cranial morphology, across dog breeds at these genes. Only Vinces et al. (2009) looked at the entire range of promoter region microsatellites in yeasts, and were able to identify the role of microsatellite length differences in causing transcription regulation by influencing nucleosome binding. Vinces et al. (2009) also generated synthetic microsatellite tracts of different lengths to understand their role regulating gene expression. However, none of these studies have focused on studying these mechanisms in natural populations. A larger study looking at naturally occurring variation in controlled setting using several candidate genes to look at the relationship between microsatellite length and gene expression level, may provide answers to the utility of the tuning knob model. Our recent bioinformatics study on the *H. annuus* transcriptome revealed that the microsatellites are enriched within genes representative of Gene Ontology terms associated with response to stress and stimulus (Chapter II). Hence, microsatellites within these genes would be good candidates for exploring their potential role as tuning knobs regulating genes in *H. annuus*.

Yet another approach to test the role of these microsatellites in gene regulation would be to use transgenics to study the effect of different allele lengths in a controlled experiment, where microsatellite tracts of different length could be cloned into *H. annuus* individuals. This study would become feasible only when the complete genome sequence of *H. annuus* becomes available, since construction of transformation vectors is otherwise not feasible. A study in *Arabidopsis thaliana* provided strong empirical evidence pointing at uncontrolled expansion of a triplet repeat microsatellite in the third intron in causing

the irregularly impaired leaf phenotype by using the transgenic approach (Sureshkumar et al. 2009). The effect of allele lengths encountered in natural populations as well as those not encountered in natural populations can be tested by using the transgenic approach. This approach could enable us to understand the observation made in this study, where allele lengths at the genes exhibited a small length range, and if these alleles are maintained in nature due to optimum "phenotypes" associated with them.

**Acknowledgements**

The authors would like to thank SG Shaak for help with sample collection, JL Martin for help with extraction and genotyping, and Dr CP Brooks for help with statistical analyses. We would also like to acknowledge support from the Department of Biological Sciences, College of Arts and Sciences and the Office of Research and Economic Development at Mississippi State University.

## References


Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics* 8:619–631. doi:10.1038/nrg2158

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: A theory. *Science* 165:349-358

Blackman BK, Michaels SD, Rieseberg LH. 2011. Connecting the sun to flowering in sunflower adaptation. *Molecular Ecology* 20: 3503-3512

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ and Burke JM. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *The Plant Cell* 20: 2931–2945

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5:435-445

Excoffier L, Smouse P and Quattro J. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479 -491

Fay JC and Wittkopp PJ. 2008. Evaluating the role of natural selection in evolution of gene regulation. *Heredity* 100:191-199

Falconer DS and Mackay TFC. 1996. Introduction to quantitative genetics (Fourth edition). ©*Addison Wesley Longman Limited, Essex, England*

Fondon JW III and Garner HR. 2004. Molecular origins of rapid and continous morphological evolution. *Proceedings of the National Academy of Sciences* 99:1991-2004

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics* 6:743–55

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* 44:445-477

Goodman SJ. 1997. $R_{ST}$-calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology* 6: 881-887

Goudet J. 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html. Updated from Goudet 1995

Kane NC, Rieseberg LH. 2007. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* 175: 1823-1834

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* 22(5): 253-259

Kashi Y, King DG, Soller M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics* 13:74-78

Levinson G, Gutman GA.1987. Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution. *Molecular Biology and Evolution* 4 (3): 203–221

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* 11: 2453-2465

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: Structure, Function and evolution. *Molecular Biology and Evolution* 21(6):991-1007

Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. *Genome Biology* 3(3):reviews0004.1-0004.10

Morgante M, Hanafey M, Powell W. 2002. Microsatellites are preferentially associated with non repetitive DNA in plant genomes. *Nature Genetics* 30:194-200

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89(3):583-590

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EHA, et al. (14 more authors). 2006. Analyses of Expressed Sequence Tags from Apple. *Plant Physiology* 141:147–166

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327: 92–94

Peakall R and Smouse PE. 2006. GENALEX 6:genetic analysis in excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288-295

Pramod S, Rasberry AR, Butler TG, Welch ME. 2011. Characterization of long transcribed microsatellites in *Helianthus annuus* (Asteraceae). *American Journal of Botany* 98(12): e388-e390

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. url [http://www.R-project.org]

Rockman MV, Wray GA. 2002. Abundant Raw Material for Cis-Regulatory Evolution in Humans. *Molecular Biology and Evolution* 19(11):1991–2004

Rousset F. 1996. Equilibrium values of measure of population subdivision for stepwise mutation processes. *Genetics* 142: 1357–1362

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18: 233–234

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457-462

Stern DL and Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62(9): 2155–2177

Sureshkumar S, Todesco M, Schneeberger K, Harilal R, Balasubramanian S, and Weigel D. 2009. A Genetic Defect Caused by a Triplet Repeat Expansion in *Arabidopsis thaliana. Science* 323 (5917):1060-1063

Tautz D and Renz M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* 12: 4127-4138

Townsend JP, Cavalieri D and Hartl DL. 2003. Population genetic variation in genome-wide gene expression. *Molecular Biology and Evolution* 20: 955-963

Trifanov EN. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation; In "*Evolutionary theory and processes: Modern Horizons, papers in honor of Eviatar Nevo"* (Ed. Wasser SP); Kluwer Academic Publishers. pp 115-138

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science* 324:1213-1216

Wang Z, Weber JL, Zhong G, Tanksley SD. 1994. Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* 88:1-6

Weir BS and Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38:1358-1370

Wittkopp PJ. 2007. Variable gene expression in eukaryotes: a network perspective. *The Journal of Experimental Biology* 210:1567-157

# CHAPTER VII

# SUMMARY

In this dissertation, I examine the role of microsatellites in bringing about adaptive evolutionary change in natural populations. Previously, several researchers had explored means by which microsatellites within genes might play a role as "tuning knobs" of phenotypic variation by either modulating gene expression or protein function in a stepwise manner (Trifanov 2004, Kashi and King 2006, Gemayel et al. 2010). Moreover, microsatellites, by virtue of their high mutation rates, may serve as reservoirs of novel heritable genetic variation that can be rapidly generated. Anecdotal evidence on the functional role of microsatellites is abundant (Gemayel et al 2010). Further, Vinces et al (2009) provide evidence that gene expression divergence results from microsatellite length variation in the yeast genome. Yet, no study had looked at naturally occurring variation at the scale presented in this dissertation, nor provided evidence that this source of variation is influenced by selection.

The hypothesis that transcribed microsatellites play a functional role in adaptive evolutionary processes was tested by using *Helianthus annuus*, the common sunflower, as a model system. The hypothesis was addressed using three distinct approaches. The first approach was a bioinformatic analysis detailed in Chapter II. Here, the entire unigene database of *H. annuus* representing the transcriptome was analyzed to characterize the distribution of transcribed microsatellites and to determine if this pattern is consistent with microsatellites playing an evolutionary significant role. The study

shows that the distribution of microsatellites non-random. Microsatellite motifs associated with gene regulation are prevalent in untranslated regions (UTRs), and gene ontologies (GO) associated with plant response to stress and environmental stimuli are significantly enriched with microsatellites. The prevalence of microsatellites within GOs associated with response to stress and stimuli indicates support for the proposed role of microsatellites as tuning knobs modulating changes to stress and stimulus. This is consistent with what is expected assuming that microsatellites aid natural populations in tracking environmental change.

Further, a population genetics study utilizing natural populations compared patterns of microsatellite variation among populations between anonymous and transcribed microsatellite loci. This study shows that anonymous and transcribed microsatellites are under different selection pressures, with anonymous microsatellites showing elevated differences in allele frequencies relative to transcribed microsatellites. This pattern is consistent with balancing selection maintaining a greater proportion of genetic variation within populations. These transcribed microsatellites are not only characterized by maintenance of similar allele frequencies across populations, but also by maintenance of a tighter range of allele lengths. The tighter regulation of allele lengths and the signal of balancing selection at the transcribed microsatellites relative to that observed at anonymous microsatellites are evidence consistent with greater selection on allele length at transcribed microsatellites. This finding is consistent with the tuning knob hypothesis.

The final study utilizing a functional genomics approach was designed to determine if microsatellite allele lengths influence gene expression in natural populations. The study looked at gene expression variation resulting from variation in lengths of

microsatellites contained in UTRs of three genes. A significant association between allele length and gene expression levels was not detected. However, possible reasons for this include the narrow length range of alleles sampled in the population, and the prevalence of similar allele frequencies across populations. These findings are not entirely inconsistent with predictions of the tuning knob model that predicts similar expression levels for similar allele lengths. At the same time, the prevalence of similar allele frequencies and a small range of allele lengths in the sampled populations suggest that purifying selection might be dictating the evolution of these microsatellites. This result is also consistent with selection maintaining universally adaptive alleles of optimum lengths across all populations. Overall, the results presented in this dissertation are consistent with the proposed role of microsatellites as evolutionary tuning knobs. However, these findings are not entirely unambiguous.

## References

Gemayel R, Vinces MD, Legendre M, and Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* 44:445-477

Kashi Y and King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* 22(5): 253-259

Trifanov EN. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation; In "*Evolutionary theory and processes: Modern Horizons, papers in honor of Eviatar Nevo"* (Ed. Wasser SP); Kluwer Academic Publishers. pp 115-138

Vinces MD, Legendre M, Caldara M, Hagihara M, and Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 324:1213-1216

# APPENDIX A

# SUPPLEMENTARY DATA TABLES

Table A.1 Gene Ontology (GO) terms within which microsatellites showed significant enrichment within each of the three regions (5'UTR, Coding or 3'UTR) are provided here. The GO terms are broadly classified into biological processes (P), cellular components (C) or molecular functions (F).

| Region | GO Term | GO term description | Aspect | P-value | Number of Genes |
|---|---|---|---|---|---|
| 5'UTR (Total of 808 genes) | GO:0050896 | response to stimulus | P | 7.47E-10 | 176 |
| | GO:0042221 | response to chemical stimulus | P | 2.48E-09 | 109 |
| | GO:0009628 | response to abiotic stimulus | P | 2.03E-07 | 77 |
| | GO:0006950 | response to stress | P | 3.76E-07 | 109 |
| | GO:0009987 | cellular process | P | 1.62E-06 | 363 |
| | GO:0006970 | response to osmotic stress | P | 2.32E-06 | 36 |
| | GO:0006810 | transport | P | 3.91E-06 | 91 |
| | GO:0051234 | establishment of localization | P | 4.75E-06 | 91 |
| | GO:0044281 | small molecule metabolic process | P | 9.11E-06 | 81 |
| | GO:0010033 | response to organic substance | P | 1.01E-05 | 68 |
| | GO:0051179 | localization | P | 1.19E-05 | 92 |
| | GO:0009651 | response to salt stress | P | 1.56E-05 | 33 |
| | GO:0032787 | monocarboxylic acid metabolic process | P | 6.64E-05 | 31 |
| | GO:0070887 | cellular response to chemical stimulus | P | 8.45E-05 | 34 |
| | GO:0071310 | cellular response to organic substance | P | 8.77E-05 | 32 |
| | GO:0009719 | response to endogenous stimulus | P | 0.000384 | 54 |
| | GO:0051716 | cellular response to stimulus | P | 0.000441 | 50 |
| | GO:0009056 | catabolic process | P | 0.00132 | 45 |
| | GO:0023033 | signaling pathway | P | 0.00241 | 45 |
| | GO:0006631 | fatty acid metabolic process | P | 0.00369 | 20 |
| | GO:0019752 | carboxylic acid metabolic process | P | 0.00515 | 43 |
| | GO:0043436 | oxoacid metabolic process | P | 0.00515 | 43 |
| | GO:0006082 | organic acid metabolic process | P | 0.00533 | 43 |
| | GO:0042180 | cellular ketone metabolic process | P | 0.00855 | 43 |
| | GO:0044282 | small molecule catabolic process | P | 0.00909 | 20 |
| | GO:0071495 | cellular response to endogenous stimulus | P | 0.00985 | 24 |
| | GO:0005886 | plasma membrane | C | 1.38E-23 | 145 |

Table A.1 (continued)

| | | | | | |
|---|---|---|---|---|---|
| | GO:0016020 | membrane | C | 4.61E-22 | 240 |
| | GO:0005623 | cell | C | 5.56E-16 | 542 |
| | GO:0044464 | cell part | C | 5.56E-16 | 542 |
| | GO:0005622 | intracellular | C | 2.16E-13 | 357 |
| | GO:0044424 | intracellular part | C | 4.06E-13 | 345 |
| | GO:0005737 | cytoplasm | C | 3.81E-12 | 269 |
| | GO:0043231 | intracellular membrane-bounded organelle | C | 5.73E-12 | 302 |
| | GO:0043227 | membrane-bounded organelle | C | 5.83E-12 | 302 |
| | GO:0043229 | intracellular organelle | C | 2.34E-11 | 311 |
| | GO:0043226 | organelle | C | 2.38E-11 | 311 |
| | GO:0044444 | cytoplasmic part | C | 3.93E-10 | 247 |
| | GO:0031090 | organelle membrane | C | 5.4E-06 | 40 |
| | GO:0005618 | cell wall | C | 5.52E-06 | 42 |
| | GO:0030312 | external encapsulating structure | C | 6.81E-06 | 42 |
| | GO:0044446 | intracellular organelle part | C | 8.26E-06 | 116 |
| | GO:0044422 | organelle part | C | 8.64E-06 | 116 |
| | GO:0031975 | envelope | C | 0.000065 | 47 |
| | GO:0031967 | organelle envelope | C | 0.000065 | 47 |
| | GO:0005773 | vacuole | C | 0.000342 | 39 |
| | GO:0009526 | plastid envelope | C | 0.000355 | 34 |
| | GO:0009941 | chloroplast envelope | C | 0.00101 | 32 |
| | GO:0009505 | plant-type cell wall | C | 0.00629 | 22 |
| | GO:0044437 | vacuolar part | C | 0.00712 | 12 |
| | GO:0005515 | protein binding | F | 2.92E-06 | 96 |
| Coding (Total of 1098 genes) | GO:0009987 | cellular process | P | 1.09E-09 | 494 |
| | GO:0042221 | response to chemical stimulus | P | 1.12E-08 | 134 |
| | GO:0009628 | response to abiotic stimulus | P | 2.77E-07 | 95 |
| | GO:0010038 | response to metal ion | P | 5.4E-06 | 42 |
| | GO:0010035 | response to inorganic substance | P | 2.66E-05 | 46 |
| | GO:0050896 | response to stimulus | P | 3.02E-05 | 204 |
| | GO:0044237 | cellular metabolic process | P | 6.82E-05 | 364 |
| | GO:0046686 | response to cadmium ion | P | 0.000416 | 33 |
| | GO:0065007 | biological regulation | P | 0.00122 | 186 |
| | GO:0008152 | metabolic process | P | 0.00144 | 423 |
| | GO:0009737 | response to abscisic acid stimulus | P | 0.00169 | 30 |
| | GO:0009651 | response to salt stress | P | 0.00227 | 35 |
| | GO:0006950 | response to stress | P | 0.00271 | 122 |

Table A.1 (continued)

| | | | | |
|---|---|---|---|---|
| GO:0044238 | primary metabolic process | P | 0.00272 | 365 |
| GO:0006807 | nitrogen compound metabolic process | P | 0.00284 | 174 |
| GO:0034641 | cellular nitrogen compound metabolic process | P | 0.00379 | 170 |
| GO:0006970 | response to osmotic stress | P | 0.00474 | 36 |
| GO:0009725 | response to hormone stimulus | P | 0.00824 | 59 |
| GO:0005622 | intracellular | C | 3.33E-37 | 543 |
| GO:0044424 | intracellular part | C | 3.36E-36 | 525 |
| GO:0043226 | organelle | C | 9.12E-35 | 483 |
| GO:0043229 | intracellular organelle | C | 2.01E-34 | 482 |
| GO:0043227 | membrane-bounded organelle | C | 1.1E-33 | 464 |
| GO:0043231 | intracellular membrane-bounded organelle | C | 2.42E-33 | 463 |
| GO:0005737 | cytoplasm | C | 1.22E-30 | 408 |
| GO:0044444 | cytoplasmic part | C | 2.12E-29 | 384 |
| GO:0005623 | cell | C | 1.63E-23 | 740 |
| GO:0044464 | cell part | C | 1.63E-23 | 740 |
| GO:0005886 | plasma membrane | C | 1.01E-19 | 168 |
| GO:0009536 | plastid | C | 1.36E-19 | 226 |
| GO:0009507 | chloroplast | C | 1.79E-18 | 219 |
| GO:0016020 | membrane | C | 6.4E-14 | 275 |
| GO:0044422 | organelle part | C | 1.01E-11 | 168 |
| GO:0044446 | intracellular organelle part | C | 4.06E-11 | 166 |
| GO:0044435 | plastid part | C | 1.1E-09 | 86 |
| GO:0044434 | chloroplast part | C | 2.71E-09 | 82 |
| GO:0005634 | nucleus | C | 2.66E-08 | 163 |
| GO:0009579 | thylakoid | C | 4.27E-07 | 47 |
| GO:0009534 | chloroplast thylakoid | C | 4.48E-06 | 39 |
| GO:0031976 | plastid thylakoid | C | 4.48E-06 | 39 |
| GO:0009535 | chloroplast thylakoid membrane | C | 4.9E-06 | 36 |
| GO:0055035 | plastid thylakoid membrane | C | 4.9E-06 | 36 |
| GO:0042651 | thylakoid membrane | C | 4.95E-06 | 37 |
| GO:0031984 | organelle subcompartment | C | 5.24E-06 | 39 |
| GO:0044436 | thylakoid part | C | 6.9E-06 | 40 |
| GO:0034357 | photosynthetic membrane | C | 9.49E-06 | 37 |
| GO:0005829 | cytosol | C | 0.000024 | 58 |
| GO:0043228 | non-membrane-bounded organelle | C | 0.000629 | 71 |
| GO:0043232 | intracellular non-membrane-bounded organelle | C | 0.000629 | 71 |

Table A.1 (continued)

| | | | | | |
|---|---|---|---|---|---|
| | GO:0005773 | vacuole | C | 0.000654 | 47 |
| | GO:0031975 | envelope | C | 0.000658 | 55 |
| | GO:0031967 | organelle envelope | C | 0.000658 | 55 |
| | GO:0009532 | plastid stroma | C | 0.000729 | 43 |
| | GO:0005730 | nucleolus | C | 0.0018 | 29 |
| | GO:0031981 | nuclear lumen | C | 0.00214 | 36 |
| | GO:0044428 | nuclear part | C | 0.00239 | 44 |
| | GO:0009570 | chloroplast stroma | C | 0.00241 | 39 |
| | GO:0031974 | membrane-enclosed lumen | C | 0.00454 | 39 |
| | GO:0009526 | plastid envelope | C | 0.00832 | 38 |
| | GO:0043233 | organelle lumen | C | 0.00832 | 38 |
| | GO:0005488 | binding | F | 1.16E-08 | 448 |
| | GO:0005199 | structural constituent of cell wall | F | 1.5E-06 | 13 |
| | GO:0005198 | structural molecule activity | F | 1.32E-05 | 49 |
| | GO:0005515 | protein binding | F | 0.00107 | 111 |
| | GO:0008135 | translation factor activity, nucleic acid binding | F | 0.00511 | 18 |
| 3'UTR (Total of 790 genes) | GO:0009628 | response to abiotic stimulus | P | 3.34E-12 | 87 |
| | GO:0050896 | response to stimulus | P | 4.36E-11 | 177 |
| | GO:0009987 | cellular process | P | 3.08E-10 | 373 |
| | GO:0042221 | response to chemical stimulus | P | 1.31E-09 | 108 |
| | GO:0006950 | response to stress | P | 4.13E-09 | 113 |
| | GO:0044237 | cellular metabolic process | P | 9.95E-06 | 277 |
| | GO:0009415 | response to water | P | 1.04E-05 | 23 |
| | GO:0046686 | response to cadmium ion | P | 3.41E-05 | 29 |
| | GO:0010038 | response to metal ion | P | 3.59E-05 | 33 |
| | GO:0010035 | response to inorganic substance | P | 4.33E-05 | 37 |
| | GO:0006970 | response to osmotic stress | P | 0.000054 | 33 |
| | GO:0009651 | response to salt stress | P | 0.000108 | 31 |
| | GO:0009414 | response to water deprivation | P | 0.000117 | 21 |
| | GO:0008152 | metabolic process | P | 0.000238 | 319 |
| | GO:0009266 | response to temperature stimulus | P | 0.00217 | 29 |
| | GO:0009409 | response to cold | P | 0.00222 | 23 |
| | GO:0071840 | cellular component organization or biogenesis | P | 0.00333 | 73 |
| | GO:0042180 | cellular ketone metabolic process | P | 0.0047 | 43 |
| | GO:0006810 | transport | P | 0.00554 | 79 |
| | GO:0051234 | establishment of localization | P | 0.00642 | 79 |

Table A.1 (continued)

| | | | | |
|---|---|---|---|---|
| GO:0019752 | carboxylic acid metabolic process | P | 0.00671 | 42 |
| GO:0043436 | oxoacid metabolic process | P | 0.00671 | 42 |
| GO:0006082 | organic acid metabolic process | P | 0.00694 | 42 |
| GO:0044424 | intracellular part | C | 1.18E-30 | 392 |
| GO:0005737 | cytoplasm | C | 1.27E-30 | 318 |
| GO:0005622 | intracellular | C | 1.68E-29 | 400 |
| GO:0044444 | cytoplasmic part | C | 7.68E-28 | 296 |
| GO:0005623 | cell | C | 1.13E-27 | 566 |
| GO:0044464 | cell part | C | 1.13E-27 | 566 |
| GO:0043229 | intracellular organelle | C | 1.12E-26 | 354 |
| GO:0043226 | organelle | C | 1.15E-26 | 354 |
| GO:0043231 | intracellular membrane-bounded organelle | C | 1.57E-21 | 328 |
| GO:0043227 | membrane-bounded organelle | C | 1.61E-21 | 328 |
| GO:0044446 | intracellular organelle part | C | 3.45E-20 | 153 |
| GO:0044422 | organelle part | C | 3.74E-20 | 153 |
| GO:0005829 | cytosol | C | 1.56E-17 | 69 |
| GO:0005886 | plasma membrane | C | 3.87E-16 | 127 |
| GO:0016020 | membrane | C | 2.95E-15 | 217 |
| GO:0032991 | macromolecular complex | C | 1.02E-13 | 114 |
| GO:0022626 | cytosolic ribosome | C | 1.5E-12 | 39 |
| GO:0005840 | ribosome | C | 4.05E-12 | 48 |
| GO:0043228 | non-membrane-bounded organelle | C | 1.41E-11 | 74 |
| GO:0043232 | intracellular non-membrane-bounded organelle | C | 1.41E-11 | 74 |
| GO:0044445 | cytosolic part | C | 2.35E-09 | 31 |
| GO:0030529 | ribonucleoprotein complex | C | 2.98E-09 | 51 |
| GO:0031975 | envelope | C | 1.61E-08 | 54 |
| GO:0031967 | organelle envelope | C | 1.61E-08 | 54 |
| GO:0033279 | ribosomal subunit | C | 4.99E-08 | 33 |
| GO:0009536 | plastid | C | 3.05E-07 | 143 |
| GO:0030312 | external encapsulating structure | C | 3.22E-07 | 44 |
| GO:0009507 | chloroplast | C | 7.5E-07 | 139 |
| GO:0005618 | cell wall | C | 2.76E-06 | 42 |
| GO:0005730 | nucleolus | C | 5.98E-06 | 28 |
| GO:0005773 | vacuole | C | 6.42E-05 | 40 |
| GO:0005634 | nucleus | C | 0.000149 | 114 |
| GO:0022627 | cytosolic small ribosomal subunit | C | 0.000261 | 15 |

Table A.1 (continued)

| | | | | | |
|---|---|---|---|---|---|
| | GO:0031981 | nuclear lumen | C | 0.000268 | 31 |
| | GO:0070013 | intracellular organelle lumen | C | 0.00059 | 33 |
| | GO:0043233 | organelle lumen | C | 0.00062 | 33 |
| | GO:0031974 | membrane-enclosed lumen | C | 0.000829 | 33 |
| | GO:0070469 | respiratory chain | C | 0.000888 | 14 |
| | GO:0005740 | mitochondrial envelope | C | 0.000907 | 21 |
| | GO:0044429 | mitochondrial part | C | 0.00124 | 23 |
| | GO:0044434 | chloroplast part | C | 0.00194 | 52 |
| | GO:0044435 | plastid part | C | 0.002 | 54 |
| | GO:0022625 | cytosolic large ribosomal subunit | C | 0.0023 | 15 |
| | GO:0015934 | large ribosomal subunit | C | 0.00259 | 18 |
| | GO:0009535 | chloroplast thylakoid membrane | C | 0.00277 | 25 |
| | GO:0055035 | plastid thylakoid membrane | C | 0.00277 | 25 |
| | GO:0030964 | NADH dehydrogenase complex | C | 0.00321 | 11 |
| | GO:0045271 | respiratory chain complex I | C | 0.00321 | 11 |
| | GO:0031966 | mitochondrial membrane | C | 0.0057 | 19 |
| | GO:0031090 | organelle membrane | C | 0.00593 | 33 |
| | GO:0042651 | thylakoid membrane | C | 0.00635 | 25 |
| | GO:0043234 | protein complex | C | 0.00664 | 60 |
| | GO:0019866 | organelle inner membrane | C | 0.00823 | 20 |
| | GO:0044428 | nuclear part | C | 0.00852 | 34 |
| | GO:0015935 | small ribosomal subunit | C | 0.00882 | 15 |
| | GO:0034357 | photosynthetic membrane | C | 0.00966 | 25 |
| | GO:0005198 | structural molecule activity | F | 9.91E-14 | 54 |
| | GO:0003735 | structural constituent of ribosome | F | 1.18E-09 | 40 |
| | GO:0005515 | protein binding | F | 3.43E-05 | 91 |
| | GO:0005488 | binding | F | 0.00335 | 308 |

Table A.2 Representative individuals from each sampled population of *Helianthus annuus* L. are noted by the collection dates and voucher information at the Mississippi State University herbarium (MISSA).

| Population | Collection date | Voucher Number |
|---|---|---|
| KS1 | 4 September 2010 | 37144 |
| KS2 | 4 September 2010 | 37139 |
| KS3 | 5 September 2010 | 37143 |

Table A.3 Forward (F) and Reverse (R) Primer sequences, microsatellite repeat motif and copy number, exonic region or location of microsatellite within the gene, expected product size, Genbank accession number of EST sequence, putative function as per BLAST analysis, F primer concentration and annealing temperature (Ta) of additional 2 microsatellite primers utilized in Chapter III.

| Primer | Primer Sequences | Repeat motif & Copy Number | Region | Product size | Genbank Accession # | Putative function | Forward Primer (μM) | Ta (°C) |
|---|---|---|---|---|---|---|---|---|
| L31 | F: 5'cac gac gtt gta aaa cga cca caa att taa agg tga aaa tca ca3'<br>R: 5'tct gtc tag aac aaa gaa taa aaa cca3' | $(TC)_{23.5}$ | 5'UTR | 504 | DY904501.1 | leucine-rich repeat family protein | 0.15 | 52 |
| L5 | F: 5'cac gac gtt gta aaa cga gtg aag ctg ccc aga atg at3'<br>R: 5'acc cac atc aag aac cca ag3' | $(GAT)_{31.3}$ | 5'UTR | 230 | BU025569.1 | Nucleosome assembly protein1 | 0.15 | 50 |