

GENOME ASSEMBLY AND CIS-REGULATORY ELEMENT DETECTION IN THE COMPACT

UTRICULARIA GIBBA GENOME

By

Lynsey Kovar

(Under the Direction of Jason Wallace)

ABSTRACT

Cis-regulatory elements (CREs) such as insulators have been demonstrated to shield genes from the effects of transcriptional misregulation due to nearby promoter elements of neighboring genes. These elements will be increasingly important for controlling expression in future multi-gene crop traits. In this study, we aim to identify CREs such as insulator, terminator, bidirectional promoter, and unidirectional promoter elements in the *Utricularia gibba* (bladderwort) genome. This organism is an exceptional model for CRE detection due to its extremely small genome size. The *U. gibba* genome used in this study was Illumina sequenced, assembled and annotated. RNA-seq data was then used to detect pairs of independently expressed genes genome-wide. Intergenic regions between independently expressed gene-pairs were subsequently used for CRE detection. Putative insulator, terminator, unidirectional promoter and bidirectional promoter sequences were identified and filtered based on the presence of conserved elements in angiosperm genome alignments. The candidate CREs will undergo *in vivo* validation by collaborators.

INDEX WORDS: Bioinformatics, genome assembly, comparative genomics, plant genome

GENOME ASSEMBLY AND CIS-REGULATORY ELEMENT DETECTION IN THE COMPACT

UTRICULARIA GIBBA GENOME

By

Lynsey Kovar

B.S., New Mexico State University, 2017

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial

Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2020

© 2020

Lynsey Kovar

All Rights Reserved

GENOME ASSEMBLY AND CIS-REGULATORY ELEMENT DETECTION IN THE COMPACT

UTRICULARIA GIBBA GENOME

By

Lynsey Kovar

Major Professor: Jason Wallace

Committee: James Leebens-Mack
Casey Bergman

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

ACKNOWLEDGEMENTS

Thanks first and foremost to Jason Wallace for the incredible amount of help and support he provided during my graduate school career and for fostering a lab environment that allows people to grow as researchers and individuals. In addition, I'd like to thank my committee members Casey Bergman and Jim Leebens-Mack for giving me guidance along the way.

Thanks also to my lab mates. Specifically, Matthew Johnson and Sahar Voghoei, who provided lots of great conversation and meaningful insight and made the lab a genuinely positive place to be the whole time I was there. Also, thanks to Corey Schultz, Kivanc Corut, Hanxia Li and Darrian Talamantes, Holly Griffis and Naomi Rodman for joining the lab. I hope I can have a work group in the future who I enjoy being around as much as you all. I also couldn't have made it through the grad school experience without fellow grad student, roommate, and friend Hallie Wright, who made my education and life here in Athens full of great memories and for always being able to make me laugh. The community here at UGA in general is amazing and everyone I've met has shaped my education and sense of self for the better. I'm thankful for the past few years I spent here.

Lastly, I'd like to thank my parents for always supporting me through my entire educational journey. I wouldn't be here if it weren't for their kindness and generosity in all years of schooling.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1 BACKGROUND AND INTRODUCTION	1
Cis-regulatory elements in plant biotechnology	2
<i>Utricularia gibba</i> as a model organism	3
Conservation of plant Cis-regulatory elements and motif detection	6
Present study	9
2 CIS-REGULATORY ELEMENT REGION DETECTION AND ANALYSIS IN	
<i>UTRICULARIA GIBBA</i>	10
Abstract	11
Introduction	11
Methods	17
Results	23
Discussion	33
3 CONCLUSION	35
REFERENCES	37

LIST OF TABLES

	Page
Table 1.1: Genome size comparison between angiosperm species.....	5
Table 2.1. Quality metrics of the newly assembled <i>U. gibba</i> genome compared to the previously published genome.	24
Table 2.2: Significant motifs as discovered by MEME for candidate bidirectional promoter, unidirectional promoter, terminator, and insulator sequences.....	30

LIST OF FIGURES

	Page
Figure 1.1. Distribution of intergenic region lengths between gene pairs in convergent, divergent, and parallel orientation	7
Figure 2.1. Unidirectional promoter, bidirectional promoter, terminator and insulator region representative gene pair orientations and expression patterns	17
Figure 2.2. Quality metrics associated with each genomic dataset used in this study.....	19
Figure 2.3. Cladogram of species used in whole genome alignments for the conservation analysis.....	22
Figure 2.4. Blobplot showing coverage, size and GC content distribution of different phylogenetic groups represented in contigs of the initial assembly.....	25
Figure 2.5. Maker annotation edit distance (AED) distributions after rounds one, two, and three of maker annotation	26
Figure 2.6. Number of mapped reads to genic regions in each sample when using the de-novo maker annotation vs. the mapped annotation.....	27
Figure 2.7. One candidate from each CRE class.....	28
Figure 2.8. Distribution of conserved sequences in each intergenic CRE candidate class	32

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The rise of plant transformation in the past 30 years has led to yield increases of 21% on average and lessened pesticide inputs by 37% (Klumper and Qaim 2014). It is widely known that agricultural production will need to further increase from 25% to as much as 70% to keep pace with the projected population growth by 2050 (Hunter et al. 2017). Currently, herbicide tolerance and insect resistance traits make up the majority of genetically modified (GM) crops being grown worldwide (ISAAA 2017). Introducing GM traits that address a wider array of concerns is paramount to increasing yields in the face of climate change. Future traits should serve to lessen environmental impacts of agriculture by incorporating those that enhance stress tolerance and decrease inputs needed. Such traits will likely require integration of complex metabolic pathways, as benefits such as enhanced stress tolerance and optimized nutrient uptake tend to require the coordination of multiple genes and regulatory elements. The number of genes needed for a trait to work properly can even differ between species. For example, the synthesis of biopolymer polyhydroxybutyrate (PHB) reached 20% dry weight (DW) PHB in *Arabidopsis* using an expression cassette containing 5 genes (Kourtz et al. 2007). However, introducing the same pathway, plus two additional genes needed for boosted efficiency in switchgrass (an economically feasible crop) is far less efficient at max 7% DW PHB (Somleva et al. 2012). Moreover, the two extra genes added were introduced on different expression cassettes, which would make future breeding efforts with this trait more challenging (Halpin 2005).

There are known to be problems with expression regulation in multigene cassettes. Issues involving coordination of expression of transgenes are thought to arise either from the chromatin environment near the insertion site, or epigenetic silencing of transgenes (such as via methylation) induced by the host organism (Que et al. 2010). Positional effects of cis-regulatory elements (CREs) within and around the multigene cassettes can also influence expression of transgenes. Enhancer bleedover can affect genes at a distance of 50kb or more away. The constitutive Cauliflower Mosaic Virus 35S Promoter (CaMV 35s) can affect the expression of genes over 78kb away and can cause tissue-specific promoters to express regardless of tissue type (Zheng et al. 2007). This enhancer-promoter crosstalk phenomenon has been observed for constitutive and tissue/organ specific promoters. For example, the pollen-specific LAT52 promoter was found to activate expression of a transgene under the control of stigma specific promoter in pollen (Liu, Zhou, and Wu 2008). Due to the harmful effects of enhancer-promoter interference on multigene cassettes, finding ways to reduce or mitigate them is imperative to advancing plant transformation technology.

Cis-regulatory elements in plant biotechnology

Elements such as insulators and matrix attachment regions (MARs) are known to reduce interference from enhancers and mitigate transgene silencing. These elements serve to create a chromatin barrier that isolates the gene-space from flanking regions (Valenzuela and Kamakaka 2006). In animal systems, insulators and MARs have been utilized successfully for maintaining stable transgene expression regardless of genomic insertion point. Currently, insulators such as CTCF (Bell, West, and Felsenfeld 1999),

su(Hw) (Spana, Harrison, and Corces 1988), and BEAF (Zhao, Hart, and Laemmli 1995) have shown potential for use in biotechnology in metazoan systems. Some of these insulators and MARs have also been tested in plants, but only a few have shown promising results. Comparatively little research has been done in plants to identify similar elements. Hiley et al. (2009) tested three plant MARs. Of all three, only the petunia TBS MAR element was found to block interactions between the CaMV 35s enhancer and a downstream transgene promoter in *Arabidopsis* plants. A similar study was carried out using three putative plant insulator/MAR elements and found that only two of three showed true enhancer blocking activity (Yang, Singer, and Liu 2010). These two insulator sequences and the one that worked in the Hiley et al. (2009) study are both longer than 1kb. Thus, their size could be a hindrance for use in engineering complex traits given that transformation efficiency decreases with increasing cassette size (S.H. Park 2000). Additionally, terminator sequence elements serve to terminate transcription after the target gene has been transcribed. Currently, terminators used for most biotechnology applications do not completely prevent read through (Xing et al. 2010). Finding elements such as insulators and strong terminators with potential for use in plant biotechnology is a primary aim of our research.

***Utricularia gibba* as a model system**

Utricularia gibba, otherwise known as humped bladderwort, is an aquatic carnivorous plant species with one of the smallest genomes sequenced to date. It is closely related to the model organisms snapdragon (*Antirrhinum*) and monkey flower (*Mimulus*).

The first *U. gibba* genome assembly was released in 2013 and was produced using a hybrid 454 and Illumina/Sanger strategy (Ibarra-Laclette et al. 2013). This produced an 82-megabase (Mb) assembly, which is ~5Mb larger than the genome size estimated using flow-cytometry. This assembly also identified ~28,500 genes, which is more than the ~27,500 in *Arabidopsis* (Cheng et al. 2017). A second, more comprehensive assembly was released in 2017 (Lan et al. 2017). This assembly was ~102Mb in length and contained ~29,600 annotated genes and was generated using PacBio sequencing on 10 flow cells. The difference in genome size observed between assemblies was mostly due to the ability of PacBio reads to incorporate longer repetitive regions, which were unable to assemble in the short-read assembly.

In Lentibulariaceae, the family encompassing *Utricularia*, 95% of species are known to have a 1C-value smaller than 1000Mbp. Species in this family have been used previously to investigate characteristics of genomic gain and loss among closely related species (Veleba et al. 2014, Lan et al. 2017, Carretero-Paulet, Chang, et al. 2015, Fleischmann et al. 2014, Vu et al. 2015, Carretero-Paulet, Librado, et al. 2015). Interestingly, in spite of its small genome, *U. gibba* has undergone at least three whole-genome duplication events, two of which have occurred since its divergence from tomato and grapevine (Ibarra-Laclette et al. 2013). Its small genome size is likely due to repression of mobile elements and a series of microdeletions through time as evidenced by loss of retrotransposon segments, intron deletions, and compressed promoter spaces (Ibarra-Laclette et al. 2013). When compared to the eudicots *Arabidopsis*, grape, *Mimulus* and tomato, which have diverse whole genome duplication histories, *U. gibba* showed a gene death rate significantly higher than all other species (Carretero-Paulet, Librado, et al. 2015).

Large-scale differences in genome size are mostly due to intergenic sequence content, which varies in size depending on activities of transposable elements (Tenaillon, Hollister, and Gaut 2010, Bennetzen and Wang 2014). *U. gibba* has highly reduced intergenic sequence content when compared to other angiosperms, at around 50% of the genome compared to 85% in tomato (Tomato Genome 2012), 77% in soybean (Schmutz et al. 2010), and 95% in corn (Jiao et al. 2017) (See Table 1) . Our preliminary investigations using the published PacBio assembly (Lan et al. 2017) indicated that 62% of gene pairs are

SPECIES	GENOME LENGTH (MB)	NUMBER OF GENES	INTERGENIC SEQUENCE LENGTH (MB)	SOURCE
<i>Utricularia gibba</i>	101	29,666	51	(Lan et al. 2017)
<i>Arabidopsis thaliana</i>	120	28,775	45	(Lamesch et al. 2012)
<i>Selaginella moellendorffii</i>	210	34,551	154	(Banks et al. 2011)
<i>Mimulus guttatus</i>	313	50,930	140	(Hellsten et al. 2013)
<i>Medicago truncatula</i>	402	55,706	278	(Young et al. 2011)
<i>Physcomitrella patens</i>	473	65,820	364	(Lang et al. 2018)
<i>Vitis vinifera</i>	486	26,346	316	(Jaillon et al. 2007)
<i>Solanum lycopersicum</i>	828	34,879	707	(Tomato Genome 2012)
<i>Glycine max</i>	978	56,044	788	(Schmutz et al. 2010)
<i>Zea mays</i>	2,066	39,656	1,972	(Jiao et al. 2017)

Table 1.1: Genome size comparison between angiosperm species.

less than 1Kb apart based on annotations from the most recent genome assembly. Of these gene pairs, many show a >50-fold difference in expression despite being spaced at a considerably short distance when compared to other angiosperms, these gene pairs show signs of strong regulatory separation.

Conservation of plant cis-regulatory elements and motif detection

Considerable interest has been placed on finding and characterizing conserved noncoding sequences (CNSs) that are responsible for tight control of gene expression (Korkuc, Schippers, and Walther 2014, Yang et al. 2011, Hettiarachchi et al. 2014, Freeling and Subramaniam 2009, Haudry et al. 2013, Van de Velde et al. 2016). CNSs are normally found through alignment-based methodology. Sequence conservation and synteny can give rise to detection of homologous and orthologous loci. *U. gibba*, because of its small genome, serves as an ideal model system for mining CNSs and assessing CNS localization and conservation in a highly dynamic genome with propensity for DNA loss.

Previous studies have identified conserved noncoding sequences from angiosperms, though attention has not been focused on identifying insulator or terminator sequences specifically. A study in Arabidopsis identified promoter-specific novel and previously characterized CNSs using whole-genome SNP data from the 1,001 genomes project (Korkuc, Schippers, and Walther 2014). Hettiarachchi et al. (2014) identified lineage specific CNSs among eudicots, monocots, grasses, and angiosperms that presumably shaped the distinct morphological characteristics of these lineages. More comprehensive studies have aimed to identify all CNSs in specific lineages utilizing large genomic datasets

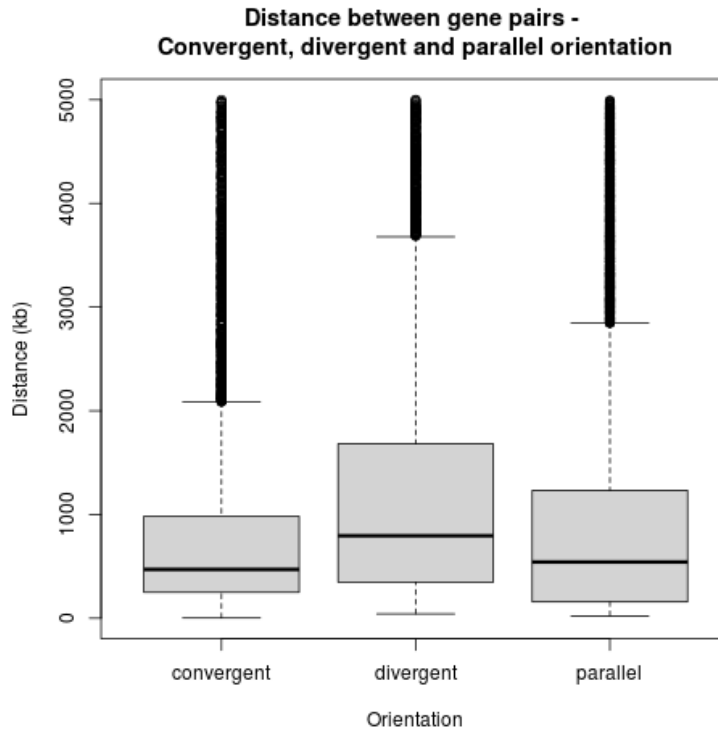


Figure 1.1. Distribution of intergenic region lengths between gene pairs in convergent, divergent, and parallel orientation. Outliers above 5k in length were not displayed.

and whole genome multi-alignments (Haudry et al. 2013, Kaplinsky et al. 2002, Casillas, Barbadilla, and Bergman 2007, Davydov et al. 2010). It is clear from these studies that the CREs outside the promoter regions of genes comprise a large portion of CNSs. However, these sequences, until recent years, have remained difficult to characterize. Large scale transcriptomic, ChIP-seq, and open chromatin datasets have provided means for elucidating functionality of these sequences.

There are two primary algorithmic approaches to detect motifs within CNS regions: enumerative detection, and alignment-based detection (MacIsaac and Fraenkel 2006). Enumerative detection, utilized in software such as Weeder, uses a list of all words up to a

predetermined size in a dataset. The statistical significance of each word is then evaluated. Alignment-based detection involves developing a probabilistic model of the input sequences and optimization to find common motifs across sequences. The commonly-used software MEME uses an alignment-based algorithm that incorporates expectation maximization to find the likelihood of the observed sequence data given a probability matrix of the motif model and treating the rest of the sequence as Markovian background (Bailey and Elkan 1995). Typically, for genome-wide studies, an alignment is produced to first gauge high-confidence intergenic regions that are conserved and then algorithmic approaches are used to hone in on motifs (Haudry et al. 2013, Hettiarachchi et al. 2014, MacIsaac and Fraenkel 2006, Prabhakar et al. 2006, Thomas et al. 2007, Van de Velde et al. 2016).

To find regulatory motifs such as insulators and terminators, candidate genomic regions must first be established. A study in *Arabidopsis* assessed intergenic space between pairs of divergent genes (genes transcribed in opposite directions) that showed large differences in expression despite sharing the same promoter region (Yang et al. 2011). These regions were thought to contain insulator sequences because they were independently expressed yet shared the same promoter region. It is widely known that enhancers in gene promoters can influence the expression of neighboring genes (Xie, He, and Gan 2001), so genes showing an independent expression pattern despite having promoters in some cases <400bp apart suggests the presence of insulator elements. In *U. gibba*, the average distance between neighboring genes is ~1500bp (Figure 1.1), making mining of sequences between independently expressed genes less computationally expensive, as motif prediction algorithms work better on sequences with less genomic noise (Hu, Li, and Kihara 2005).

Present study

This project aims to find candidate loci containing insulators, short terminators, and promoters using tissue-specific expression data in *U. gibba* and sequence conservation across angiosperms. After candidate regulatory sequences are identified, they will be tested *in vivo* by a collaborator. In parallel, CNSs will be analyzed for conservation within *U. gibba* and across the asterid, rosid, and monocot clades. Though the primary aim of this study is to find putative insulator sequences, this will also result in a comprehensive list of sequences containing candidate insulators, terminators, and promoters. It is our hope that these datasets and the corresponding analyses will pave the way for enhanced transformation of multigene cassettes in plants and identify a suite of potential cis-regulatory sequences conserved in angiosperm genomes.

CHAPTER 2

CIS-REGULATORY ELEMENT CANDIDATE DETECTION IN *UTRICULARIA GIBBA*¹

¹Kovar, L.K. and J. Wallace. To be submitted to *The Plant Cell*.

Abstract

Finding regulatory elements to maintain consistent gene expression patterns in multi-gene cassettes is essential to the future of plant biotechnology. Currently, the most widely used way to separate genes from regulatory effects of their neighbors is by using long intergenic sequences to act as spacers. *Utricularia gibba* has a highly compressed intergenic space compared to other sequenced plant genomes while maintaining expression patterns consistent with independent regulation. The existence of insulator sequences in animal lineages has inspired an effort to find similar sequences in plants. An accession-specific *Utricularia gibba* genome was assembled and annotated. RNA-Seq data then was used to find intergenic sequences flanked by genes with expression patterns consistent with regulatory element presence. Novel insulator sequences are the main priority of this study, but terminators, bidirectional promoters, and unidirectional promoters that act independently of their surrounding genomic environment were also sought out. Intergenic sequences potentially containing each of the regulatory element classes were found. Many of these intergenic regions contained conserved noncoding sequences when aligned with other asterid genomes. Some conserved noncoding elements located distally to the transcription start site were even conserved in lineages as distant as monocots.

Introduction

Engineering of quantitative traits in plants will require integration of multiple genes on a single cassette. Traits such as stress tolerance, nutrient uptake, and biopolymer

synthesis will require many genes to confer worthwhile benefits (Halpin 2005, Ye et al. 2000, Wang et al. 2018). Currently, most transgenic plants are monogenic, or have arisen from multiple single-gene transformation events (Halpin 2005, Naqvi et al. 2010, Que et al. 2010). This is mostly due to the difficulty in coordinating expression of multiple genes, and the lack of availability of effective cis-regulatory elements that can be used in transformation. Insulators, terminators, bidirectional promoters and unidirectional promoters that confer independent expression patterns and are not sensitive to genomic context and proximity to enhancer elements are especially needed for future plant biotechnology and will make multi-gene transformation more manageable.

Unpredictability of transgene expression is a well-known phenomenon in plant biotechnology and is often referred to as the position effect, which can be caused by enhancer bleedover (Zheng et al. 2007). Expression patterns of transgenes and their neighbors can also vary based on species, insertion site in the genome and presence regulatory elements located on the expression cassette itself. These effects can result in unpredictable expression patterns, silencing of genes, and variability between transformants (Singer, Liu, and Cox 2012). This issue has been demonstrated in constitutive promoters such as the Cauliflower Mosaic Virus 35S promoter (35S CaMV) and nopaline synthase promoter (*nos*), which were first used in 1984 and 1985, respectively (Odell, Nagy, and Chua 1985, Shaw et al. 1984), and are still widely used today. The first mention of the enhancer bleedover phenomenon was in a 1988 study that found when the 35S CaMV promoter was positioned upstream of the *nos* promoter, the expression of the *nos* promoter was enhanced (Odell et al. 1988). Similarly, expression patterns of tissue specific promoters can become unspecific in the presence of the 35S CaMV promoter,

mimicking a constitutive pattern of expression (Zheng et al. 2007, Hily et al. 2009). This off-target expression can be mitigated by using a weaker promoter such as *nos*, but often strong expression of the transgene is required for the trait to be effective. In these cases, it becomes necessary to find a promoter that will provide strong expression but will not affect the expression patterns of nearby genes. The effectiveness of this strategy can vary from species to species depending on the promoter used. It seems there is no “catch-all” strong constitutive promoter that can be used without issue regardless of species.

The enhancer bleedover phenomenon is not limited to strong constitutive promoters. Tissue and timepoint specific promoters and their corresponding enhancers have been shown to produce off-target expression in proximal genes (Gudynaite-Savitch, Johnson, and Miki 2009). A previous study tested four flower-specific promoters in tobacco for their ability to cause mis-regulation of a nearby pollen- and stigma-specific promoter for *Pps* and found that three out of four caused off-target *Pps* expression in flowers (Wen et al. 2014). Additionally, weaker constitutive promoters like *nos* and *mas* have been shown to affect expression of nearby genes. When these promoters were located proximal to the seed-specific *napin* promoter, expression was observed in multiple tissues (Gudynaite-Savitch, Johnson, and Miki 2009). In this same study, the misexpression caused by proximal promoter elements could be mitigated by placing spacer DNA (in this case, 2.7 kb of LacZ coding sequence), between the head-to-head promoters or by flipping one of the genes so that they were in head-to-tail orientation and thus separated by a transcribed gene. However, when they used a sequence that was only ~1kb in length, misexpression was still observed. This means that the spacer DNA is effective in mitigating misexpression, but the length needed for promoter insulation will likely vary based on enhancer strength and

promoter sensitivity (Gudynaite-Savitch, Johnson, and Miki 2009, Jagannath et al. 2001). Also, this strategy presents problems for cassettes containing many genes, where the amount of spacer DNA needed would likely decrease transformation efficiency (S.H. Park 2000).

Sequence content has also been implicated as a factor in insulation effectiveness. A 2-kb sequence from *Petunia hybrida* and a 1-kb *EcoRI/Sall* fragment from bacteriophage lambda were found to insulate a flower specific promoter from the CaMV 35S promoter in head-to-head orientation. However, a 4-kb sequence from bacteriophage lambda did not confer the same effect under identical conditions (Hily et al. 2009, Singer, Hily, and Liu 2009). This implies some sequences could contain elements conferring better insulation efficiency.

In metazoans, there are a few well-studied insulator sequences and corresponding DNA-binding proteins being used to prevent misexpression in transformation. The proteins su(Hw) (Suppressor of hairy wing), BEAF, and Zw5 were characterized in *Drosophila melanogaster*, while CTCF is the most widely studied enhancer blocking protein in vertebrates. Binding of su(Hw) mostly occurs in regions containing the gypsy retrotransposon in *D. melanogaster* (Spana, Harrison, and Corces 1988). Binding of BEAF and Zw5 occurs near transcription start sites (Nègre et al. 2010, Jiang et al. 2009).

Interestingly, there is very little similarity in the binding sequences or protein characteristics of these three *D. melanogaster* insulators. For Zw5, binding motifs are difficult to discern since its zinc finger domain is able to recognize a diverse array of short nucleotide sequences (Gaszner, Vazquez, and Schedl 1999). BEAF, on the other hand, appears to bind a short, five nucleotide motif but its binding distribution depends more on

genomic context than the occurrence of the motif itself (Yang, Ramos, and Corces 2012, Zhao, Hart, and Laemmli 1995). In vertebrates, all of the well-studied enhancer blocking elements bind the CTCF protein (Bell, West, and Felsenfeld 1999); Half of the binding sites of which are located in intergenic regions, while the others are distributed in intragenic, and promoter regions (Chen et al. 2008) and the binding mostly occurs in CpG-rich regions (Wang et al. 2012). The lack of similarity between the known insulator protein structures and their binding sites leads to difficulty in identification of new insulators based on prior sequence information alone.

There are two types of insulators according to the literature: enhancer blockers and barriers. Enhancer blockers function to block the interaction between an enhancer and a promoter, while barriers serve to prevent the spread of heterochromatin which could potentially silence genes. Both of these types of insulators can be used for enhancement of transgene expression. For example, matrix attachment regions of the chicken lysozyme gene (*chiMARs*) and copies of the chicken HS4 (*cHS4*) insulator as well as the *scs/scs'* domain boundaries from *Drosophila* have been found to increase transgene expression and minimize variation in expression among transformants, even in distant species (Stief et al. 1989, Kellum and Schedl 1992, Ciana et al. 2001, Taboit-Dameron et al. 1999). The *cHS4*, gypsy, and *scs* insulators have been the most widely-used enhancer blockers for transgenic studies in metazoan systems (Cai and Levine 1995), but other insulators serving this function are beginning to emerge such as sea urchin *Ars1* and *sns5* in addition to human BEAD-1 (reviewed in Emery 2011). In this study, we are particularly interested in enhancer-blocking insulators.

The same insulators that have been proven effective in metazoan systems have shown mixed results in plants (Perez-Gonzalez and Caro 2019). A few different studies have tested the effect of using different insulator sequences in plant transformation cassettes.

Currently, the use of long sequences to separate genes in expression cassettes is the most widely used way to prevent misexpression in plants (Gudynaite-Savitch, Johnson, and Miki 2009). There is a need to identify a wider array of insulating sequences due to the unpredictability of their effectiveness in different contexts.

The *U. gibba* (bladderwort) genome is a highly compact plant genome. At approximately 100 Mb in length with an average amount of genes for a diploid plant genome, it contains a smaller than average amount of intergenic sequence relative to other sequenced angiosperm genomes such as maize, soybean, mimulus, and tomato to name a few. Despite this small amount of intergenic sequence, regulation of genes is tightly controlled, with many genes separated by <1000bp showing expression differences >100 fold (Lan et al. 2017). Our goal is to mine the *U. gibba* genome for putative regulatory elements that can be used on plant transformation vectors. We are particularly interested in insulator sequences but will also be looking for unidirectional promoters, bidirectional promoters, and terminators that can be used for fine-tuning gene expression in compressed intergenic spaces. In this study, we have re-sequenced the *U. gibba* genome, built upon the existing annotation, and used 3' RNA-Seq data to mine for intergenic regions associated with the presence of the four regulatory element classes. Resequencing and de-novo assembly of the genome was done to minimize error due to an unknown divergence between our genotype and the previously sequenced genotype, which we were unable to obtain. Inspection of intergenic regions for conserved sequences and motifs across the published PacBio genome

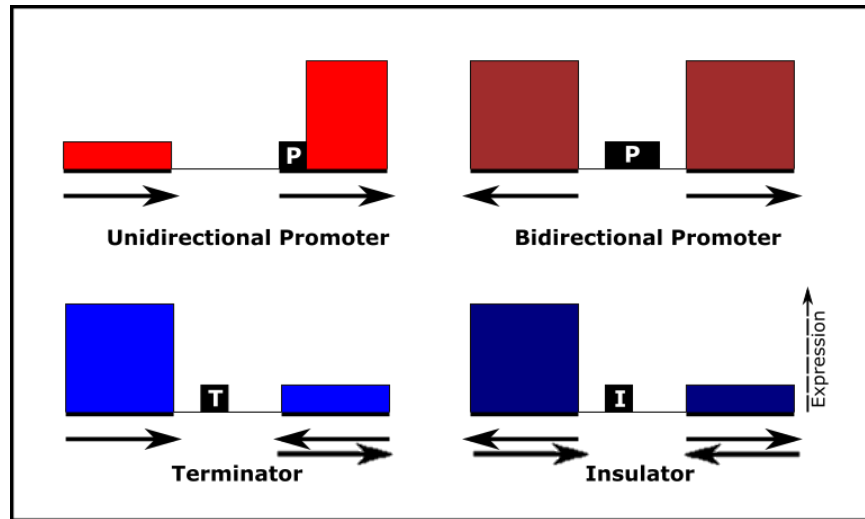


Figure 2.1. Representative gene pair orientations and expression patterns for unidirectional promoter, bidirectional promoter, terminator and insulator regions. Arrows represent possible orientations of genes around regions containing putative CREs.

was also carried out. The top 10 candidate sequences of each class will be validated by a collaborator in soybean hairy roots to confirm regulatory element activity. Once confirmed and characterized *in vivo*, these sequences will pave the way for enhanced multi-gene transformation vectors in plants.

Methods

DNA isolation and sequencing of Utricularia gibba

Utricularia gibba plugs were obtained from pitcherplant.org and propagated in three separate tanks. Collection site and genotype information were unknown at the time of propagation. Plugs were anchored in a layer of sand above a layer of peat moss under approximately six inches of water. DNA was isolated using the Promega Wizard Genomic

DNA Purification Kit and libraries were prepped using a KAPA library prep kit (#KK8231) and sequenced on two Illumina MiSeq flowcells at the Georgia Genomics and Bioinformatics Core, producing paired end 300bp reads.

RNA isolation and sequencing of Utricularia gibba

Tissues were separated by dissection under ice water and RNA was isolated using a Trizol/chloroform extraction from stem, leaf, rhizoid, bladder, and whole plant samples. After RNA extraction, a stranded KAPA kit (#KK8420) and Lexogen QuantSeq[®]™ 3' mRNA-Seq FWD kit were used to generate full transcript (whole plant) and tissue specific 3' RNA data (stem, leaf, rhizoid, bladder), respectively. The full transcript data was used to annotate the newly assembled *U. gibba* genome, while the 3' data was used for gene expression quantification. Full transcript KAPA libraries were sequenced on two Illumina MiSeq flowcells, producing paired end 250bp reads. 3' Libraries were sequenced on a mid-output NextSeq flowcell, producing single-end 150bp reads.

Genome assembly and annotation

Before genome assembly, R1 and R2 files from each flow cell were merged and then trimmed in Trimmomatic version 0.36 using the “ILLUMINACLIP” option with default options (max adapter mismatch count equal to two, palindrome clip threshold equal to 30, and match accuracy to 10) (Bolger, Lohse, and Usadel 2014). Additionally, the minimum length of a trimmed read in order to be kept was 36. After this, FastQC was used to evaluate read quality before proceeding with assembly (Andrews 2010). Common quality metrics

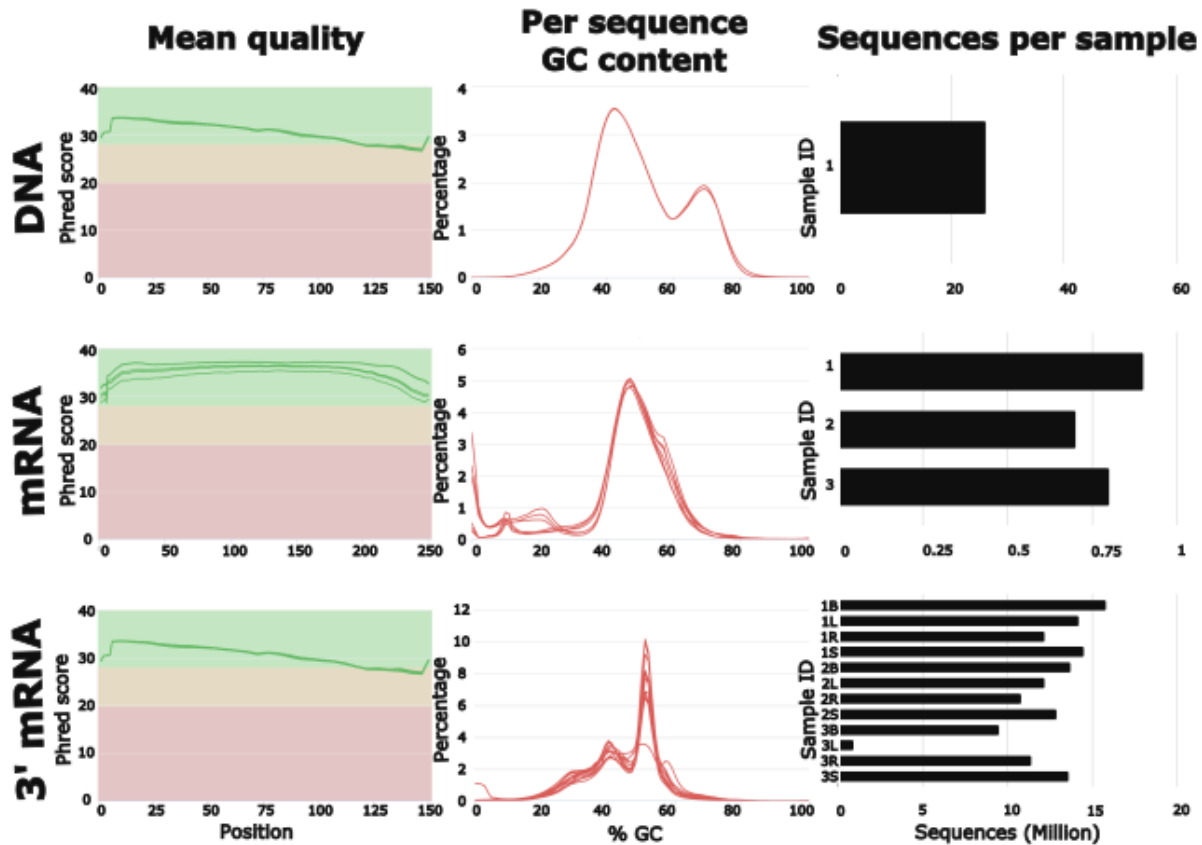


Figure 2.2. Quality metrics associated with each genomic dataset used in this study: DNA-seq, mRNA-seq and 3' mRNA-seq. The plots represent (left to right) mean per-base quality score of each read per sample, per-sequence GC content of each sample, and number of million reads per sample.

for our DNA samples were calculated using FastQC version 1.8.0 and MultiQC version 1.5, and are shown in Figure 2.2. The genome was assembled using SPAdes (Bankevich et al. 2012) version 3.13.1 with default parameters. Quality metrics such as length, N50, number of contigs, and BUSCO% were determined using QUAST version 5.0.2 (Gurevich et al. 2013). Contigs originating from contaminant, mitochondrial, and chloroplast genomes were filtered out using Kraken version 2.0.7 to search for sequence similarity in a database

of bacterial, fungal, plant, and published *Utricularia gibba* genome sequences (Wood and Salzberg 2014). Annotation was completed using Maker version 2.31.10 and Augustus 3.2.3 (Cantarel et al. 2008, Korf 2004). The first round of annotation was done using Trinity version 2.6.6 (Haas et al. 2014) assembled transcripts from the whole transcript mRNA-Seq data in addition to protein models from other sequenced asterid genomes: *Daucus carota* (GCF_001625215.1), *Helianthus annuus* (GCF_002127325.1), *Nicotiana attenuate* (GCF_001879085.1), and *Solanum lycopersicum* (GCF_000188115.4). After the initial round of annotation, putative genes were pulled out and used to train a gene prediction model in Augustus. Gene models were refined in the next rounds of annotation using Augustus gene prediction models trained after each Maker run. Maker was run for three rounds (until the AED score distribution stopped improving). Genes missing from the Maker annotation were added in using a BLAST (version 2.9.0) sequence similarity search of the published *U. gibba* PacBio genome annotations to the newly annotated genome. Mapped annotations had to have $\geq 95\%$ sequence similarity, $\geq 90\%$ sequence length and not overlap a Maker annotation in the new genome in order to be kept. Overlaps were found using BedTools intersect version 2.29.2.

Mining for regulatory regions using 3' RNA-Seq data

Raw 3' RNA-Seq reads were trimmed using Trimmomatic version 0.36 using the “ILLUMINACLIP” option with max adapter mismatch count equal to two, palindrome clip threshold equal to 30, and match accuracy to 10. Quality metrics were then visualized using FastQC version 1.8.0 and MultiQC version 1.5 and shown in Figure 2.2. Reads were mapped to the genome using the STAR aligner and reads per gene were quantified using htseq-

count version 0.9.1. Once counts were obtained, they were normalized in DESeq2 using the standard “DESeq” function (Love, Huber, and Anders 2014). Pairs of genes were extracted and classified as either divergent, convergent, or parallel orientation and their expression levels were used to mine for intergenic sequences containing unidirectional promoters, bidirectional promoters, insulators and terminators.

The criteria for selecting putative regulatory regions differed based on element class. For unidirectional promoters, terminators and insulators, intergenic regions were selected if they had a fold change in expression between gene pairs that was greater than the 90% quantile fold change for the dataset. In addition to a high fold change, the genes both had to be expressed (no zero expression values), be less than 1000bp apart, and the higher expressed gene needed to be greater than the median expression level for the dataset. For bidirectional promoter region selection, the fold change needed to be less than the 10% quantile for the dataset, less than 1000bp apart, and one or both genes needed to be expressed greater than the median expression level for the dataset.

Once a list of candidate regions meeting these criteria were selected, they were prioritized by fold change (either high or low depending on element class), sequence length (shorter rather than longer), and consistency across datasets. Consistency was based on Pearson correlation coefficient of expression of both genes in respective gene pairs across datasets. For bidirectional promoters, they needed to be highly correlated (>0.6). For unidirectional promoters, terminators and insulators, they needed to be uncorrelated (<0.5 and >-0.5). All code used to find candidate regulatory regions can be found in a publicly accessible GitHub repository: <https://github.com/lkov0/bladderwort-analysis>.

Intergenic sequence conservation and presence of putative insulator sequences

Conserved intergenic sequences were found using whole genome multiple alignments of different clades of angiosperms to the newly assembled *U. gibba* genome. The four clades comprising whole genome alignments were asterids, rosids, and monocots. Representative species from each clade can be seen in Figure 2.3.

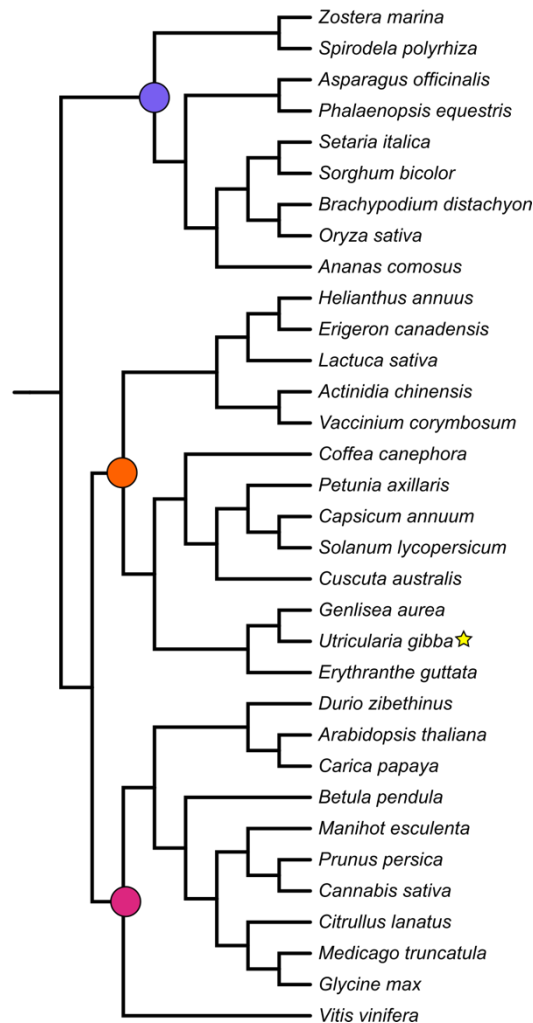


Figure 2.3. Cladogram of species used in whole genome alignments for the conservation analysis. Colored circles represent clades corresponding to different alignments. Purple = monocots, orange = asterids, and pink = rosids. The yellow star indicates *Utricularia gibba*. Generated using PhyloT (<https://phylot.biobyte.de/>).

Alignments of each clade to *Utricularia gibba* were completed using the UCSC genome browser full genome alignment tutorial (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto), but using LastZ instead of MultiZ for the initial alignment step. After alignments were completed and combined maf files were generated, PhastCons was used to generate genome wide sets of conserved coordinates for each clade with options target-coverage = 0.125 and expected-length = 20 (Siepel et al. 2005). All other options were set to default. Analysis of conserved region distribution in gene pairs was carried out in R (Scripts available at <https://github.com/lkov0/bladderwort-analysis>). MEME was used to find over-represented motifs in sets of co-expressed genes and was also used to find genome-wide distribution of those motifs (Bailey et al. 2009). In putative promoter and non-promoter regions.

Results

Genome assembly and annotation

SPAdes produced an initial assembly of around 705Mb. After using Kraken2 to examine the phylogenetic classification of the contigs, we found that there was a high amount of contamination in our dataset. Contaminant contig size, guanine/cytosine content (GC content) and corresponding taxonomic classification can be seen in Figure 2.4. The contigs not classified as *U. gibba* were filtered out using Kraken. Only contigs which were classified as NCBI Tax ID 13748 were kept which led to an assembly size of around 101Mb. Chloroplast and mitochondrial contigs made up around 300Kb of the assembly. Assembly

Table 2.1. Quality metrics of the newly assembled *U. gibba* genome compared to the previously published genome.

QUALITY METRIC	PACBIO ASSEMBLY	MY ASSEMBLY
Length	102Mb	101Mb
# Contigs	581	13033
# Genes	29,666	27,166
BUSCO (%)	88%	84%

metrics for the final assembly compared to the existing PacBio assembly can be seen in Table 2.1. The new assembly is much less contiguous than the PacBio assembly, which is in line with expectations given the read lengths of Illumina vs. PacBio technology. The overarching goal for our assembly was that enough pairs of genes would be assembled on the same contig so that expression data could be used to pull out intergenic sequences containing putative regulatory elements.

The genome annotation was run in three iterations. The first iteration produced 69.9% complete BUSCOs and ~22,000 genes, the second iteration produced 70.3% complete BUSCOs and ~15,500 genes, and the third iteration produced 73.7% complete BUSCOs and ~17,400 genes. We chose to stop at the third round of annotation because that is when the annotation edit distance distribution stopped improving (See Figure 2.5).

When verifying the completeness of the annotation, the gene length distribution was compared to the gene length distributions of the existing genomes. Interestingly, our annotation contained a higher proportion of longer genes and appeared to be missing a large proportion of short genes compared to the other genomes. To add missing genes, the

PacBio genome annotations were mapped to our genome using BLASTn and the best hit was kept. New gene coordinates were added to the annotation if they did not overlap with any Maker-annotated genes. This led to a total of ~27,000 genes in the final annotation. We also compared the ability of our annotation to capture our mapped 3' RNA-Seq reads when

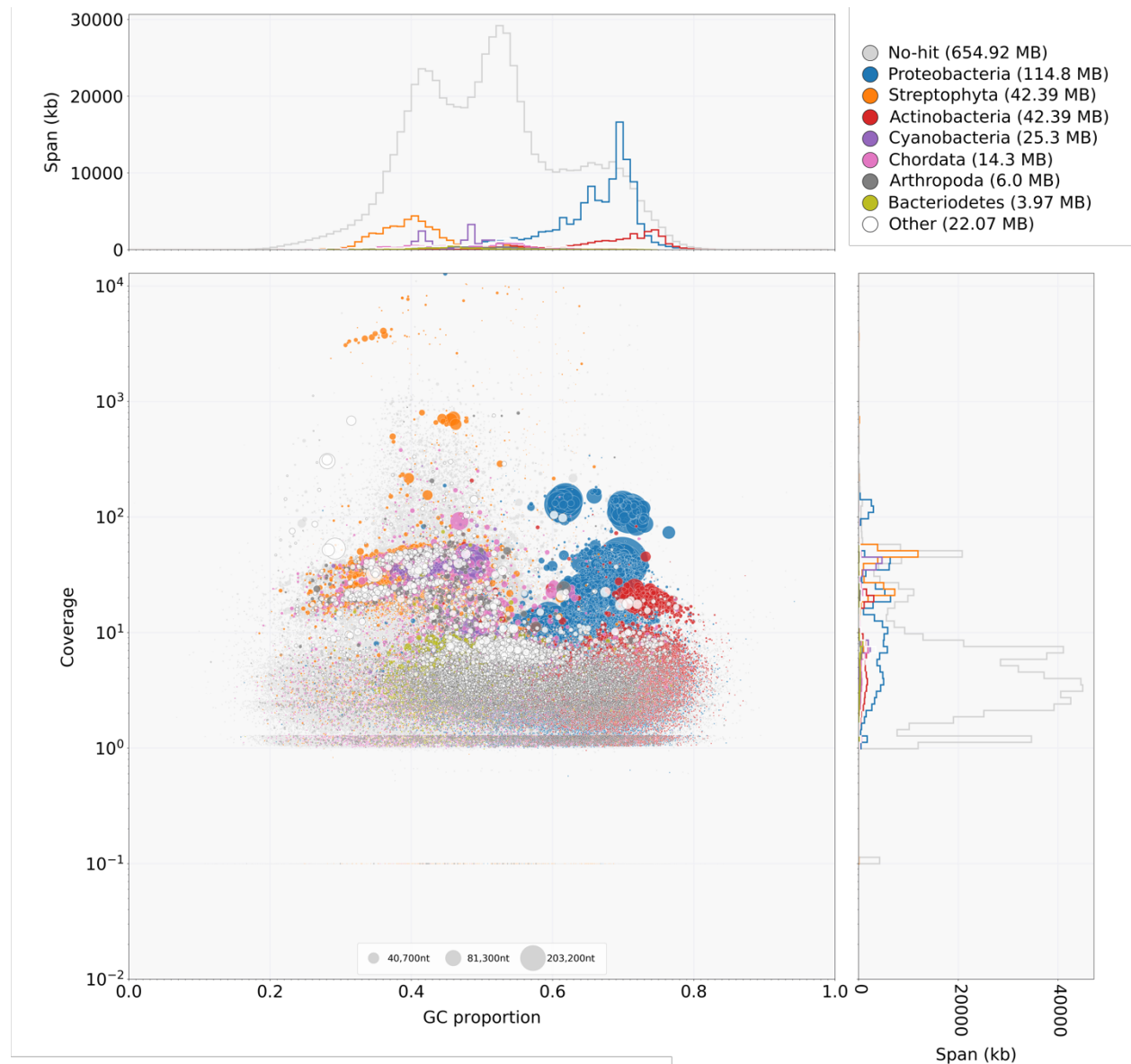


Figure 2.4. Blobplot showing coverage, size and GC content distribution of different phylogenetic groups represented in contigs of the initial pre-filtered *U. gibba* assembly.

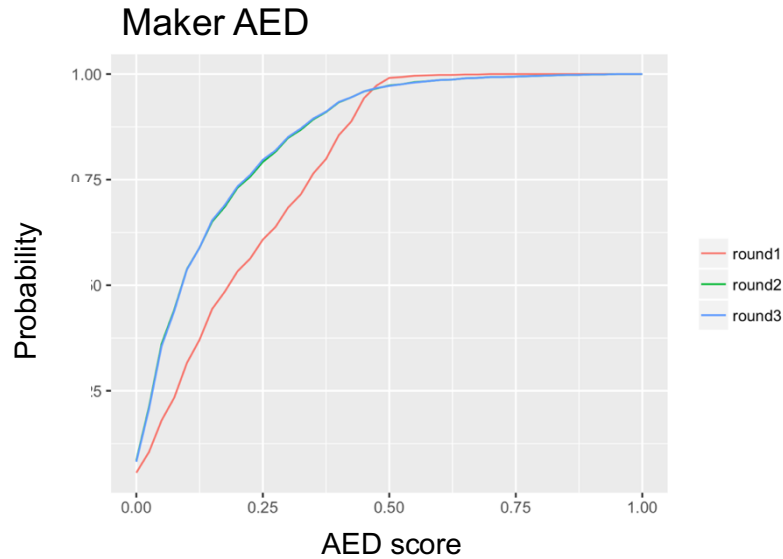


Figure 2.5. Maker annotation edit distance (AED) distributions after rounds one, two, and three of maker annotation.

compared to the previous annotation. A higher proportion of reads were counted per gene when using our de-novo Maker annotation than when using the mapped PacBio annotation coordinates (see Figure 2.6). This shows that annotating the newly assembled genome using transcripts from the same genotype and extensive protein evidence from closely related genomes was necessary for establishing proper gene boundaries.

Mining for insulators, terminators, and promoters

We were interested in finding unidirectional promoters, bidirectional promoters, insulators and terminators using our 3' RNA-Seq expression data. Particularly, we wanted

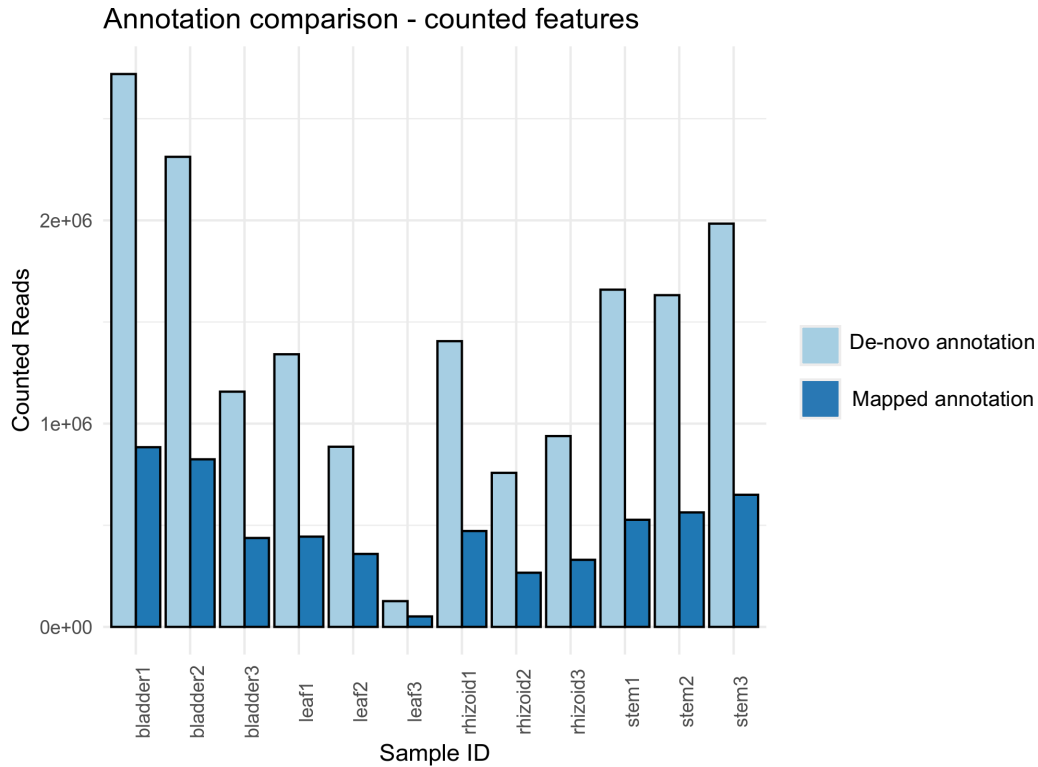
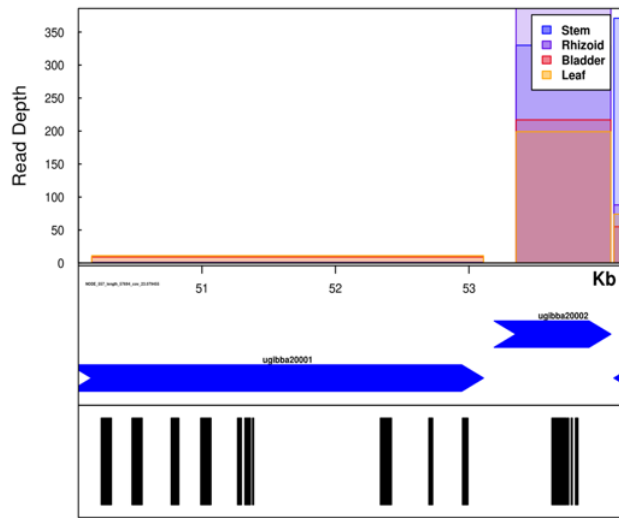
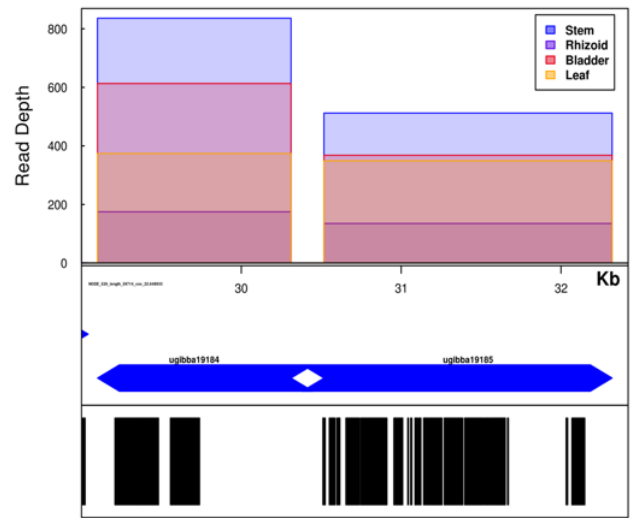


Figure 2.6. Number of mapped reads to genic regions in each sample when using the de-novo maker annotation vs. the mapped annotation.

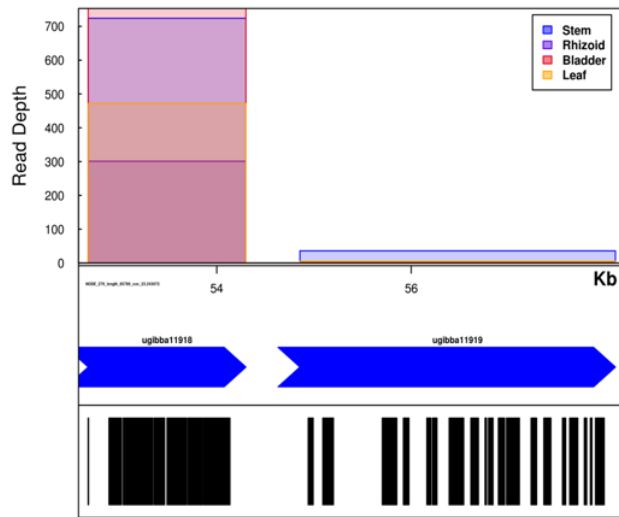
to pull out whole intergenic regions between gene pairs showing expression patterns of interest as candidates for containing regulatory elements. The first step was to quantify expression for each gene. When performing alignment, we found that one of the leaf samples was an outlier with an extremely low amount of reads so that sample was not included in the analysis. We therefore had three bladder samples, two leaf samples, three rhizoid samples, and three stem samples. For insulators, terminators, and unidirectional promoters we were interested in gene pairs that showed strong regulatory separation despite being close together. Putative regulatory elements in those intergenic regions



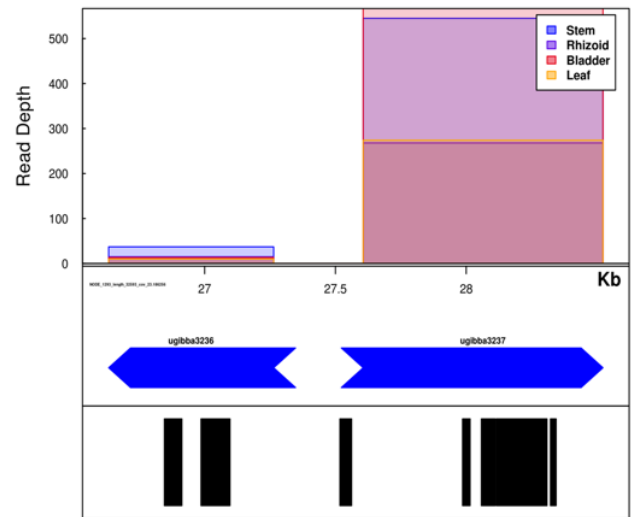
Unidirectional Promoter



Bidirectional Promoter



Terminator



Insulator

Figure 2.7. Example candidates from each CRE class. The top panel in each graph shows median expression levels of the four sampled tissues. The second panel shows gene annotations in blue arrows to represent directionality. Black bars indicate conserved regions based on an asterid whole-genome alignment.

would be useful for biotechnology due to their apparent shielding from enhancer bleedover. Only gene pairs that showed a high fold change in expression, low correlation across datasets, and short distance were considered as candidates. For bidirectional promoters, gene pairs showing low fold change in expression, high correlation across datasets and short distance were used to find candidate intergenic regions. A whole genome alignment of *U. gibba* and other asterid genomes was also used to find conserved regions in candidate intergenic sequences. Figure 2.7 shows a representative element of each class. In total there were 43 unidirectional promoters, 161 bidirectional promoters, 75 terminators, and 43 insulators meeting the aforementioned criteria. Of these, 74 bidirectional promoters, 17 insulators, 29 terminators, and 20 unidirectional promoters showed conservation based on the asterid whole-genome alignment.

Intergenic sequence motif analysis

For unidirectional promoters, bidirectional promoters, terminators and insulators there were 2, 7, 2, and 1 overrepresented motifs found, respectively. Their sequence logos can be found in Table 2.2. When searching the sequences against the same promoter database mentioned above, there was significant similarity in two of the bidirectional promoter motifs to a MADS box and an AP2-related TF family, respectively. Aside from these, there were no other motifs with significant similarity to known promoters in this database. We also checked for similarity to our putative promoters from the CEMiTool analysis and 11 out of 12 motifs contained significant hits to those. There was no significant

difference found with regard to GC content between the different putative CRE classes ($p = 0.16$).

Intergenic sequences were further analyzed for presence of non-promoter sequence motifs. Since little is known about insulators and terminators in plants, our goal was to seek out motifs that were associated primarily with non-promoter regions. To do that, a list of putative promoter-affiliated motifs was needed. The CEMiTools package in R (Russo et al. 2018) was used to find groups of co-expressed genes and identified 10 modules. Sequences










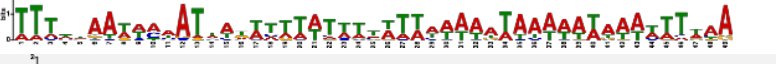

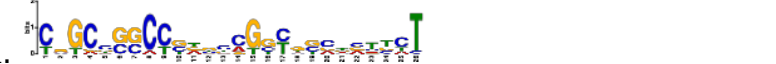
Sequence logo	p value	Prop. of sequences	CRE type
*1. 	2.9e-83	.43	BP
*2. 	6.1e-17	.53	BP
*3. 	2.7e-13	.40	BP
*4. 	7.8e-9	.30	BP
*5. 	9.9e-6	.20	BP
*6. 	1.0e-5	.38	BP
*7. 	2.0e-5	.39	BP
8. 	9.4e-7	.58	UP
*9. 	1.9e-8	.49	UP
*10. 	3.1e-14	.40	T
*11. 	2.0e-3	.17	T
*12. 	8.2e-3	.30	I

Table 2.2: Significant motifs as discovered by MEME for candidate bidirectional promoter (BP), unidirectional promoter (UP), terminator (T), and insulator (I) sequences. An asterisk indicates significant similarity to a putative promoter found using the CEMiTools analysis.

200bp upstream of the transcription start site for each module were used as inputs to MEME to find motifs affiliated with promoters. Motifs found were then searched against the Plant PAN database (Chow et al. 2019). In total, there were 25 motifs found and 12 had significant similarity ($q < 0.005$) to a motif in Plant PAN.

Alignment based conservation in intergenic regions

Whole-genome alignments of the newly sequenced *U. gibba* genotype and groups of asterid, rosid, and monocot genomes respectively were also performed to gauge conservation patterns within candidate CRE-containing sequences. A visual representation of conserved regions in each intergenic sequence type, normalized by intergenic sequence length can be seen in Figure 2.8. A majority of conserved regions are directly proximal to genes, likely indicating portions of promoters, terminators, upstream ORFs or similar. To inspect the proportion of conserved sequences likely originating from promoters and non-promoters, conserved sequences in each element class were extracted and searched for similarity against the Plant PAN promoter database. We found that 44%, 38%, 47%, and 31% of conserved sequences in putative insulator, terminator, unidirectional promoter and bidirectional promoter intergenic regions respectively show similarity to known promoters, and that these are widely distributed across the intergenic space in each putative CRE class. This demonstrates a possibility for protein binding outside of the probable promoter region in many of our candidate sequences.

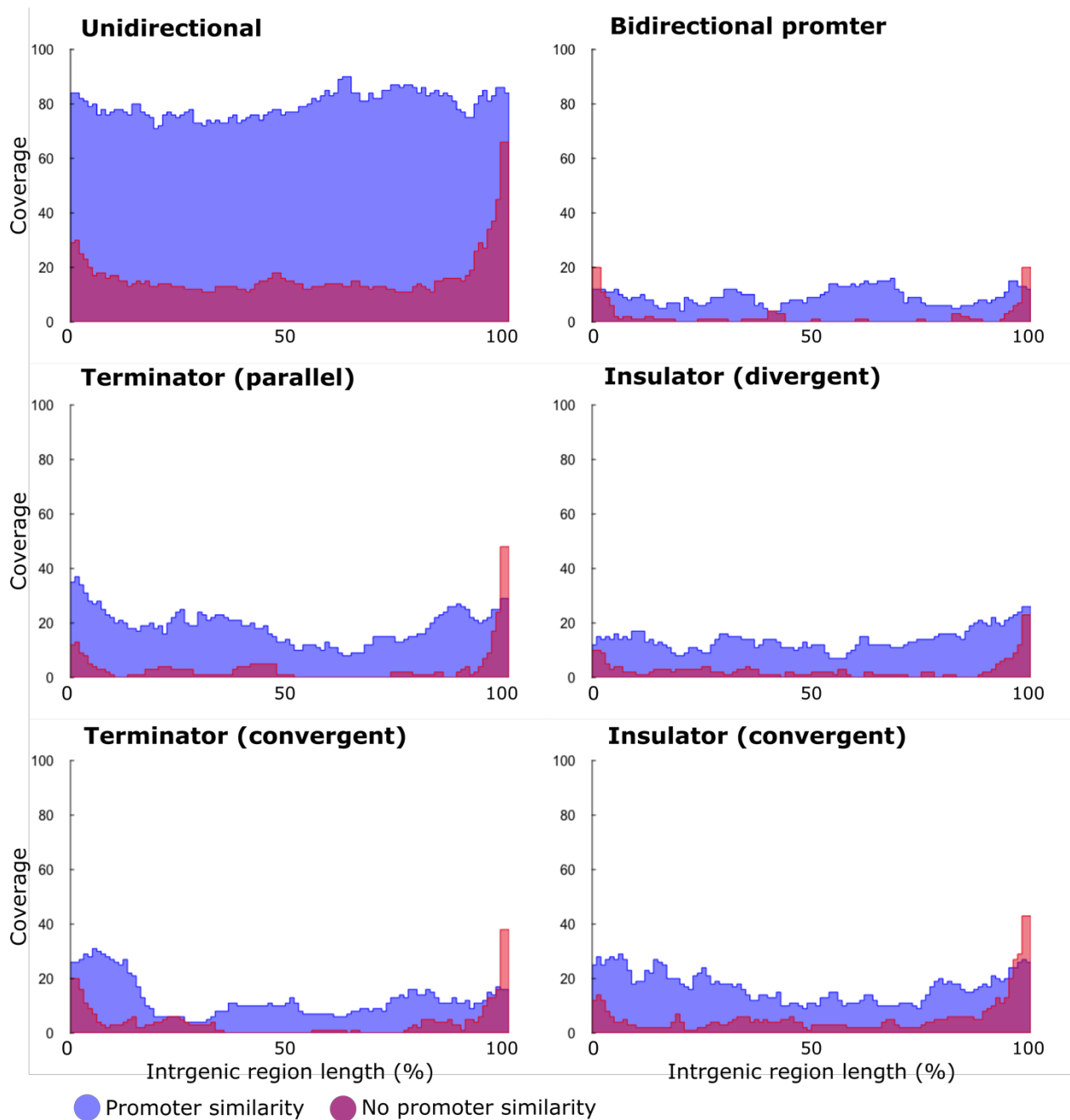


Figure 2.8. Locations of conserved sequences in each type of CRE candidate. X axis represents conserved sequence coverage, colored by if sequence similarity to a promoter in the Plant PAN database exists there. (normalized by percent of region length)

Discussion

This study aimed to find novel regulatory sequences for use in plant biotechnology. First, the specific genotype used in this study was assembled and annotated. Although it is not as contiguous as the previous published PacBio assembly, the genotype-specific assembly yielded an adequate genomic percentage when compared to the previous assembly for us to proceed with mining for regulatory regions. A more contiguous assembly would be possible only through a change in sequencing strategy, such as using long read (e.g. PacBio or Nanopore) or insert size variation methodology. In total, 43 unidirectional promoter regions, 161 bidirectional promoter regions, 75 terminator regions, and 43 insulator regions were found based on using the most stringent cutoffs.

Twelve significant motifs were found using MEME on all four sets of sequences from our candidate CRE classes. Of those, 11 showed significant similarity to promoters found using a co-expression network approach and a sequence similarity search to the Plant PAN database. In addition to this, 7 out of 12 of the motifs found were in the bidirectional promoter CRE class. This class had by far the most CRE candidates: 161 compared to 43, 43, and 75 of each of the other classes. There is a possibility that our approach lacked the power needed to find motifs other than promoters. Relaxing our criteria could remedy this, but since the purpose of this study was to identify high confidence intergenic regions containing each CRE class for testing *in vivo*, stringency was our top priority.

There is a high probability that *in silico* methods alone will not be enough to identify the mechanism behind insulation activity. Sequence motifs themselves have yet to be implicated in plant insulation (Singer, Liu, and Cox 2012). If the insulation is due to protein

binding or three-dimensional structure similar to in metazoan systems, wet-lab based methods will likely be needed to implicate motifs if they play an important role (Lunyak 2008, Perez-Gonzalez and Caro 2019, Singer, Liu, and Cox 2012, Yang, Ramos, and Corces 2012, Bell, West, and Felsenfeld 1999). Many of the intergenic sequences in our candidate regions show conservation with asterids, rosids and monocots, and in many cases these sequences also show similarity to known promoters, even outside of putative promoter regions. In this early stage, it is difficult to assign importance to specific sequences in intergenic regions. Experimental evidence will be needed to determine if protein binding is actually occurring, and if this is important for enhancer blocking. The top ten intergenic regions of each class identified in this study will be experimentally validated using a reporter-assay based system in soybean hairy roots. If predicted regulatory activity is conferred by any of the candidates, they will be added to the repertoire of publicly available sequences for use in plant biotechnology. Additionally, lab-based and computational analyses will hopefully help us gain a deeper understanding of why enhancer-blocking activity is conferred and how that mechanism functions in plant systems.

CHAPTER 3

CONCLUSION

The compact *U. gibba* genome, which is smaller than that of *Arabidopsis* but contains more genes, served as a useful model for mining for sequences containing cis-regulatory elements due to its sparse intergenic sequence space. Using replicated tissue-specific RNA-Seq data, it is possible to find genes that show independent regulation and drastically different expression patterns despite being separated by short distances. Using patterns specific to each regulatory element class, we were able to find intergenic regions containing putative unidirectional promoters, bidirectional promoters, terminators, and insulators. Whole genome alignments with three different angiosperm clades revealed that many of these intergenic sequences contained conserved regions outside of the putative promoter space. Interestingly, some sequences even showed conservation with distant monocot species. Evolutionary conservation provides evidence of these sequences having a potentially meaningful biological role. This role, however, will need to be experimentally validated and explored to elucidate mechanisms leading to enhancer blocking in candidates such as insulators and terminators. Apart from classifying conserved sequences as containing known protein binding motifs, little is known about the functionality of these sequences.

The first step of the characterization process will be confirming the regulatory activity of these sequences in reporter assays, which are being carried out for the top ten candidates in each class by the Parrott Lab at the University of Georgia. If validated, these sequences

will serve to ease future multi-gene transformation efforts in the plant science community by providing a means of preventing misregulation in tightly packed genic spaces.

Additionally, the workflow outlined here, if successful, could be applied to other species for the mining of putative CRE-containing sequences. If any of these sequences contain enhancer-blocking insulator activity, more experiments to characterize them will provide some of the first insight into mechanisms of action of these sequences in plants.

REFERENCES

- Andrews, Simon. 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. 2009. "MEME SUITE: tools for motif discovery and searching." *Nucleic acids research* no. 37 (suppl_2):W202-W208.
- Bailey, Timothy L., and Charles Elkan. 1995. "The Value of Prior Knowledge in Discovering Motifs with MEME." *ISMB-95 Proceedings*.
- Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. 2012. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *J Comput Biol* no. 19 (5):455-77. doi: 10.1089/cmb.2012.0021.
- Banks, Jo Ann, Tomoaki Nishiyama, Mitsuyasu Hasebe, John L. Bowman, Michael Gribskov, Claude dePamphilis, Victor A. Albert, Kaoki Aono, Tsuyoshi Aoyama, Barbara A. Ambrose, Niel W. Ashton, Michael J. Axtell, Elizabeth Barker, Michael S. Barker, Jerrery L. Bennetzen, Nicholas D. Bonawitz, Clint Chapple, Chaoyang Cheng, Luiz Gustavo Guedes Correa, Michael Darce, Jeremy DeBarry, Ingo Dreyer, Marek Elias, Eric M. Engstrom, Mark Estrella, Liang Feng, Cedric Finet, Sandra K. Floyd, Wolf B. Frommer, Tomomichi Fujita, Lydia Gramzow, Michael Gutensohn, and Jasper Harholt. 2011. "The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants." *Science* no. 332:960-963.
- Bell, Adam C., Adam G. West, and Gary Felsenfeld. 1999. "The Protein CTCF is Required for the Enhancer Blocking Activity of Vertebrate Insulators." *Cell* no. 98:387-396.
- Bennetzen, J. L., and H. Wang. 2014. "The contributions of transposable elements to the structure, function, and evolution of plant genomes." *Annu Rev Plant Biol* no. 65:505-30. doi: 10.1146/annurev-arplant-050213-035811.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* no. 30 (15):2114-20. doi: 10.1093/bioinformatics/btu170.
- Cai, Haini, and Michael Levine. 1995. "Modulation of enhancer-promoter interactions by insulators in the Drosophila embryo." *Nature* no. 376 (6540):533-536.
- Cantarel, Brandi L, Ian Korf, Sofia MC Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes." *Genome research* no. 18 (1):188-196.
- Carretero-Paulet, L., T. H. Chang, P. Librado, E. Ibarra-Laclette, L. Herrera-Estrella, J. Rozas, and V. A. Albert. 2015. "Genome-wide analysis of adaptive molecular evolution in the carnivorous plant *Utricularia gibba*." *Genome Biol Evol* no. 7 (2):444-56. doi: 10.1093/gbe/evu288.

- Carretero-Paulet, L., P. Librado, T. H. Chang, E. Ibarra-Laclette, L. Herrera-Estrella, J. Rozas, and V. A. Albert. 2015. "High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant *Utricularia gibba*." *Mol Biol Evol* no. 32 (5):1284-95. doi: 10.1093/molbev/msv020.
- Casillas, S., A. Barbadilla, and C. M. Bergman. 2007. "Purifying selection maintains highly conserved noncoding sequences in *Drosophila*." *Mol Biol Evol* no. 24 (10):2222-34. doi: 10.1093/molbev/msm150.
- Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. 2008. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." *Cell* no. 133 (6):1106-17. doi: 10.1016/j.cell.2008.04.043.
- Cheng, C. Y., V. Krishnakumar, A. P. Chan, F. Thibaud-Nissen, S. Schobel, and C. D. Town. 2017. "Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome." *Plant J* no. 89 (4):789-804. doi: 10.1111/tpj.13415.
- Chow, C. N., T. Y. Lee, Y. C. Hung, G. Z. Li, K. C. Tseng, Y. H. Liu, P. L. Kuo, H. Q. Zheng, and W. C. Chang. 2019. "PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants." *Nucleic Acids Res* no. 47 (D1):D1155-D1163. doi: 10.1093/nar/gky1081.
- Ciana, Paolo, Giovanni Di Luccio, Silvia Belcredito, Giuseppe Pollio, Elisabetta Vegeto, Laura Tatangelo, Cecilia Tiveron, and Adriana Maggi. 2001. "Engineering of a mouse for the in vivo profiling of estrogen receptor activity." *Molecular Endocrinology* no. 15:1104-1113.
- Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. 2010. "Identifying a high fraction of the human genome to be under selective constraint using GERP++." *PLoS Comput Biol* no. 6 (12):e1001025. doi: 10.1371/journal.pcbi.1001025.
- Fleischmann, A., T. P. Michael, F. Rivadavia, A. Sousa, W. Wang, E. M. Temsch, J. Greilhuber, K. F. Muller, and G. Heubl. 2014. "Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms." *Ann Bot* no. 114 (8):1651-63. doi: 10.1093/aob/mcu189.
- Freeling, M., and S. Subramaniam. 2009. "Conserved noncoding sequences (CNSs) in higher plants." *Curr Opin Plant Biol* no. 12 (2):126-32. doi: 10.1016/j.pbi.2009.01.005.
- Gaszner, Miklos, Julio Vazquez, and Paul Schedl. 1999. "The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction." *Genes and Development* no. 13:2098-2107.
- Gudynaite-Savitch, L., D. A. Johnson, and B. L. Miki. 2009. "Strategies to mitigate transgene-promoter interactions." *Plant Biotechnol J* no. 7 (5):472-85. doi: 10.1111/j.1467-7652.2009.00416.x.
- Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. "QUAST: quality assessment tool for genome assemblies." *Bioinformatics* no. 29 (8):1072-5. doi: 10.1093/bioinformatics/btt086.
- Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, D Philip, Joshua Bowden, Matthew Brian Couger, D Eccles, B Li, and MD Macmanes. 2014. "Reference generation and analysis with trinity." *Nat Protoc* no. 8 (8):1-43.

- Halpin, C. 2005. "Gene stacking in transgenic plants--the challenge for 21st century plant biotechnology." *Plant Biotechnol J* no. 3 (2):141-55. doi: 10.1111/j.1467-7652.2004.00113.x.
- Haudry, A., A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, Z. Joly-Lopez, J. G. Steffen, K. M. Hazzouri, K. Dewar, J. R. Stinchcombe, D. J. Schoen, X. Wang, J. Schmutz, C. D. Town, P. P. Edger, J. C. Pires, K. S. Schumaker, D. E. Jarvis, T. Mandakova, M. A. Lysak, E. van den Bergh, M. E. Schranz, P. M. Harrison, A. M. Moses, T. E. Bureau, S. I. Wright, and M. Blanchette. 2013. "An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions." *Nat Genet* no. 45 (8):891-8. doi: 10.1038/ng.2684.
- Hellsten, U., K. M. Wright, J. Jenkins, S. Shu, Y. Yuan, S. R. Wessler, J. Schmutz, J. H. Willis, and D. S. Rokhsar. 2013. "Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing." *Proc Natl Acad Sci U S A* no. 110 (48):19478-82. doi: 10.1073/pnas.1319032110.
- Hettiarachchi, Nilmini, Kirill Kryukov, Kenta Sumiyama, and Naruya Saitou. 2014. "Lineage-Specific Conserved Noncoding Sequences of Plant Genomes: Their Possible Role in Nucleosome Positioning." *Genome Biology and Evolution* no. 6 (9):2527-2542. doi: 10.1093/gbe/evu188.
- Hily, J. M., S. D. Singer, Y. Yang, and Z. Liu. 2009. "A transformation booster sequence (TBS) from *Petunia hybrida* functions as an enhancer-blocking insulator in *Arabidopsis thaliana*." *Plant Cell Rep* no. 28 (7):1095-104. doi: 10.1007/s00299-009-0700-8.
- Hu, J., B. Li, and D. Kihara. 2005. "Limitations and potentials of current motif discovery algorithms." *Nucleic Acids Res* no. 33 (15):4899-913. doi: 10.1093/nar/gki791.
- Hunter, Mitchell C., Richard G. Smith, Meagan E. Schipanski, Lesley W. Atwood, and David A. Mortensen. 2017. "Agriculture in 2050: Recalibrating Targets for Sustainable Intensification." *BioScience* no. 67 (4):386-391. doi: 10.1093/biosci/bix010.
- Ibarra-Laclette, E., E. Lyons, G. Hernandez-Guzman, C. A. Perez-Torres, L. Carretero-Paulet, T. H. Chang, T. Lan, A. J. Welch, M. J. Juarez, J. Simpson, A. Fernandez-Cortes, M. Arteaga-Vazquez, E. Gongora-Castillo, G. Acevedo-Hernandez, S. C. Schuster, H. Himmelbauer, A. E. Minoche, S. Xu, M. Lynch, A. Oropeza-Aburto, S. A. Cervantes-Perez, M. de Jesus Ortega-Estrada, J. I. Cervantes-Luevano, T. P. Michael, T. Mockler, D. Bryant, A. Herrera-Estrella, V. A. Albert, and L. Herrera-Estrella. 2013. "Architecture and evolution of a minute plant genome." *Nature* no. 498 (7452):94-8. doi: 10.1038/nature12132.
- ISAAA. 2017. "Global Status of Commercialized Biotech/GM Crops: 2017 ISAAA Brief No. 53."
- Jagannath, Arun, Panchali Bandyopadhyay, N. Arumugam, Vibha Gupta, Pradeep Kumar Burma, and Deepak Pental. 2001. "The use of a Spacer DNA fragment insulates the tissue specific expression of a cytotoxic gene (barnase) and allows high-frequency generation of transgenic male sterile lines in *Brassica juncea* L." *Molecular Breeding* no. 8:11-23.
- Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Huguene, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G.

- Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pe, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quetier, P. Wincker, and Characterization French-Italian Public Consortium for Grapevine Genome. 2007. "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." *Nature* no. 449 (7161):463-7. doi: 10.1038/nature06148.
- Jiang, Nan, Eldon Emberly, Olivier Cuvier, and Craig M Hart. 2009. "Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in *Drosophila melanogaster* links BEAF to transcription." *Molecular and cellular biology* no. 29 (13):3556-3568.
- Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C. S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. L. Schneider, T. K. Wolfgruber, M. R. May, N. M. Springer, E. Antoniou, W. R. McCombie, G. G. Presting, M. McMullen, J. Ross-Ibarra, R. K. Dawe, A. Hastie, D. R. Rank, and D. Ware. 2017. "Improved maize reference genome with single-molecule technologies." *Nature* no. 546 (7659):524-527. doi: 10.1038/nature22971.
- Kaplinsky, N. J., D. M. Braun, J. Penterman, S. A. Goff, and M. Freeling. 2002. "Utility and distribution of conserved noncoding sequences in the grasses." *Proc Natl Acad Sci U S A* no. 99 (9):6147-51. doi: 10.1073/pnas.052139599.
- Kellum, Rebecca, and Paul Schedl. 1992. "A group of scs elements function as domain boundaries in an enhancer blocking assay." *Molecular and Cellular Biology*:2424-2431.
- Klumper, W., and M. Qaim. 2014. "A meta-analysis of the impacts of genetically modified crops." *PLoS One* no. 9 (11):e111629. doi: 10.1371/journal.pone.0111629.
- Korf, Ian. 2004. "Gene finding in novel genomes." *BMC bioinformatics* no. 5 (1):59.
- Korkuc, P., J. H. Schippers, and D. Walther. 2014. "Characterization and identification of cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism information." *Plant Physiol* no. 164 (1):181-200. doi: 10.1104/pp.113.229716.
- Kourtz, Lauralynn, Kevin Dillon, Sean Daughtry, Oliver P. Peoples, and Kristi D. Snell. 2007. "Chemically inducible expression of the PHB biosynthetic pathway in *Arabidopsis*." *Transgenic Research* no. 16:759-769.
- Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala. 2012. "The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools." *Nucleic Acids Res* no. 40 (Database issue):D1202-10. doi: 10.1093/nar/gkr1090.
- Lan, Tianying, Tanya Renner, Enrique Ibarra-Laclette, Kimberly M. Farr, Tien-Hao Chang, Sergio Alan Cervantes-Pérez, Chunfang Zheng, David Sankoff, Haibao Tang, Rikky W. Purbojati, Alexander Putra, Daniela I. Drautz-Moses, Stephan C. Schuster, Luis Herrera-Estrella, and Victor A. Albert. 2017. "Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome." *Proc Natl Acad Sci U S A* no. 114 (27):E5483. doi: doi.org/10.1073/pnas.1702072114.
- Lang, D., K. K. Ullrich, F. Murat, J. Fuchs, J. Jenkins, F. B. Haas, M. Piednoel, H. Gundlach, M. Van Bel, R. Meyberg, C. Vives, J. Morata, A. Symeonidi, M. Hiss, W. Muchero, Y. Kamisugi, O. Saleh, G. Blanc, E. L. Decker, N. van Gessel, J. Grimwood, R. D. Hayes, S. W. Graham, L. E. Gunter, S. F. McDaniel, S. N. W. Hoernstein, A. Larsson, F. W. Li, P. F.

- Perroud, J. Phillips, P. Ranjan, D. S. Rokshar, C. J. Rothfels, L. Schneider, S. Shu, D. W. Stevenson, F. Thummler, M. Tillich, J. C. Villarreal Aguilar, T. Widiez, G. K. Wong, A. Wymore, Y. Zhang, A. D. Zimmer, R. S. Quatrano, K. F. X. Mayer, D. Goodstein, J. M. Casacuberta, K. Vandepoele, R. Reski, A. C. Cuming, G. A. Tuskan, F. Maumus, J. Salse, J. Schmutz, and S. A. Rensing. 2018. "The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution." *Plant J* no. 93 (3):515-533. doi: 10.1111/tpj.13801.
- Liu, Z., C. Zhou, and K. Wu. 2008. "Creation and analysis of a novel chimeric promoter for the complete containment of pollen- and seed-mediated gene flow." *Plant Cell Rep* no. 27 (6):995-1004. doi: 10.1007/s00299-008-0522-0.
- Love, M. I., W. Huber, and S. Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol* no. 15 (12):550. doi: 10.1186/s13059-014-0550-8.
- Lunyak, V. V. 2008. "Boundaries. Boundaries...Boundaries???" *Curr Opin Cell Biol* no. 20 (3):281-7. doi: 10.1016/j.ceb.2008.03.018.
- MacIsaac, K. D., and E. Fraenkel. 2006. "Practical strategies for discovering regulatory DNA sequence motifs." *PLoS Comput Biol* no. 2 (4):e36. doi: 10.1371/journal.pcbi.0020036.
- Naqvi, S., G. Farre, G. Sanahuja, T. Capell, C. Zhu, and P. Christou. 2010. "When more is better: multigene engineering in plants." *Trends Plant Sci* no. 15 (1):48-56. doi: 10.1016/j.tplants.2009.09.010.
- Nègre, Nicolas, Christopher D Brown, Parantu K Shah, Pouya Kheradpour, Carolyn A Morrison, Jorja G Henikoff, Xin Feng, Kami Ahmad, Steven Russell, and Robert AH White. 2010. "A comprehensive map of insulator elements for the Drosophila genome." *PLoS genetics* no. 6 (1).
- Odell, Joan T, Ferenc Nagy, and Nam-Hai Chua. 1985. "Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter." *Nature* no. 313:810-812.
- Odell, Joan T., Susan Knowlton, Willy Lin, and Jeffry Mauvais. 1988. "Properties of an isolated transcription stimulating sequence derived from the cauliflower mosaic virus 35S promoter." *Plant Mol Biol* no. 10:263-272.
- Perez-Gonzalez, A., and E. Caro. 2019. "Benefits of using genomic insulators flanking transgenes to increase expression and avoid positional effects." *Sci Rep* no. 9 (1):8474. doi: 10.1038/s41598-019-44836-6.
- Prabhakar, S., J. P. Noonan, S. Paabo, and E. M. Rubin. 2006. "Accelerated evolution of conserved noncoding sequences in humans." *Science* no. 314 (5800):786. doi: 10.1126/science.1130738.
- Que, Q., M. D. Chilton, C. M. de Fontes, C. He, M. Nuccio, T. Zhu, Y. Wu, J. S. Chen, and L. Shi. 2010. "Trait stacking in transgenic crops: challenges and opportunities." *GM Crops* no. 1 (4):220-9. doi: 10.4161/gmcr.1.4.13439.
- Russo, P. S. T., G. R. Ferreira, L. E. Cardozo, M. C. Burger, R. Arias-Carrasco, S. R. Maruyama, T. D. C. Hirata, D. S. Lima, F. M. Passos, K. F. Fukutani, M. Lever, J. S. Silva, V. Maracaja-Coutinho, and H. I. Nakaya. 2018. "CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses." *BMC Bioinformatics* no. 19 (1):56. doi: 10.1186/s12859-018-2053-1.

- S.H. Park, B.-M. Lee, M.G. Salas, M. Srivatanakul, R.H. Smith. 2000. "Shorter T-DNA or additional virulence genes improve *Agrobacterium*-mediated transformation." *Theoretical and Applied Genetics* (101):1015-1020.
- Schmutz, J., S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, and S. A. Jackson. 2010. "Genome sequence of the palaeopolyploid soybean." *Nature* no. 463 (7278):178-83. doi: 10.1038/nature08670.
- Shaw, C.H., G.H. Carter, M.D. Watson, and C.H. Shaw. 1984. "A functional map of the nopaline synthase promoter." *Nucleic Acids Res* no. 12:7831-7846.
- Siepel, Adam, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, and Stephen Richards. 2005. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." *Genome research* no. 15 (8):1034-1050.
- Singer, S. D., Z. Liu, and K. D. Cox. 2012. "Minimizing the unpredictability of transgene expression in plants: the role of genetic insulators." *Plant Cell Rep* no. 31 (1):13-25. doi: 10.1007/s00299-011-1167-y.
- Singer, Stacy D., Jean-Michel Hily, and Zongrang Liu. 2009. "A 1-kb Bacteriophage Lambda Fragment Functions as an Insulator to Effectively Block Enhancer–Promoter Interactions in *Arabidopsis thaliana*." *Plant Molecular Biology Reporter* no. 28 (1):69-76. doi: 10.1007/s11105-009-0122-3.
- Somleva, Mariya, Himani Chinnapen, Aminat Ali, Kristi D. Snell, Oliver P. Peoples, Nii Patterson, Jihong Tang, and Karen Bohmert-Tatarev. 2012. Increasing carbon flow for polyhydroxybutyrate production in biomass crops. In *Patent Application Publication*, edited by United States. United States.
- Spana, Carl, Douglas A. Harrison, and Victor G. Corces. 1988. "The *Drosophila melanogaster* suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon." *Genes and Development* no. 9:1414-1423.
- Stief, Aribert, D'iana M. Winter, Wolf H. Stratling, and Albrecht E. Sippel. 1989. "A nuclear DNA attachment element mediates elevated and position independent gene activity." *Nature* no. 341:343-345.
- Taboit-Dameron, Frederique, Benoit Malassagne, Celine Viglietta, Claudine Puissant, Mathieu Leroux-Coyau, Christiane Chereau, Joe Attal, Bernard Weill, and Louis-Marie Houdebine. 1999. "Association of the 5' HS4 sequence of the chicken β -globin locus control region with human EF1 α gene promoter induces ubiquitous and high expression of human CD55 and CD59 cDNAs in transgenic rabbits." *Transgenic Research* no. 8:223-235.
- Tenaillon, M. I., J. D. Hollister, and B. S. Gaut. 2010. "A triptych of the evolution of plant transposable elements." *Trends Plant Sci* no. 15 (8):471-8. doi: 10.1016/j.tplants.2010.05.003.
- Thomas, B. C., L. Rapaka, E. Lyons, B. Pedersen, and M. Freeling. 2007. "Arabidopsis intragenomic conserved noncoding sequence." *Proc Natl Acad Sci U S A* no. 104 (9):3348-53. doi: 10.1073/pnas.0611574104.

- Tomato Genome, Consortium. 2012. "The tomato genome sequence provides insights into fleshy fruit evolution." *Nature* no. 485 (7400):635-41. doi: 10.1038/nature11119.
- Valenzuela, L., and R. T. Kamakaka. 2006. "Chromatin insulators." *Annu Rev Genet* no. 40:107-38. doi: 10.1146/annurev.genet.39.073003.113546.
- Van de Velde, J., M. Van Bel, D. Vaneechoutte, and K. Vandepoele. 2016. "A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants." *Plant Physiol* no. 171 (4):2586-98. doi: 10.1104/pp.16.00821.
- Veleba, Adam, Petr Bures, Lubomir Adamec, Petr Smarda, Ivana Lipnerova, and Lucie Horova. 2014. "Genome size and genomic GC content evolution in the miature genome-size family Lentibulariaceae." *New Phytologist* no. 203:22-28.
- Vu, Giang T. H., Thomas Schmutzer, Fabian Bull, Hieu X. Cao, Jörg Fuchs, Trung D. Tran, Gabriele Jovtchev, Klaus Pistrick, Nils Stein, Ales Pecinka, Pavel Neumann, Petr Novak, Jiri Macas, Paul H. Dear, Frank R. Blattner, Uwe Scholz, and Ingo Schubert. 2015. "Comparative Genome Analysis Reveals Divergent Genome Size Evolution in a Carnivorous Plant Genus." *The Plant Genome* no. 8 (3). doi: 10.3835/plantgenome2015.04.0021.
- Wang, H., M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers, and J. A. Stamatoyannopoulos. 2012. "Widespread plasticity in CTCF occupancy linked to DNA methylation." *Genome Res* no. 22 (9):1680-8. doi: 10.1101/gr.136101.111.
- Wang, Z., C. Yang, H. Chen, P. Wang, P. Wang, C. Song, X. Zhang, and D. Wang. 2018. "Multi-gene co-expression can improve comprehensive resistance to multiple abiotic stresses in *Brassica napus* L." *Plant Sci* no. 274:410-419. doi: 10.1016/j.plantsci.2018.06.014.
- Wen, Z., Y. Yang, J. Zhang, X. Wang, S. Singer, Z. Liu, Y. Yang, G. Yan, and Z. Liu. 2014. "Highly interactive nature of flower-specific enhancers and promoters, and its potential impact on tissue-specific expression and engineering of multiple genes or agronomic traits." *Plant Biotechnol J* no. 12 (7):951-62. doi: 10.1111/pbi.12203.
- Wood, Derrick E, and Steven L Salzberg. 2014. "Kraken: ultrafast metagenomic sequence classification using exact alignments." *Genome biology* no. 15 (3):R46.
- Xie, Mingtang, Yuehui He, and Susheng Gan. 2001. "Bidirectionalization of polar promoters in plants." *Nature Biotechnology* no. 19:677-679.
- Xing, A., B. P. Moon, K. M. Mills, S. C. Falco, and Z. Li. 2010. "Revealing frequent alternative polyadenylation and widespread low-level transcription read-through of novel plant transcription terminators." *Plant Biotechnol J* no. 8 (7):772-82. doi: 10.1111/j.1467-7652.2010.00504.x.
- Yang, J., E. Ramos, and V. G. Corces. 2012. "The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species." *Genome Res* no. 22 (11):2199-207. doi: 10.1101/gr.142125.112.
- Yang, Xiaohan, Cara M. Winter, Xiuying Xia, and Shusheng Gan. 2011. "Genome-wide analysis of the intergenic regions in *Arabidopsis thaliana* suggests the existence of bidirectional promoters and genetic insulators." *Curent Topics in Plant Biology* no. 12.
- Yang, Yazhou, Stacy D. Singer, and Zongrang Liu. 2010. "Evaluation and comparison of the insulation efficiency of three enhancer-blocking insulators in plants." *Plant Cell*,

- Tissue and Organ Culture (PCTOC)* no. 105 (3):405-414. doi: 10.1007/s11240-010-9880-8.
- Ye, Xudong, Salim Al-Babili, Andreas Kloti, Jing Zhang, Paola Lucca, Peter Beyer, and Ingo Potrykus. 2000. "Engineering the provitamin A biosynthetic pathway into (carotenoid-free) rice endosperm." *Science* no. 287:303-305.
- Young, N. D., F. Debelle, G. E. Oldroyd, R. Geurts, S. B. Cannon, M. K. Udvardi, V. A. Benedito, K. F. Mayer, J. Gouzy, H. Schoof, Y. Van de Peer, S. Proost, D. R. Cook, B. C. Meyers, M. Spannagl, F. Cheung, S. De Mita, V. Krishnakumar, H. Gundlach, S. Zhou, J. Mudge, A. K. Bharti, J. D. Murray, M. A. Naoumkina, B. Rosen, K. A. Silverstein, H. Tang, S. Rombauts, P. X. Zhao, P. Zhou, V. Barbe, P. Bardou, M. Bechner, A. Bellec, A. Berger, H. Berges, S. Bidwell, T. Bisseling, N. Choisne, A. Couloux, R. Denny, S. Deshpande, X. Dai, J. J. Doyle, A. M. Dudez, A. D. Farmer, S. Fouteau, C. Franken, C. Gibelin, J. Gish, S. Goldstein, A. J. Gonzalez, P. J. Green, A. Hallab, M. Hartog, A. Hua, S. J. Humphray, D. H. Jeong, Y. Jing, A. Jocker, S. M. Kenton, D. J. Kim, K. Klee, H. Lai, C. Lang, S. Lin, S. L. Macmil, G. Magdelenat, L. Matthews, J. McCorrison, E. L. Monaghan, J. H. Mun, F. Z. Najar, C. Nicholson, C. Noirot, M. O'Bleness, C. R. Paule, J. Poulain, F. Prion, B. Qin, C. Qu, E. F. Retzel, C. Riddle, E. Sallet, S. Samain, N. Samson, I. Sanders, O. Saurat, C. Scarpelli, T. Schiex, B. Segurens, A. J. Severin, D. J. Sherrier, R. Shi, S. Sims, S. R. Singer, S. Sinharoy, L. Sterck, A. Viollet, B. B. Wang, K. Wang, M. Wang, X. Wang, J. Warfsmann, J. Weissenbach, D. D. White, J. D. White, G. B. Wiley, P. Wincker, Y. Xing, L. Yang, Z. Yao, F. Ying, J. Zhai, L. Zhou, A. Zuber, J. Denarie, R. A. Dixon, G. D. May, D. C. Schwartz, J. Rogers, F. Quetier, C. D. Town, and B. A. Roe. 2011. "The Medicago genome provides insight into the evolution of rhizobial symbioses." *Nature* no. 480 (7378):520-4. doi: 10.1038/nature10625.
- Zhao, Keji, Craig M. Hart, and Ulrich K. Laemmli. 1995. "Visualization of Chromosomal Domains with Boundary Element-Associated Factor BEAF-32." *Cell* no. 81:879-889.
- Zheng, X., W. Deng, K. Luo, H. Duan, Y. Chen, R. McAvoy, S. Song, Y. Pei, and Y. Li. 2007. "The cauliflower mosaic virus (CaMV) 35S promoter sequence alters the level and patterns of activity of adjacent tissue- and organ-specific gene promoters." *Plant Cell Rep* no. 26 (8):1195-203. doi: 10.1007/s00299-007-0307-x.