2016-05-01

The Path to Understanding Salt Tolerance: Global Profiling of Genes Using Transcriptomics of the Halophyte *Suaeda fruticosa*

Joann Diray Arce
Brigham Young University

The Path to Understanding Salt Tolerance: Global Profiling of Genes Using Transcriptomics of the Halophyte *Suaeda fruticosa*

Joann Diray Arce

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Brent L. Nielsen, Chair
Mark J. Clement
R. Paul Evans
Joel S. Griffitts
Peter J. Maughan

Department of Microbiology and Molecular Biology

Brigham Young University

May 2016

ABSTRACT

The Path to Understanding Salt Tolerance: Global Profiling of Genes Using Transcriptomics of the Halophyte *Suaeda fruticosa*

Joann Diray Arce
Department of Microbiology and Molecular Biology, BYU
Doctor of Philosophy

Salinity is a major abiotic stress in plants that causes significant reductions in crop yield. The need for improvement of food production has driven research to understand factors underlying plant responses to salt and mechanisms of salt tolerance. The aim of improving tolerance in traditional crops has been initiated but most crops can only tolerate a limited amount of salt in their systems to survive and produce biomass. Studies of naturally occurring high salt-tolerant plants (halophytes) are now being promoted for economic interests such as food, fodder or ecological reasons. *Suaeda fruticosa*, a member of the family Chenopodiaceae, belongs to a potential model halophyte genus for studying salt tolerance. However, published reports on the identification of genes, expression patterns and mechanisms of salinity tolerance in succulent halophytes are very limited. Next generation RNA-sequencing techniques are now available to help characterize genes involved in salinity response, along with expression patterns and functions of responsive genes. In this study, we have optimized the assembly of the transcriptome of *S. fruticosa*. We have annotated the genes based on their gene ontology characteristics and analyzed differential expression to identify genes that are up- and down-regulated in the presence of salt and have grouped the genes based on their putative functions. We also have provided evidence for groups of transcription factors that are involved in salt tolerance of this species and have identified those that may affect the regulation of salt tolerance. This work elucidates the characterization of genes involved in salinity tolerance to increase our understanding of the regulation of salt in a succulent halophyte.

Keywords: *Suaeda fruticosa*, halophytes, salt tolerance, transcriptome, RNA-seq, salinity

ACKNOWLEDGEMENTS

Words cannot fully express my appreciation to my Heavenly Father for all the talents, guidance and blessings He has given me and my family as I go through this graduate experience.

I would like to express my gratitude and forever appreciation to my mentor, Dr. Brent L. Nielsen for his full support, words of wisdom and exemplary dedication to my graduate experience. He has helped me succeed despite of my failures, limitations and challenges. He has given me an outstanding example of humility, hard work and faith that everything will work out in the end.

I would like to thank my excellent committee members, Dr. Griffitts, Dr. Maughan, Dr. Evans, Dr. Prince and Dr. Clement for their invaluable critiques and for allowing me to grow and become a better scientist. I would like to specially thank Dr. Mark J. Clement for his expertise and encouragement to learn Bioinformatics.

I would like to thank my BYU colleagues and collaborators that made significant contributions to my projects: Drs. Ajmal Khan and Bilquees Gul for the halophyte project, Bin Liu for the Twinkle project, Huan Kang for the proteomics analysis, Justin Page, Paul Bodily and Stanley Fujimoto for their bioinformatics expertise, Anton Suvorov for his evolutionary biology expertise, Collin Hansen, Stewart Morley and all Nielsen lab undergraduates for their help in the lab.

I would also want to express my gratitude to my parents, Jorge and Reneca Diray for teaching me that learning does not stop, and to my siblings and in-laws for their support and encouragement despite being far away from them.

Finally, I would like to dedicate this dissertation to the love of my life: my husband Carlo and my children born in-between writing papers and grants, Travis and Arabella.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Halophyte Transcriptomics: Understanding Mechanisms of Salinity Tolerance

Joann Diray-Arce[1], Bilquees Gul[2], M. Ajmal Khan[2,3], and Brent Nielsen[1]

[1]Department of Microbiology and Molecular Biology, Brigham Young University, Provo, Utah
[2]Institute of Sustainable Halophyte Utilization, University of Karachi, Pakistan
[3]Centre for Sustainable Development, College of Arts and Sciences, Qatar University, Doha, Qatar

© 2016 M.A. M.A. Khan, M. Ozturk, B. Gul, & M.Z. Ahmed (Eds):
Halophytes for Food Security in Dry Lands. DOI: http://dx.doi.org/10.1016/B978-0-12-801854-5.00010-8. Elsevier Inc.

ABSTRACT

Salinity stress represses biological processes that inhibit crop production. Most crops are glycophytes that can only tolerate low concentrations of salt. Halophytes are promising alternatives and have the ability to maintain productivity when grown with saline water. Studies of these naturally occurring salt-tolerant plants are being conducted to determine their economic potential as crops. Transcriptome analysis provides a promising approach for identifying gene expression patterns and functions of candidate genes involved in salt tolerance. This chapter reviews the current status of transcriptome analysis of halophytes, the implications on mechanisms of salt tolerance with the identified genes and proteins, and future prospects.

KEYWORDS

Halophyte; molecular biology; next-generation sequencing; salinity tolerance; RNA-seq; *Suaeda fruticosa*

INTRODUCTION

Studies of genomes and transcriptomes have rapidly advanced with next-generation sequencing (NGS) approaches. NGS technologies are utilized for single nucleotide polymorphism-based markers and draft sequencing of species without a reference genome (Wicker et al., 2006). These approaches have led to the discovery of markers that can be used to study genetic variations, population genetics, transcript profiling, mutations, and genetic associations for plant breeding (Qin et al., 2010). As NGS technologies have matured, RNA sequencing has become a preferred method for gene expression profiling (McGettigan, 2013), as it has the ability to identify transcripts and their expression over time and under different conditions. Transcriptome sequencing is less expensive than genome sequencing since only transcribed regions are investigated (Brautigam and Gowik, 2010).

Problems caused by high soil salinity for plants include the lowering of water potential leading to osmotic stress caused by cellular dehydration, toxicity of absorbed Na^+ and Cl^- ions which inhibits enzymatic activities and various cellular processes and the restriction of uptake of essential nutrients (Flowers and Colmer, 2008 and Abideen et al., 2014). Plant salinity tolerance involves mechanisms at the physiological and molecular levels. Physiological response involves the adaptation of plants as the concentration of salt in the soil increases or the availability of water in the soil decreases (Hasegawa et al., 2000). Molecular mechanisms vary among halophyte species and involve a number of metabolites, genes, and pathways. In this chapter we discuss halophytes that have been characterized by NGS and implications for salt tolerance.

Transcriptome Sequencing Overview

Initial transcriptome studies relied on microarray analysis, qPCR, or real-time PCR techniques to measure gene expression. The development of NGS techniques provides high speed and throughput and projects can now be completed in weeks or days at lower costs. NGS technologies allow gene expression profiling, genome annotation, and discovery of non-coding RNA (Mutz et al., 2013). NGS technology obtains short sequence tags, 20–35 bases long, from each transcript in the sample. This allows detection of low-abundance RNAs, small RNAs, or other elements (Ansorge, 2009). The transcriptomics variant based on sequencing by synthesis is called short-read massively parallel sequencing or RNA-seq.

Applications of RNA Studies

Applications using RNA-seq data include mapping of short reads, detection of intron splicing junctions, isoform expression quantification, and differential expression analysis (Chen et al., 2011). For mapping-first methods, sequenced reads are mapped to the genome or transcriptome sequences for guided assembly. Low-quality reads are removed to prevent incorrect mapping. The accuracy is determined by the mapping, therefore the best way to quantify genes or isoforms is to directly map the RNA-seq reads to the transcriptome sequences. The bioinformatics community is continually developing software to more effectively analyze RNA-seq (Trapnell et al., 2012).

Another application of RNA-seq is detection of differentially expressed genes and isoforms to compare conditions or samples at given time points. The expression level of transcripts is related to the number of reads mapped on them. Differences in read counts between two different experimental conditions at a statistically significant value can be regarded as differentially expressed. Several biases must be considered including sequencing depth, count distribution, library size, and length of transcripts. Approaches include probability distributions used by different pipelines and software packages for detecting differential expression between samples (Seyednasrollah et al., 2013).

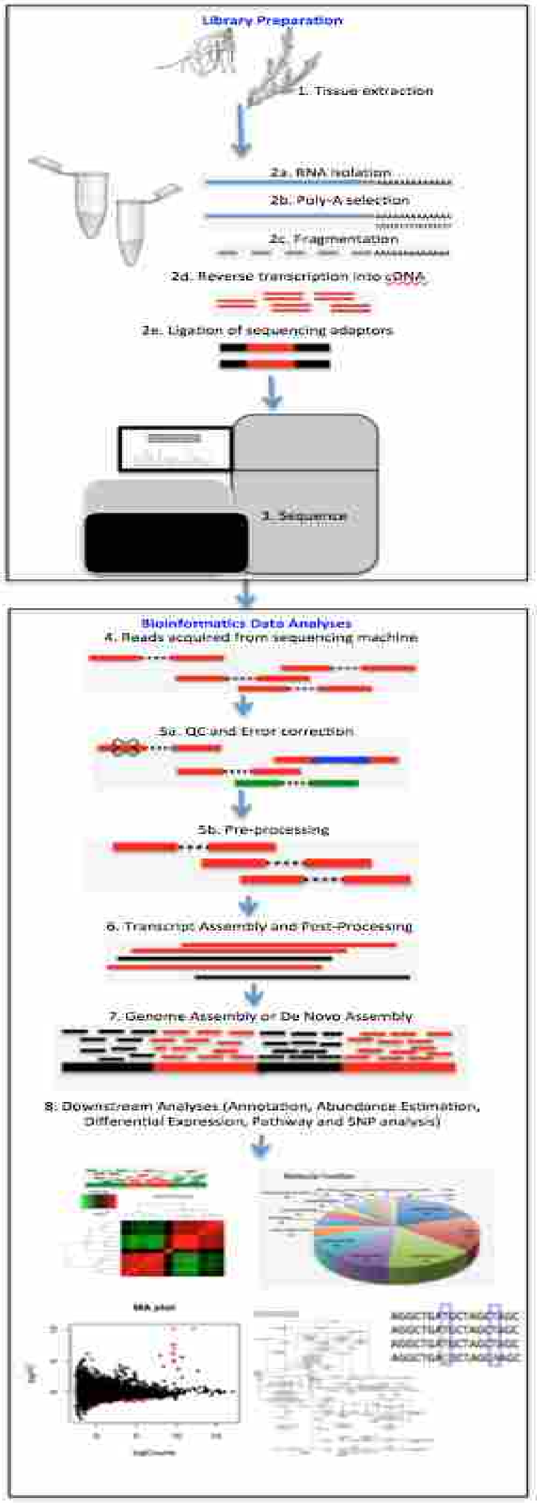

Library Preparation
1. Tissue extraction
2a. RNA isolation
2b. Poly-A selection
2c. Fragmentation
2d. Reverse transcription into cDNA
2e. Ligation of sequencing adaptors
3. Sequence
Bioinformatics Data Analyses
4. Reads acquired from sequencing machine
5a. QC and Error correction
5b. Pre-processing
6. Transcript Assembly and Post-Processing
7. Genome Assembly or De Novo Assembly
8: Downstream Analyses (Annotation, Abundance Estimation, Differential Expression, Pathway and SNP analysis)
MA plot

Figure 1.1 Methods for RNA Sequencing
Initial library preparation involves the isolation of RNA, which is converted to cDNA fragments with adaptors attached to one or both ends. The molecules are amplified, the libraries are quantified, analyzed for quality control and sequenced by high-throughput sequencers (Roche 454, Illumina, ABI SOLiD sequencing, PacBio, Ion Torrentor Helicos BioSciences) (Morozova and Marra, 2008). Bioinformatics is applied to the sequences generated. Pre-processing of data includes trimming of the sequencing adapters, error corrections, and elimination of poor-quality reads.

Box 1.1 Transcriptome Assembly Features

Transcriptome (RNA-seq) sequencing: This technology analyzes RNA presence and measures the levels of transcripts and their isoforms using NGS technologies (Clarke et al., 2013).
De novo versus reference-based assembly: For species that do not have a reference genome, de novo reconstruction of transcriptomes using RNA-seq data is performed. Reference-based assembly uses the genome sequence to serve as a guide for transcriptome reconstruction (Clarke et al., 2013).

Assembly, Alignment, and Visualization
Overlap layout consensus: Assemblers developed for Sanger reads use an overlap layout consensus method which computes pairwise overlaps and captures the information in a graph. This method constructs a read graph and assigns reads as nodes and then creates a link between two nodes when the reads overlap is larger than a cutoff length. The computation of reads and consensus sequence of contigs is determined by the overlap graph (Kumar and Blaxter, 2010, Li et al., 2011 and Miller et al., 2010).
De Bruijn graph approach: Reads are broken into smaller sequences or *k*-mers where *k* is the length in bases of the sequences. The *k* value is defined over a finite alphabet span, where *k* is a cyclic string where all words of length *k* appear exactly once in the sequence (Clarke et al., 2013 and Compeau et al., 2011).
Sequence aligners: Alignments of transcriptome sequences reveal novel splice forms and sequence polymorphisms. Choosing an aligner is necessary to accurately detect transcripts expressed in a given cell or tissue type. Most aligners can increase accuracy by prioritizing alignments in which read pairs map consistently (Engstrom et al., 2013).
Gene annotation: Different approaches are used to predict biological information: structural annotation by identification of genomic elements (gene structure, coding regions, motifs, ORFs) and functional annotation (molecular function, biological processes, cellular component, regulations and interactions, and expression) (Garber et al., 2011 and Stein, 2001).
Differential expression: RNA-seq measures the expression of specific gene products. Poorly replicated conditions, insufficient depths, or sequencing quality errors can lead to artifacts during differential analysis of the number of genes and transcripts showing significant fold changes in overall gene expression. A comparison between true replicates can reveal differences in gene transcripts from each condition while different tissues can show thousands of differentially expressed genes (Anders and Huber, 2010 and Tarazona et al., 2011).

Transcriptome reconstruction is another application of RNA-sequencing reads, and includes the genome-guided approach, which maps all the sequencing reads back to the reference genome, and genome-independent approach, which does not need a reference genome and directly assembles the reads into transcripts (Miller et al., 2010). Assembly using de novo techniques often uses de Bruijn graphs or the use of *k*-mers to assemble the reads into contigs. If a species already has a high-quality, complete reference genome, the genome-dependent approach is appropriate. The genome-independent approach is used for species that have no available reference genome (Miller et al., 2010). It is best to construct the transcriptome using de novo assembly to capture reads that cannot be obtained by genome-guided methods and then combine the results to produce a more comprehensive transcriptome (Box 1).

NGS Approaches for Salt-Tolerance Studies

Genomic technologies have been applied to study plant stress tolerance in some halophytes (Table 1), which have been compared with the *Arabidopsis* genome (Kant et al., 2008). *Thellungiella* spp. share many characteristics with *Arabidopsis* and are tolerant to salt and drought stresses (Griffith et al., 2007 and Wong et al., 2006). Draft genomes for *Thellungiella parvula* and *Thellungiella salsuginea* were constructed with NGS to understand adaptation to abiotic stresses (Dassanayake et al., 2011 and Wu et al., 2012).

Table 1.1 Known Halophytes with Analyzed Transcriptomes or Genomes[1]

Species	Genome/Transcriptome Information	Technology	Software Used	Purpose
Ceriops tagal (Liang et al., 2012)	432 DE transcripts 59 unigenes assembled	Microarray	LOWESS, SAM, BLASTX	Gene identification, differential expression, functional annotation
Eutrema salsugineum (Brassicaceae) (Yang et al., 2013)	241 Mb-genome 26,531 genes 137,652 bp exons	Sanger	Arachne, FGENESH, Genome Scan, BLAST	Phylogenetic analysis, genome assembly, synteny analysis, orthologue identification
Leymus chinensis (Gramineae) (Sun et al., 2013)	104,105 unigenes	454 FLX	LuCY, TagDust MIRA	Differential expression, annotation
Mesembryanthemum crystallinum (Kore-eda et al., 2004)	9733 expressed sequenced tags	cDNA library-dideoxy chain termination method	PHRED, CROSS-MATCH, PHRAP, BLASTX	EST assembly, functional categorization
Millettia pinnata (Huang et al., 2012)	54,596 unisequences, 65.8 Mb transcriptome	Illumina GA	SOAPdenovo	Gene annotation, differential expression
Populus euphratica (Zhang et al., 2013 and Qiu et al., 2011)	86,777 unigenes	Illumina GA	SOAPdenovo, TGICL	De novo assembly, annotation, differential expression
Populus pruinosa (Zhang et al., 2013)	114,866 unique sequences	Illumina GA	SOAPdenovo, TGICL	De novo assembly, ortholog identification, annotation
Porteresia coarctata (Garg et al., 2014)	152,367 unique transcripts	Illumina GA II	Velvet, Oases, ABySS, Trinity, CLC Genomics, CDHIT	De novo assembly, gene ontology, pathway analysis

Reaumuria trigyna (Tamaricaceae) (Dang et al., 2013)	65,340 unigenes	Illumina Hi Seq 2000	SOAPdenovo, Blast2GO	De novo assembly, gene ontology, expression pattern analysis
Salicornia europaea (Fan et al., 2013)	57,151 unigenes	Illumina Hi Seq 2000	SOAPdenovo, ESTScan	De novo assembly, gene ontology, digital gene expression tag sequencing, differential expression
Salicornia europaea (Ma et al., 2013)	109,712 unigenes	Illumina Hi Seq 2000	Trinity, Blast2GO	De novo assembly, GO annotation, differential expression
Spartina maritima (Ferreira de Carvalho et al., 2013)	114,857 singletons	454 GS XLR70	GS Assembler v 2.3	De novo assemblies, GO annotation, polymorphism analysis
Spartina alterniflora (Ferreira de Carvalho et al., 2013)	58,298 singletons	454 GS XLR70	GS Assembler v 2.3	De novo assemblies, GO annotation, polymorphism analysis
*Suaeda fruticosa (*Diray-Arce et al., 2015	54,526 unigenes	Illumina Hi seq 2000	Trinity, Oases, Velvet, CDHIT-EST, Blast2Go, Transdecoder	De novo assembly, GO annotation, differential expression
Suaeda maritima (Sahu and Shaw, 2009)	429 ESTs	PCR-based suppression subtractive hybridization	BLASTX, TIGR	SSH library construction, functional categorization
Schrenkiella parvula (*Thellungiela parvula*) (Oh et al., 2010)	21,619 contigs	454 GS FLX Titanium	Newbler, FGENESG, Repeat Masker, Pip maker	De novo assembly, annotation, synteny, comparative analyses of transcription, repeat identification
Thellungiella parvula (Dassanayake et al., 2011)	140 Mb genome	454 GS FLX Titanium, Illumina GA II	Newbler, ABySS, FGENESH, GENSCAN, BLAST, Blast2GO	Genome assembly, macrosynteny, ORF prediction and annotation
Thellungiella salsuginea (Lee et al., 2013)	42,810 unigenes	454 GS FLX Titanium	SFF Tools, MIRA, BLAST, MUSCLE, UGENE	De novo assembly, functional annotation, microRNA prediction, gene identification
Thellungiella salsuginea (Wu et al., 2012)	233.7 Mb genome	Illumina	ABySS, SOAPdenovo, Minimus2, MAUVE	Genome assembly, repetitive sequences identification, phylogenetic analyses, pathway analyses

[1] As of March 2015

A number of genes are involved in the response to salinity, and have been grouped in the following categories (Xiong and Zhu, 2002): (i) genes that encode enzymes, transcription factors, hormones, detoxifiers, osmolytes, and those responsible for general metabolism; (ii) genes that function in water and ion uptake such as ABC transporters, ion transporters, aquaporins, ATP binding cassette transporters, antiporters, and those involved in the SOS pathway; (iii) those that are involved in regulation, such as protein kinases and phosphatases; and (iv) genes that function to protect the cells against abiotic stress, such as late embryogenesis abundant (LEA) proteins, heat shock proteins, and osmoprotectants, such as dehydrin and osmotins.

Genes Involved in General Metabolism

This group includes genes that encode proteins for biosynthesis of osmolytes, hormones, and detoxification (Aslam et al., 2011). Some genes are responsible for abscisic acid signaling, which regulates plant germination, dormancy, and seed development. Others include antioxidants and enzymes that maintain the level of reactive oxygen species (ROS) to protect the cells from oxidative damage. Salt tolerance involves osmotic adjustment to maintain turgor, cell expansion, adjustments in photosynthesis and stomatal mechanisms, and plant growth. The sequestration of salt ions in the vacuole minimizes toxicity. Osmotic adjustment requires the accumulation of enzymes or osmolytes in the cytoplasm. The chemical nature of osmolytes varies from carbohydrates, polyols, and amino acids, and they are synthesized by halophytes and glycophytes in response to stress (Flowers and Colmer, 2008 and Grigore et al., 2011).

Genes for Cell Maintenance

Genes responsible for transcription, translation, and post-translational modifications play a role in salt tolerance. Transcription factors in *Suaeda maritima* include ethylene-responsive element-binding protein, ethylene-responsive element, jasmonate and ethylene response factor (Sahu and Shaw, 2009). HDZip genes are involved in abscisic acid-related responses, such as water deficiency in *Arabidopsis* (Ariel et al., 2007). HDZip genes ATH -7, -12, -6, -21, -40, and -53 are overexpressed upon salt treatment (Söderman et al., 1996). Genes encoding pectin methyl-esterase inhibitor protein, glutathione S-transferases, and RNA transcription factors are up-regulated after NaCl treatment in *Salicornia*. Enzymes for cell wall metabolism and peroxidase are decreased at early stages of the treatment. There is an up-regulation after salt treatment of pectin methyl-esterase inhibitor family protein, aminotransferase, and unspecific anion channel. Down-regulated genes in the roots are involved in cell wall precursor synthesis and cellulose synthesis reducing plant lignifications (Fan et al., 2013).

A comparison study of salt-tolerant species of *Festuca rubra* ssp. *litoralis* and rice identified a differentially regulated WRKY-type transcription factor and a SUI homologous translation initiation factor in response to salinity (Diédhiou et al., 2009). WRKY transcription factor was also differentially regulated in *Suaeda fruticosa* in response to abiotic stress (Diray-Arce et al., 2015). Phosphorylation and O-linked β-N-acetylglucosamine (O-GlcNAc) modification of proteins are found in *S. maritima* under salt stress (Sahu and Shaw, 2009).

Schrenkiella parvula expresses genes encoding tetratricopeptide repeat protein 1 involved in flowering, glycine-rich protein for cell wall structure, phosphoenolpyruvate carboxykinase, carbonic anhydrase for C_4 assimilation, acyl coA-binding protein for fatty acid metabolism, and other genes that are involved in cell organization and plant growth (Jarvis et al., 2014). In *S. fruticosa* we have found f-box kelch protein for actin filament interaction, ribosomal proteins for translation, DNA-binding protein escarola-like for late flowering and leaf development, catepsin b-like cysteine protease for disease resistance, and glutathione S-transferase tau for increased protection against toxins to be up-regulated (Diray-Arce et al., 2015). Xyloglucan endotrans glycosylase/hydrolase (XTH) and expansin-3 are overexpressed in *S. maritima* (Sahu and Shaw, 2009) and *Ceriops tagal* (Liang et al., 2012) upon salt treatment. XTH catalyzes molecular grafting to maintain cell wall thickness and promote cell wall formation and elongation (Jan et al., 2004).

Stress Genes

High concentrations of ions are toxic to plants because of their effect on cell homeostasis, cytosolic enzyme activities, and photosynthetic and cellular metabolism. Salt stress leads to the closure of stomata, reducing carbon fixation and photosynthesis, loss of cell turgor due to hyperosmotic shock, inhibition of cell division and expansion, toxicity, and plant yield reduction (Aslam et al., 2011).

Millettia pinnata, a halophytic mangrove, has 21.9% of its genes differentially expressed. In roots, most of these genes are involved in gene expression, sulfur metabolic processes, redox, and secondary metabolic processes. In leaves, induced genes are involved in redox, cellular amino acid derivative metabolism, and cellular aromatic compound metabolic processes. Stress response genes are also activated, which might serve as protection from salt-induced deleterious effects (Huang et al., 2012). In *S. parvula*, differentially expressed genes include ABA insensitive-5, D1-pyrroline-5-carboxylate synthase1, repressor of silencing 1, calcineurin B-like10, and are responsible for signaling under salt stress (Jarvis et al., 2014). In the root transcriptome of *S. maritima*, zeaxanthine epoxidase, a precursor of ABA, and chaperone protein DNA J genes are up-regulated (Sahu and Shaw, 2009).

Photosynthetic Genes

Photosystem II family protein-coding genes (protein Z, d2 protein, cp43 chlorophyll protein) are up-regulated in salt-treated *S. fruticosa* (Diray-Arce et al., 2015). In *Salicornia europaea* photosynthetic genes, PSI and PSII pigment-binding proteins, b6f complex, and ATPase synthase CF1 were significantly induced (Fan et al., 2013). *Populus euphratica* expression of psbA proteins, D2 protein, and Rubisco large unit were decreased after 12 h of salt shock. Genes for plastidic and nuclear protein synthesis, genes with undefined functions, genes pointing to glycolysis and stress (a putative glutathione S-transferase and COBRA protein precursor) suggest the relationship of salinity with decreased photosystem II activity. Restored water potential after salinity shock causes an increase in calcineurin-like protein CLB activity, 1 aminocyclopropane-1-carboxylic acid oxidase, root organelle-specific genes psbA, and mitochondrial ATPase (Brinker et al., 2010).

Mitochondrial and ROS Related Genes

Salinity stress increases ROS that cause oxidative damage to cellular components (Dang et al., 2013). Thioredoxin, glutaredoxin, glutathione S-transferase family genes were found in *Reaumuria trigyna* (Dang et al., 2013), *S. maritima* (Sahu and Shaw, 2009*)* and *S. fruticosa* (Diray-Arce et al., 2015). The thioredoxin gene is involved in redox regulation in the apoplast, which regulates cell division, cell differentiation, pollen germination, and stress responses (Zhang et al., 2011). Superoxide dismutase is highly induced in halophytes, which rapidly dismutates superoxide radicals into oxygen and hydrogen peroxide. In *R. trigyna*, there is increased transcription of glutathione disulfide-reductase and glutathione S-transferases, enzymes for resisting oxidative stress and maintaining the reducing environment of the cell (Dang et al., 2013).

Proline and Other Amino Acids

Proline is concentrated in the cytosol, chloroplast, and vacuoles for osmotic adjustment in many species and also contributes to detoxification of ROS (Ketchum et al., 1991, Khan et al., 2000 and Sucre and Suárez, 2011). Amino acid permease and proline transporter (ProT) were both up-regulated in the absence of salt and down-regulated at 10–500 mM salt concentration in *S. europaea* (Ma et al., 2013).

Glycinebetaine (GB) is up-regulated in plants exposed to dehydration (Lokhande and Suprasanna, 2012). The synthesis and accumulation of GB protects the cytoplasm from ion toxicity, dehydration, and temperature stress. It functions by stabilizing macromolecule structures and protecting photosystem II, and has been reported in many species (Khan et al., 2000 and Lokhande et al., 2010). In *Atriplex nummularia*, GB is accumulated under salt stress and the transcript levels of S-adenosyl-l-methionine co-regulate with that of phosphoethanolamine N-methyl transferase (PEAMT) in response to salinity (Nedjimi and Daoud, 2009). In *S. maritima* the most overexpressed gene encodes PEAMT that is responsible for synthesis and accumulation of GB (Sahu and Shaw, 2009).

Genes Encoding Plant Hormones

There is a significant increase in plant biomass in some halophytes while there is a decreasing biomass in others at different salt conditions. Gibberellic acid (GA) genes are involved in the synthesis of gibberellin hormone, which regulates many aspects of the growth and development of plants. In *S. europaea*, GA genes were regulated at 200 mM NaCl, similar to the homologues of gibberellin 3-oxidase and gibberellin 20-oxidase in *Populus trichocarpa*. Two DELLA domain GRAS family transcription factors, inhibitors of plant growth, were down-regulated in plants with 200 mM salt (Ma et al., 2013). In *Arabidopsis*, bioactive GA is reduced through an increase in gibberellin 2-oxidase 7 (GA2ox7) that accumulates DELLA, which inhibits plant growth (Magome et al., 2008). However, down-regulation of GA2ox at 300 mM salt treatment in *S. fruticosa* deactivates bioactive GA. A decrease in GA2ox and DELLA in *S. fruticosa* favors plant growth upon salt treatment (Diray-Arce et al., 2015).

Genes Encoding Ion Transporters

ABC Transporters

Ion homeostasis involves the transport of ions, cellular uptake, sequestration of salt, and ion export. Plant cells require high K^+ (100–200 mM) and lower Na^+ (1 mM) to maintain osmotic balance. A large influx of extracellular Na^+ occurs in halophytes (Lokhande and Suprasanna, 2012). Several ion transporters such as high-affinity potassium transporters (HKT), low-affinity cation transporters, nonselective cation channels, cyclic nucleotide-gated channels, and glutamate-activated channels have been identified in halophytes (Horie and Schroeder, 2004). In *R. trigyna*, five vacuolar H^+ pumping pyrophosphatases (PPases) were detected and may generate a proton electrochemical gradient to compartmentalize excess Na^+ ions. Genes associated with K^+ transport composed the largest proportion of genes suggesting their importance in Na^+/K^+ homeostasis. Seven HKT1 genes for Na^+ influx were also salt-responsive. Other genes encode plasma membrane H^+-ATPases, vacuolar H^+-ATPases, and H^+-pyrophosphatases (Dang et al., 2013 and Ahmed et al., 2013).

Most abundant transcripts in *S. parvula* under salt stress encode 17 transport-related proteins, including sodium and potassium ion transmembrane transporters, chloride channels, and ABC transporters. This halophyte and its relative *Eutrema salsugineum* highlighted the HKT1 Na^+/K^+ transporter (Wu et al., 2012). Highly enhanced expression of genes for cation-efflux transporters was observed in *Arabidopsis thaliana* (Jarvis et al., 2014). Studies in *Thellungiella* showed genes encoding transporters such as chloride channels and P-type H^+-ATPase. Chloride channels are groups of voltage-gated Cl^- channels that function in stabilizing cell membrane potential, regulating cell volume, and transcellular chloride transport (Hechenberger et al., 1996).

Antiporters

Ionic and osmotic equilibrium are necessary for plant salinity tolerance. Genes providing ionic stress protection are more abundant in *T. salsuginea* than in *Arabidopsis*. Studies have associated high Salt Overly Sensitive 1 (SOS1) expression levels with increased salt tolerance (Jarvis et al., 2014 and Maughan et al., 2009). SOS1 is required for salt tolerance in *Arabidopsis* and encodes a plasma membrane Na^+/H^+ antiporter (Shi et al., 2000).

Studies showed that a plasma membrane Na^+/H^+ antiporter (SOS1), vacuolar Na^+/H^+ antiporter (NHX1), and a plasma membrane Na^+ transporter (HKT1) are important for salt tolerance (Bassil et al., 2011 and Vera-Estrella et al., 2005). NHK1 is responsible for Na^+ sequestration and is up-regulated. Four genes that have strong homology to *A. thaliana* NHX2, *Mesembryanthemum crystallinum* and *Tetragoniate tragoniodes* NHX1 were slightly down-regulated and suggest that they play a role in mitigating the deleterious effects of high Na^+ levels in the cytosol and regulate intravacuolar K^+ and pH (Bassil et al., 2011). Halophytes have the ability to sequester large quantities of Na^+ into vacuoles. Cation/H^+ antiporters mediate these processes by vacuolar H^+-ATPase and H^+-PPase (Gaxiola et al., 2007).

Aquaporins

Aquaporin are intrinsic membrane proteins that serve as water-selective channels, and are involved in compartmentalization of water molecules. They likely play a role in maintaining osmosis and turgor of halophyte cells under salt stress (Dibas et al., 1998). *S. parvula* contains differentially expressed aquaporin genes, NOD26-like intrinsic protein (NIP) 5,1, and NIP 6,1 (Jarvis et al., 2014 and Martínez-Ballesta et al., 2008). In *Poplar*, suppression of these genes prevents water loss during salt stress (Brinker et al., 2010).

Regulatory Molecules

Osmotic stress induces transmembrane histidine protein kinases and stretch-activated channels. Mitogen-activated protein kinases and phosphatases transduce signals for compatible osmolyte synthesis and ROS detoxification by antioxidants and regulate stress response (Senadheera and Maathuis, 2009). Brassinosteroid insensitive-1-associated receptor kinase acts synergistically with auxins and gibberellins by promoting cell elongation while protein phosphatase 2C (PP2C) regulates signal transduction pathways (Senadheera and Maathuis, 2009). In *Thellungiella*, A-type PP2C phosphatases are generally up-regulated in response to abscisic acid (ABA). SOS2, a protein kinase that phosphorylates SOS1 in *Thellungiella*, interacts directly with V-ATPase as part of its salt-tolerance mechanism (Lee et al., 2013). Serine-threonine protein kinase HT1, responsible for a reduced response to ABA or light, is decreased in salt-treated *S. fruticosa* (Diray-Arce et al., 2015).

LEA Protein Coding Genes

Late embryogenesis abundant (LEA) protein coding genes have been found to have a protective effect against desiccation or osmotic stresses due to water loss. They may function as chaperones to prevent denaturation of important proteins (Vinocur and Altman, 2005). Most genes encoding LEA proteins have abscisic acid response and/or low-temperature response elements in their promoters (Aslam et al., 2011). In *T. salsuginea*, the RAV (Related to ABI3 and VP1) gene family responds to high-salt and cold stresses (Wu et al., 2012). Osmotins are required for homeostasis by maintaining cell functions at low osmotic potentials and high ionic stress. Genes encoding cold-circadian rhythm RNA binding like protein and two isoforms of carbonic anhydrase are overexpressed in *Suaeda maritima* after salt treatment (Sahu and Shaw, 2009).

Other Genomic Elements

Several stresses can activate transposable elements (TE). A dramatic expansion of pericentromeric heterochromatin in *E. salsugineum* is hypothesized to be a result of stress-induced activation of TEs (Hundertmark and Hincha, 2008). There is a prevalence of CT-rich regions and a pyrimidine-rich region close to ATG initiation codon in *Thellungiella* 5′ UTR sequences, cytosolic cyclophilin ROC3, and transcription factor B3. Cyclophilins are abundant proteins induced under abiotic stress and transcription factor B3 is induced in specific developmental stages (Yang et al., 2013).

Pathways

Gene targets in *E. salsugineum* include four copies reported to post-transcriptionally regulate transcription factor NAC required for an ABA-independent pathway (Oh et al., 2007). In *Thellungiella*, genes involved in hormone pathways which include ZEP, AAO, and CYP707A families are all involved in ABA biosynthesis pathway contributing to salt tolerance (Kim et al., 2009). Calcium serves as a messenger in developmental processes in plants and the main mechanism for Na^+ extrusion is through the plasma membrane H^+-ATPase and Ca^{2+}-ATPase, which pump H^+ and Na^+ into the cell. This action removes a single calcium ion in exchange for the import of three sodium ions (Wu, 2012).

In *P. euphratica*, 40 metabolic pathways were changed under salt stress including carbohydrate pathway, amino acid, energy, lipid, secondary metabolite, cofactor and vitamin, terpenoid, and polyketide metabolism. ABA signaling and synthesis pathways exhibited highly induced genes under salt stress. At ZEP homologue zeaxanthine epoxidase and 9-cis-epoxycarotenoid dioxygenase increases ABA to improve drought and salt tolerance (Sun et al., 2013). Sodium accumulation induced genes involved in stress and signal transduction pathway with the involvement of calcium, ethylene, ABA signaling regulation, and biosynthesis, which play a role in drought and salinity responses (Qiu et al., 2011).

The elevation of sodium content increases root osmotic potential due to dehydration (Brinker et al., 2010). Calcium-signaling pathways were triggered after salt treatment as calcium-binding and calmodulin-binding proteins were enriched. This indicates that salt promotes auxin-signaling pathways to facilitate growth of *S. europaea* (Fan et al., 2013). The auxin-signaling pathway was considered to be critical during salt treatment because most differentially expressed genes showed increased expression.

CONCLUSIONS AND FUTURE DIRECTIONS

The study of halophyte transcriptomes is still in its infancy. Many differentially expressed genes have been identified, and the results show that different species of halophytes utilize a variety of genes and pathways to establish salinity tolerance. Additional work in this area is warranted to increase our understanding of halophyte responses to salinity stress.

Acknowledgment

Support for halophyte research in the Nielsen, Khan, and Gul laboratories has been supported in part by a grant from the Pakistan—US Science & Technology Cooperation Program.

BIBLIOGRAPHY

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. New Biotechnology 25.

Ariel, F., Manavella, P., Dezar, C., Chan, R., 2007. The true story of HD-Zip family. Trends in Plant Science 12, 419-426.

Aslam, R., Bostan, N., Amen, N., Maria, M., Safdar, W., 2011. A critical review on halophytes: Salt tolerant plants. Journal of Medicinal Plants Research 5, 7108-7118.

Bassil, E., Tajima, H., Liang, Y., Ohto, M., Ushijima, K., Nakano, R., Esumi, T., Coku, A., Belmonte, M., Blumwald, E., 2011. The *Arabidopsis* Na+/H+ Antiporters NHX1 and NHX2 Control Vacuolar pH and K+ Homeostasis to Regulate Growth, Flower Development, and Reproduction. The Plant Cell Online 23, 3482-3497.

Brautigam, A., Gowik, U., 2010. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biology 12, 831 - 841.

Brinker, M., Brosche, M., Vinocur, B., Abo-Ogiala, A., Fayyaz, P., Janz, D., Ottow, E.A., Cullmann, A.D., Saborowski, J., Kangasjarvi, J., Altman, A., Polle, A., 2010. Linking the Salt Transcriptome with Physiological Responses of a Salt-Resistant *Populus* Species as a Strategy to Identify Genes Important for Stress Acclimation. Plant Physiology 154, 1697-1709.

Chen, G., Wang, C., Shi, T., 2011. Overview of available methods for diverse RNA-Seq data analyses. Science China Life Sciences 54, 1121-1128.

Dang, Z., Zheng, L., Wang, J., Gao, Z., Wu, S., Qi, Z., Wang, Y., 2013. Transcriptomic profiling of the salt-stress response in the wild recretohalophyte *Reaumuria trigyna*. BMC Genomics 14, 29.

Dassanayake, M., Oh, D., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D., Bressan, R.A., Zhu, J., Bohnert, H.J., Cheeseman, J.M., 2011. The genome of the extremophile crucifer *Thellungiella parvula*. Nature Genetics 43, 913-918.

Dibas, A.I., Mia, A.J., Yorio, T., 1998. Aquaporins (Water Channels): Role in Vasopressin-Activated Water Transport. Experimental Biology and Medicine 219, 183-199.

Diédhiou, C.J., Popova, O.V., Golldack, D., 2009. Comparison of salt-responsive gene regulation in rice and in the salt-tolerant *Festuca rubra ssp. litoralis*. Plant Signaling & Behavior 4, 533-535.

Diray-Arce, J., Clement, M., Gul, B., Khan, M.A., Nielsen, B.L., 2015. Transcriptome assembly, profiling and differential gene expression analysis of the halophyte Suaeda fruticosa provides insights into salt tolerance. BMC Genomics 16.

Engstrom, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., The, R.C., Ratsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigo, R., Bertone, P., 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10, 1185-1191.

Fan, P., Nie, L., Jiang, P., Feng, J., Lu, S., Chen, X., Bao, H., Guo, J., Tai, F., Wang, J., Jia, W., Li, Y., 2013. Transcriptome analysis of *Salicornia europaea* under saline conditions revealed the adaptive primary metabolic pathways as early events to facilitate salt adaptation. PLoS One 8.

Ferreira de Carvalho, J., Poulain, J., Da Silva, C., Wincker, P., Michon-Coudouel, S., Dheilly, A., Naquin, D., Boutte, J., Salmon, A., Ainouche, M., 2013. Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Heredity 110, 181-193.

Flowers, T.J., Colmer, T.D., 2008. Salinity tolerance in halophytes. New Phytologist 179, 945 - 963.

Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8, 469-477.

Garg, R., Verma, M., Agrawal, S., Shankar, R., Majee, M., Jain, M., 2014. Deep Transcriptome Sequencing of Wild Halophyte Rice, Porteresia coarctata, Provides Novel Insights into the Salinity and Submergence Tolerance Factors. DNA Research 21, 69-84.

Gaxiola, R.A., Palmgren, M.G., Schumachner, K., 2007. Plant proton pumps. FEBS letters 581, 2204-2214.

Griffith, M., Timonin, M., Wong, A.C.E., Gray, G.R., Akhter, S.R., Saldanha, M., Rogers, M.A., Weretilnyk, E.A., Moffatt, B.A., 2007. *Thellungiella*: an *Arabidopsis*-related model plant adapted to cold temperatures. Plant Cell Environment 30, 529 - 538.

Grigore, M.N., Boscaiu, M., Vicente, O., 2011. Assessment of the Relevance of Osmolyte Biosynthesis for Salt Tolerance of Halophytes under Natural Conditions. The European Jounal of Plant Science and Biotechnology 5, 12-19.

Hasegawa, P.M., Bressan, R.A., Zhu, J., Bohnert, H.J., 2000. Plant Cellular and Molecular Responses to High Salinity. Annual Review of Plant Physiology and Plant Molecular Biology 51, 463-499.

Hechenberger, M., Schwappach, B., Fischer, W.N., Frommer, W.B., Jentsch, T.J., Steinmeyer, K., 1996. A Family of Putative Chloride Channels from *Arabidopsis* and Functional Complementation of a Yeast Strain with a CLC Gene Disruption. Journal of Biological Chemistry 271, 33632-33638.

Horie, T., Schroeder, J.I., 2004. Sodium transporters in plants. Diverse genes and physiological functions. Plant Physiology 136, 2457-2462.

Huang, J., Lu, X., Yan, H., Chen, S., Zhang, W., Huang, R., Zheng, Y., 2012. Transcriptome characterization and sequencing-based identification of salt-responsive genes in *Millettia pinnata*, a semi-mangrove plant. DNA Research 19, 195-207.

Hundertmark, M., Hincha, D.K., 2008. LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. BMC Genomics 9, 118.

Jan, A., Yang, G., Nakamura, H., Ichikawa, H., Kitano, H., Matsuoka, M., Matsumoto, H., Komatsu, S., 2004. Characterization of a Xyloglucan Endotransglucosylase gene that is Up-regulated by Gibberelin in Rice. Plant Physiology 136, 3670-3681.

Jarvis, D.E., Ryu, C., Beilstein, M.A., Schumaker, K.S., 2014. Distinct Roles for SOS1 in the Convergent Evolution of Salt Tolerance in *Eutrema salsugineum* and *Schrenkiella parvula*. Molecular Biology and Evolution 31, 2094-2107.

Kant, S., Bi, Y.M., Weretilnyk, E., Barak, S., Rothstein, S.J., 2008. The *Arabidopsis* halophytic relative *Thellungiella halophila* tolerates nitrogen-limiting conditions by maintaining growth, nitrogen uptake, and assimilation. Plant Physiology 147, 1168 - 1180.

Ketchum, R.E.B., Warren, R.S., Klima, L.J., Lopez-Gutiérrez, F., Nabors, M.W., 1991. The mechanism and regulation of proline accumulation in suspension cell cultures of the halophytic grass *Distichlis spicata L*. Journal of Plant Physiology 137, 368-374.

Khan, M.A., Ungar, I.A., Showalter, A.M., 2000. The effect of salinity on the growth, water status, and ion content of a leaf succulent perennial halophyte, *Suaeda fruticosa (L.) Forssk*. Journal of Arid Environments 45, 73-84.

Kim, J.H., Woo, H.R., Kim, J., Lim, P.O., Lee, I.C., Choi, S.H., 2009. Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in *Arabidopsis*. Science 323, 1053-1057.

Kore-eda, S., Cushman, M.A., Akselrod, I., Bufford, D., Fredrickson, M., Clark, E., Cushman, J.C., 2004. Transcript profiling of salinity stress responses by large-scale expressed sequence tag analysis in *Mesembryanthemum crystallinum*. Gene 341, 83-92.

Lee, Y., Giorgi, F., Lohse, M., Kvederaviciute, K., Klages, S., Usadel, B., Meskiene, I., Reinhardt, R., Hincha, D., 2013. Transcriptome sequencing and microarray design for functional genomics in the extremophile *Arabidopsis* relative *Thellungiella salsuginea (Eutrema salsugineum)*. BMC Genomics 14, 793.

Liang, S., Fang, L., Zhou, R., Tang, T., Deng, S., Dong, S., Huang, T., Zhong, C., Shi, S., 2012. Transcriptional Homeostasis of a Mangrove Species, *Ceriops tagal*, in Saline Environments, as Revealed by Microarray Analysis. PLoS One 7.

Lokhande, V., Nikam, T., Penna, S., 2010. Differential osmotic adjustment to iso-osmotic NaCl and PEG stress in the in vitro cultures of *Sesuvium portulacastrum (L.) L*. Journal of Crop Science and Biotechnology 13, 251-256.

Lokhande, V., Suprasanna, P., 2012. Prospects of Halophytes in Understanding and Managing Abiotic Stress Tolerance, in: Ahmad, P., Prasad, M. (Eds.), Environmental Adaptations and Stress Tolerance of Plants in the Era of Climate Change. Springer, Maharashtra, India.

Ma, J., Zhang, M., Xiao, X., You, J., Wang, J., Wang, T., Yao, Y., Tian, C., 2013. Global Transcriptome Profiling of *Salicornia europaea* L. Shoots under NaCl Treatment. PLoS One 8.

Martínez-Ballesta, M.d.C., Bastías, E., Carvajal, M., 2008. Combined effect of boron and salinity on water transport: The role of aquaporins. Plant Signaling & Behavior 3, 844-845.

Maughan, P.J., Turner, T.B., Coleman, C.E., Elzinga, D.B., Jellen, E.N., Morales, J.A., Udall, J.A., Fairbanks, D.J., Bonifacio, A., 2009. Characterization of Salt Overly Sensitive 1 (SOS1) gene homoelogs in quinoa (*Chenopodium quinoa Willd*.). Genome 52, 647-657.

McGettigan, P.A., 2013. Transcriptomics in the RNA-seq era. Current Opinion in Chemical Biology 17, 4-11.

Morozova, O., Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255-264.

Mutz, K., Heilkenbrinker, A., Lönne, M., Walter, J.-G., Stahl, F., 2013. Transcriptome analysis using next-generation sequencing. Current Opinion in Biotechnology 24, 22-30.

Nedjimi, B., Daoud, Y., 2009. Cadmium accumulation in *Atriplex halimus subsp. schweinfurthii* and its influence on growth, proline, root hydraulic conductivity and nutrient uptake. Flora - Morphology, Distribution, Functional Ecology of Plants 204, 316-324.

Oh, D., Dassanayake, M., Haas, J.S., Kropornika, A., Wright, C., d'Urzo, M.P., Hong, H., Ali, S., Hernandez, A., Lambert, G.M., Inan, G., Galbraith, D.W., Bressan, R.A., Yun, D., Zhu, J., Cheeseman, J.M., Bohnert, H.J., 2010. Genome Structures and Halophyte-Specific Gene Expression of the Extremophile *Thellungiella parvula* in Comparison with *Thellungiella salsuginea (Thellungiella halophila)* and *Arabidopsis*. Plant Physiology 154, 1040-1052.

Oh, D., Gong, Q., Ulanov, A., Zhang, Q., Li, Y., Ma, W., Yun, D., Bressan, R., Bohnert, H., 2007. Sodium stress in the halophyte *Thellungiella halophila* and transcriptional changes in a thsos1-RNA interference line. Journal of Integrative Plant Biology 49, 1484 - 1496.

Qin, Q.P., Zhang, L.L., Li, N.Y., Cui, Y.Y., Xu, K., 2010. Optimizing of cDNA preparation for next generation sequencing. Yi Chuan 32, 974-977.

Qiu, Q., Ma, T., Hu, Q., Liu, B., Wu, Y., Zhou, H., Wang, Q., Wang, J., Liu, J., 2011. Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. Tree Physiology 31, 452-461.

Sahu, B.B., Shaw, B., 2009. Isolation, identification and expression analysis of salt-induced genes in *Suaeda maritima*, a natural halophyte, using PCR-based suppression subractive hybridization. BMC Plant Biology 9.

Senadheera, P., Maathuis, F.J.M., 2009. Differentially regulated kinases and phosphatases in roots may contribute to inter-cultivar difference in rice salinity tolerance. Plant Signaling & Behavior 4, 1163-1165.

Seyednasrollah, F., Laiho, A., Elo, L.L., 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics.

Shi, H., Ishitani, M., Kim, C., Zhu, J.-K., 2000. The *Arabidopsis thaliana* salt tolerance gene SOS1 encodes a putative Na+/H+ antiporter. Proceedings of the National Academy of Sciences 97, 6896-6901.

Söderman, E., Mattsson, J., Engström, P., 1996. The *Arabidopsis* homeobox gene ATHB-7 is induced by water deficit and by abscisic acid. The Plant Journal 10, 375-381.

Stein, L., 2001. Genome annotation: from sequence to biology. Nature Reviews Genetics 2, 493-503.

Sucre, B., Suárez, N., 2011. Effect of salinity and PEG-induced water stress on water status, gas exchange, solute accumulation, and leaf growth in *Ipomoea pescaprae*. Environmental and Experimental Botany 70, 192-203.

Sun, Y., Wang, F., Wang, N., Dong, Y., Liu, Q., Zhao, L., Chen, H., Liu, W., Yin, H., Zhang, X., Yuan, Y., Li, H., 2013. Transcriptome Exploration in *Leymus chinensis* under Saline-Alkaline Treatment Using 454 Pyrosequencing. PLoS One 8.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7, 562-578.

Vera-Estrella, R., Barkla, B.J., Garcia-Ramirez, L., Pantoja, O., 2005. Salt stress in *Thellungiella halophila* activates Na+ transport mechanisms required for salinity tolerance. Plant Physiology 139, 1507-1517.

Vinocur, B., Altman, A., 2005. Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Current Opinion in Biotechnology 16, 123-132.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T., Keller, B., Stein, N., 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7, 275.

Wong, C., Li, Y., Labbe, A., Guevara, D., Nuin, P., Whitty, B., Diaz, C., Golding, G., Gray, G., Weretilnyk, E., 2006. Transcriptional profiling implicates novel interactions between abiotic stress and hormonal responses in *Thellungiella*, a close relative of *Arabidopsis*. Plant Physiology 140, 1437 - 1450.

Wu, H., Zhang, Z., Wang, J., Oh, D., Dassanayake, M., Liu, B., Huang, Q., Sun, H., Xia, R., Wu, Y., 2012. Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proceedings of the National Academy of Sciences 109, 12219 - 12224.

Wu, Y., 2012. Unwinding and rewinding: double faces of helicase? Journal of nucleic acids 2012, 140601.

Yang, R., Jarvis, D.J., Chen, H., Beilstein, M., Grimwood, J., Jenkins, J., Shu, S., Prochnik, S., Xin, M., Ma, C., Schmutz, J., Wing, R.A., Mitchell-Olds, T., Schumaker, K., Wang, X., 2013. The reference genome of the halophytic plant *Eutrema salsugineum*. Front Plant Sci 4.

Zhang, C., Zhao, B., Ge, W., Zhang, Y., Song, Y., Sun, D., Guo, Y., 2011. An Apoplastic H-Type Thioredoxin Is Involved in the Stress Response through Regulation of the Apoplastic Reactive Oxygen Species in Rice. Plant Physiology 157, 1884-1899.

Zhang, J., Xie, P., Lascoux, M., Meagher, T.R., Liu, J., 2013. Rapidly Evolving Genes and Stress Adaptation of Two Desert Poplars, *Populus euphratica* and *P. pruinosa*. PLoS One 8, e66370.

CHAPTER 2: *Suaeda fruticosa*, a Potential Model Halophyte for Salt Tolerance Research

Halophytes have different strategies to tolerate and maintain productivity while growing with saline water. Salt tolerance of some halophytes involves biochemical adaptation in electrolyte accumulation to maintain protoplasm viability. Others have specialized salt tolerance mechanisms through ion exclusion [1] in root membranes [2-5]. In some species, roots have a thick epidermis that is impervious to salt and an endodermis with a waxy layer allowing water to pass through cells to filter the salts (Fig 2.1, left) [6, 7]. Ion exclusion in leaves may be achieved through cuticle diffusion at leaf surfaces [8, 9], through secretion in structures such as glands or trichomes [10, 11], or ejection through stomatal guttation [12]. Some succulent halophytes have evolved salt bladders on the leaf surface, which eliminates excess salt from active tissues [2]. Other halophytes have stress mechanisms that are able to handle salts through turgor pressure and control accumulation and sequestration of ions to adjust osmosis to salinity [13].

The focus of this study, *Suaeda fruticosa*, a succulent shrub in the family Chenopodiaceae, can grow optimally at 300 mM NaCl and has the adaptation to reduce sodium build up for long term survival [14]. This obligate halophyte sequesters sodium and chloride in shoot vacuoles and synthesizes osmoprotectants such as glycinebetaine, which maintains a water potential gradient and protects cellular structures [15]. Glycinebetaine, found in halophytic members of Poaceae and Chenopodiaceae, is a stabilizing osmolyte that can offset the high salinity concentration in the vacuole [16]. It protects cells from environmental stresses indirectly via its role in signal transduction and it has been shown to play a role in Na^+/K^+ discrimination, an important factor that contributes to plant salt tolerance [17]. However, the pathway of how glycinebetaine affects expression of genes responsible for and its relation to plant salinity stress is still scarce. More information about the pathway and regulation of genes by glycinebetaine may lead to new approaches for the improvement of plant stress tolerance.

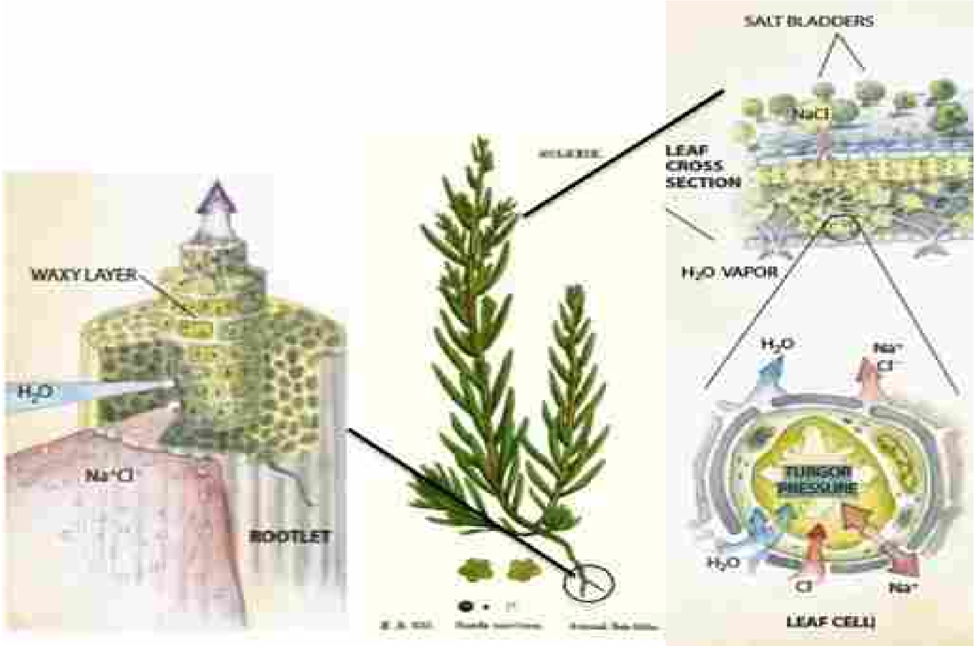


Figure 2.1 Physiological Adaptation of Halophytes During Salt Treatments
Halophytes have specialized organs that can sequester salt to keep homeostasis inside the cell. Illustration by Scientific American, 1998.

When soils in arid regions are irrigated, solutes in irrigation water are accumulated, increasing salinity levels to a point that have an adverse effect on plant growth [1]. An introduction of alternate crops such as halophytes is advisable in areas with reduced water availability and increased soil salinity. Halophytes can then be used to evaluate the overall feasibility of high saline agriculture. Researchers have started working on the development of salt-tolerant crops through breeding and domestication of wild halophytes [18, 19]. Field trials confirmed high-yield potential of halophytes from the Chenopodiaceae family, which produced a biomass mean of 18 tons per hectare using 40 g/L NaCl as irrigation source (Puerto Penasco, Sonora, Mexico), comparable to yields from forage crops [20]. Some *Atriplex* species produced 12.6 to 20.9 tons per hectare of biomass on full-strength seawater. Halophytes can be grown similar to traditional crops with high productivity and good quality biomass under full seawater irrigation. However, most productive halophytes have optimum growth in the salinity range of 200 to 340 mM NaCl [13, 20].

Halophytes as non-traditional crops have great potential to be utilized for oil, food, fodder or other purposes. *Suaeda fruticosa* is a good source of high quality edible oil [21], has potential for antiophthalmic, hypolipidaemic and hypoglycemic medicinal purposes [22], and has economic usage as forage for animals [23]. *S. fruticosa* also could help in bioremediation and reclamation of soils contaminated with toxic metals [24] and salinity [25]. Cattle raised on a diet supplemented with salt-tolerant plants such as *Suaeda* species have gained at least as much weight and yield meat of the same quality as control cattle that are fed with conventional grass hay, although they convert less of the feed to meat and drink almost twice as much water [6]. The potential of halophytes has been under limited examination until recently and their utilization may allow production and economic value to farmers in traditionally poor regions of the world.

While a number of halophytes have been studied to characterize their basic properties, and some plant genes that contribute to salt tolerance have been identified, there is still much to learn about halophytes. With additional halophytes examined recently, it has been found that different species utilize many of the same genes, but some also express novel genes and their protein products to allow them to grow in salty soils. Some model halophyte species have been suggested to enhance future research. *Thellungiella halophila*, a close relative species of *Arabidopsis*, has been regarded as a valuable model halophyte because of the copious information available for *Arabidopsis*. *T. halophila* survives at 500 mM NaCl but some research shows that growth is inhibited at 150 mM, which is lower than many other halophytes [26, 27].

This plant also belongs to the family Brassicaceae, which has only has few halophytic species (ca. 19 in five genera), compared to halophytes belonging to Chenopodiaceae (380 halophytes) and Poaceae (140 halophytes)[28]. Another species well-studied in halophyte research is *Mesembryanthemum crystallinum*, which can tolerate high concentrations of salt and other metabolites and is also able to switch from C3 to CAM metabolism. This plant possesses bladder cells that can store ions. However, the photosynthetic mechanisms are not similar to most halophytes and also because the majority of halophytes do not have glands or external bladders. Hence, the genus *Suaeda* has been suggested as a potential model halophyte for understanding salt tolerance mechanism of halophytes [28].

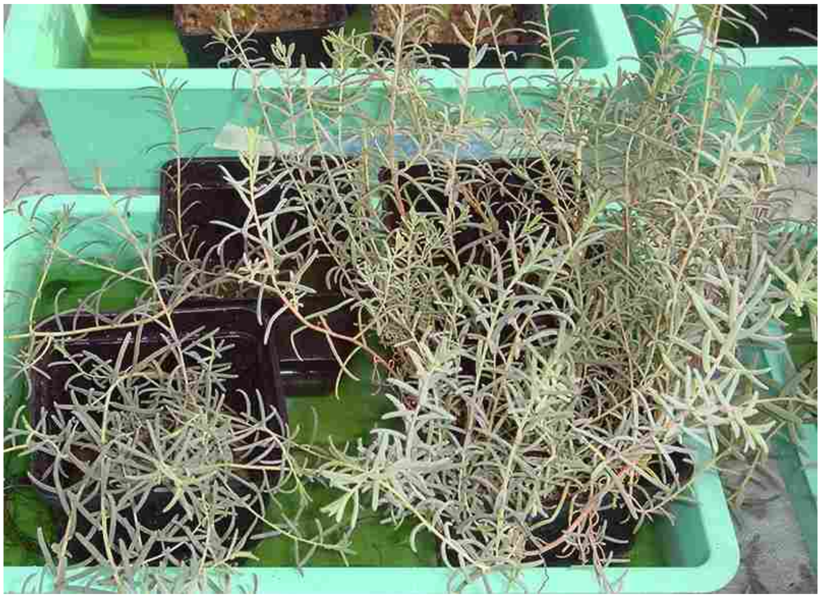


Figure 2.2 Halophyte Species *Suaeda fruticosa*

Physiological properties of *Suaeda fruticosa* have been studied with its unique characteristics in accommodating ions without the need for secretion via salt glands [29]. Optimal growth conditions, such as salt concentration in the watering solution, temperature and light conditions have been characterized. The biochemical basis of salt tolerance has been studied in this plant using different exogenous treatment under different levels of salinity [29].

Earlier reports proposed to study regulation of plant genes under salt stress using RNA [30] or protein analysis [31] but was considered implausible at the time. Because of the emergence of new technologies such as next generation sequencing and proteomic analysis, the characterization of the aforementioned species, *Suaeda fruticosa,* will provide an understanding of salt tolerance and information for the improvement of halophyte species into cash crops. As part of this, differential expression analysis is necessary to identify genes involved in salt stress tolerance in *S. fruticosa* grown under optimal salt conditions (300 mM NaCl) or compared to growth in the absence of salt (0 mM NaCl) conditions.

REFERENCES

1. Koryo HW, Khan M, Lieth H: Halophytic crops: A resource for the future to reduce the water crisis? *Emir J Food Agric* 2011, 23(1):001-016.

2. Marschner H: Mineral nutrition of higher plants. London: Academic Press; 1995.

3. Mengel K, Kirkby EA: Principles of Plant Nutrition. London: Kluwer Academic Published; 2001.

4. Munns R: Physiological processes limiting plant growth in saline soils: some dogmas and hypotheses. *Plant, Cell, and Environment* 1993, 16:15-24.

5. Koryo HW, Huchzermeyer B: Ecophysiological needs of potential biomass crop *Spartina townsendii* GROV. *Tropical Ecology* 2004, 45:123-139.

6. Glenn E, Brown JJ, O'Leary JW: Irrigating Crops with Seawater. In: *Scientific American.* 1998.

7. Tomlinson PB: The Botany of Mangroves. Cambridge, London: Cambridge University Press; 1986.

8. Schonherr J: Characterization of aqueous pores in plant cuticles and permeation of ionic solutes. *J Exp Bot* 2006, 57(11):2471-2491.

9. Schönherr J: A mechanistic analysis of penetration of glyphosate salts across astomatous cuticular membranes. *Pest Management Science* 2002, 58(4):343-351.

10. Koyro H-W, Geissler N, Hussin S, Huchzermeyer B: Survival at extreme locations: Life strategies of halophytes - The long way from system ecology, whole plant physiology, cell biochemistry and molecular aspects back to sustainable utilization at field sites. In: *Bios Agric High Sal Tol.* Edited by Abdelly C, Ashraf M, Grignon C; 2008: 1-20.

11. Lefevre I, Marchal G, Meerts P, Correal E, Lutts S: Chloride salinity reduces cadmium accumulation by the Mediterranean halophyte species Atriplex halimus L. *Env Exp Bot* 2009, 65(1):142-152.

12. Shi-qing LI, Chun-rong JI, Ya-ning F, Xiao-li C, Sheng-xiu LI: Advances in nitrogen loss leached by precipitation from plant canopy. *Agric Sci China* 2008, 7:480-486.

13. Flowers TJ, Yeo AR: Ion relations of plants under drought and salinity. *Aus J of Plant Physiol* 1986, 13:75-91.

14. Hameed A, Hussain T, Gulzar S, Aziz I, Gul B, Khan M: Salt tolerance of a cash crop halophyte *Suaeda fruticosa*: biochemical responses to salt and exogenous chemical treatments. *Acta Physiol Plant* 2012.

15. Ajmal Khan M, Ungar IA, Showalter AM: The effect of salinity on the growth, water status, and ion content of a leaf succulent perennial halophyte, Suaeda fruticosa (L.) Forssk. *Journal of Arid Environments*, 45(1):73-84.

16. Aslam R, Bostan N, Amen N, Maria M, Safdar W: A critical review on halophytes: Salt tolerant plants. *Journal of Medicinal Plants Research* 2011, 5(33):7108-7118.

17. Ashraf M, Foolad MR: Roles of glycine betaine and proline in improving plant abiotic stress resistance. *Environmental and Experimental Botany* 2007, 59(2):206-216.

18. Somers G: Seed-Bearing Halophytes as Food Plants. In. University of Delaware Sea Grant Program: College of Marine Studies, Newark, Delaware; 1975.

19. Felger RS: Ancient crops for the twenty-first century. Boulder, Colorado: Westview Press; 1979.

20. Glenn EP, O'Leary JW: Productivity and irrigation requirements of halophytes grown with seawater in the Sonoran Desert. *J Arid Environment* 1985, 9:81-91.

21. Weber DJ: Adaptive mechanisms of halophytes in deser regions. *Salinity and water stress* 2008, 44:179-185.

22. Bennani-Kabachi N, El-Bouayadi F, Kehel L, Fdhil H, Marquie G: Effect of Suaeda fruticosa aqueous extract in the hypercholesterolaemic and insulin-resistant sand rat. *Therapie* 1999, 54:725-730.

23. Towhidi A, Saberifar T, Dirandeh E: Nutritive value of some herbage for dromedary camels in the central arid zone of Iran. *Tropical Animal Health Pro* 2011, 43:617-622.

24. Bareen F, Tahira SA: Metal accumulation potential of wild plants in tannery effluent contaminated soil of Kasur, Pakistan: field trials for toxic metal cleanup using *Suaeda fruticosa. J Hazard Mater* 2011, 186:443-450.

25. Khan MA, Ansari R, Ali H, Gul B, Nielsen BL: Panicum turgidum: a sustainable feed alternative for cattle in saline areas. *Agric Eco Env* 2009, 129:542-546.

26. Inan G, Zhang Q, Li P: Salt cress. A halophyte and cryophyte *Arabidopsis* relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant Physiol* 2004, 135:1718-1737.

27. Gong Q, Li P, Ma S, Indu Rupassara S, Bohnert HJ: Salinity stress adaptation competence in the extremophile *Thellungiella halophila* in comparison with its relative Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* 2005, 44:826-839.

28. Flowers T, Colmer T: Salinity tolerance in halophytes. *New Phytol* 2008, 179:945 - 963.

29. Hameed A, Hussain T, Gulzar S, Aziz I, Gul B, Khan M: Salt tolerance of a cash crop halophyte *Suaeda fruticosa*: biochemical responses to salt and exogenous chemical treatments. *Acta Physiol Plant* 2012, 34(6):2231-2340.

30. Kawasaki S, Borchert C, Deyholos M, Wang H, Brazille S, Kawai K, Galbraith D, Bohnert HJ: Gene expression profiles during the initial phase of salt stress in rice. *The Plant Cell* 2001, 13:889-905.

31. Salekdeh GH, Siopongco LJ, Wade NG, Bennett J: A proteomic approach to analyzing drought- and salt responsiveness in rice. *Field Crops Res* 2002, 76:199-219.

CHAPTER 3: Optimization of *de novo* Transcriptome Assembly of the Halophyte *Suaeda fruticosa* Using Clustering Methods

Joann Diray-Arce[1], Mark Clement[2], Bilquees Gul[3], M. Ajmal Khan[3] and Brent L. Nielsen[1*]

[1]Department of Microbiology & Molecular Biology, Brigham Young University,
Provo, Utah 84602 U.S.A.
[2]Department of Computer Science, Brigham Young University, Provo, Utah 84602
[3]Institute of Sustainable Halophyte Utilization, University of Karachi, Pakistan

E-mail addresses:
Joann Diray-Arce, joann.diray@gmail.com
Mark Clement, clement@cs.byu.edu
Bilquees Gul, bilqueesgul@uok.edu.pk
Ajmal Khan, ajmal.khan@qu.edu.qa

Corresponding Author: Brent L. Nielsen, brentnielsen@byu.edu

ABSTRACT

Background

RNA-seq analysis has the potential to identify transcriptomes and elucidate the functions of genes and the interactions among them. Current tools required to filter, assemble and cluster next generation sequencing (NGS) reads are often difficult to use and it is not easy to determine which tools will work best for identification and analysis of genes. This study reports optimization of *de novo* transcriptome assemblies of *Suaeda fruticosa,* which is an obligate halophyte that sequesters salts in its shoots to reduce sodium buildup for long-term survival.

Results

This paper contains a thorough examination of the algorithms necessary to convert RNA-seq data into a usable de novo transcriptome. We trimmed and normalized the reads for pre-assembly quality check. We used Trinity and Velvet-Oases to generate multiple assemblies and compared their mapping efficiencies using GSNAP. Optimization of the assemblies was performed and compared using clustering methods (CAP3, CDHIT-EST and Isofuse) that reduced the number of transcripts while retaining the mapping coverage. A new algorithm (Isofuse) was developed as part of this work, which provides superior clustering of splice variants and can be used to improve the usability of a transcriptome. This maximizes the coverage of reads while reducing the number of transcripts without losing important information needed for de novo transcriptome assembly.

Conclusions

The tools are applied to *Suaeda fruticosa*, a succulent halophyte species, which has recently been investigated as a potentially important crop in developing nations with salty soils. This work provides a reference genome for other succulent halophytes and an outline of tools to use for *de novo* analysis of transcriptomes.

Keywords

Halophytes, RNA-seq, de novo assembly, transcriptome, clustering methods, salt-tolerance

BACKGROUND

Next generation sequencing (NGS) is a powerful tool that is readily available and widely used for transcriptome studies. NGS technologies have reduced the time and costs associated with genotyping large eukaryotic genomes when compared with microarray platforms. The early generation of reference genomes provided genetic analysis for model organisms as well as a diverse variety of other species [1]. Since the completion of the human genome, NGS technologies have developed and emerged into a wide variety of applications. NGS is used for whole-genome re-sequencing for species for which the reference genome is available. The short reads are mapped back to a reference genome for identification of single-nucleotide polymorphisms (SNPs), insertions or deletions (indels), structural variants and copy number variation so that these differences can be associated with phenotypes [2]. NGS has also assisted in the discovery of single nucleotide polymorphism markers in plants that can be used to study genetic variations, population genetics, transcript profiling, and genetic associations necessary for plant breeding [3, 4]. NGS of plant genomes has led to mapping of mutations responsible for many phenotypes of interest [5]. As next generation technologies have matured, RNA sequencing has become a preferred method for gene expression profiling [6]. The development of plant genome sequencing has facilitated the identification and tracking of genetic populations that are expected to advance the understanding of crop genetics, leading to crop improvement [5, 7-9].

Halophytes (salt tolerant plants) have great potential to be utilized for oil, food, fodder, or other purposes. As a non-traditional crop, *Suaeda fruticosa,* a succulent shrub in the family Chenopodiaceae, is an obligate halophyte that can grow optimally at 300 mM NaCl and has the adaptation to reduce sodium build up for long term survival [10]. It is a good source of high quality edible oil [11], has potential for antiophthalmic, hypolipidaemic and hypoglycemic medicinal purposes [12], and has potential economic usage as forage for animals [13]. *S. fruticosa* also may help in bioremediation and reclamation of soils contaminated with toxic metals [14] and salinity [15]. Cattle raised on a diet supplemented with salt-tolerant plants such as *Suaeda* species have gained at least as much weight and yield meat of the same quality as control cattle that are fed with conventional grass hay, although they convert less of the feed to meat and drink almost twice as much water [16].

Studies of halophytes have been limited until recent investigations have shown that their utilization may allow production and economic value to farmers in traditionally poor regions of the world. While a number of halophytes have been studied to characterize their basic properties, and some plant genes that contribute to salt tolerance have been identified, there is still much to learn about halophytes. Initially, *Thellungiela halophila* was preferred as a model halophyte due to its genetic proximity to *Arabidopsis*, which is the traditional plant model organism and has a completely sequenced genome. However, the salt tolerance of *T. halophila*, although higher than

Arabidopsis, is still significantly lower than most halophytes. It may survive in 500 mM NaCl but is shown to exhibit chlorosis and growth inhibition at that concentration [17-19]. The transcriptome response of the halophyte *Poplar euphratica* under salt stress has been previously studied, and identified upregulated genes related to transport, transcription, cellular communication, and metabolism. These responses exhibit permanent activation of control mechanisms for osmotic adjustment, ion compartmentalization and detoxification of reactive oxygen species [20-22]. It is important to consider that a significant number of halophyte species do not have glands or external bladders to modulate their tissue ion concentration [23]. *Suaeda* is a good potential representative of this group of very tolerant halophytes with succulent leaves that are able to accommodate ions without the need for secretion via salt glands. Analysis of a limited number of expressed sequence tags from a cDNA library of a closely related species, *Suaeda asparagoides,* has previously identified genes whose expression is altered under stress conditions and may include genes responsible for signal transduction, transcription, metabolism, redox, transport and protein synthesis that could be involved in adaptations to salt stress [24]. Results of the EST sequencing may provide a good background of some salt responsive genes that might be functional in *Suaeda fruticosa*.

RNA sequencing generates an enormous amount of data that can be used to analyze the transcriptome; however, significant challenges must be overcome. One of the major problems is the development of expression metrics that will allow comparisons of different expression levels and provide identification of differentially expressed genes. Different approaches have been developed, but there is no set protocol for the most preferred method of RNA sequencing analysis. Another major problem is with organisms that lack a reference genome. Researchers are currently analyzing the best algorithms for de novo assembly of transcripts for organisms without a known reference. The reference transcriptome assembly is not available for *Suaeda fruticosa* and it does not have any close relative plants that can serve as a complete reference for the expression analysis. This paper provides a comparison of different assembly algorithms for transcriptome data and analyzes the most preferred assembly to be used for annotation and other downstream analyses for *Suaeda fruticosa.*

RESULTS AND DISCUSSIONS

Sequencing Method and Quality Assessment of the Reads

To prepare for the transcriptome assembly and analysis, total RNA was extracted from shoots and roots of *Suaeda fruticosa* for generation of cDNA libraries. These include triplicates of cDNA libraries for *S. fruticosa* roots from plants grown without salt (R000), roots with 300 mM optimal salt (R300), shoots with no salt (S000) and shoots with 300mM optimal salt (S300). Purification of mRNA was achieved with oligodT and transcribed into cDNA libraries using the TruSeq RNA Sample Prep Kit for Illumina paired-end sequencing. A total of 335.3 million reads of 100 bp were generated by the Illumina Hi-seq platform. The reads were filtered using the FASTX toolkit to remove low quality reads, Trimmomatic to remove adapters and the Sickle program to trim low quality ends of reads so that only high quality sequences were used in the assembly. A total of 84.58% of the reads were trimmed and filtered, resulting in 283,587,292 high quality reads.

Normalizing Reads by *k*-mer Coverage

To normalize and assemble RNA-seq reads for de novo assembly, digital normalization was used for 283.6 million reads. A *K*-mer hash of 21 with coverage of 30X was built from a set of reads to correct redundancy issues, variations in sequences, and potential errors among the reads. Since some reads with sequencing errors may escape quality score-based filtering steps, the reads with potential errors are flagged for removal from the dataset to improve the de novo assembly. Sequencing errors can affect the assembly algorithms so it is best to eliminate the reads that have non-uniform *k*-mer coverage. Reads that have non-uniform *k*-mer coverage create a problem with the assembly, therefore it is necessary to normalize the reads to a certain threshold. This threshold represents the approximate minimum for de novo assembly to work optimally and efficiently. Digital normalization [25] was applied to the total of 283,587,292 paired end reads with *k*-mer size of 21 and *k*-mer coverage cutoff of 30X. The retained reads were normalized to 99,577,045 to remove overabundant reads, reduce the noise of the sequenced sample and decrease the overall percentage of errors. High-coverage reads from shotgun data sets are removed after sequencing data has been generated. The effect of digital normalization is to retain nearly all real *k*-mers while discarding the majority of erroneous *k*-mers. This step reduces the number of reads and makes transcriptome assembly much faster than and superior to the assembly based on the full data set without affecting the quality of the assembly. This error reduction results in decreasing computational requirements for de novo assembly. Because the genes in the transcriptome have different levels of expression, *k*-mer distribution will not show a peak at any *k* value.

De novo Transcriptome Assembly

To assemble the *Suaeda fruticosa* transcriptome, we utilized a genome-independent reconstruction approach. The strategy involved building a de Bruijn graph made of overlapping subsequences or *k*-mers. These overlapping bases are used to build a graph that is used to construct contiguous sequences (contigs) which can be combined using read and paired-end coverage [26]. The contigs resulting from this de novo assembly are reported as transcripts. We used two de Bruijn graph assemblers, Trinity [27] and Oases [28], using single and multiple *k*-mer methods of assembling the transcriptome. Oases is a software package designed to assemble RNA-seq reads in the absence of a reference genome and in the presence of alternative isoforms. It uses an array of hash lengths to filter the noise and recognize alternative splicing using multiple *k*-mer values [29]. This software is run after a preliminary assembly with the Velvet assembler to produce a preliminary fragmented assembly of reads mapped into a set of contigs [28, 30, 31].

The Trinity assembler is used to reconstruct transcripts from the sample. This assembly algorithm partitions the sequence data into de Bruijn graphs, which represent the complexity of a gene or locus. The graphs are used to process full-length splicing isoforms and to straighten out transcripts derived from paralogous genes [27]. Trinity includes three independent software modules: Inchworm, Chrysalis and Butterfly/Pasafly/Cufffly. Inchworm assembled the RNA-Seq data into unique transcripts generating full-length sequences for dominant isoforms while accounting for unique portions of alternatively spliced transcripts. Chrysalis combines and clusters the contigs from Inchworm and creates a de Bruijn graph for each cluster. This partitions the full read set to prepare for the final process, in which the partitions are used to create the final transcripts for spliced isoforms corresponding to paralogous genes. Trinity version 20131110 provides three selections (Butterfly, Pasafly, Cufffly) for the final process depending on the stringency of the path. Butterfly reconstructs the reads into graph node extensions. Pasafly undergoes a PASA-like algorithm for maximally supported isoforms that will report conservative reconstructions and fewer isoforms and Cufffly follows a Cufflinks-like algorithm that will report the minimum number of transcripts or fewest isoforms. Oases-Velvet and Trinity assembly methods were used for the de novo assembly and are compared in Figure 3.1.

Velvet-Oases

We used Velvet to assemble normalized reads into contigs followed by Oases to produce scaffolds. Different *k*-mer sizes from 35 to 99 were chosen to generate the assemblies. The quality of scaffolds assembled by Oases was assessed based on the total number of transcripts and open-reading frames (ORF) (Figure 3.1A) that were predicted using Transdecoder. Transdecoder identifies coding regions with the following criteria: the transcripts using the minimum length open reading frame found in a transcript sequence, a log-likelihood score that is similar to the score computed by the Gene ID software. The coding score is greatest when the ORF is scored in the 1st reading frame and compared to the other 5 reading frames and if the candidate ORF is found in the coordinates of another candidate ORF, the longer one is reported. A single transcript can contain multiple ORFs due to alternative splicing and start sites. The total base pair length and N50 (computed by sorting the contigs from largest to smallest and then determining the minimum set whose sizes total 50% of the assembly) for both transcripts and predicted open reading frames were determined (Figure 3.1B). The resulting sequences were defined as unigenes. Among the individual *k*-mer values, transcript numbers associated with different *k*-mer values vary from 1 to 450,588. *K*-mer length is related inversely to the number of transcripts generated. The highest N50 generated for the transcripts is 1,755 generated by a *k*-mer length of 59. Predicted open reading frames range from 1 to 108,112 bp with the highest number of ORFs being generated with a *k*-mer of 41. The highest N50 for open reading frames belongs to an assembly with a *k*-mer of 65 with 1,272 bp.

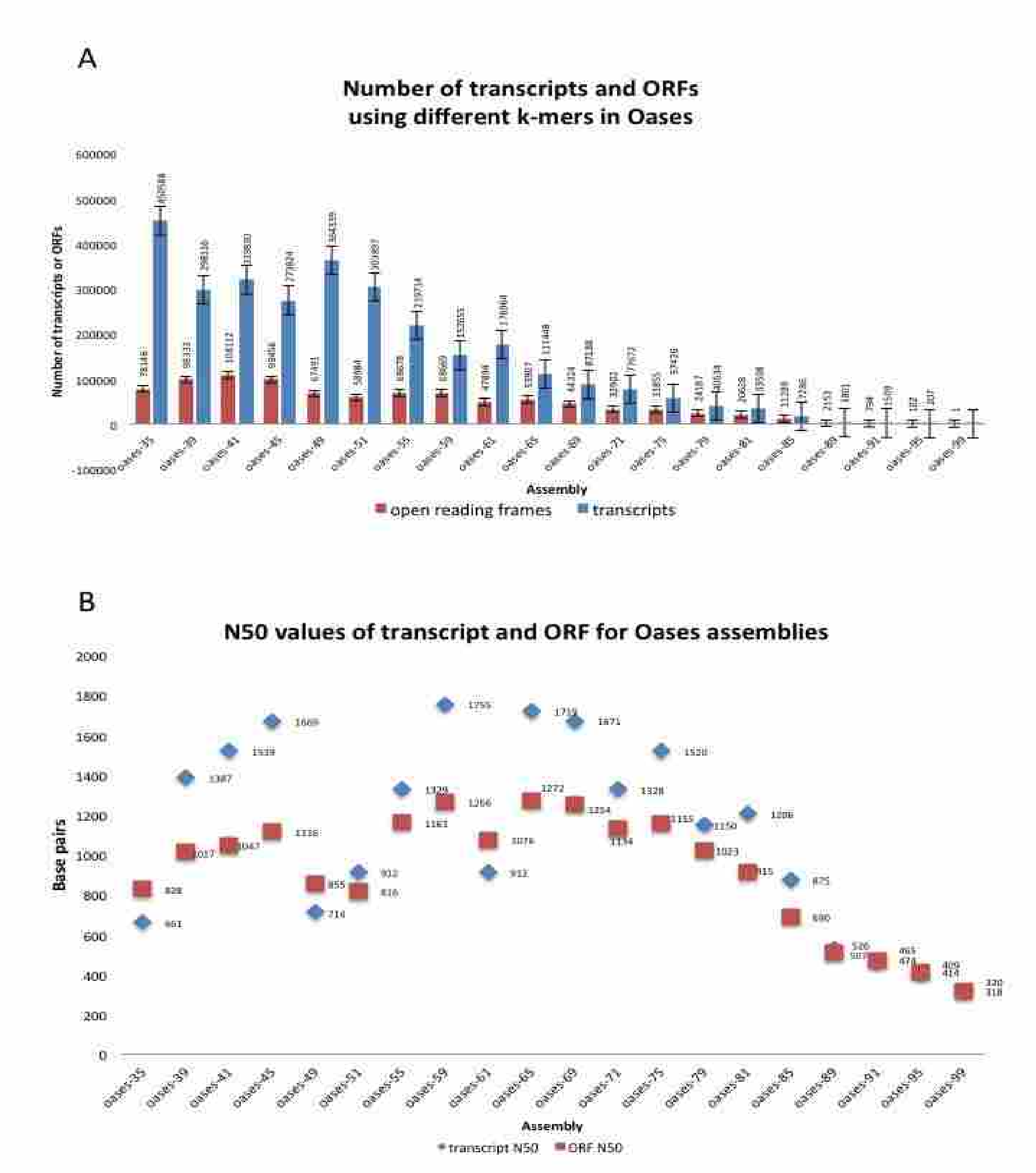


Figure 3.1 A. The Number of Transcripts and Open Reading Frames (ORF) Produced by Each Assembly Using Oases-Velvet, B. The N50 Value of Both the Transcripts Produced by the Assemblies and Their ORFs
Error bars represent the standard error for each number of transcripts and ORFs. The values indicate the counts of transcripts and ORFs produced by the Oases assemblies. N50 value is computed by sorting the contigs from largest to smallest then determining the minimum set whose sizes total 50% of the assembly. Transdecoder was used for open reading frame prediction. The values indicate how many base pairs correspond to the N50 values of the transcripts and the ORFs.

Trinity

We ran Trinity for the single *k*-mer de Bruijn graph approach which uses a set *k*-mer value of 25 in its processing. The final modes used in addition Butterfly, Pasafly and Cufffly. A merge of the three modes were run separately to compare the differences in each assembly. We assembled the reads and assessed each result including the total number of transcripts generated (with at least 200 bp), the N50 value and total length (Figure 3.2A). Similarly, the open reading frames were predicted using Transdecoder. Trinity-Butterfly mode, which is usually run as the default setting for Trinity reported 934,896 transcripts and 185,867 open reading frames. Trinity-Pasafly reported 974,952 transcripts and 220,370 open reading frames. Trinity-Cufffly reported 935,336 transcripts and 185,978 open reading frames and Trinity-merge reported 1,004,011 transcripts and 212,365 open reading frames. N50 values are computed based on the minimum set of contigs whose sizes total 50% of the assembly. The highest N50 for the transcripts were shown by Trinity-Pasafly with 1,109 bp and for its ORF, 876 bp. Trinity-Cufffly followed with transcript and ORF N50 both at 862 bp. Trinity-merge reported a transcript N50 of 940 bp and ORF N50 of 825 bp followed by Trinity-Butterfly with transcript N50 of 861 bp and ORF N50 of 810 bp (Figure 3.2B).

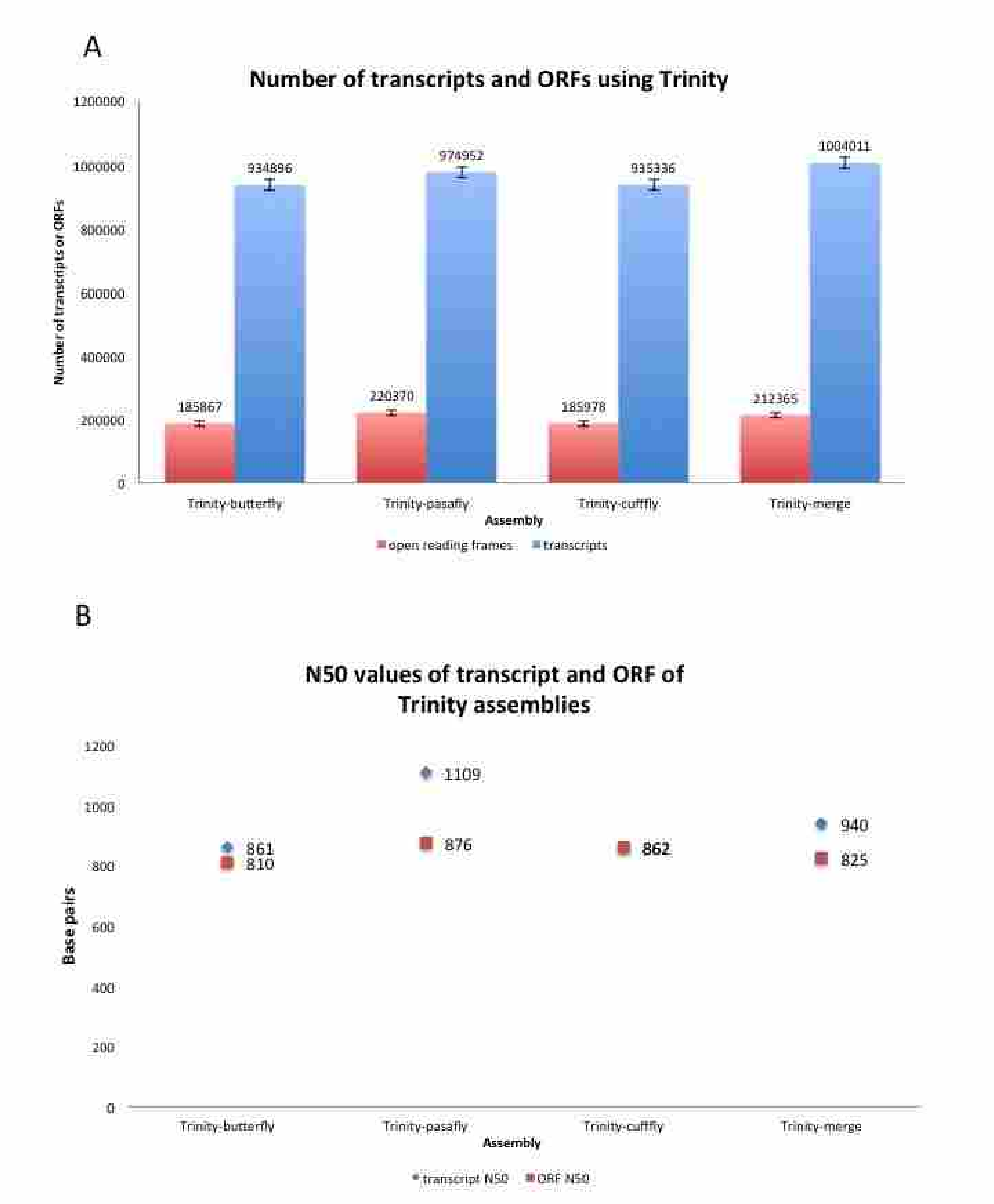


Figure 3.2 A. The Number of Transcripts and Open Reading Frames (ORF) Produced by Each Assembly Using Trinity, B. The N50 Values of the Transcripts Produced by the Trinity Assemblies and Their Open Reading Frames

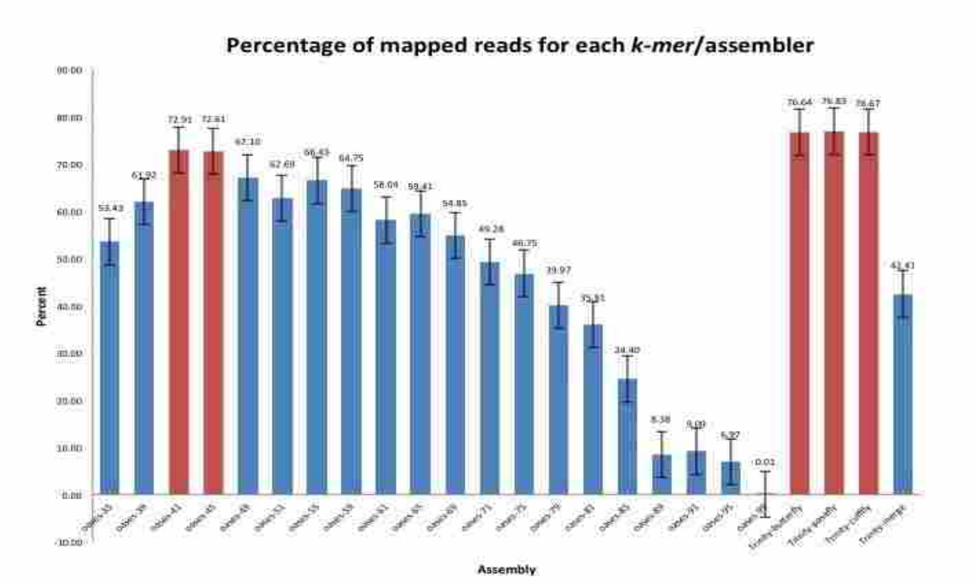


Figure 3.3 The Percentage of Reads Aligning Back to the Assembly
Good assemblies should have most of the reads aligning back to the assembled transcriptome. The Oases assemblies with *k*-mer 41 and 45 and Trinity assemblies (shown in red) achieve the highest values. GSNAP (Genomic Short-read Nucleotide Alignment Program) was used for mapping.

Assessment of the Assemblies

To determine the quality of the assemblies, reads were mapped back to the assembled transcripts using GSNAP (Genomic Short Read Nucleotide Alignment Program). Highlighted in red are the top 5 assemblies ranked according to the number of reads aligning back to the transcript assembly. Trinity assemblies have a consistently high percentage of mapped reads with 76.83% for Pasafly, 76.67% for Cufffly, and 76.64% for the Butterfly mode. Oases assemblies with *k*-mers 41 and 45 also have a high percentage of reads mapping back to the assembly with 72.91% and 72.61% mapped (Figure 3.3). We measured the ratio of the length of predicted open-reading frames to the length of the transcripts to determine how many of the transcripts are considered protein coding genes with annotations (Figure 3.4). Although Oases-99 showed the highest percentage (99.38%), only one transcript is generated and a single open-reading frame is predicted. Oases-79 contains the second highest ORF-transcript ratio with 65.75%. The Trinity assemblies resulted in an ORF-transcript length ratio from 22-23%.

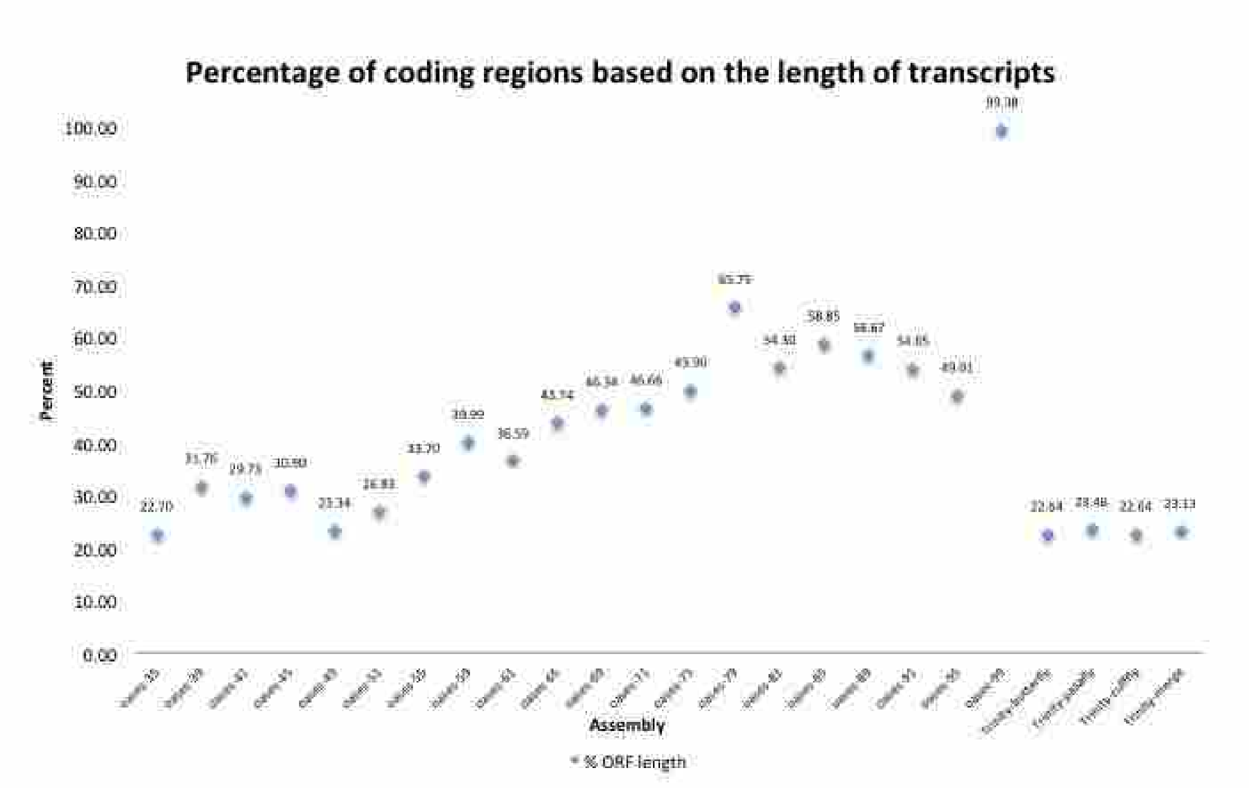


Figure 3.4 Percentage of Predicted Protein Coding Genes
Analysis is calculated by dividing the total length of the coding regions by the total length of the transcripts assembled. This indicates how much of the transcript assembly covers predicted open reading frames. Percent for each data point is shown.

To determine the redundancy and mapping coverage based on the number of transcripts, we plotted the number of transcripts generated by each assembly and the percentage of reads that mapped to the assembly (Figure 3.5). The three Trinity modes achieved high mapping coverage and a high number of transcripts generated. Trinity assemblies report 4% higher read coverage (76.83% Pasafly, 76.67% for Cufffly and 76.64% for Butterfly) than Oases assemblies; however, Oases assemblies, particularly *k*-mer 41 and 45 with 72.91% and 72.61% coverage, produce about one-third as many transcripts as Trinity. Trinity-Pasafly, which achieved the highest percentage of mapped reads, has 974,952 transcripts compared to Oases-41, which has 319,830 transcripts.

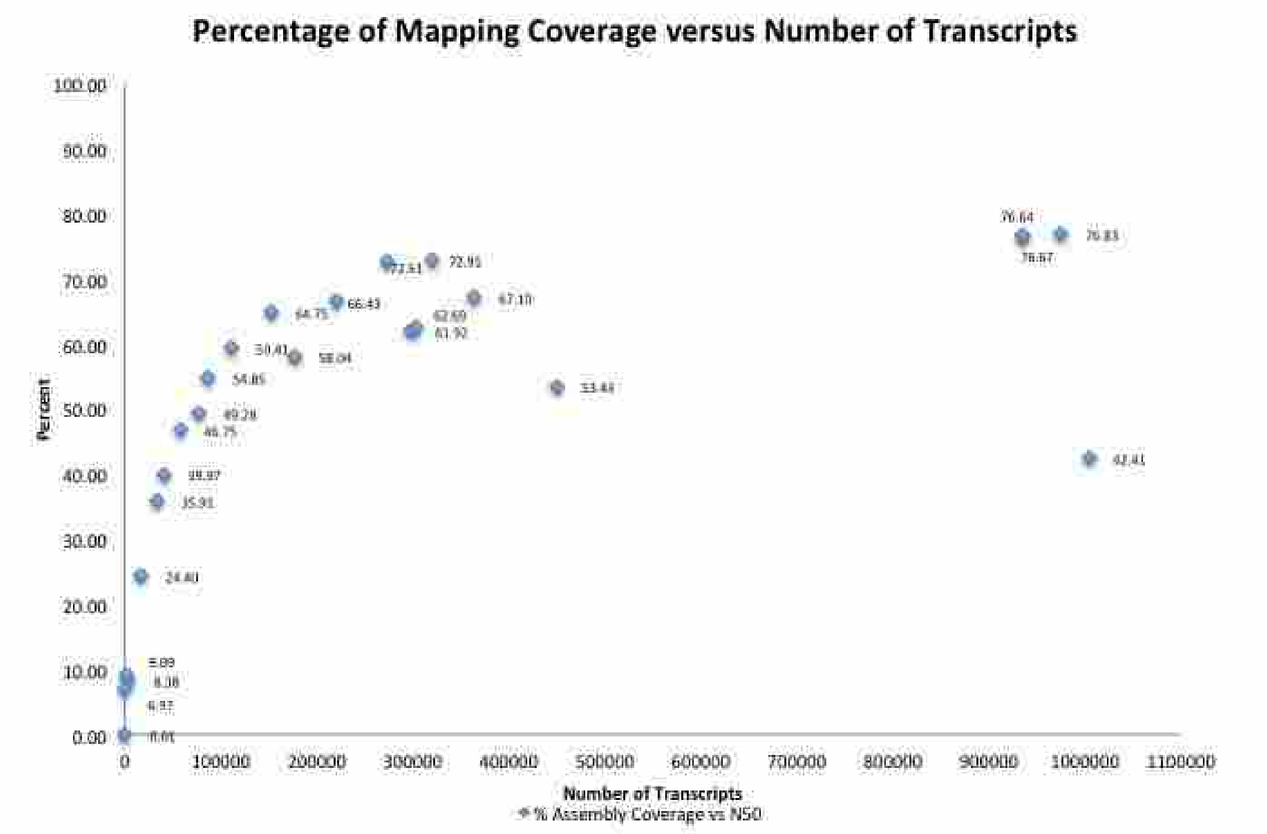


Figure 3.5 The Coverage of the Assemblies and the Corresponding Number of Transcripts Generated. This shows the number of transcripts produced by the assemblies plotted against the percent mapping coverage. The best assembly should have high mapping coverage with low number of transcripts. Percent for each data point is shown.

The mapping coverage of the assembly was plotted against the number of predicted open reading frames (Figure 3.6). The three Trinity modes produced the higher number of ORFs with high mapping coverage. Although Trinity-merge has the highest number of transcripts produced, it has only 42.41% mapping coverage. Oases *k*-mer 41 resulted in 108,112 transcripts and 72.91% mapping coverage. This means that although transcripts produced by Trinity have the higher coverage, there might be a large amount of redundancy in the assembly.

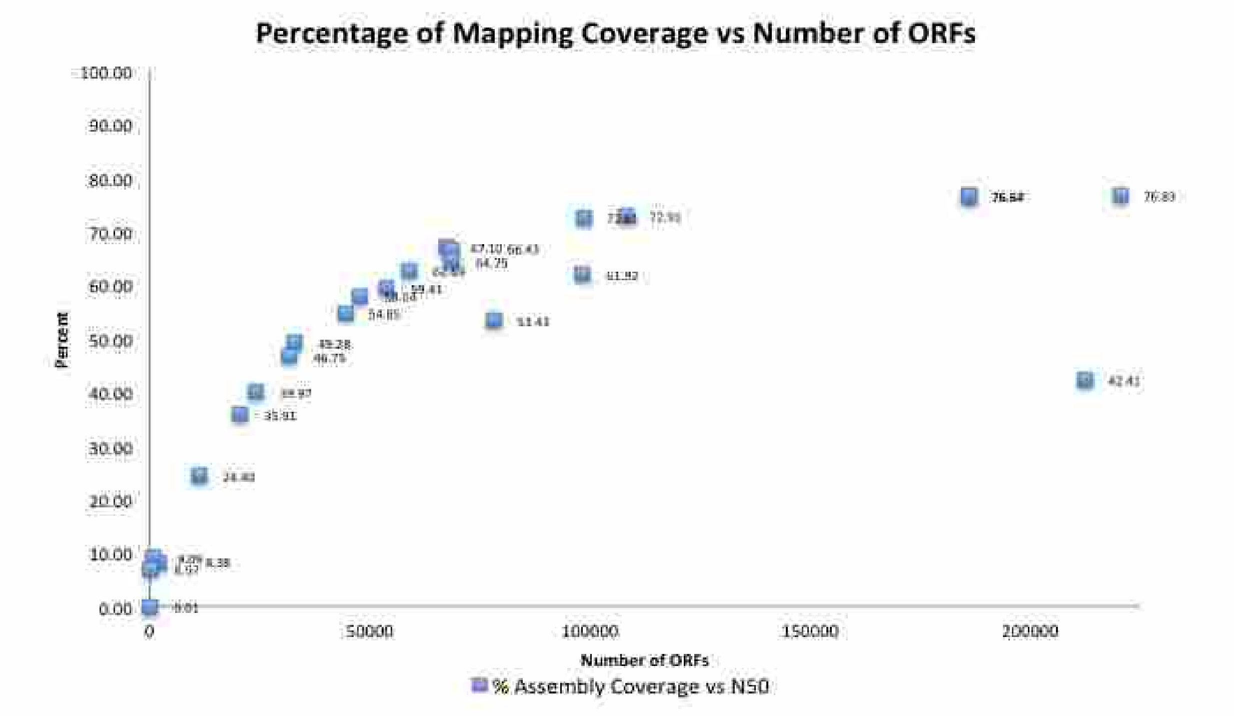


Figure 3.6 The Coverage of the Assemblies and the Number of Predicted Open Reading Frames Generated
The number of predicted open reading frames are plotted against the percent mapping coverage. GSNAP was used for mapping. Percent for each data point is shown.

Clustering using different methods

Although several of the assemblies appeared to incorporate a large number of ORFs in their transcriptome, many of the transcripts are just splice variants. It is difficult to analyze or annotate a genome with hundreds of thousands of transcripts when there are probably only tens of thousands of genes. Clustering methods were used to reduce the number of contigs and to attempt to eliminate splice variants.

To reduce splice variants, we used three clustering methods: CAP3, CDHIT-EST and Isofuse. We selected the assemblies that have the highest percentage of raw reads aligning back to the transcripts which are the three independent Trinity runs with Butterfly, Pasafly and Cufffly modes, Trinity-Merge and Oases assemblies for *k*-mers 41 and 45. CAP3 follows an overlap consensus alignment of the transcripts that links contigs and corrects assembly errors. CDHIT-

EST clusters similar DNA sequences according to a user-defined similarity threshold and Isofuse selects the best hit according to an E-value threshold of e^{-10} using local BLAST [32] and keeps the longest transcript to be used for downstream analyses. These methods will reduce the number of transcripts and compress similar transcripts to decrease redundancy without affecting the quality of the assemblies.

CAP3

We used the CAP3 software to perform a multiple consensus alignment of all the reads. CAP3 is a DNA sequence assembly program that uses base quality values for the computation of overlaps between reads. It constructs multiple alignments of reads for consensus sequence generation. CAP3 also corrects assembly errors and links contigs to produce longer transcripts with fewer errors. The number of transcripts generated, the length of the produced assembly after clustering, the number of predicted open reading frames, and the N50 values are determined. The number of contigs in the transcriptome was significantly reduced with the highest number of transcripts coming from the Trinity-merge assembly (Table 3.1). Trinity assemblies produced 10,709 transcripts for Butterfly, 8,905 transcripts for Pasafly and 10,716 transcripts for Cufffly. Oases assemblies for *k*-mer 41 yielded 29,094 transcripts and 26,254 transcripts for *k*-mer 45. The N50 values range from 1,711 to 2,719 bp. The numbers of predicted open reading frames (ORF) for Trinity assemblies are 33,610 sequences for Trinity-merge, 9,169 for Butterfly, 8,388 for Pasafly and 9179 for Cufffly with N50 values from 983 bp to 1194 bp.

Table 3.1 Summary of CAP3 Clustering

Assembly	# transcripts	Length (bp)	N50 (bp)	# ORF	N50 (bp)
Oases 41	29,094	35,536,336	1,834	13,895	1,152
Oases 45	26,254	36,253,975	2,016	14,571	1,194
Trinity Merge	68,358	84,290,012	1,711	33,610	983
Trinity Butterfly	10,709	19,147,399	2,404	9,169	1,179
Trinity Pasafly	8,905	17,640,047	2,719	8,388	1,185
Trinity Cufffly	10,716	19,169,675	2,407	9,179	1,179

This table shows the summary of the clustering using CAP3 of Oases and Trinity assemblies for multiple consensus sequences.

CD-HIT-EST

To optimize the nucleotide dataset and reduce the number of transcripts, we used CD-HIT-EST (Cluster Database at High Identity with Tolerance- EST), which is a program that clusters the dataset depending on a user-defined similarity threshold such as the sequence identity. This uses the longest sequence first to remove those above a certain threshold[33]. The longest sequence then becomes the seed of the first cluster and the remaining sequences are compared to the existing seed. This also finds high identity segments between sequences to avoid costly full alignments. The objective of CDHIT-EST is to produce DNA sequences from a non-redundant database to be used for downstream analysis. The output file is reported in Table 3.2.

Table 3.2 Summary of CDHIT-EST Clustering

Assembly	# transcripts	Length (bp)	N50	ORF	N50
Oases 41	237,634	202,366,857	1,350	71,948	957
Oases 45	206,999	188,766,870	1,464	64,581	1,011
Trinity Merge	756,976	453,278,702	862	145,570	750
Trinity Butterfly	752,400	419,039,072	744	132,413	726
Trinity Pasafly	730,138	438,462,086	871	137,499	759
Trinity Cufffly	752,699	419,284,426	744	132,497	726

This shows the summary of output for clustering of transcript assemblies using CDHIT-EST [33]. CDHIT-EST clusters using a greedy algorithm by sorting the sequences into length, then takes the longest one to compare to the rest of the similar cluster. A similarity cutoff is used to provide sequence identity that will generate a non-redundant DNA sequence for downstream analysis.

Isoform Fusion (Isofuse)

To reduce transcript number output by fusing isoforms, we created our own algorithm we called Isofuse. The step involves creating a BLAST database of the assembler's output. We then perform nucleotide BLAST using the same assembler's output as the query with a threshold of an expected value of 10^{-10}. The output produces a BLAST archive format indicating query and subject accession number. We then execute a script called Isofuse, which screens all the matches of the query and saves the longest possible sequence into an output file. The output of the file and summary of results are reported in Table 3.3.

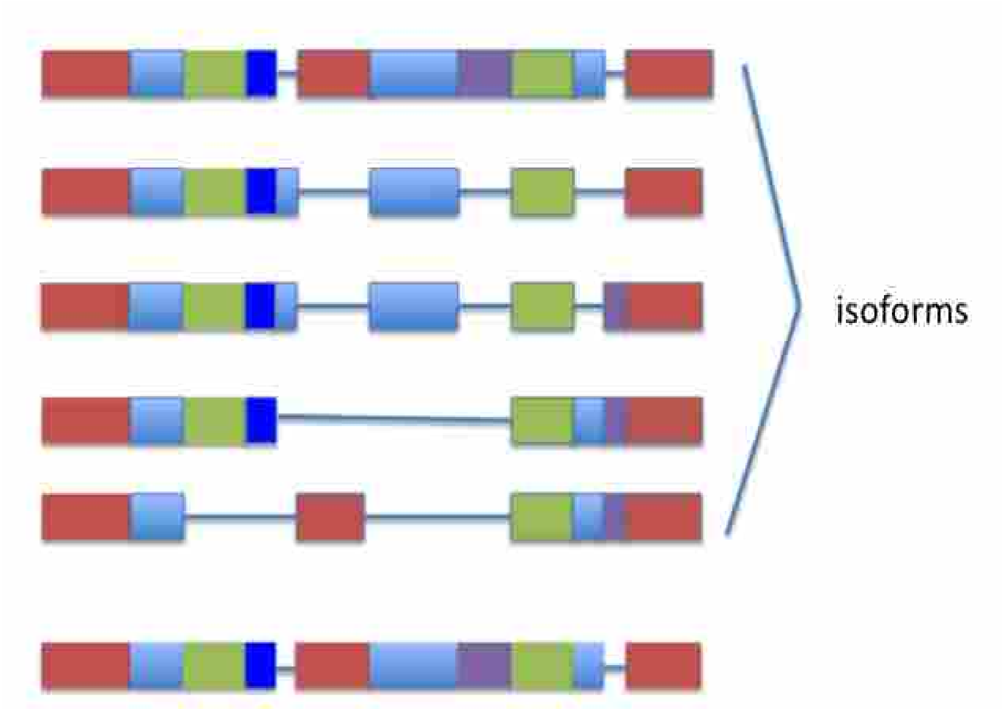


Figure 3.7 Illustration of Alternative Splicing Isoforms
One gene can be spliced in multiple ways, which makes analysis more difficult. This is addressed by the Isofuse algorithm which compresses the amount of isoforms while keeping the longest possible hit.

Table 3.3 Summary of Isofuse Clustering

Assembly	# transcripts	length	N50	ORF	N50
Oases 41	87,798	70,137,188	1,239	22,415	1,026
Oases 45	75,215	64,845,846	1,366	21,690	1,071
Trinity Merge	737,535	426,074,382	796	133,830	783
Trinity Butterfly	487,402	228,360,490	535	62,314	687
Trinity Pasafly	484,413	230,501,366	549	63,718	675
Trinity Cufffly	487,406	228,396,046	535	62,333	687

This shows the results of clustering of transcript assemblies using Isofuse. A BLAST database is created using the output of the chosen assembly then the output is used to BLAST back to itself depending on user-defined E-value threshold. This allows isoforms at closest identity to be clustered together. Isofuse script is then run to select the longest sequence and keep it into an output file.

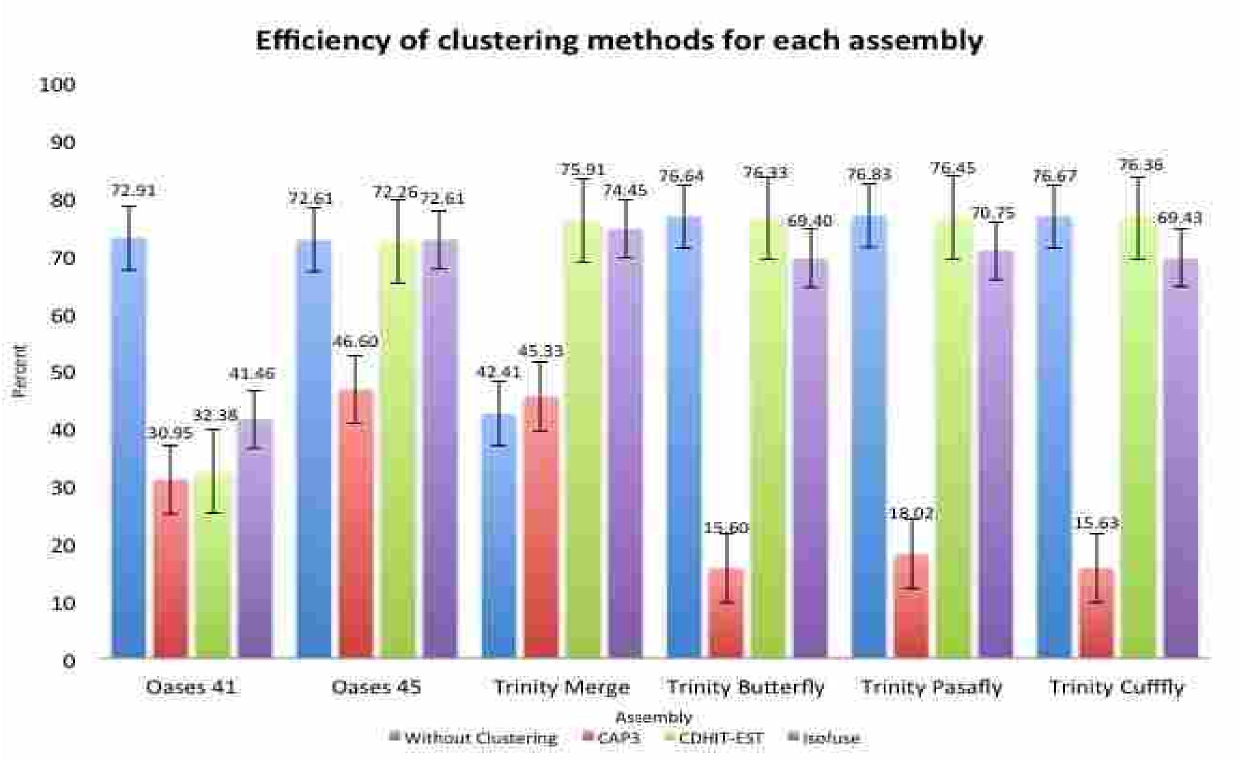


Figure 3.8 Percent Mapping for Each Clustering Algorithm
Selected assemblies were checked for efficiency using the clustering methods by assessing the mapping coverage. The assemblies without using any clustering method have high mapping coverage, however, they yield a large amount of transcripts. CDHIT-EST and Isofuse have proven effective in clustering the assemblies based on the results in comparison to CAP3. Numbers above each bar indicate the corresponding percentage for mapping coverage. Error bars represent standard error.

Analysis of the Different Clustering Methods

To examine the accuracy of the clustering methods CAP3, CDHIT-EST and Isofuse, the reads were mapped back to the clustered assemblies using GSNAP (Figure 3.8). Oases k-41 lost more information after doing the clustering assembly methods. Oases-41 aligns 72.91% without clustering, 30.95% with CAP3, 32.38% with CDHIT-EST and 41.46% with Isofuse. Oases k-45 outperforms Oases k-41 in the degree to which mapping coverage is retained after clustering. Oases k-45 aligns 72% after clustering with Isofuse and CDHIT-EST. These values are similar to its mapping without clustering. Trinity-merge has increased its mapping efficiency after clustering from 42.41% without clustering, 45.33% using CAP3, 61.82% using CDHIT-EST and 61.02% using Isofuse. The three modes of Trinity share the same trends with the results after clustering. Trinity-Butterfly aligns 76.64% without clustering, 15.60% after CAP3, 76.33% after CDHIT-EST and 69.40% after Isofuse. Trinity-Cufffly maps 76.67% without clustering, 15.63% with CAP3, 76.36% using CDHIT-EST and 69.43% with Isofuse. Trinity-Pasafly produced the highest percentage of mapping in most algorithms; without clustering, it aligns 76.83% of the reads, 18.02% using CAP3, 76.45% using CDHIT-EST and 70.75% using Isofuse.

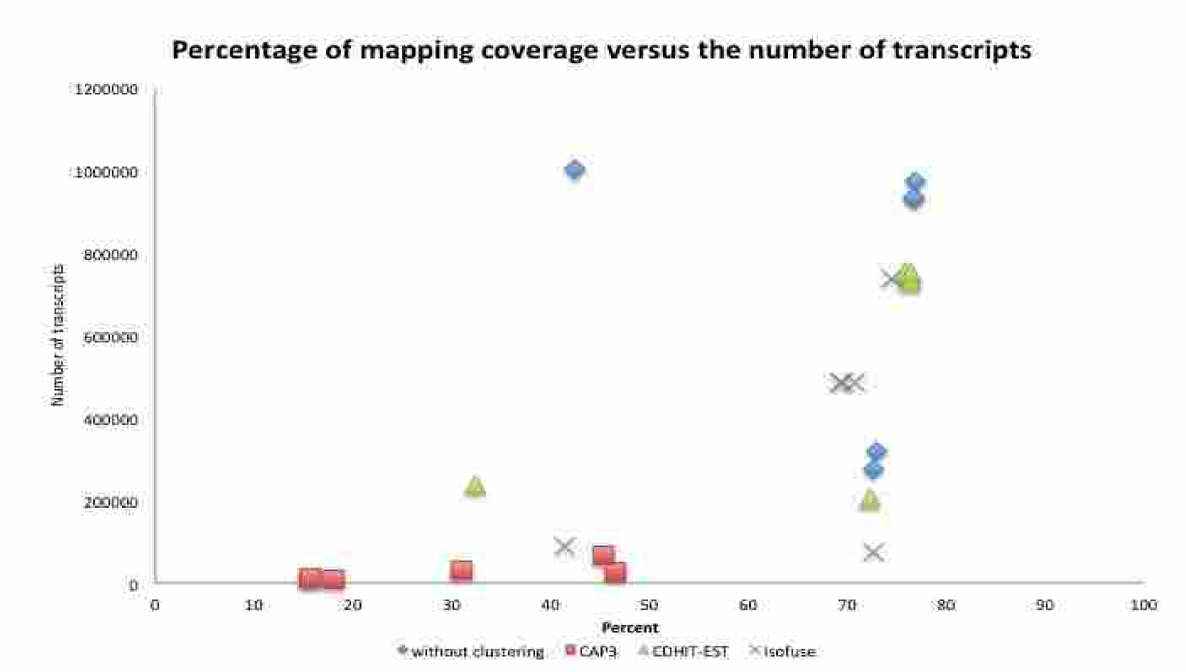


Figure 3.9 Percent Mapping Coverage of Each Algorithm Versus the Number of Transcripts Generated. CAP3 works well in reducing the number of transcript but shows less efficiency on the mapping coverage. The clustering methods (CDHIT-EST and Isofuse) show high efficiency on mapping coverage while reducing the number of transcripts.

To define which algorithms worked well for clustering, we plotted the percent mapping coverage of each of the clustered assemblies using the different algorithms and compared it with the number of transcripts generated (Figure 3.9). The aim is to compress the number of transcripts into a reasonable amount and retain the percentage of reads that map back to a transcript for each assembly without losing information needed for downstream analyses. The number of transcripts without clustering is higher than when a clustering algorithm is applied. The CAP3 method reduces the number of transcripts, however it also reduces the percentage mapping for the reads. CDHIT-EST retains nearly all the information, reduced the transcripts of Oases k-41 and k-45 by nearly 100,000 transcripts and 300,000 transcripts for the Trinity runs and retains read mappings at higher than 70% except for Oases k-41. Isofuse reduces Oases k-41 transcripts from 319,830 to 87,798 but results in 41.46% mapping. Oases k-45 transcripts were reduced from 273,824 to 75,215 while retaining 72% read mapping, Trinity-merge had 1,004,011 transcripts and was reduced to 737,535 while aligning 74.45% of the total reads. Trinity Butterfly, Pasafly and Cufffly had almost one million transcripts but were all reduced to about 480,000 transcripts by Isofuse and still had 69.40%, 70.75% and 69.43% mapping of total reads.

CONCLUSIONS

In this study, we have compared methods for assembly of the *Suaeda fruticosa* transcriptome using short read RNA-seq data. Since *Suaeda fruticosa* does not have a closely related reference genome, a genome-independent reconstruction approach was used with de Bruijn graph methods utilized by Trinity and Velvet-Oases. Pre-assembly methods performed in this study included quality assessments of reads and digital normalization to preserve reads that contain usable information without affecting the assembly process. This greatly eliminates redundancy in reads and reduces computational requirements. The de novo transcriptome assembly algorithm using Velvet-Oases *k*-41 and *k*-45, and all three Trinity modes assemblies reflected high mapping coverage. These assemblies mapped about 70% or more of the raw reads but yielded a high number of transcripts. We optimized the assemblies using the publicly available algorithm CAP3 for multiple consensus alignment, CDHIT-EST clustering using sequence identity and the Isofuse algorithm using the longest and best hit E-value threshold to merge the transcripts while preserving mapping coverage. CAP3 assemblies reduced the number of transcripts but also decreased the quality of the assembly. CDHIT-EST worked well in clustering the assemblies because it retained most of the information especially with the Trinity assemblies.

We have developed a new algorithm for dealing with genes that have multiple alternative splicing isoforms. We recommend the Isofuse method in efficiently clustering the reads, which significantly reduces the total number of transcripts while retaining the mapping coverage as we can observe in Oases *k*-45. Isofuse addresses alternative splicing issues that increase the number of transcripts produced by transcriptome assembly software. These clustering methods can be optimized depending on user-defined thresholds and parameters. This can be applied to annotation of genes, differential expression analysis and other downstream analyses. Numerous assembly packages are publicly available for the processing of RNA-sequencing data. The assembly pipeline must be optimized to produce a transcriptome that can be effectively used by downstream analyses such as differential expression studies and gene annotations. These downstream analyses are impacted by several factors that can be used to evaluate the quality of the transcriptome assembly. One of these important metrics is the percentage of reads that map back to the assembly. This parameter reflects the degree to which read information is retained in the transcriptome. Another metric is the number of open-reading frames found in the transcriptome assembly. These ORFs will be used for gene annotation later in the analysis and are important to generate a high quality transcriptome assembly.

The transcriptome assemblies conducted here provide coverage of a considerable proportion of the *Suaeda fruticosa* transcriptome. These data provide a genetic resource for discovery of potential genes for salt tolerance in this species and may serve as a reference sequence for study of other succulent halophytes.

METHODS

Plant Materials and RNA Isolation

Seeds of *Suaeda fruticosa* were planted and grown according to Hameed et al. [10]. Plant samples of 100 mg of frozen plant tissue from roots and shoots of low (0 mM NaCl) and medium (300 mM NaCl) salt conditions were ground in liquid nitrogen to a fine powder. Total RNA was extracted from these tissues using a Trizol-based method and further cleaned up using the QIAGEN RNeasy Mini kit. The RNA was analyzed for quality and concentration using the Agilent Technologies 2100 Bioanalyzer. High Quality total RNA samples should give two distinct peaks and yield an RNA Integrity Number (RIN) value greater than 8.

Illumina Sequencing Platform

The Illumina RNA-Seq library preparation protocol includes poly-A RNA isolation, RNA fragmentation, reverse transcription to cDNA using random primers, adapter ligation, size-selection from a gel and PCR enrichment [34]. The resulting cDNA library preparation is placed in one of the eight lanes of a flow-cell. Fragments of individual cDNA samples are amplified and converted into clusters of double-stranded DNA. The flow-cell is then placed in the Illumina machine where each cluster is sequenced in parallel. Four fluorescently labeled nucleotides are added at each cycle recording the signals emitted at each cluster. For each flow-cell, this process is repeated for a given number of cycles. The fluorescence intensities are then converted into base-calls. The number of cycles determines the length of the reads; the number of clusters determines the number of reads[35]. The batch of libraries was sequenced using Illumina Hi-seq 2000 sequencer and this includes cDNA libraries of *Suaeda* 0 mM NaCl-treated shoot and roots in triplicates and cDNA libraries of Suaeda 300 mM NaCl-treated shoot and roots in triplicates.

The paired-end library was developed according to the protocol of the Paired-End Sample Preparation kit (Illumina, USA). Illumina RNA-sequencing was performed by Otogenetics (Norcross, GA).

Quality Trimming and Digital Normalization

The raw RNA sequence data was filtered and trimmed using the FastX toolkit and Trimmomatic v.0.27 to utilize only high quality reads prior to the assembly. Sickle paired end trimmer then was used to trim low quality bases towards the 3' and 5' ends of the reads. Digital normalization was used to reduce the number of reads. It removes high abundance reads but retains the read complexity and low abundance transcripts. The software for digital normalization is available electronically through http://ged.msu.edu/papers/2012-diginorm/ *webcite*. A python script is used to interleave the paired-end reads files (http://github.com/ged-lab/khmer/tree/2012-paper-diginorm/sandbox *webcite)*. Khmer software package available at http://github.com/ged-lab/khmer/ *webcite* is used to perform three-pass normalization steps. Loading sequences needed for khmer software works with screed packages through http://github.com/ged-lab/screed/ *webcite* (khmer and screed are ©2010 Michigan State University, and are free software available for distribution, modification, and redistribution under the BSD license). The details of quality trimming and digital normalization are available in Additional File 1.

De novo Assembly

The high quality concatenated reads of shoots and roots were assembled using two software packages. (1) Trinity v. 20131110 was used with a fixed *k*-mer size of 25. Inchworm, Chrysalis and modes of Butterfly, Pasafly and Cufffly were run on a server of 256 Gb of RAM. Parameters of Trinity were set for CPU of 32, running jellyfish on 120 Gb with minimum contig length of 100 bases and average fragment size of 300 bases. (2) Velvet v. 1.2.10 and Oases v. 0.2.08 was used with *k*-mer sizes from ranging from values of 31, 35, 39 up to 99 with increments of 10 and an average insert length 300 bp and minimum contig length of 200. Only assembled transcripts longer than 200 bp were kept. De novo assembly scripts are available in Additional file 2.

Clustering Methods

We ran the clustering methods using CAP3 v.10/15/07, CDHIT-EST v.4.5.4-2011-03-07 and our own Isofuse on the selected assemblies Oases k-41, Oases k-45, Trinity-Merge, Trinity-Butterfly, Trinity-Pasafly and Trinity-Cufffly. We ran CAP3 with its default setting for the file of reads [36]. For CDHIT-EST, we used a sequence identity threshold of 95%, which is the number of identical amino acids in the alignment divided by the full length of the shorter sequence for CDHIT-EST [33].

For Isofuse, a database is made from the output files of the assemblies selected using BLAST. Nucleotide BLAST with megablast task is then performed on the created database and the query becomes the similar output files from the assemblies with a set threshold of E-value of 10^{-10}. The output is then formatted into a tabular format showing the query and the subject accession ID. A python script isofuse.py (Additional File 3) is executed to create a dictionary with the lengths for all the contigs. The script screens the result of the BLAST file and keeps only the longest in each group into an output file.

Sequence Analysis

The assemblies were transferred to an open-reading frame prediction software package called Transdecoder in the Trinity Package which reports candidate coding regions within the transcripts. For each assembly, the number of transcripts, N50 values and the total length of the assemblies are identified. The analysis of the efficiency of assemblies is performed using GMAP and GSNAP v. 2013-11-27. GMAP maps and aligns cDNA sequences originally used for genomic mapping, then GSNAP aligns single-end or paired-end reads. It can detect short and long distance splicing using probabilistic models or databases of known splice sites.

Authors' Contributions

JDA designed and carried out experiments, including RNA isolation, cDNA library preparation, and RNA-seq analysis and data interpretation, and wrote the manuscript. MC provided guidance on computational analysis of the sequence data and interpretation. BG provided information and seeds for growing the plants and helped with RNA isolation and cDNA preparation. MAK provided expertise in halophyte growth and analysis, obtained funding, and assisted in preparation of the manuscript. BLN provided guidance on the project, obtained funding, and helped format and edit the manuscript.

ACKNOWLEDGEMENTS

This research has been supported by a grant from the Pakistan-U.S. Science and Technology Cooperation Program, U.S. Dept. of State and Higher Education Commission of Pakistan, and by the Department of Microbiology and Molecular Biology at Brigham Young University.

Additional Files

```
Additional File 1
Quality trimming and Digital Normalization

Trimmomatic
java -jar trimmomatic -0.27.jar PE leftuntrimmed.fastq rightuntrimmed.fastq lefttrimmed.fastq
s1_se.fastq righttrimmed.fastq s2_se.fastq ILLUMINACLIP: illuminaClipping.fa:2:30:10

Interleave paired end reads
python interleave-reads.py left.pe.fq right.pe.fq | gzip -9c > reads.pe.fq.gz

Fastq quality trimmer and Fastx trimmer
for i in *.pe.fq.gz *.se.fq.gz
do
    echo working with $i
    newfile="$(basename $i .fq.gz)"
    gunzip -c $i | fastq_quality_filter -Q33 -q 30 -p 50 | gzip -9c > "${newfile}.qc.fq.gz"
done


Digital Normalization
#Extracting paired ends from the interleaved files
for i in *.pe*.qc.fq.gz
do
    python strip-and-split-for-assembly.py $i
done

#Digital Normalization
python normalize-by-median.py -p -k 21 -C 30 -N 4 -x 3e9 --savehash normC30k21.kh *.pe.qc.fq.gz

#Trim erroneous k-mers
python filter-abund.py -V normC30k21.kh *.keep


#Strip and split orphaned and paired end- reads
for i in *.pe*.qc.fq.gz
do
    python strip-and-split-for-assembly.py $i
done


De Novo Assembly
Assembly with Velvet
velveth velvet.41 41 -fastq -short reads.se.qc.keep.abundfilt.fq.gz -shortPaired
reads.pe.qc.keep.abundfilt.fq.gz \
velvetg velvet.41 -read_trkg yes -ins_length 300 -min_contig_lgth 200 -cov_cutoff 5 \

Scaffolding with Oases
oases velvet.41 -scaffolding yes -unused_reads yes -ins_length 300 -min_trans_lgth 200 -
cov_cutoff 5 \

#performed for k-mers 35 to 99


Assembly with Trinity
ulimit -s unlimited
ulimit -a
Trinity.pl \
--seqType fq \
--JM 120G \
--output Trinityresults \
--SS_lib_type FR \
--CPU 20 \
--min_kmer_cov 2 \
--left left.fastq \
--right right.fastq \
--single single.fastq \

1> trinity.out \
2> trinity.err \
```

```
#isofuse.py

from Bio import SeqIO
from collections import defaultdict
import sys

if len(sys.argv) < 4:
        print "Usage: ",sys.argv[0]," contig.fasta blast.txt output.fasta"
        print "The blast.txt file should have lines with two contigs that match"
        exit(0)
lengths = defaultdict(int)
sequences = {}
unique = []
input_handle = open(sys.argv[1])

#first create a dictionary with the lengths for all of the contigs
#and add the contigs to a sequences dictionary
for record in SeqIO.parse(input_handle, "fasta") :
        print record.id,",",record.seq
        lengths[record.id] = len(record.seq)
        sequences[record.id] = record;

input_handle.close()

#now read the blast file, and keep only the longest in each group
blast_handle = open(sys.argv[2])
for line in blast_handle:
        words = line.split()
        #compare the lengths of the two contigs and delete the shorter
        # dont do anything unless they are both in the dictionary
        if words[0] == words[1]:
                continue
        if ((words[0] in lengths) and (words[1] in lengths)):
                if lengths[words[0]] > lengths[words[1]]:
                        #delete words[1] because is it shorter
                        lengths.pop(words[1])
                else:
                        lengths.pop(words[0])

print "Results"
for key, value in lengths.iteritems():
        print key," Length ",value
        unique.append(sequences[key])

output_handle = open(sys.argv[3], "w")
SeqIO.write(unique, output_handle, "fasta")
output_handle.close()
```

REFERENCES

1. Beltran JM, Manzur CL: Overview of salinity problems in the world and FAO strategies to address the problem. *Proceedings of the International Salinity Forum* 2005:311-313.

2. Hameed A, Hussain T, Gulzar S, Aziz I, Gul B, Khan M: Salt tolerance of a cash crop halophyte *Suaeda fruticosa*: biochemical responses to salt and exogenous chemical treatments. *Acta Physiol Plant* 2012, 34(6):2231-2340.

3. Labidi N, Ammari M, Mssedi D, Benzerti M, Snoussi S, Abdelly C: Salt excretion in *Suaeda fruticosa*. *Acta biologica Hungarica* 2010, 61:299-312.

4. Weber DJ: Adaptive mechanisms of halophytes in desert regions. *Salinity and Water Stress* 2008, 44:179-185.

5. Bennani-Kabachi N, El-Bouayadi F, Kehel L, Fdhil H, Marquie G: Effect of *Suaeda fruticosa* aqueous extract in the hypercholesterolaemic and insulin-resistant sand rat. *Therapie* 1999, 54:725-730.

6. Towhidi A, Saberifar T, Dirandeh E: Nutritive value of some herbage for dromedary camels in the central arid zone of Iran. *Tropical Animal Health Pro* 2011, 43:617-622.

7. Bareen F, Tahira SA: Metal accumulation potential of wild plants in tannery effluent contaminated soil of Kasur, Pakistan: field trials for toxic metal cleanup using *Suaeda fruticosa. J Hazard Mater* 2011, 186:443-450.

8. Khan MA, Ansari R, Ali H, Gul B, Nielsen BL: *Panicum turgidum*: a sustainable feed alternative for cattle in saline areas. *Agric Eco Env* 2009, 129:542-546.

9. Chaudhri II, Shah BH, Naqvi N, Mallick IA: Investigations on the role of Suaeda fruticosa Forsk in the reclamation of saline and alkaline soils in West Pakistan plains. *Plant and Soil* 1964, 21(1):1-7.

10. Anil VS, Rahjkumar P, Kumar P, Mathew MK: A plant Ca2+ pump, ACA2, relieves salt hypersensitivity in yeast. Modulation of cytosolic calcium signature and activation of adaptive Na+ homeostasis. *J Biol Chem* 2008, 283:3497-3506.

11. Khan MA, Ungar IA, Showalter AM: The effect of salinity on the growth, water status, and ion content of a leaf succulent perennial halophyte, *Suaeda fruticosa (L.) Forssk. Journal of Arid Environments* 2000, 45:73-84.

12. Flowers TJ, Colmer TD: Salinity tolerance in halophytes. *New Phytologist* 2008, 179:945 - 963.

13. Brown CT, Howe A, Zhang Q, Pyrkosz A, Brom T: A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv: 12034802* 2012.

14. Zerbino DR: Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2010, Chapter 11:Unit 11 15.

15. Garber M, Grabherr MG, Guttman M, Trapnell C: Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* 2011, 8(6):469-477.

16. Schulz MH, Zerbino DR, Vingron M, Birney E: Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012, 28(8):1086-1092.

17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: Gene Ontology: tool for the unification of biology. *Nat Genet* 2000, 25(1):25-29.

18. Wu TD, Nacu S: Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010, 26(7):873-881.

19. Page JT: BamBam: Tools for genomic analysis. In. sourceforge; 2013.

20. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practival and powerful approach to multiple testing. *J Royal Statistical Soc Series* 1995, 57:289-300.

21. Benjamini Y, Yekutieli D: The control of the false discovery rate in multiple testing under dependency. *Ann Statistics* 2001, 29:1165-1188.

22. Rushton PJ, Somssich IE, Ringler P, Shen QJ: WRKY transcription factors. *Trends in Plant Science* 2010, 15(5):247-258.

23. Garg R, Verma M, Agrawal S, Shankar R, Majee M, Jain M: Deep Transcriptome Sequencing of Wild Halophyte Rice, *Porteresia coarctata*, Provides Novel Insights into the Salinity and Submergence Tolerance Factors. *DNA Research* 2014, 21(1):69-84.

24. Diédhiou CJ, Popova OV, Golldack D: Comparison of salt-responsive gene regulation in rice and in the salt-tolerant *Festuca rubra ssp. litoralis*. *Plant Signaling & Behavior* 2009, 4(6):533-535.

25. Xie Z, Zhang Z-L, Zou X, Huang J, Ruas P, Thompson D, Shen QJ: Annotations and Functional Analyses of the Rice WRKY Gene Superfamily Reveal Positive and Negative Regulators of Abscisic Acid Signaling in Aleurone Cells. *Plant Physiology* 2005, 137(1):176-189.

26. Robatzek S, Somssich IE: Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes & Development* 2002, 16(9):1139-1149.

27. Achard P, Gong F, Cheminant S, Alioua M, Hedden P, Genschik P: The Cold-Inducible CBF1 Factor–Dependent Signaling Pathway Modulates the Accumulation of the Growth-Repressing DELLA Proteins via Its Effect on Gibberellin Metabolism. *The Plant Cell Online* 2008, 20(8):2117-2129.

28. Alvey L, Harberd NP: DELLA proteins: integrators of multiple plant growth regulatory inputs? *Physiologia Plantarum* 2005, 123(2):153-160.

29. Ma J, Zhang M, Xiao X, You J, Wang J, Wang T, Yao Y, Tian C: Global Transcriptome Profiling of *Salicornia europaea* L. Shoots under NaCl Treatment. *PLoS One* 2013, 8(6).

30. Magrane M, consortium U: UniProt Knowledgebase: a hub of integrated protein data. In: *Database, 2011: bar009.* 2011.

31. Brinker M, Brosche M, Vinocur B, Abo-Ogiala A, Fayyaz P, Janz D, Ottow EA, Cullmann AD, Saborowski J, Kangasjarvi J *et al*: Linking the Salt Transcriptome with Physiological Responses of a Salt-Resistant *Populus* Species as a Strategy to Identify Genes Important for Stress Acclimation. *Plant Physiology* 2010, 154(4):1697-1709.

32. Dang Z, Zheng L, Wang J, Gao Z, Wu S, Qi Z, Wang Y: Transcriptomic profiling of the salt-stress response in the wild recretohalophyte *Reaumuria trigyna*. *BMC Genomics* 2013, 14(1):29.

33. Fan P, Nie L, Jiang P, Feng J, Lu S, Chen X, Bao H, Guo J, Tai F, Wang J *et al*: Transcriptome analysis of *Salicornia europaea* under saline conditions revealed the adaptive primary metabolic pathways as early events to facilitate salt adaptation. *PLoS One* 2013, 8(11).

34. Sahu BB, Shaw B: Isolation, identification and expression analysis of salt-induced genes in *Suaeda maritima*, a natural halophyte, using PCR-based suppression subractive hybridization. *BMC Plant Biology* 2009, 9(69).

35. Qi YC, Liu WQ, Qiu LY, Zhang SM, Ma L, Zhang H: Overexpression of glutathione S-transferase gene increases salt tolerance of Arabidopsis. *Russ J Plant Physiol* 2010, 57(2):233-240.

36. VP R, RH S, ER A, RD A: Overexpression of glutathione S-transferase/glutathione peroxidase enhances the growth of transgenic tobacco seedlings during stress. *Nat Biotechnol* 1997, 15:988.

37. Hurkman WJ, Tanaka CK: Effect of Salt Stress on Germin Gene Expression in Barley Roots. *Plant Physiology* 1996, 110(3):971-977.

38. Lane BG, Dunwell JM, Ray JA, Schmitt MR, Cuming AC: Germin, a protein marker of early plant development, is an oxalate oxidase. *Journal of Biological Chemistry* 1993, 268(17):12239-12242.

39. Busov V, Johannes E, Whetten R, Sederoff R, Spiker S, Lanz-Garcia C, Goldfarb B: An auxin-inducible gene from loblolly pine (Pinus taeda L.) is differentially expressed in mature and juvenile-phase shoots and encodes a putative transmembrane protein. *Planta* 2004, 218(6):916-927.

40. Dana MdlM, Pintor-Toro JA, Cubero B: Transgenic Tobacco Plants Overexpressing Chitinases of Fungal Origin Show Enhanced Resistance to Biotic and Abiotic Stress Agents. *Plant Physiology* 2006, 142(2):722-730.

41. Mullen RT, Lisenbee CS, Miernyk JA, Trelease RN: Peroxisomal Membrane Ascorbate Peroxidase Is Sorted to a Membranous Network That Resembles a Subdomain of the Endoplasmic Reticulum. *The Plant Cell Online* 1999, 11(11):2167-2185.

42. Ondrasek G: The Responses of Salt-Affected Plants to Cadmium. In: *Salt Stress in Plants.* Edited by Ahmad P, Azooz MM, Prasad MNV: Springer New York; 2013: 439-463.

43. Nedjimi B, Daoud Y: Cadmium accumulation in *Atriplex halimus subsp. schweinfurthii* and its influence on growth, proline, root hydraulic conductivity and nutrient uptake. *Flora - Morphology, Distribution, Functional Ecology of Plants* 2009, 204(4):316-324.

44. Lefevre I, Marchal G, Meerts P, Correal E, Lutts S: Chloride salinity reduces cadmium accumulation by the Mediterranean halophyte species *Atriplex halimus L. Environmental and Experimental Botany* 2009, 65(1):142-152.

45. Guan C, Rosen ES, Boonsirichai K, Poff KL, Masson PH: The ARG1-LIKE2 Gene of Arabidopsis Functions in a Gravity Signal Transduction Pathway That Is Genetically Distinct from the PGM Pathway. *Plant Physiology* 2003, 133(1):100-112.

46. Kroczyńska B, Coop NE, Miernyk JA: AtJ6, a unique J-domain protein from *Arabidopsis thaliana*. *Plant Science* 2000, 151(1):19-27.

47. Pardo J, Reddy M, Yang S, Maggio A, Huh G-H, Matsumoto T, Coca M, Paino-D'Urzo M, Koiwa H, Yun D-J *et al*: Stress signaling through Ca2+/calmodulin-dependent protein phosphatase calcineurin mediates salt adaptation in plants. *Proceedings of the National Academy of Sciences* 1998, 95(16):9681-9686.

48. Galon Y, Aloni R, Nachmias D, Snir O, Feldmesser E, Scrase-Field S, Boyce J, Bouché N, Knight M, Fromm H: Calmodulin-binding transcription activator 1 mediates auxin signaling and responds to stresses in *Arabidopsis*. *Planta* 2010, 232(1):165-178.

49. Cheong YH, Kim K-N, Pandey GK, Gupta R, Grant JJ, Luan S: CBL1, a Calcium Sensor That Differentially Regulates Salt, Drought, and Cold Responses in *Arabidopsis*. *The Plant Cell Online* 2003, 15(8):1833-1845.

50. Nakajima S, Ito H, Tanaka R, Tanaka A: Chlorophyll b Reductase Plays an Essential Role in Maturation and Storability of *Arabidopsis* Seeds. *Plant Physiology* 2012, 160(1):261-273.

51. Lokhande V, Suprasanna P: Prospects of Halophytes in Understanding and Managing Abiotic Stress Tolerance. In: *Environmental Adaptations and Stress Tolerance of Plants in the Era of Climate Change.* Edited by Ahmad P, Prasad M. Maharashtra, India: Springer; 2012.

52. Martínez-Ballesta MdC, Bastías E, Carvajal M: Combined effect of boron and salinity on water transport: The role of aquaporins. *Plant Signaling & Behavior* 2008, 3(10):844-845.

53. Jarvis DE, Ryu C, Beilstein MA, Schumaker KS: Distinct Roles for SOS1 in the Convergent Evolution of Salt Tolerance in *Eutrema salsugineum* and *Schrenkiella parvula*. *Molecular Biology and Evolution* 2014, 31(8):2094-2107.

54. Hechenberger M, Schwappach B, Fischer WN, Frommer WB, Jentsch TJ, Steinmeyer K: A Family of Putative Chloride Channels from *Arabidopsis* and Functional Complementation of a Yeast Strain with a CLC Gene Disruption. *Journal of Biological Chemistry* 1996, 271(52):33632-33638.

55. Qiu Q, Guo Y, Dietrich M, Schumaker K, Zhu J: Regulation of SOS1, a plasma membrane Na+/H+ exchanger in Arabidopsis thaliana, by SOS2 and SOS3. *Proc Natl Acad Sci USA* 2002, 99:8436 - 8441.

56. Hill CB, Jha D, Bacic A, Tester M, Roessner U: Characterization of Ion Contents and Metabolic Responses to Salt Stress of Different *Arabidopsis* AtHKT1.1 Genotypes and Their Parental Strains. *Molecular Plant* 2012, 6(2):350-368.

57. Senadheera P, Maathuis FJM: Differentially regulated kinases and phosphatases in roots may contribute to inter-cultivar difference in rice salinity tolerance. *Plant Signaling & Behavior* 2009, 4(12):1163-1165.

58. Lee Y, Giorgi F, Lohse M, Kvederaviciute K, Klages S, Usadel B, Meskiene I, Reinhardt R, Hincha D: Transcriptome sequencing and microarray design for functional genomics in the extremophile *Arabidopsis* relative *Thellungiella salsuginea (Eutrema salsugineum)*. *BMC Genomics* 2013, 14(1):793.

59. Hashimoto M, Negi J, Young J, Israelsson M, Schroeder JI, Iba K: Arabidopsis HT1 kinase controls stomatal movements in response to CO2. *Nat Cell Biol* 2006, 8(4):391-397.

60. Ho S-L, Huang L-F, Lu C-A, He S-L, Wang C-C, Yu S-P, Chen J, Yu S-M: Sugar starvation- and GA-inducible calcium-dependent protein kinase 1 feedback regulates GA biosynthesis and activates a 14-3-3 protein to confer drought tolerance in rice seedlings. *Plant Molecular Biology* 2013, 81(4-5):347-361.

61. Cui MH, Yoo KS, Hyoung S, Nguyen HTK, Kim YY, Kim HJ, Ok SH, Yoo SD, Shin JS: An *Arabidopsis* R2R3-MYB transcription factor, AtMYB20, negatively regulates type 2C serine/threonine protein phosphatases to enhance salt tolerance. *FEBS Letters* 2013, 587(12):1773-1778.

62. Hundertmark M, Hincha DK: LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 2008, 9:118.

63. Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Hosokawa D, Shinozaki K: Role of *Arabidopsis* MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *The Plant Cell Online* 1997, 9(10):1859-1868.

64. Swindell W, Huebner M, Weber A: Transcriptional profiling of *Arabidopsis* heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics* 2007, 8(1):125.

65. Illumina: Ilumina, Inc.; 2009.

66. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26(1):139-140.

67. Haddad F, Baldwin K: Reverse transcription of the ribonucleic acid: the first step in RT-PCR assay. *Methods in molecular biology* 2010, 630:261-270.

68. Livak K, Schmittgen T: Analysis of relative gene expression data using real-time quantitative PCR and the 2-(delta delta C(T) method. *Methods* 2001, 25:402 - 408.

69. Mutz K, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F: Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology* 2013, 24(1):22-30.

70. Nowrousian M: Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryotic cell* 2010, 9(9):1300-1310.

71. Qin QP, Zhang LL, Li NY, Cui YY, Xu K: Optimizing of cDNA preparation for next generation sequencing. *Yi chuan* 2010, 32(9):974-977.

72. Salgotra RK, Gupta BB, Stewart CN, Jr.: From genomics to functional markers in the era of next-generation sequencing. *Biotechnology letters* 2013.

73. Jimenez-Gomez JM: Next generation quantitative genetics in plants. *Frontiers in plant science* 2011, 2:77.

74. McGettigan PA: Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology* 2013, 17(1):4-11.

75. Dang Z-h, Zheng L-l, Wang J, Gao Z, Wu S-b, Qi Z, Wang Y-c: Transcriptomic profiling of the salt-stress response in the wild recretohalophyte Reaumuria trigyna. *BMC Genomics* 2013, 14:29.

76. Kamle S, Ali S: Genetically modified crops: detection strategies and biosafety issues. *Gene* 2013, 522(2):123-132.

77. Kujur A, Saxena MS, Bajaj D, Laxmi, Parida SK: Integrated genomics and molecular breeding approaches for dissecting the complex quantitative traits in crop plants. *Journal of biosciences* 2013, 38(5):971-987.

78. Glenn E, Brown JJ, O'Leary JW: Irrigating Crops with Seawater. In: *Scientific American.* 1998.

79. Gong Q, Li P, Ma S, Indu Rupassara S, Bohnert HJ: Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* 2005, 44(5):826-839.

80. Inan G, Zhang Q, Li P, Wang Z, Cao Z, Zhang H, Zhang C, Quist TM, Goodwin SM, Zhu J *et al*: Salt cress. A halophyte and cryophyte Arabidopsis relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant Physiol* 2004, 135(3):1718-1737.

81. Vera-Estrella R, Barkla BJ, Garcia-Ramirez L, Pantoja O: Salt stress in *Thellungiella halophila* activates Na+ transport mechanisms required for salinity tolerance. *Plant Physiology* 2005, 139(3):1507-1517.

82. Ding M, Hou P, Shen X, Wang M, Deng S, Sun J, Xiao F, Wang R, Zhou X, Lu C *et al*: Salt-induced expression of genes related to Na+/K+ and ROS homeostasis in leaves of salt-resistant and salt-sensitive poplar species. *Plant Molecular Biology* 2010, 73(3):251-269.

83. Qiu Q, Ma T, Hu Q, Liu B, Wu Y, Zhou H, Wang Q, Wang J, Liu J: Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. *Tree Physiology* 2011, 31(4):452-461.

84. Ayarpadikannan S, Chung E, Cho C-W, So H-A, Kim S-O, Jeon J-M, Kwak M-H, Lee S-W, Lee J-H: Exploration for the salt stress tolerance genes from a salt-treated halophyte, Suaeda asparagoides. *Plant Cell Rep* 2012, 31(1):35-48.

85. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* 2011, 29(7):644-652.

86. Clarke K, Yang Y, Marsh R, Xie L, Zhang KK: Comparative analysis of de novo transcriptome assembly. *Sci China Life Sci* 2013, 56(2):156-162.

87. Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 2008, 18:821-829.

88. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *Journal of Molecular Biology* 1990, 215(3):403-410.

89. Fu L, Niu B, Zhu Z, Wu S, Li W: CDHIT-accelerated for clustering the next generation sequencing data. *Bioinformatics* 2012, 28:3150-3152.

90. Bullard J, Purdom E, Hansen K, Dudoit S: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* 2010, 11(1):94.

91. Huang X, Madan A: CAP3: A DNA sequence assembly program. *Genome Res* 1999, 9:868 - 877.

CHAPTER 4: Transcriptome Assembly, Profiling and Differential Gene Expression Analysis of the Halophyte *Suaeda fruticosa* Provides Insights into Salt Tolerance

Joann Diray-Arce[1], Mark Clement[2], Bilquees Gul[3], M. Ajmal Khan[4] and Brent L. Nielsen[1]
[1]Department of Microbiology and Molecular Biology, Brigham Young University,
Provo, Utah 84602 U.S.A.
[2]Department of Computer Science, Brigham Young University, Provo, Utah, 84602
[3]Institute of Sustainable Halophyte Utilization, University of Karachi, Pakistan
[4]College of Arts and Sciences, Qatar University, Doha, Qatar

E-mail addresses:

Joann Diray-Arce, joann.diray@gmail.com
Mark Clement, clement@cs.byu.edu
Bilquees Gul, bilqueesgul@uok.edu.pk
M. Ajmal Khan, ajmal.khan@qu.edu.qa

Corresponding Author: Brent L. Nielsen, brentnielsen@byu.edu

© 2015 BMC Genomics 16:253 DOI: 10.1186/s12864-015-1553-x Diray-Arce et al.;
licensee BioMed Central. 6 May 2015

ABSTRACT

Background

Improvement of crop production will be required in order to feed the growing world population as the amount and quality of agricultural land decreases and salinity in soil increases. This has stimulated research to understand mechanisms of salt tolerance in plants. Most crops can only tolerate a limited amount of salt to survive and produce biomass. Halophytes (salt-tolerant plants) have the ability to maintain productivity and biomass while growing with saline water utilizing specific biochemical mechanisms. However, little is known about the genes and proteins involved in salt tolerance. We have characterized the transcriptome of *Suaeda fruticosa*, a halophyte that has the ability to sequester salts in its leaves. *Suaeda fruticosa* is an annual shrub in the family Chenopodiaceae found in coastal and inland regions of Pakistan and throughout Mediterranean shores. This plant is an obligate halophyte that has the capacity for bioremediation of toxic metals and saline soils. It grows optimally from 200-400 mM NaCl and can grow at up to 1000 mM NaCl. High throughput sequencing technology was performed to provide understanding of genes involved in the salt tolerance mechanism. *De novo* assembly of the transcriptome and analysis is presented for identification of differentially expressed and unique genes present in this non-conventional crop.

Results

Twelve sequencing libraries prepared from control (0 mM NaCl treated) and optimum (300 mM NaCl treated) plants were sequenced using Illumina Hiseq 2000 to investigate differential gene expression between shoots and roots of *Suaeda fruticosa*. The transcriptome was assembled de novo using Velvet and Oases k-45 and clustered using CDHIT-EST. There are

65,880 unigenes generated from the assembly. Among these genes, 475 genes are downregulated and 44 genes are upregulated when compared with samples from plants grown under optimal salt and those grown with no salt. These results correlate closely with the physiological data of *Suaeda fruticosa* where the plant grows optimally at 300 mM NaCl. BLAST analysis identified the differentially expressed genes and they have been annotated with a cutoff E-value of 10^{-10} . The genes were categorized in gene ontology terms and their pathways.

Conclusions

This work has identified potential genes that are involved in mechanisms of salt tolerance in *Suaeda fruticosa* and has provided an outline of tools to use for de novo analysis of transcriptomes. The assemblies that were used provide coverage of a considerable proportion of the transcriptome, which allows analysis of differential gene expression and identification of specific genes that may be involved in salt tolerance in this plant. These data provide a genetic resource for discovery of potential genes for salt tolerance in this species and may serve as a reference sequence for study of other succulent halophytes.

Keywords

Halophytes, *Suaeda*, RNA-seq, differential expression, transcriptome profiling, de novo assembly, transcriptome, salt tolerance

BACKGROUND

Salinity affects about 400 million hectares of land worldwide due to excessive irrigation and continues to increase in parallel with the population. Salinity in soil and water has caused substantial economic losses, including an estimated $230 million for the Indus Basin in Pakistan and $2 billion for the Colorado River basin in the U.S. [37][2]. An estimated total of 200 million hectares of new cropland is needed to feed the rapidly expanding population but only 93 million hectares are available for expansion and farming of traditional crops [1]. Attempts have been made with conventional crops to breed salt tolerance; however, these crops can only tolerate limited amounts of salt in their systems. The potential of halophytes, the natural flora of saline habitats, has been under-examined until recently and their utilization may allow production of useful crops on salty soils.

Suaeda fruticosa, a succulent shrub in the family Chenopodiaceae, is an obligate halophyte that grows optimally at 300 mM NaCl and has the adaptation to reduce sodium buildup for long term survival [10]. This perennial halophyte has a strong ability to accumulate and sequester Na^+ and Cl^- without the aid of salt glands, bladder or trichomes [38]. It is a good source of high quality edible oil [11], has potential for antiophthalmic, hypolipidaemic and hypoglycemic medicinal purposes [12], and has economic usage as forage for animals [13]. *S. fruticosa* also could help in bioremediation and reclamation of soils contaminated with toxic metals [14] and salinity [15]. Field studies showed that this plant can remove about 2646 kg of NaCl per hectare from the soil each year [39]. At optimum (300 mM NaCl) salt treatment of this species antioxidant enzymes trigger stress response through the activation of $H_2O_2^-$ mediated Ca^{2+} uptake to maintain Na^+ homeostasis at the cellular or tissue level [10]. Calcium ions, responsible for the overall signaling network of growth and development of the plant, are accumulated in the cell cytosol with the increase of Na^+ [40]. At higher salinities, a significant reduction in growth is prevalent which might be due to the maximum threshold of the plant's ability to adjust to specific ion toxicity and osmotic capability. Physiological data analysis has led to reports of ion accumulation, osmotic adjustments, maintenance of pressure potential and growth and production of glycinebetaine as part of a salt tolerance mechanism [41]. Previous studies of the impact of salinity on *S. fruticosa* have linked salt tolerance to its ability to uptake K^+ in order to maintain a higher K^+/Na^+ ratio in the shoots. Higher sequestration of sodium and chloride in the shoot vacuoles together with the ability to synthesize osmoprotectants such as glycinebetaine has been suggested to maintain a favorable water potential gradient and protect cellular structures. Similar to *Suaeda fruticosa*, the majority of halophytes do not have glands or external bladders to modulate their tissue ion concentration therefore it has been seen to be a good model genus for the study of salt tolerance [23].

Next generation sequencing allows differential gene expression analysis of gene alleles and spliced transcripts, non-coding RNA and others, which will lead to identification of differentially expressed and/or unique genes. In this transcriptome paper, we report the identification of genes that are induced or repressed in plants grown under optimal salt conditions in comparison to low salt conditions. We generated a data set of transcript sequences from the roots and the shoots of *Suaeda fruticosa*. The genes were compared for differential expression under the indicated treatments using the assembled transcriptome, and common and tissue-specific patterns of transcriptomic responses were also analyzed. This first transcriptome study of *Suaeda fruticosa* expands our knowledge on global gene expression data for salt-accumulating halophytes that do not have external bladders.

RESULTS AND DISCUSSIONS

De Novo Transcriptome Assembly and Assessments of Expressed Sequenced Tags

Experimental Design

To prepare for the transcriptome assembly and analysis, total RNA was extracted from shoots and roots of *Suaeda fruticosa*. These include biological triplicates of cDNA libraries for *S. fruticosa* roots from plants grown without salt (R000), roots with 300 mM optimal salt (R300), shoots with no salt (S000) and shoots with 300 mM optimal salt (S300). Total mRNA was purified using oligo dT and transcribed into cDNA libraries using TruSeq RNA Sample Prep Kit for Illumina 100 bp paired-end sequencing.

Sequencing Method and Quality Assessment of the Reads

A total of 335.3 million reads of 100 bp were generated by Illumina Hi-seq platform. The reads were filtered using Trimmomatic to remove adapters, FASTX toolkit and Sickle program to remove low quality reads and discard reads based upon the threshold of length. A total of 84.58% of the reads were trimmed and filtered totaling to 283,587,292 reads.

To normalize and assemble RNA-seq reads for *de novo* assembly, digital normalization was used for 283.6 million reads. *K*-mer hash of 21 with coverage of 30X was built from a set of reads to correct redundancy issues, variations in sequences, and potential errors among the reads. Since some reads with sequencing errors may escape quality score-based filtering steps, the reads with potential errors are flagged for removal from the dataset to improve the de novo assembly. Sequencing errors can affect the assembly algorithms so it is best to eliminate the reads that have non-uniform *k*-mer coverage. Reads that have non-uniform k-mer coverage create a problem with the assembly therefore it is necessary to normalize the reads to a certain threshold. This threshold represents the approximate minimum for de novo assembly to work optimally and efficiently. Digital normalization [25] was applied to the total of 283,587,292 paired end reads with *k*-mer size of 21 and k-mer coverage cutoff of 30X. The retained reads were normalized to 99,577,045 (Table 4.1) to remove overabundant reads, reduce noise of the sequenced sample and decrease the overall percentage of errors (Figure 4.1). The effect of digital normalization is to retain nearly all real *k*-mers while discarding the majority of erroneous and redundant k-mers. This step allows reducing the reads and obtaining a transcriptome assembly much faster than and superior to the assembly based on the full data set without affecting the quality of the assembly. Because the genes in the transcriptome have different levels of expression, k-mer distribution will not show any peak at any k value.

Table 4.1 Statistics of Reads

Reads preparation	Libraries	Number of reads	Total reads
Raw reads	R000	95,248,764	335,271,656
	S000	75,414,804	
	R300	84,162,958	
	S300	80,445,130	
FastX toolkit and Trimmomatic	R000	68,444,064	292,898,120
	S000	68,872,348	
	R300	79,313,812	
	S300	76,267,896	
Sickle Trimmed	All		283,587,292
Digital Normalization	All		99,577,045

The summaries of the pre-assembly methods are indicated. R000 represents roots in 0 mM NaCl treatment, S000 are shoots in 0 mM NaCl treatment, R300 are roots in 300 mM NaCl treatment and S300 are shoots in 300 mM NaCl treatment.

To assemble the *Suaeda fruticosa* transcriptome we utilized a genome-independent reconstruction approach. The strategy involved building a de Bruijn graph made of overlapping subsequences or *k*-mers using Velvet [30]. The overlapping bases allow building a graph of all the sequences that then traverse a path guided by read and paired-end coverage [26]. The path through the graph is reported as transcripts. To assemble the contigs into scaffolds, we used a de Bruijn graph software, Oases [28]. *K*-mer sizes from 35 to 99 were chosen to generate the assemblies. We assessed the quality of the assemblies based on the total number of transcripts, open reading frames predicted using Transdecoder and highest mapping percentage of the reads using GSNAP. The number of sequences, N50 values, mean length of the sequences and total base pair length for the contigs, scaffolds and unigenes were also determined (Table 4.1). Among the individual *k*-mer values, transcript numbers associated with different *k*-mer values vary from 1 to 450,588. *K*-mer length is related inversely to the number of transcripts generated. The highest N50 (computed by sorting the contigs from largest to smallest and then determining the minimum set whose sizes total 50% of the assembly) generated for the transcripts is 1,755 generated by a *k*-mer length of 59. Predicted open reading frames range from 1 to 108,112 bp with the highest number of ORFs being generated with a *k*-mer of 41. The highest N50 for open reading frames belongs to an assembly with a *k*-mer of 65 with 1,272 bp. Mapping coverage for the assemblies range from 30.39% to 72.91%. The highest percentage of reads mapping back to the assembly belongs to assemblies with k-mers 41 and 45 with 72.91% and 72.61% mapped.

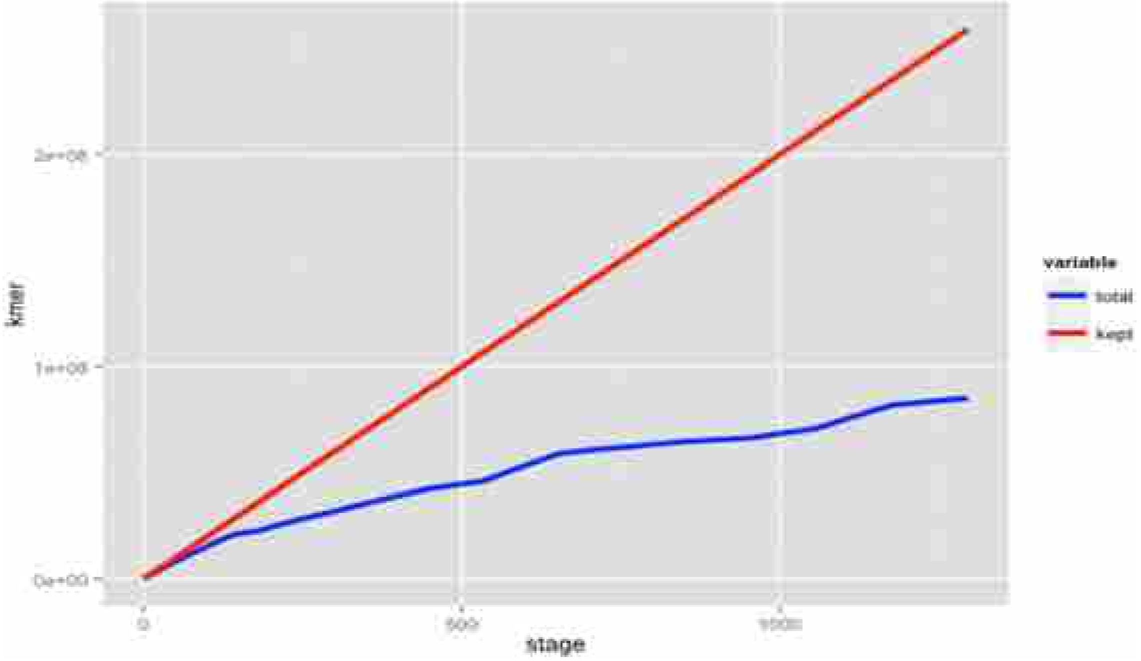


Figure 4.1 Plot of Total Read Pairs Versus Kept Read Pairs After Digital Normalization Algorithms
The true k-mer counts are kept using digital normalization to reduce computational memory and correct redundancy.

Assembly *k*-45 contains a higher percentage of proper pairs aligned with 59.56% and higher N50 compared to Assembly *k*-41, therefore it was chosen to be the assembly for the succeeding steps. Assembly k-45 contained 296,776 contigs from Velvet with a N50 length of 1548 bp and mean size of 928 bp. We selected contigs that were greater than 200 bp in length. The contigs were assembled into scaffolds using Oases and yielded 273,824 contigs with an N50 length of 1669 bp and mean size of 1012 bp. The shortest scaffold is 152 bp and the longest one is 14,046 bp. Using CDHIT-EST, scaffold sequences were assembled into clusters and Transdecoder was used to predict open reading frames. We obtained 65,880 unigenes with an N50 of 1002 bp. The size range of the unigenes is between 297 to 6639 bp. There are 16,778 unigenes comprising 25.5% of the total that have lengths of more than 1000 bp. The mean size of the unigenes is 795 bp (Table 4.2).

Table 4.2 Statistics of Sequence Assembly

	Contigs	Scaffolds	Unigenes
Number of sequences	296,766	273,824	54,526
N50 (bp)	1,548	1,669	957
Mean length (bp)	928	1,012	764
Total length (bp)	275,319,083	277,056,733	41,651,347

The table shows the summary of de novo sequence assembly after using Velvet for contig assembly, Oases for Scaffolds then CDHIT-EST and Transdecoder for the unigenes determination.

Functional Annotation, Gene Ontology Assignments and Analysis

The unigenes assembled were used as query for annotation using BlastX searches based on sequence homologies to the databases of the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein database, RefSeq, SwissProt UniProt and the Kyoto Encyclopedia of Genes and Genomes (KEGG) using BLAST2GO. The summary of top hit distribution similar to *Suaeda fruticosa* unigenes is illustrated in Figure 4.2A. The species distribution with the lowest e-value matching the best sequence alignment result showed that the

S. fruticosa transcriptome sequences have 8697 unigenes (13%) matching to *Vitis vinifera* (grapes), 3818 unigenes (5.7%) matching to *Theobroma cacao* (cacao tree) and 3127 unigenes (4.7%) similar to *Beta vulgaris* (beet). The closest halophyte species matching with *Suaeda fruticosa* is *Populus trichocarpa* (poplar tree) with 2327 unigenes (3.5% matching). For poplar only the initial analysis of the draft genome has been completed; additional mapping and sequencing is ongoing. Some of the halophytes mentioned in this paper do not have full annotation of genes submitted to NCBI database and some only contain partial transcriptome information. Figure 4.2B summarizes the data distribution summary from the sequences from the assembled transcriptome.

Gene names and GO terms were assigned to the transcripts based on homologies with an E-value threshold of 10^{-10}. The data distribution summary for these sequences is shown in Figure 2B. Annotated sequences utilize assigned functional terms to query sequences from GO terms based on the gene ontology vocabulary. Mapped sequences are those with retrieved GO terms associated with the hits obtained after a BLAST search. The search produced 36,668 annotated sequences among 65,870 total transcripts, comprising 55.67% of total sequences. There are 8972 sequences comprising 13.62% of the total transcripts that did not surpass the annotation threshold and 6881 sequences or 10.45% had hits in the databases but lack functional information. A large proportion has no significant sequence alignment or hits in any of the databases, comprising 13,349 sequences or 20.2% of total transcripts which suggests that they may contain novel sequences or a high number of *Suaeda fruticosa* specific transcripts or transcript portions such as orphan untranslated regions.

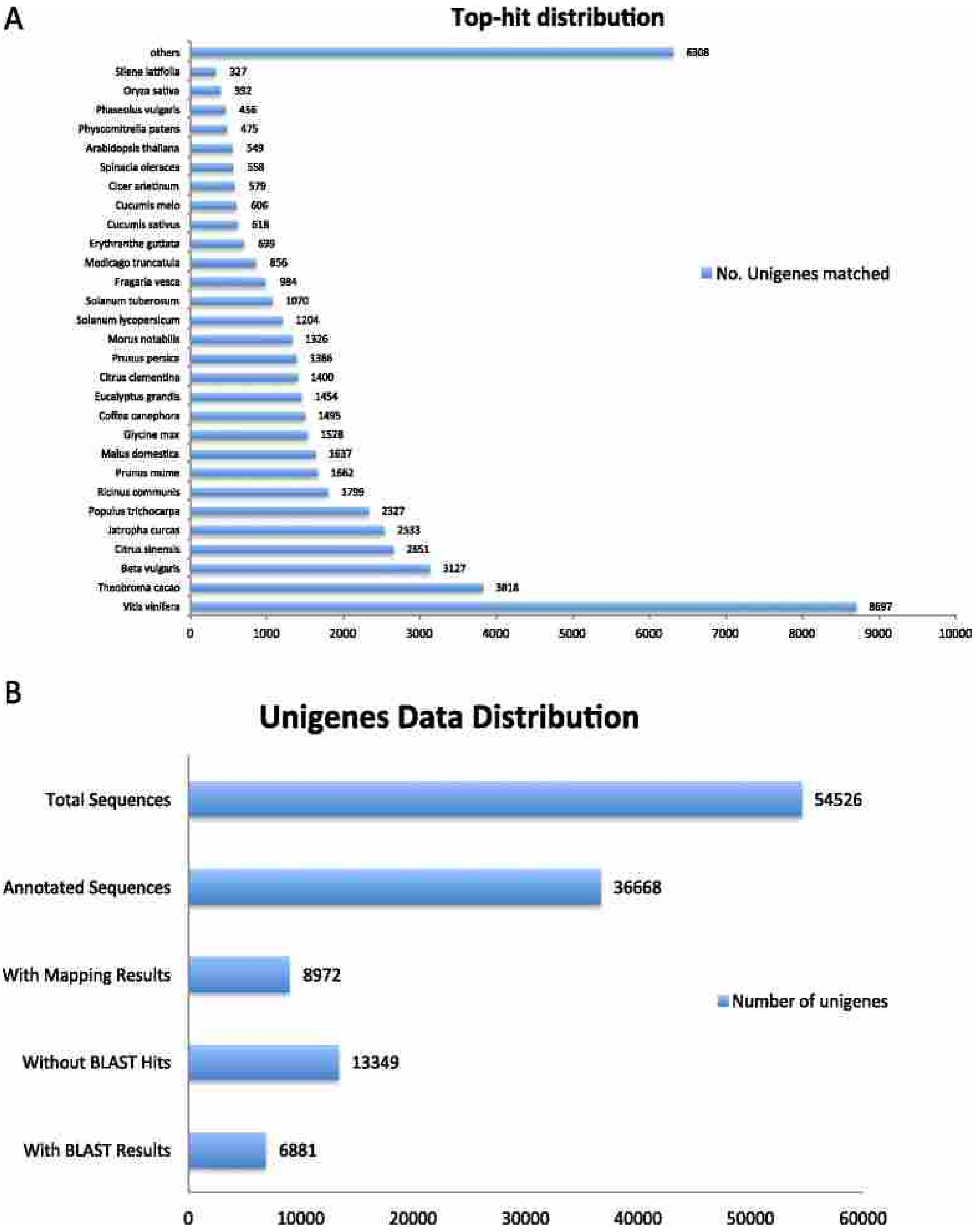


Figure 4.2 A. Top Hit Distribution of Matched Unigenes Among Different Species Generated from BLASTX, B. Data Distribution Summary from BLAST2GO Shows BLAST Hits, Mapping Results and Annotated Sequences

Gene ontology encompasses a dynamic library for gene and protein roles in cells. This includes three main categories: Biological process, referring to the biological objective of the genes or gene products. Molecular function is defined by the biochemical activity of the genes or gene products; and Cellular components, referring to the place in the cell where the gene product is active [42]. Figure 4.3 illustrates the gene ontology annotation of the total assembled unigenes from the de novo assembled transcriptome of *Suaeda fruticosa* using BLAST2GO.

In the biological process category, genes related to stress make up 12% or 1229 of the total unigenes annotated (Figure 4.3A). The other dominant subcategories were protein modification (933 unigenes or 9%), structural development (930 unigenes or 9%) and DNA metabolic process (923 unigenes or 9%). The following subcategories include unigenes involved in carbohydrate metabolism (798 unigenes or 8%), nucleobase-containing compound catabolic process (443 unigenes or 4%), organelle organization (348 unigenes or 3%), reproduction (513 unigenes or 5%), ribosome biogenesis (319 unigenes or 3%), signal transduction (594 unigenes or 6%), single organism development (457 unigenes or 5%), translation (346 unigenes or 3%), transmembrane transport (556 unigenes or 6%), lipid metabolism (513 unigenes or 5%), cofactor metabolism (314 unigenes or 3%), and cellular amino acid metabolism (850 unigenes or 8%). Figure 4.3B illustrates the cellular component category, which has a dominant subcategory of plastid (1374 unigenes or 17%), plasma membrane (1003 unigenes or 12%) and protein complex (941 unigenes or 11%). The molecular function category was comprised of protein coding genes involved in ion binding (3061 unigenes or 31%), oxidoreductase (961 unigenes or 10%), and those responsible for redox reactions of the cell and kinases (809 unigenes or 8%) (Figure 4.3C).

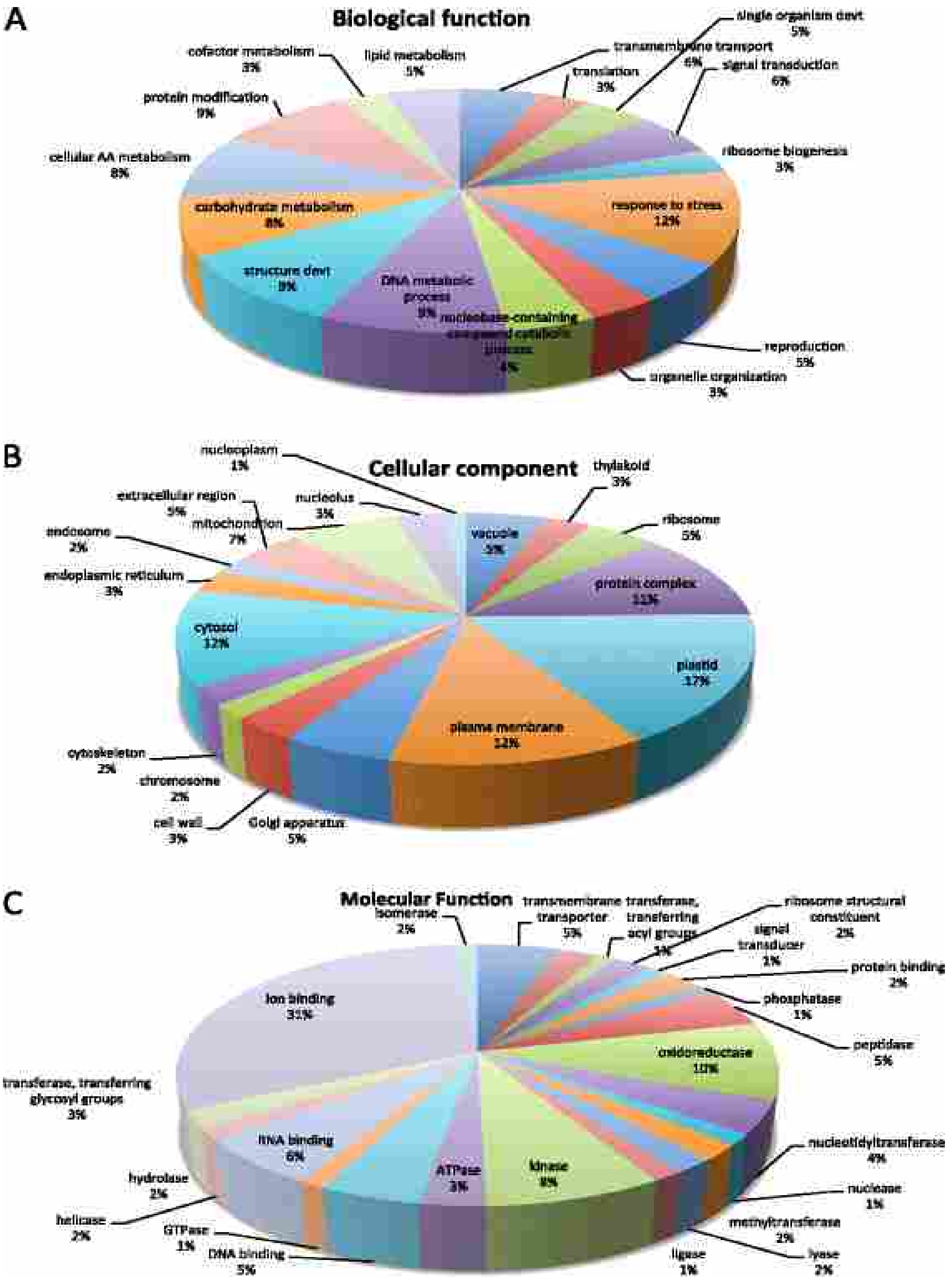


Figure 4.3 Gene Ontology Summary of Total Aassembled ESTs Using BLAST2GO
Distribution of Gene Ontology Annotation of *Suaeda fruticosa* transcriptome. The results are summarized as follows: **(A)** Biological Process, **(B)**. Cellular component **(C)** Molecular Function.

These gene ontology annotations represent a profile for gene expression of *Suaeda fruticosa* suggesting that this species has diverse protein coding genes comprising its structural, regulatory, metabolic and stress response mechanisms.

Differential Expression Analysis

To acquire counts data for differential expression analysis, samples of different treatments (0 mM and 300 mM NaCl treatments) were mapped to the newly generated reference transcriptome using GSNAP (Genomic Short-Read Nucleotide Alignment Program) which utilizes computational methods to detect variants and splicing isoforms in short reads through merging and filtering position lists from a genomic index. It also detects short and long-distance splicing including interchromosomal splicing using probability models or a database of known splice sites [43]. Conversion of bam files into count data was performed using BamBam [44] to summarize the number of reads mapped to each annotated feature. Differential expression calls were made using the EdgeR package. Normalization is applied to the treatments and tissue types to provide accurate differential expression rather than individual quantification. The EdgeR package adjusted the analysis taking into account sequencing depths represented by library sizes. Variations between biological replicates were clustered closely using a multidimensional scaling (MDS) plot similar to that shown in Figure 4 to check for variations among replicates and samples.

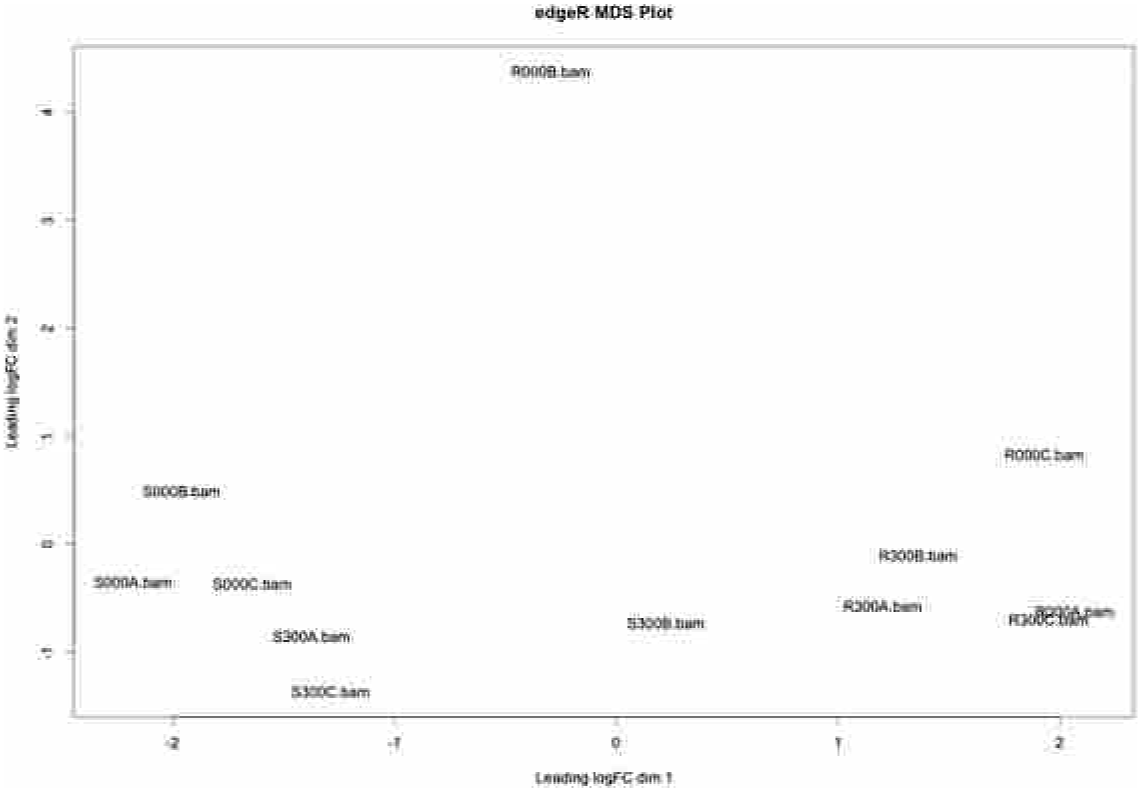


Figure 4.4 Multidimensional Scaling Plot for the Sequencing Libraries
Multidimensional Scaling Plot (MDS) is designed to indicate sample relationship similarity. Shoots and roots of 0 mM and 300 mM NaCl with their biological replicates are analyzed. (Key: S000A- shoots 0 mM replicate A, S000B- shoots 0 mM replicate B, S000C- shoots 0 mM replicate C, S300A- shoots 300 mM replicate A, S300B- shoots 300 mM replicate B, S300C- shoots 300 mM replicate C, R000A- roots 0 mM replicate A, R000B- roots 0 mM replicate B, R000C- roots 0 mM replicate C, R300A- roots 300 mM replicate A, R300B- roots 300 mM replicate B, R300C- roots 300 mM replicate C, bam (bam files)).

The replicates of treatments and their tissue types from transcriptome analysis were used to produce a multidimensional scaling plot (Fig. 4.4), which allows us to see a spatial configuration of how similar or dissimilar the different treatments and biological replicates of *S. fruticosa* shoot and root samples are. The relationship of shoot treatments is more closely clustered together in comparison to root treatments. The tight clustering of the shoot data points means there are fewer variations among biological replicates in comparison to the root treatments. Root samples, however, have greater variations among the treatments and their biological replicates. This indicates that root tissues show less consistency with expression of genes among treatments. Common dispersions were then estimated on the distributions of reads across genes. Each gene gets an assignment of a unique dispersion estimate, which is to be compared to a common dispersion. The biological coefficients of variation versus the abundance were plotted (Fig. 4.5). This specifies relative abundance of each gene variation between RNA samples and also measurement error estimated by the sequencing technology. From this sample, it shows a common dispersion of 0.37 and BCV of 61.09%. This means that common variation shows overall variability across the genome for this dataset and the common variation square root indicates high coefficient of biological variation.

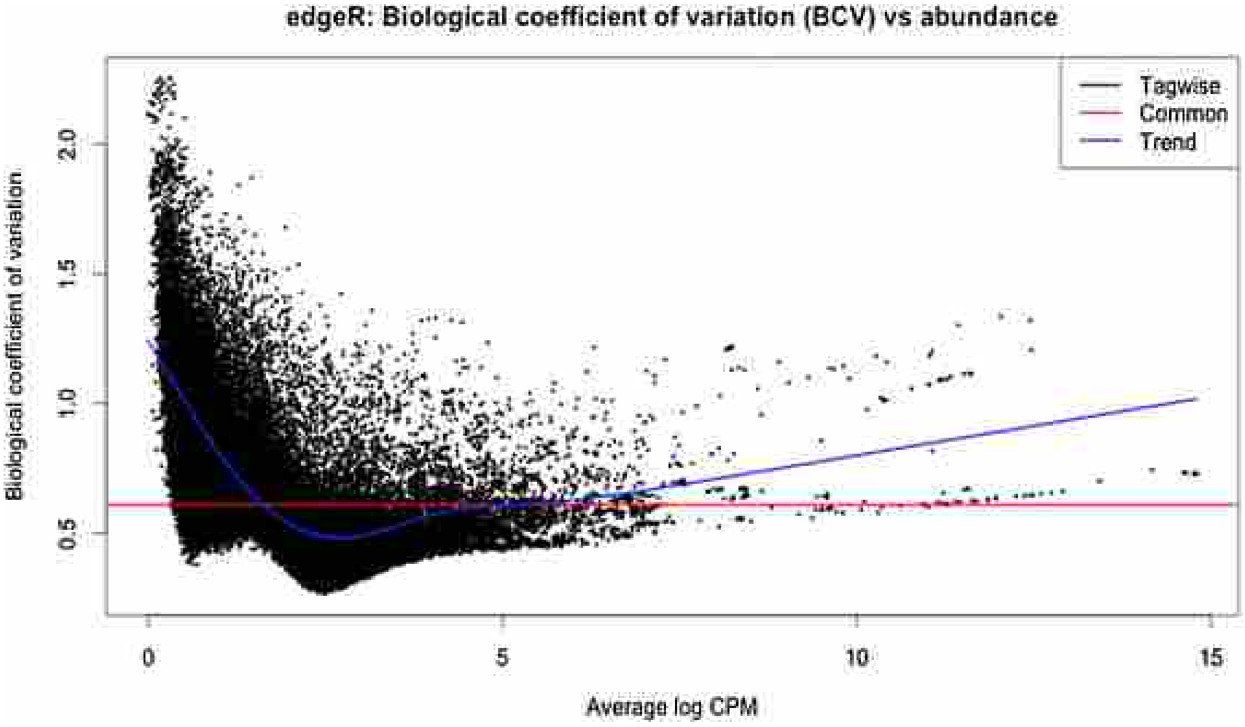


Figure 4.5 Biological Coefficient of Variation Plot
Genewise dispersion plot for twelve libraries is indicated. Estimation of genewise BCV allows observation of changes for genes that are consistent between biological replicates and giving less priority to those with inconsistent results. Generalized linear model is used to determine the evidence of significant difference of counts for a transcript or exon across conditions. The BH method is used in this dataset to control false discovery rate.

The genewise dispersions show a decrease at low average log counts per million. It indicates that at low expression level of genes or transcripts, the variability of gene abundance is high. The analyses were concentrated on genes that are significantly different in expression levels in the optimum salt transcriptome as compared to the low condition transcriptome. Genes whose adjusted p-values were less than 0.05 using the BH method were considered differentially expressed [45, 46]. The BH method known also as FDR (false discovery rate) by Benjamini, Hochberg, and Yekutieli enables the user to control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error rate, so these methods are more powerful than the others. RNA-seq gene expression for *Suaeda fruticosa* is visualized as an MA plot (log ratio versus abundance plot) in Figure 4.6. The red dots highlight transcripts that are differentially expressed among biological replicates and treatments. There are 475 genes that are downregulated and 44 genes are upregulated with a p-value <0.05 and false discovery rate <0.05. The results are consistent with the physiological data of *Suaeda fruticosa* [41] where at 0 mM NaCl treatment, more genes are downregulated in comparison to optimal growth of 300 mM NaCl.

Gene Annotation and Identification of Differentially Expressed Genes

The differentially expressed genes were annotated using Blast2GO software against NCBI non-redundant protein database with a cut-off E-value of 10^{-10}. Enrichment analysis was performed for the biological functions of the identified DEGs. Among 519 differentially expressed unigenes, 44 of them are upregulated upon salt treatment and 475 are downregulated. These genes were identified from BLAST nr, SwissProt and UniProt databases and assigned with Gene Ontology terms in biological process, molecular function and cellular component categories.

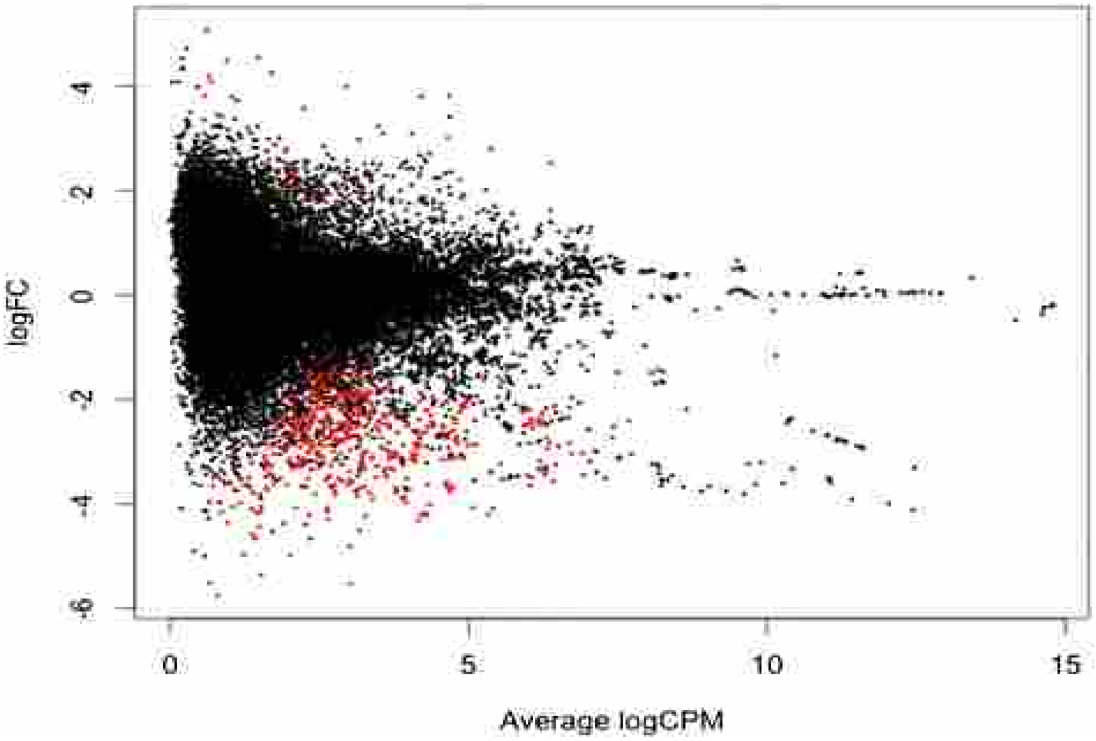


Figure 4.6 Differential Expression Genes Plot.
Plot of LogFCs against average count size, highlighting the differentially expressed genes in red. From the samples and the replicates, there are 475 genes identified to be downregulated and 44 genes that are upregulated with p-value of <0.05 and FDR rate of <0.05.

The top hit species distribution of these differentially expressed genes included grapes (*Vitis vinifera*) with 48 unigenes, orange (*Citrus sinensis*) with 35 unigenes, and *Theobroma cacao* with 29 genes. The closest halophyte is *Populus trichocarpa* with 13 unigenes and *Mesembryanthemum crystallinum* with 9 unigenes. Draft genome projects for both *P. trichocarpa* and *M. crystallinum* are currently ongoing while other halophytes only have partial transcriptome information available in the NCBI database.

From 519 differentially expressed genes, 391 unigenes have significant BLAST hits (75%) and the remaining 25% do not have any significant sequence alignments, which suggests that they might be genes that are novel or have not been reported in any other plant databases. There were 371 annotated sequences (71.5%), and 282 have InterProScan matches from the

European Bioinformatics Institute (EBI) that can be mapped to GO terms and annotations while 162 of them have been assigned to gene ontology IDs. The summary in Figure 4.7 shows how many genes are assigned to at least one GO term and grouped into three main GO categories: biological processes (A), cellular component (B), and molecular function (C). Direct GO terms from Blast2GO were performed by counting annotated sequences in each term and suggesting the top terms. Among these sequences, 177 total unigenes are identified in the molecular function category of GO annotation. The top hits included genes functioning in ion binding (147 unigenes), kinase activity (43 unigenes) and DNA binding (40 unigenes). In the cellular component category, the top hits are genes found to be active in the nucleus (122 unigenes), protein complex component (121 unigenes) and plasma membrane (90 unigenes). For the biological process category, 205 unigenes have been assigned with GO terms and GO IDs. There are 174 unigenes that are important in biosynthetic process, 139 unigenes responding to stress and 124 unigenes involved in cellular nitrogen compound metabolic process. The differentially expressed genes were assigned to KEGG to identify pathways that these genes might be involved in related to salt tolerance. Among the annotated differentially expressed unigenes, the top hit included 6 sequences that are involved in both nitrogen and histidine metabolism. Others function in lysine degradation, glycerolipid metabolism and linoleic acid metabolism. Other pathways are illustrated in Additional File 1.

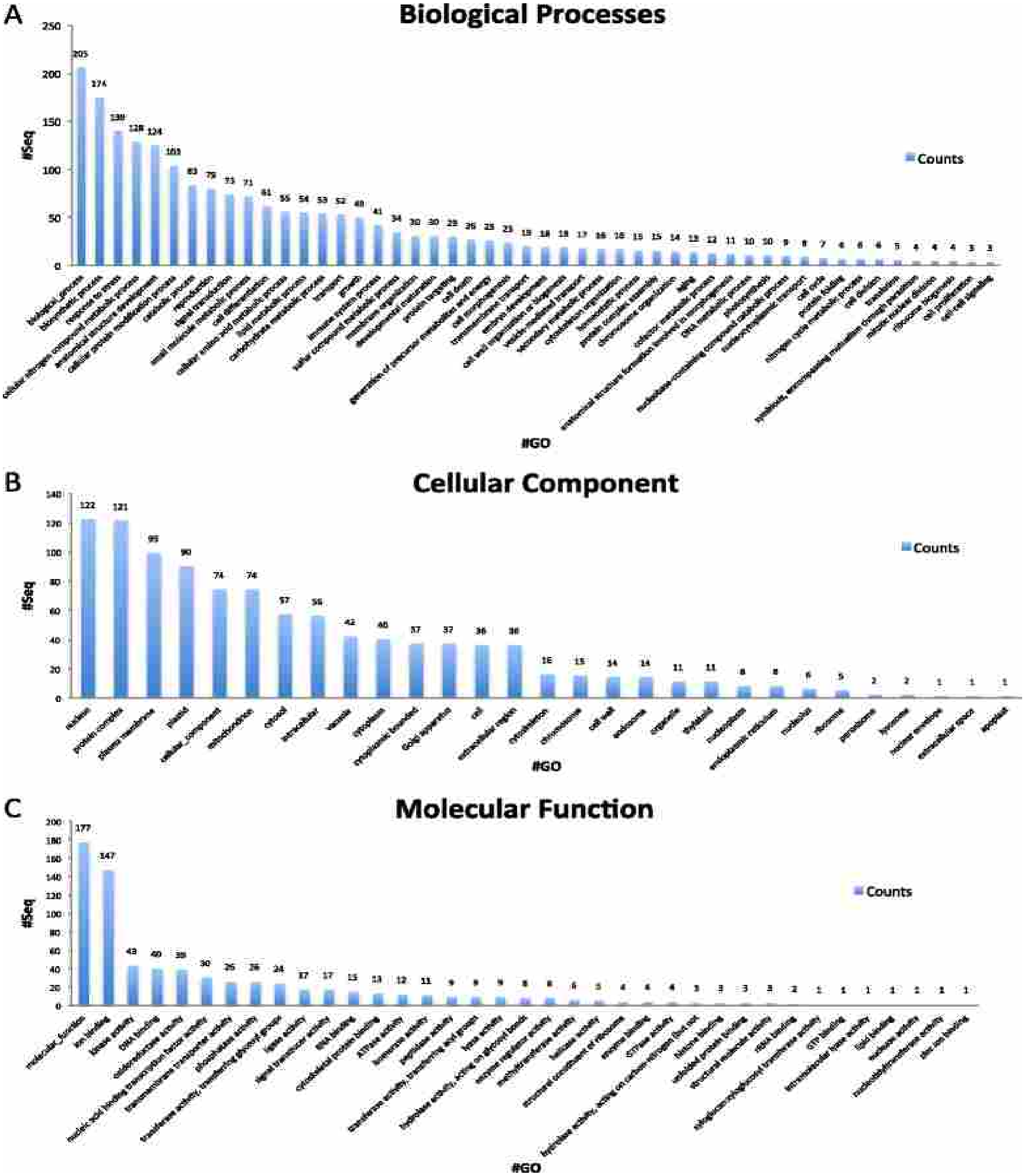


Figure 4.7 Summary of Differentially Expressed ESTs Using BLAST2GO
Differentially expressed transcripts were classified into 3 main GO annotations: Biological Processes (A), Cellular Component (B) and Molecular Functions (C). There are 25 GO terms for biological processes, 37 GO for molecular function and 15 GO for cellular component. A majority assigns the GO from biological process as stress response genes, genes responsible for oxidation-reduction and structure development. A few transcripts reflect oxidoreductase and kinase activity for molecular function. A majority of the transcripts is distributed to the nucleus and plasma membrane.

Relative Gene Expression Validation using qRTPCR Analysis

To validate the results from the transcriptome analysis, we selected seven differentially expressed genes with putative functions related to salt tolerance. Specific primers were designed and optimized using PCR for the selected DE genes and for alpha tubulin as the endogenous control (Supplementary File 4). We amplified a cDNA library from six samples of 0 mM treated plants and six 300 mM treated samples. Analysis of transcript levels by qRTPCR showed that expression for all seven gene targets selected correspond with the differential expression patterns

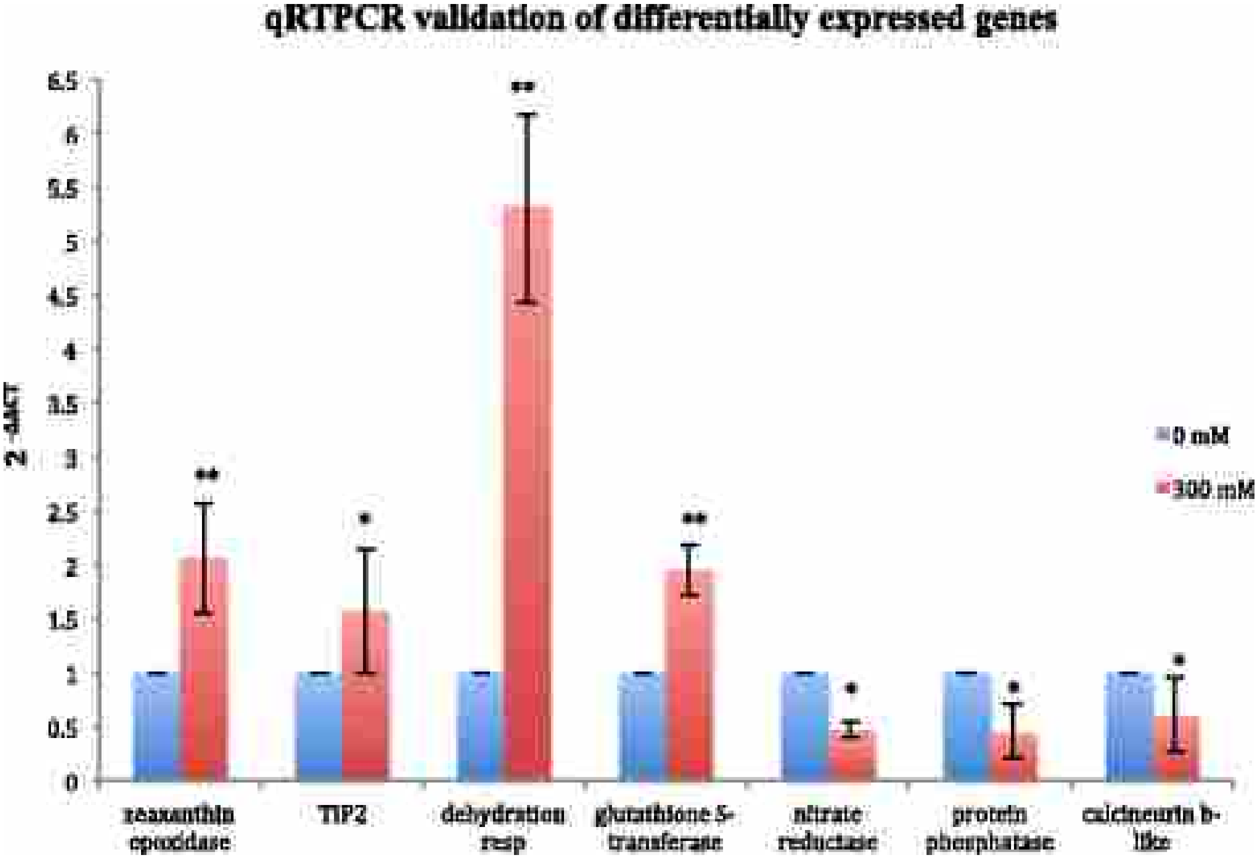


Figure 4.8 QRTPCR Validation of the Transcriptome Data
Each panel shows the qRTPCR results for seven test genes. The annotated putative genes are listed on the x-axis and the mean fold change represented by the 2-ΔΔCT method relative to 0 mM treated samples are shown on the y axis. Error bars depict the standard error of the mean for 3 biological replicates. Significant differences ($p < 0.05$) are denoted with an asterisk and highly significant differences with p-value of <0.005 are represented with double asterisks.

determined from the transcriptome analysis. Four targets (zeaxanthin epoxidase, aquaporin TIP2, dehydration responsive protein and glutathione S-transferase) show upregulation of mRNA expression while the other three targets (nitrate reductase, putative protein phosphatase and calcineurin B-like (CBL 4-1) show downregulation upon 300 mM NaCl treatment compared to the absence of salt treatment (Figure 4.8).

Putative Salt Tolerance-Related Genes

BLAST analysis data identified a large number of differentially expressed genes and we have grouped them in the following categories: 1. Genes responsible for enzymes, transcription factors, hormones, photosynthetic genes, detoxifiers and osmolytes for general metabolism, 2. Genes functioning as transporters for water and ion uptake, 3. Genes involved in regulation such as kinases and phosphatases, and 4. Genes that function to protect the cells against abiotic stress such as late embryogenesis abundant protein, heat shock proteins, osmoprotectants such as dehydrins and osmotins. The number of transcripts reported to be differentially regulated or expressed depends on the conditions being compared. In this study, we are comparing transcript expression between 0 and 300 mM NaCl treatment and their biological replicates. Upregulated genes are those with significant increased expression when treated with salt (300 mM). Those downregulated are annotated sequences with decreased expression with salt treatment. A summary of these sequences, their definitions and putative functions, and references from other halophytes or plants is shown in Table 4.3.

Table 4.3 Summary of Selected Differentially Expressed Genes in *Suaeda fruticosa*

Differentially expressed Protein-coding genes[1]	Putative function or role in salt tolerance	Expression upon 300 mM salt treatment	Plants with orthologous genes	References
General metabolism genes				
Probable WRKY transcription factor 72	Sequence specific DNA binding transcription factor; Activators of ABA signaling Repressors of aleurone cells	Upregulated	*Festuca rubra ssp litoralis* *Glycine soja* *Glycine max* *Oryza sativa* *Porteresia coarctata*	[22,25,69]
WRKY transcription factor 6--like	Influence senescence and pathogen defense---associated PR1 promoter activity; mediates arsenate/phosphate transporter gene expression	Downregulated	*Arabidopsis thaliana* *Festuca rubra ssp litoralis* *Glycine soja* *Glycine max* *Porteresia coarctata*	[22,23,25,26]
WRKY DNA binding protein isoform 2	Transcription factors involved in various regulations; crucial to salinity tolerance	Downregulated	*Festuca rubra ssp litoralis* *Glycine soja* *Glycine max* *Porteresia coarctata*	[22-25,69]
Gibberellin 2--beta---dioxygenase 2 family (GA2OX2)	Gibberellin catabolic, response to jasmonic acid and red light	Downregulated	*Arabidopsis thaliana*	[28,70,71]
40S ribosomal protein S4 (RPS4)	Disease resistance; SRP---dependent cotranslational protein	Upregulated	*Arabidopsis thaliana* *Mesembryanthemum crystallinum*	[72,73]
60S acidic ribosomal protein P2	Elongation step of protein synthesis	Upregulated	*Zea mays*	[30,74]
60S ribosomal protein L18–2---like	Plastidic and nuclear protein synthesis	Upregulated	*Populus euphratica* *Suaeda maritima*	[31,34]
Pre--mRNA processing protein 40C	Co-activator involved in the regulated transcription of nearly all RNA polymerase II---dependent genes	Downregulated	*Arabidopsis thaliana*	[30]
DNA--binding protein escarola---like	Late flowering and leaf development Leaf senescence	Upregulated	*Arabidopsis thaliana*	[30]
MADS--box transcription factor AGL24	Transcription activator that mediates floral transition in response to vernalization promotes inflorescence fate in apical meristem	Downregulated	*Arabidopsis thaliana* *Porteresia coarctata*	[23,30]
DNA binding protein with zinc finger isoform1	Binds DNA; structural regulation	Downregulated	*Glycine max* *Malus zumi* *Arabidopsis thaliana*	[30,69,75]

F--box protein AT1g78280---like transferase	Regulation of transcription; defense response by callose deposition	Downregulated	*Arabidopsis thaliana* *Eutrema salsugineum* *Suaeda maritima*	[30,34,76]
Glutathione--S---transferase tau 1	Glutathione metabolism; and production; promoted a higher level of salt tolerance	Upregulated	*Arabidopsis thaliana* *Glycine max* *Glycine soja* *Nicotiana tabacum* *Populus euphratica* *Reaumuria trigyna* *Salicornia europaea* *Suaeda maritima* *Suaeda fruticosa* *Suaeda salsa*	[2,35,36,69,77,78]
Germin--like protein	Salt---stress regulation marker	Upregulated	*Arabidopsis thaliana* *Hordeum vulgare*	[37,38]
Flowering promoting factor 1--like protein 3	Regulates flowering time	Upregulated	*Arabidopsis thaliana*	[30]
Auxin--induced protein 5NG4---like	Transport of molecules functioning downstream of the auxin response; Root formation	Upregulated	*Arabidopsis thaliana*	[30]
Pathogenesis--related protein	Defense response; Response to water deprivation;	Upregulated	*Arabidopsis thaliana*	[30]
Chitinase	Enhance biotic and abiotic stress tolerance; reduce chitin in the cell wall contributing to salt sensitivity	Upregulated	*Glycine max* *Glycine soja* *Nicotiana tabacum*	[40,69]
Peroxisomal ascorbate peroxidase (APX)	Response to oxidative stress; Regeneration of NAD+; induced by high temperature	Upregulated	*Nicotiana tabacum* *Zea mays*	[30,41]
Plant cadmium resistance 2--like	Reduces cadmium accumulation	Upregulated	*Atriplex halimus*	[42-44]
Chaperone protein DNAJ–16 like	Protein folding; protein partitioning into organelles; signal transduction; directly interacts with HSP70; induced by heat shock and prevents apoptosis	Downregulated	*Arabidopsis thaliana* *Atriplex nummularia* *Suaeda maritima* *Spartina maritima* *Spartina alterniflora*	[34,45,46,79,80]
Ethylene--responsive transcription factor rap2---7 like isoform	Transcriptional activator; GCC box binding; pathogenesis related promoter; Involved in gene expression by stress factors; negatively regulates transition to flowering time	Downregulated	*Arabidopsis thaliana* *Suaeda maritima*	[30,34]

Senescence--associated protein	Induced by abscisic acid; regulated during natural and artificially induced leaf senescence	Downregulated	*Arabidopsis thaliana* *Mesembryanthemum crystallinum*	[72,81]
Stem specific protein TSJT1--like	Stem-specific (active at lower levels in other organs)	Downregulated	*Nicotiana tabacum*	[30]
Protein F3H11–7	Positive regulation of transcription; leaf morphogenesis	Downregulated	*Arabidopsis thaliana*	[30]
Cell wall protein AWA1--like	Cell wall organization and biosynthesis	Downregulated	*Theobroma cacao*	[30]
Callose synthase 7	Callose synthesis at forming cell plate during cytokinesis; transitory component of the cell plate in dividing cells	Downregulated	*Arabidopsis thaliana*	[30]
Calcineurin B-like protein (CBL) 4-1	SOS like-gene; Acts as a calcium sensor involved in regulatory pathway for Na + and K+ homeostasis and salt tolerance; Activates in synergy with CIPK24/SOS2 to activate Na+/H+ antiporter SOS1	Downregulated	*Arabidopsis thaliana* *Eutrema salsugineum*	[47,82,83]
Calmodulin binding isoform 1	Regulates transcriptional activity in response to calcium signals; activates the expression of the V-PPase proton pump in pollen	Downregulated	*Arabidopsis thaliana* *Glycine max* *Glycine soja* *Leymus chinensis*	[47,48,69,84]
Photosystem II protein z (PsbZ)	Controls photosystem II cores with the light--harvesting antenna	Upregulated	*Arabidopsis thaliana* *Malus zumi* *Mesembryanthemum crystallinum*	[30,72,75]
Photosystem D2 protein chloroplastic (psbD)	One of the two reaction center proteins of photosystem II; needed for assembly of a stable PSII complex	Upregulated	*Arabidopsis thaliana*	[30]
Photosystem II CP43 chlorophyll apoprotein (psbC)	Core component of the antenna complex of photosystem II; binds chlorophyll and catalyze the primary PII light-induced processes	Upregulated	*Arabidopsis thaliana*	[30]

Hypothetical Chloroplast RF19 (ycf1)	Unknown; may have a function not related to photosynthesis.	Upregulated	*Arabidopsis thaliana*	[30]
Zeaxanthin epoxidase, Chloroplastic--like isoform X2	Abscisic acid precursor, involved in salt and heavy metal tolerance; required for resistance to osmotic and drought stresses, ABA-dependent stomatal closure, seed development and dormancy, modulation of defense gene expression;	Upregulated	*Arabidopsis thaliana* *Spartina maritima* *Spartina alterniflora*	30,80]
Sufe-like chloroplastic like	Cysteine desulfurization in chloroplast and mitochondria; Fe-S cluster biosynthesis	Upregulated	*Arabidopsis thaliana*	[30]
High Light--induced chloroplastic protein	Possible role in chlorophyll and/or carotenoid binding	Downregulated	*Arabidopsis thaliana*	[30]
CRS2--associated factor chloroplastic like	Required for the group IIB intron splicing in chloroplast; mRNA processing; intron specificity	Downregulated	*Arabidopsis thaliana* *Zea mays*	[30]
Thioredoxin-like protein z chloroplastic like	Apoplast redox regulation; cell division and differentiation; stress responses	Downregulated	*Reaumuria trigyna* *Spartina alterniflora* *Spartina maritima*	[32,80]
Triose phosphate chloroplastic like isoform X2	Exports photoassimilates from chloroplast; transports inorganic phosphate, 3-phosphoglycerate and triose phosphate	Downregulated	*Arabidopsis thaliana*	[30]
Probable chlorophyll b reductase chloroplastic--like	Chlorophyll B degradation	Downregulated	*Arabidopsis thaliana*	[30]
Phosphate chloroplastic like	Hypothetical protein	Downregulated	*Arabidopsis thaliana*	[30]
Ion transporters				
Aquaporin tonoplast intrinsic protein 1	H2O channel; facilitates the transport of water across cell membrane; osmoregulation; hydrogen peroxide transmembrane transport	Upregulated	*Arabidopsis thaliana* *Glycine max* *Glycine soja* *Malus zumi* *Oryza sativa* *Populus euphratica* *Schrenkiella parvula*	[30,31,52,53,69,75,85]

High affinity nitrate transporter 3.1 like	High-affinity nitrate transport and assimilation; repressor of lateral root initiation; wounding response	Upregulated	*Arabidopsis thaliana* *Oryza sativa*	[30]
Nitrate transporter 1.5	Transmembrane nitrate transporter; xylem transport of nitrate from root to shoot; induced response to nitrate	Upregulated	*Arabidopsis thaliana*	[30]
Aluminum--activated malate transporter 10	Malate transporter for aluminum tolerance	Upregulated	*Arabidopsis thaliana*	[30]
Flavonol 4'--sulfotransferase, putative	Auxin transport; catalyze the sulfate conjugation	Upregulated	*Flaveria chlorifolia*	[30]
Bidirectional sugar transporter SWEET3	Mediates low affinity uptake, sugar efflux across the plasma membrane	Upregulated	*Arabidopsis thaliana*	[30]
Glucosyltransferase	Catalyzes the glycosylation of flavonoids from UDP glucose	Upregulated	*Arabidopsis thaliana* *Populus euphratica* *Populus pruinosa*	[30,31]
Seed storage/lipid transfer protein	Bifunctional inhibitor/lipid transfer protein/seed storage 2S albumin	Upregulated	*Arabidopsis thaliana* *Populus euphratica* *Thellungiella halophila*	[30,31,86]
ATPase subunit 1 (chloroplast)	Maintenance of the pH of endomembrane compartments	Upregulated	*Arabidopsis thaliana* *Leymus chinensis* *Thellungiella halophila*	[30]
Sodium HKT1--like	Plant salt tolerance and osmotic stress; involves in Na + recirculation; K+ ion transmembrane transporter	Downregulated	*Arabidopsis thaliana* *Populus trichocarpa* *Reaumuria trigyna* *Salicornia europaea* *Schrenkiella parvula* *Thellungiella halophila* *Thellungiella salsuginea*	[30,32,33,53,86]
Sodium pyruvate chloroplastic cotransporter	Pyruvate transport across chloroplast envelope	Downregulated	*Arabidopsis thaliana*	[30]
Magnesium transporter NIPA2	Magnesium ion/ other divalent cations transmembrane transport	Downregulated	*Arabidopsis thaliana*	[30]
Vacuolar Iron transporter family	Regulation of iron distribution; cellular response to ethylene stimulus; cellular response to nitric oxide; iron ion homeostasis; ion transport	Downregulated	*Arabidopsis thaliana*	[30]

Vacuolar proton ATPase A1--like	Essential component of the vacuolar proton pump; cell expansion; ATP hydrolysis	Downregulated	*Arabidopsis thaliana* *Leymus chinensis*	[30,84]
ATPase ASNA1 homolog	Required for the post---translational delivery of tail anchored proteins to the ER; binds the transmembrane domain of tail-anchored proteins in the cytosol	Downregulated	*Arabidopsis thaliana*	[30]
Glutamyl-tRNA amidotransferase subunit chloroplastic mitochondrial-like	Allows the formation of correctly charged Gln-tRNA; ATP binding; glutaminyl-tRNAGln biosynthesis; mitochondrial translation	Downregulated	*Arabidopsis thaliana*	[30]
Tonoplast dicarboxylate transporter--like protein	Malate transmembrane transport; critical for pH homeostasis; indirectly involved in the uptake of malate and fumarate to the vacuole	Downregulated	*Arabidopsis thaliana*	[30]
Probable Galacturonosyltransferase 12-like	Involved in pectin assembly and/or distribution; cell wall organization	Downregulated	*Arabidopsis thaliana*	[30]
Regulatory molecules				
Cysteine rich receptor like protein kinase	ATP binding; defense responses; disease resistance	Upregulated	*Arabidopsis thaliana*	[30]
Phosphatase 2C family protein	Stress responses; metal ion binding; protein dephosphorylation; Serine/threonine phosphatase activity	Downregulated	*Arabidopsis thaliana* *Ceriops tagal* *Glycine max* *Glycine soja* *Populus trichocarpa* *Spartina maritima* *Spartina alterniflora* *Thellungiella salsuginea*	[30,31,58,69,80,87]
Phosphatase 2C 76 isoform 1	Metal ion binding; Binds 2 magnesium or manganese ions	Downregulated	*Arabidopsis thaliana*	[30]
CDPK related kinase 1	Signal transduction pathways that involve calcium as second messenger; ATP binding; Ca2 + binding; protein autophosphorylation	Downregulated	*Arabidopsis thaliana* *Eutrema salsugineum* *Malus zumi* *Suaeda maritima*	[30,34,75,76]

PERK1 receptor protein kinase	Protein autophosphorylation; response to wounding; ATP binding	Downregulated	*Arabidopsis thaliana*	[30]
Casein kinase I--2---like protein	ATP binding; protein serine/threonine kinase activity	Downregulated	*Arabidopsis thaliana*	[30]
Serine--threonine protein kinase (histidine transporter) HT1	Control stomatal movement; shows a reduced response to ABA or light	Downregulated	*Arabidopsis thaliana* *Oryza sativa*	[30,57]
Phosphotidylinositol 4--kinase gamma 4	Phosphatidylinositol phosphorylation; Response to salt stress; Protein autophosphorylation	Downregulated	*Arabidopsis thaliana*	[30]
Serine threonine protein phosphatase pp1--like	Binds 2 manganese ions per subunit; protein dephosphorylation; serine/threonine phosphatase activity	Downregulated	*Arabidopsis thaliana*	[30]
Serine threonine--protein phosphatase PP2A catalytic subunit	Metal ion binding; serine/threonine phosphatase activity	Downregulated	*Arabidopsis thaliana*	[30]
Late embryogenesis abundant proteins				
Dehydration--responsive RD22---like	Induced by salt stress; stress response	Upregulated	*Malus zumi* *Populus euphratica* *Populus pruinosa*	[88]
HSP20--like chaperones superfamily protein	Associated with stress and other abiotic factors	Downregulated	*Glycine max* *Oryza sativa*	[89,90]

The selected genes are identified and annotated using BLAST nr database using BLAST2GO[1]

General Metabolism Genes

Genes that are involved in transcription, translation and post-translational modification have been seen to play roles in salt tolerance processes. WRKY transcription factors are important regulators for signaling mechanisms that modulate various plant processes. It has been found to interact with protein partners, MAP kinases, calmodulin, histone deacetylases, resistance proteins for autoregulation and transcriptional reprogramming [47]. It has also been suggested to be crucial for salinity tolerance [48]. From the differential expression analysis of the transcriptome, we have found WRKY transcription factor 72 to be significantly upregulated while WRKY transcription factor 6-like and WRKY DNA-binding protein isoform 2 are downregulated. The salt tolerant grass *Festuca rubra ssp litoralis* was found to have a differentially regulated WRKY-type transcription factor in response to salinity [49]. Transient expression studies have also found OsWRKY72 and OsWRKY77 to be activators of ABA signaling but repressors of gibberellic acid signaling in aleurone cells [50]. Moreover, in *Arabidopsis*, AtWRKY6 negatively autoregulates its own promoter to influence senescence and pathogen defense-associated PR1 promoter activity. This targets SIRK, a gene encoding a receptor-like protein kinase that is strongly induced during leaf senescence. The activation of SIRK is dependent on WRKY6 function [51]. These studies suggest that WRKY72 transcription factor is upregulated to respond to ABA signaling, important for stress tolerance while downregulating protein-coding genes involved in senescence for protection and defense. Gibberellic acid (GA) genes, which regulate many aspects of growth and development of plants, are involved in the synthesis of gibberellin hormone. In *Arabidopsis*, reduction of bioactive GA is shown via an increase in gibberellin 2-oxidase 7 (GA2ox7). This leads to accumulation of DELLA proteins, which are transcriptional regulators that repress GA-responsive growth and development, inhibiting plant growth [52]. Downregulation of GA2ox2 is observed at 300 mM salt treatment in *Suaeda fruticosa*. This suggests that the decrease deactivates bioactive GA [53]. GA genes were regulated at 200 mM NaCl in *S. europaea* similar to homologues of gibberellin 3-oxidase and gibberellin 20-oxidase in *P. trichocarpa*. Two DELLA domain GRAS family transcription factors were downregulated in plants treated with 200 mM salt [54].

Both 40S ribosomal protein S4 and 60S ribosomal protein L18-2-like that are upregulated in *S. fruticosa* are part of a group of SRP-dependent co-translational proteins targeting to membranes responsible for translation and protein binding [55]. Ribosomal protein 40S and 60S and RNA binding family protein are also highly upregulated in this transcriptome study. Similar studies were performed in *Poplar euphratica,* which found that ribosomal 60S rRNA, important for plastidic and nuclear protein synthesis, is increased in response to salinity [22]. 60S acidic ribosomal protein P2, known to play an important role in the elongation step of protein synthesis and other RNA-binding family proteins are upregulated in 300 mM NaCl treated *S. fruticosa*. However, the gene encoding pre-mRNA processing protein 40C undergoes downregulation in salt treated plants. This protein has been found to be a coactivator involved in regulated transcription of RNA polymerase II-dependent genes important in transcription and other regulatory mechanisms [55]. Some DNA binding proteins also show concerted regulation upon salt treatment. DNA-binding escarola-like protein responsible for late flowering and leaf development and F-box kelch repeat protein AT1g80440-like are upregulated while MADS-box transcription factor AGL24, an early target of transcriptional repression at floral transitional stage, DNA-binding protein with zinc finger isoform1 and F-box protein AT1g78280-like transferase involved in regulation of transcription are downregulated.

An increase in reactive oxygen species causing damage to cellular components is evident when salinity increases. Genes that are responsible in regulating redox reactions are usually involved in protecting the cell environment during these stresses [56]. Upregulation of glutathione-S-transferase tau 1 (GST) and glutathione transferase were seen to be differentially expressed in *S. fruticosa*. Similarly, glutathione S-transferases were greatly increased upon salt treatment in roots of the halophyte *Salicornia europaea* [57], *Suaeda maritima* and *Reaumuria trigyna* [56, 58]. The *Suaeda salsa* GST gene was introduced into *Arabidopsis* and improved salt tolerance after overexpression in transgenic plants. Glutathione content increased in salt-stressed *Arabidopsis* and promoted a higher level of salt tolerance [59]. The level of glutathione is increased at 0 mM and 900 mM treatment and decreased at the optimal condition of 300 mM NaCl in *S. fruticosa* [10].

A similar trend of higher salt tolerance is seen in tobacco seedlings upon overexpression of GST and these genes have been found to be responsible for increased protection against toxins [60]. Some proteins important for seed production and growth show differential expression in *S. fruticosa*. Germin-like protein, found to be an important plant marker for salt stress regulation and suggested to undergo change when salt-tolerant plants are subjected to salt stress has been found to be significantly upregulated upon salt treatment [61, 62]. An ortholog of flowering promoting factor 1-like protein 3, which promotes flowering in *Arabidopsis*, and auxin-induced protein 5NG4-like gene involved in transport of molecules functioning downstream of the auxin response and responsible for root formation are also upregulated [63]. Some genes encoding proteins involved in protection such as pathogenesis-related protein, chitinase, peroxisomal ascorbate peroxidase (APX), and plant cadmium resistance 2-like are also increased. Plant chitinase plays an important role in plant defense and enhances resistance and tolerance to heat, salt and drought [55]. Overexpression of chitinases in transgenic tobacco has been shown to enhance biotic and abiotic stress tolerance [64]. In tobacco cells, APX functions in the regeneration of NAD+ and is usually induced by high temperature stress and functions against toxic reactive oxygen species [65]. In the halophyte *Atriplex halimus L.*, chloride salinity reduces cadmium accumulation as salinity resistance is found to be closely associated with the gene loci responsible for cadmium extraction [66-68]. Proteins containing chaperone domains and DNAJ-16 like chaperon protein are also decreased upon salt treatment. The DNAJ protein family is included in the group of heat shock proteins functioning as molecular chaperones, and is associated with HSP70 and involved in resisting environmental stresses in *Suaeda maritima* [58].

Specifically DNA-J16 in *Arabidopsis* is encoded by the gene known as Altered Response to Gravity 1 (ARG1), and mediates gravity signal transduction and hypocotyl gravitropism [69, 70]. Other genes that are downregulated include ethylene-responsive transcription factor rap2-7 like isoform, senescence-associated protein, stem specific protein TSJT1-like, root hair protein F3H11-7, cell wall protein AWA1-like isoform X1 for cell wall organization, and callose synthase 7, a major component of pollen tubes and pollen cell walls. Molecular mechanisms of cellular calcium changes have been seen with the downregulation of calcineurin B-like protein (CBL) and calmodulin binding isoform 1 upon salt treatment suggesting their potential role as regulators of salt and drought responses [71]. Calmodulin mediates auxin signaling and responds to stresses in *Arabidopsis* [72]. CBL interacts with CIPK serine-threonine protein kinases and mediates activation of AKT1 in response to low potassium conditions and stomatal movement [73].

Various photosynthetic genes have been found to be differentially upregulated upon salt treatment in *S. fruticosa*. These include genes encoding photosystem II protein z, d2 protein, cp43 chlorophyll apoprotein, chloroplast RF19, zeaxanthin epoxidase, chloroplastic like isoform X2 and sufe-like chloroplastic protein. Significant induction has also been found in the halophyte *Salicornia europaea* in which photosynthetic genes, PSI and PSII pigment binding proteins, b6f complex and ATPase synthase CF1 are upregulated in salt treated plants [57]. Some genes encoding light-induced chloroplastic protein, CRS2-associated factor, thioredoxin-like protein chloroplastic like, triose phosphate chloroplastic-like isoform X2, probable chlorophyll b reductase chloroplastic-like and phosphate chloroplastic-like are downregulated in *S. fruticosa*. While some of these proteins have no definite functions determined yet, chlorophyll b reductase has been found to play a role in maturation and storability of seeds in *Arabidopsis*. *Arabidopsis* plants lacking chlorophyll b show a stay-green phenotype in leaves [74]. This suggests that as chlorophyll b reductase decreases in plants, they tend to prevent chlorophyll degradation.

Ion Transporters (Transporters and Aquaporins)

Homeostasis of the cellular environment involves the maintenance of cellular uptake to control ionic balance. Since a large influx of extracellular Na+ occurs in halophytes, plants require high amounts of K+ (100-200 mM) to lower the amount of Na+ and maintain osmosis [75]. Aquaporin tonoplast intrinsic proteins showed upregulation in salt treated *S. fruticosa*. Aquaporins are membrane proteins that facilitate uptake of soil water and mediate regulation of root hydraulic conductivity. They are also involved in compartmentalization of water and are found in halophytes to play a role in maintaining osmosis and turgor of plant cells [76]. The halophyte *Schrenkiella parvulla* contains high numbers of aquaporins for tolerance to boron toxicity [77]. In *Poplar* species, some aquaporins are decreased to prevent water loss during salt stress [22]. Some other transporters that are upregulated upon salt treatment include high-affinity nitrate transporter 3.1-like and nitrate transporter 1.5 important for nitrate uptake, aluminum-activated malate transporter 10 for increased aluminum tolerance, flavonol 4-sulfotransferase for auxin transport, bidirectional sugar transporters and glucosyltransferase for glucose and other sugar transport, seed storage/lipid transfer protein responsible for metabolism and transport, and ATPase subunit 1. Other halophytes such as *Schrenkiella parvula* and *Thellungiella* showed upregulation of genes encoding for ATPases that are necessary for large influx of ions [77, 78].

Studies have shown vacuolar Na+/H+ antiporter to be important for salt tolerance through Na+ sequestration [79]. However, in *Suaeda fruticosa*, sodium transporter HKT1-like is shown to be downregulated. In *Arabidopsis*, HKT1 knockouts accumulate the highest concentration of Na+ in the shoots suggesting a role in maintenance of Na+ concentration [80]. Some other ion transporters are also downregulated such as sodium pyruvate chloroplastic co-transporter, magnesium transporter NIPA2, vacuolar iron transporter, vacuolar proton ATPase A1-like and ASNA1 (arsenic pump driving ATPase). Glutamyl-tRNA amidotransferase involved in carbon-nitrogen ligase activity, tonoplast dicarboxylate transporter-like protein for malate transmembrane transport and regulation of intracellular pH and galacturonosyltransferase 12-like for glycan and pectin biosynthesis are also decreased with salt treatment.

Regulatory Molecules (Kinases and Phosphatases)

Differentially regulated molecules such as kinases and phosphatase are involved in regulation of proteins involved in osmolyte synthesis and detoxification by oxidants. They are suggested to play a role in ionic and osmotic homeostasis and modulate ion transport for salt tolerance [81]. Cysteine-rich receptor like protein kinase, phosphatase 2C family protein including phosphatase 2C 15-like isoform X1 and purple acid phosphatase 27-like are upregulated at 300 mM NaCl treatment. Protein phosphatase 2C (PP2C) regulates signal transduction pathways. In *Thellungiella*, A-type PP2C phosphatases are generally upregulated in response to abscisic acid [82]. Moreover, there are other kinases that are downregulated in this study such as CDPK-related kinase 1, PERK1 kinases, casein kinase I2-like protein, and serine-threonine protein kinase HT1 and phosphoinositide 4-kinase gamma 4. Serine threonine protein kinase HT1 is important for regulation of stomatal movement in response to carbon dioxide [83] while CDPK-kinase 1 has been shown to play an important role in mediating signal transduction of growth and development [55]. In rice, OsCDPK1 negatively regulates the expression of enzymes for gibberellic acid biosynthesis. This also transduces post-germination of Ca^{2+} signal from sugar starvation and gibberellic acid to prevent drought stress injury [84]. Some phosphatases are also downregulated such as serine-threonine protein phosphatase pp1-like, phosphatase 2C 76 isoform 1 and PP2A catalytic subunit. In *Arabidopsis*, transcription factor MYB20 negatively regulates 2C serine-threonine protein phosphatases to enhance salt tolerance [85].

LEA Genes

Late embryogenesis abundant (LEA) proteins comprise a group of proteins that have crucial roles in cellular dehydration tolerance. They have been associated with tolerance to dehydration caused by freezing, salinity or drying. During stress conditions such as salinity, plant hormone abscisic acid (ABA) is produced to develop tolerance against drought. Some genes are induced to trigger the production of ABA [86].

Overexpression of LEA proteins can improve stress tolerance of transgenic plants. In this transcriptome study, salt treatment causes upregulation of dehydration-responsive RD22-like protein. RD22 expression in *Arabidopsis* is mediated by abscisic acid (ABA). This is also induced by salt stress and dehydration [87] and is expressed during early and middle stages of seed development. Housekeeping gene HSP20 chaperone superfamily is found to be downregulated upon salt treatment. HSP20 family has been associated with the most stress-general expression pattern including salt stress in *Arabidopsis* [88].

CONCLUSIONS

This study provides an overview of the genes present in a non-model plant species and identifies the genes associated with salt tolerance. The assembled transcriptome was used for differential expression studies and gene annotations. We have identified 519 genes that are differentially expressed based on p-value and adjusted false discovery rate of less than 0.05. The same pattern of differential expression for seven of these genes was confirmed by qRT-PCR analysis, which each showed similar levels of up- or down-regulation (Fig. 4.8).

The annotation of genes using next generation sequencing is more readily available through the advancement of technology. Analysis of predicted genes allows assumptions to be made on the complexity of genetic mechanisms for this plant. RNA sequencing generates an enormous amount of data in terms of identifying the transcripts, however the challenges remain with the analysis. One of the major problems is the development of expression metrics that will allow comparisons of different expression levels and also provide identification of differentially expressed genes. We have utilized a combination of approaches to conduct this analysis for the *Suaeda fruticosa* transcriptome. The reference transcriptome assembly was not previously available and this species does not have any close relative plant that can serve as a basis for the expression analysis.

This study reports comprehensive information about the transcriptome of the succulent halophyte *S. fruticosa*. This will provide a basis for further study of the mechanism of salt tolerance, discovery of novel genes involved and comparison of expression profiles with no salt and optimal salt concentration. The de novo transcriptome generated in this study provides a useful source of reference sequence for succulent halophytes.

METHODS

Plant Materials and RNA Isolation

Seeds of *Suaeda fruticosa* obtained from the Institute of Sustainable Halophyte Utilization, University of Karachi, Pakistan were planted and grown at Brigham Young University, Provo, Utah, U.S.A. according to protocol [10]. Plant samples of 100 mg of frozen plant tissue from roots and shoots of low (0 mM NaCl) and optimal (300 mM NaCl) salt conditions were ground in liquid nitrogen to a fine powder. Total RNA was extracted from these tissues using a Trizol-based method or QIAGEN RNeasy Mini kit. The RNA was analyzed for quality and concentration using the Agilent Technologies 2100 Bioanalyzer. High quality total RNA samples should give two distinct peaks and yield an RNA Integrity Number (RIN) value greater than 8.

Illumina Sequencing Platform

The Illumina RNA-Seq library preparation protocol includes poly-A RNA isolation, RNA fragmentation, reverse transcription to cDNA using random primers, adapter ligation, size-selection from a gel and PCR enrichment [34]. The batch of libraries was sequenced at the BYU sequencing center and by Otogenetics (Norcross, GA) using Illumina Hi-seq 2000 sequencer. This includes cDNA libraries of Suaeda 0 mM NaCl-treated shoot and roots in triplicates and cDNA libraries of Suaeda 300 mM NaCl-treated shoot and roots in triplicates. The paired-end library was developed according to the protocol of the Paired-End Sample Preparation kit (Illumina, USA).

Bioinformatics Analysis

Quality Trimming and Digital Normalization

The adapters of raw RNA sequence data were trimmed using Trimmomatic v.0.27. FastX toolkit and Sickle paired-end trimmers were used to determine low quality reads towards the 3' and 5' ends of the reads. The software for digital normalization is available electronically through http://ged.msu.edu/papers/2012-diginorm/. A python script was used to interleave the paired-end reads file http://github.com/ged-lab/khmer/tree/2012-paper-diginorm/sandbox. Khmer software package available at http://github.com/ged-lab/khmer/ was used to perform three-pass normalization steps. Loading sequences needed for khmer software works with screed packages through http://github.com/ged-lab/screed/ (khmer and screed are ©2010 Michigan State University, and are free software available for distribution, modification, or redistribution under the BSD license). The details of quality trimming and digital normalization are available in Additional File 1.

De Novo Assembly and Gene Ontology

The high quality concatenated reads of shoots and roots were assembled using Velvet v. 1.2.10 and Oases v. 0.2.08 with optimized determined *k*-mer size of 45 with an average insert length of 300 bp and minimum contig length of 200. Only assembled transcripts longer than 200 bp were kept. De novo assembly scripts are available in Additional File 2. We ran the clustering methods using CDHIT-EST v.4.5.4-2011-03-07 on the assembly. All Illumina assembled unigenes were searched against nr database in NCBI, Swiss-Prot, UniProt, and Kyoto Encyclopedia of Genes and Genome (KEGG) with the BLASTX algorithm. The E-value cut-off was set to 10^{-10}. Genes were identified according to best hits against known sequences.

Prediction of gene ontology (GO) terms, sequences functions, metabolic pathways in KEGG databases were performed.

Sequence Analysis

The assemblies were transferred into Transdecoder, an open-reading frame predictor software under the Trinity package, which reports candidate coding regions within the transcripts. For each assembly, the number of transcripts, N50 values and the total length of the assemblies are identified. The analysis of the efficiency of assemblies is performed using GMAP and GSNAP v. 2013-11-27. GMAP maps and aligns cDNA sequences originally used for genomic mapping then GSNAP aligns single-end or paired-end reads. It can detect short and long distance splicing using probabilistic models or database of known splice sites.

Differential Expression Analysis

To determine the DEGs (differentially expressed genes) between different treatments of shoots and roots of *Suaeda fruticosa*, gene expression level analysis was performed using the EdgeR package from R [89]. Calculated gene expression can be directly used for comparing the differences in gene counts between treatments and tissue types. Generalized Linear Models were used for data analysis to take account of different salt conditions and tissue types of biological replicates. This determines the evidence of significant difference of counts for a transcript or exon across experimental conditions. The estimation for biological variation is measured. DEGs were identified and subject to further annotation using BLAST2GO.

Validation of Differentially Expressed Genes Through qRTPCR

Several putative annotated genes were selected for validation of differential expression using qRTPCR. These include aquaporin TIP2, protein phosphatase, calcineurin b-like protein (CBL) 4-1, zeaxanthin epoxidase, dehydration responsive protein, glutathione S-transferase and nitrate reductase. We selected alpha tubulin as an endogenous control. The primers for these genes were designed from the *Suaeda fruticosa* transcriptome sequences and optimized for PCR (Supplementary File 4).

For each qRTPCR reaction, 1 ug of RNA of 0 mM and 300 mM NaCl treated samples were reverse transcribed into cDNA using oligo(dT) primers, and the cDNA libraries produced were used for qRTPCR using the method of Haddad et al [90]. To assess validation for each gene, qRTPCR data were analyzed based on ΔΔCT and $2^{-\Delta\Delta CT}$ method [91]. The ΔCT value of each gene was calculated by subtracting the CT value of the endogenous control from the CT value of the target gene. Each gene's mean ΔΔCT value,
$2^{-\Delta\Delta CT}$ and standard error of the mean were calculated using the data analysis package in Microsoft Excel. Data were plotted as mean fold change ($2^{-\Delta\Delta CT}$). Significant differences ($p<0.05$) were determined using a one-tailed two sample t-test assuming equal variances for comparison of the fold change values between groups using GraphPad software.

AVAILABILITY OF SUPPLEMENTARY FILES

Supplementary Files and Figures are available here in our published paper.

http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1553-x

LIST OF ABBREVIATIONS USED

R000-roots treated without salt; R300-roots treated with 300 mM NaCl; S000- shoots treated without salt; S300- shoots treated with 300 mM salt; ORF- open reading frame; nr- non-redundant database; GO- gene ontology; KEGG- Kyoto Encyclopedia of Genes and Genomes; BLAST- Basic Local Alignment Search Tool; GSNAP- Genomic Short Read Nucleotide Alignment Program; MDS- multidimensional scaling plot; BCV- biological coefficient of variation; FDR- false discovery rate; BH- Benjamini, Hochberg; GA- gibberellic acid; GST- glutathione S-transferase tau 1; APX- peroxisomal ascorbate peroxidase; PP2C- protein phosphate 2C; LEA- late embryogenesis abundant proteins; ABA- abscisic acid; DEG- differentially expressed genes.

COMPETING INTERESTS

The authors declare no competing interests.

AUTHOR'S CONTRIBUTION

JDA designed and carried out experiments, including RNA isolation, cDNA library preparation, RNA-seq analysis, gene annotation, differential expression and data interpretation, and wrote the manuscript. MC provided guidance on computational analysis of the sequence data and interpretation. BG provided information and seeds for growing the plants and helped with RNA isolation and cDNA preparation. MAK provided expertise in halophyte growth and analysis, obtained funding, and assisted in preparation of the manuscript. BLN provided guidance on the project, obtained funding, and helped format and edit the manuscript.

ACKNOWLEDGEMENTS

Special thanks to Justin Page and Dr. Joshua Udall for guiding transcriptome analysis and experimental design and for Nielsen Lab members Collin Hansen, Daniel Ricks and Kevin Prier for helping with the wet lab preparations. This research has been supported by a grant from the Pakistan-U.S. Science and Technology Cooperation Program, U.S. Dept. of State and Higher Education Commission of Pakistan, and by the Department of Microbiology and Molecular Biology at Brigham Young University.

REFERENCES

1. Mutz K, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F: Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology* 2013, 24(1):22-30.

2. Nowrousian M: Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryotic cell* 2010, 9(9):1300-1310.

3. Qin QP, Zhang LL, Li NY, Cui YY, Xu K: Optimizing of cDNA preparation for next generation sequencing. *Yi chuan* 2010, 32(9):974-977.

4. Salgotra RK, Gupta BB, Stewart CN, Jr.: From genomics to functional markers in the era of next-generation sequencing. *Biotechnology letters* 2013.

5. Jimenez-Gomez JM: Next generation quantitative genetics in plants. *Frontiers in plant science* 2011, 2:77.

6. McGettigan PA: Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology* 2013, 17(1):4-11.

7. Dang Z-h, Zheng L-l, Wang J, Gao Z, Wu S-b, Qi Z, Wang Y-c: Transcriptomic profiling of the salt-stress response in the wild recretohalophyte Reaumuria trigyna. *BMC Genomics* 2013, 14:29.

8. Kamle S, Ali S: Genetically modified crops: detection strategies and biosafety issues. *Gene* 2013, 522(2):123-132.

9. Kujur A, Saxena MS, Bajaj D, Laxmi, Parida SK: Integrated genomics and molecular breeding approaches for dissecting the complex quantitative traits in crop plants. *Journal of biosciences* 2013, 38(5):971-987.

10. Hameed A, Hussain T, Gulzar S, Aziz I, Gul B, Khan M: Salt tolerance of a cash crop halophyte *Suaeda fruticosa*: biochemical responses to salt and exogenous chemical treatments. *Acta Physiol Plant* 2012, 34(6):2231-2340.

11. Weber DJ: Adaptive mechanisms of halophytes in desert regions. *Salinity and Water Stress* 2008, 44:179-185.

12. Bennani-Kabachi N, El-Bouayadi F, Kehel L, Fdhil H, Marquie G: Effect of *Suaeda fruticosa* aqueous extract in the hypercholesterolaemic and insulin-resistant sand rat. *Therapie* 1999, 54:725-730.

13. Towhidi A, Saberifar T, Dirandeh E: Nutritive value of some herbage for dromedary camels in the central arid zone of Iran. *Tropical Animal Health Pro* 2011, 43:617-622.

14. Bareen F, Tahira SA: Metal accumulation potential of wild plants in tannery effluent contaminated soil of Kasur, Pakistan: field trials for toxic metal cleanup using *Suaeda fruticosa. J Hazard Mater* 2011, 186:443-450.

15. Khan MA, Ansari R, Ali H, Gul B, Nielsen BL: *Panicum turgidum*: a sustainable feed alternative for cattle in saline areas. *Agric Eco Env* 2009, 129:542-546.

16. Glenn E, Brown JJ, O'Leary JW: Irrigating Crops with Seawater. In: *Scientific American.* 1998.

17. Gong Q, Li P, Ma S, Indu Rupassara S, Bohnert HJ: Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* 2005, 44(5):826-839.

18. Inan G, Zhang Q, Li P, Wang Z, Cao Z, Zhang H, Zhang C, Quist TM, Goodwin SM, Zhu J *et al*: Salt cress. A halophyte and cryophyte Arabidopsis relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant Physiol* 2004, 135(3):1718-1737.

19. Vera-Estrella R, Barkla BJ, Garcia-Ramirez L, Pantoja O: Salt stress in *Thellungiella halophila* activates Na+ transport mechanisms required for salinity tolerance. *Plant Physiology* 2005, 139(3):1507-1517.

20. Ding M, Hou P, Shen X, Wang M, Deng S, Sun J, Xiao F, Wang R, Zhou X, Lu C *et al*: Salt-induced expression of genes related to Na+/K+ and ROS homeostasis in leaves of salt-resistant and salt-sensitive poplar species. *Plant Molecular Biology* 2010, 73(3):251-269.

21. Qiu Q, Ma T, Hu Q, Liu B, Wu Y, Zhou H, Wang Q, Wang J, Liu J: Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. *Tree Physiology* 2011, 31(4):452-461.

22. Brinker M, Brosche M, Vinocur B, Abo-Ogiala A, Fayyaz P, Janz D, Ottow EA, Cullmann AD, Saborowski J, Kangasjarvi J *et al*: Linking the Salt Transcriptome with Physiological Responses of a Salt-Resistant *Populus* Species as a Strategy to Identify Genes Important for Stress Acclimation. *Plant Physiology* 2010, 154(4):1697-1709.

23. Flowers TJ, Colmer TD: Salinity tolerance in halophytes. *New Phytologist* 2008, 179:945 - 963.

24. Ayarpadikannan S, Chung E, Cho C-W, So H-A, Kim S-O, Jeon J-M, Kwak M-H, Lee S-W, Lee J-H: Exploration for the salt stress tolerance genes from a salt-treated halophyte, Suaeda asparagoides. *Plant Cell Rep* 2012, 31(1):35-48.

25. Brown CT, Howe A, Zhang Q, Pyrkosz A, Brom T: A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv: 12034802* 2012.

26. Garber M, Grabherr MG, Guttman M, Trapnell C: Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* 2011, 8(6):469-477.

27. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* 2011, 29(7):644-652.

28. Schulz MH, Zerbino DR, Vingron M, Birney E: Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012, 28(8):1086-1092.

29. Clarke K, Yang Y, Marsh R, Xie L, Zhang KK: Comparative analysis of de novo transcriptome assembly. *Sci China Life Sci* 2013, 56(2):156-162.

30. Zerbino DR: Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2010, Chapter 11:Unit 11 15.

31. Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 2008, 18:821-829.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *Journal of Molecular Biology* 1990, 215(3):403-410.

33. Fu L, Niu B, Zhu Z, Wu S, Li W: CDHIT-accelerated for clustering the next generation sequencing data. *Bioinformatics* 2012, 28:3150-3152.

34. Illumina: Ilumina, Inc.; 2009.

35. Bullard J, Purdom E, Hansen K, Dudoit S: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* 2010, 11(1):94.

36. Huang X, Madan A: CAP3: A DNA sequence assembly program. *Genome Res* 1999, 9:868 - 877.

37. Beltran JM, Manzur CL: Overview of salinity problems in the world and FAO strategies to address the problem. *Proceedings of the International Salinity Forum* 2005:311-313.

38. Labidi N, Ammari M, Mssedi D, Benzerti M, Snoussi S, Abdelly C: Salt excretion in *Suaeda fruticosa*. *Acta biologica Hungarica* 2010, 61:299-312.

39. Chaudhri II, Shah BH, Naqvi N, Mallick IA: Investigations on the role of Suaeda fruticosa Forsk in the reclamation of saline and alkaline soils in West Pakistan plains. *Plant and Soil* 1964, 21(1):1-7.

40. Anil VS, Rahjkumar P, Kumar P, Mathew MK: A plant Ca2+ pump, ACA2, relieves salt hypersensitivity in yeast. Modulation of cytosolic calcium signature and activation of adaptive Na+ homeostasis. *J Biol Chem* 2008, 283:3497-3506.

41. Khan MA, Ungar IA, Showalter AM: The effect of salinity on the growth, water status, and ion content of a leaf succulent perennial halophyte, *Suaeda fruticosa (L.) Forssk*. *Journal of Arid Environments* 2000, 45:73-84.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: Gene Ontology: tool for the unification of biology. *Nat Genet* 2000, 25(1):25-29.

43. Wu TD, Nacu S: Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010, 26(7):873-881.

44. Page JT: BamBam: Tools for genomic analysis. In. sourceforge; 2013.

45. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practival and powerful approach to multiple testing. *J Royal Statistical Soc Series* 1995, 57:289-300.

46. Benjamini Y, Yekutieli D: The control of the false discovery rate in multiple testing under dependency. *Ann Statistics* 2001, 29:1165-1188.

47. Rushton PJ, Somssich IE, Ringler P, Shen QJ: WRKY transcription factors. *Trends in Plant Science* 2010, 15(5):247-258.

48. Garg R, Verma M, Agrawal S, Shankar R, Majee M, Jain M: Deep Transcriptome Sequencing of Wild Halophyte Rice, Porteresia coarctata, Provides Novel Insights into the Salinity and Submergence Tolerance Factors. *DNA Research* 2014, 21(1):69-84.

49. Diédhiou CJ, Popova OV, Golldack D: Comparison of salt-responsive gene regulation in rice and in the salt-tolerant *Festuca rubra ssp. litoralis*. *Plant Signaling & Behavior* 2009, 4(6):533-535.

50. Xie Z, Zhang Z-L, Zou X, Huang J, Ruas P, Thompson D, Shen QJ: Annotations and Functional Analyses of the Rice WRKY Gene Superfamily Reveal Positive and Negative Regulators of Abscisic Acid Signaling in Aleurone Cells. *Plant Physiology* 2005, 137(1):176-189.

51. Robatzek S, Somssich IE: Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes & Development* 2002, 16(9):1139-1149.

52. Achard P, Gong F, Cheminant S, Alioua M, Hedden P, Genschik P: The Cold-Inducible CBF1 Factor–Dependent Signaling Pathway Modulates the Accumulation of the Growth-Repressing DELLA Proteins via Its Effect on Gibberellin Metabolism. *The Plant Cell Online* 2008, 20(8):2117-2129.

53. Alvey L, Harberd NP: DELLA proteins: integrators of multiple plant growth regulatory inputs? *Physiologia Plantarum* 2005, 123(2):153-160.

54. Ma J, Zhang M, Xiao X, You J, Wang J, Wang T, Yao Y, Tian C: Global Transcriptome Profiling of *Salicornia europaea* L. Shoots under NaCl Treatment. *PLoS One* 2013, 8(6).

55. Magrane M, consortium U: UniProt Knowledgebase: a hub of integrated protein data. In: *Database, 2011: bar009.* 2011.

56. Dang Z, Zheng L, Wang J, Gao Z, Wu S, Qi Z, Wang Y: Transcriptomic profiling of the salt-stress response in the wild recretohalophyte *Reaumuria trigyna*. *BMC Genomics* 2013, 14(1):29.

57. Fan P, Nie L, Jiang P, Feng J, Lu S, Chen X, Bao H, Guo J, Tai F, Wang J *et al*: Transcriptome analysis of *Salicornia europaea* under saline conditions revealed the adaptive primary metabolic pathways as early events to facilitate salt adaptation. *PLoS One* 2013, 8(11).

58. Sahu BB, Shaw B: Isolation, identification and expression analysis of salt-induced genes in *Suaeda maritima*, a natural halophyte, using PCR-based suppression subractive hybridization. *BMC Plant Biology* 2009, 9(69).

59. Qi YC, Liu WQ, Qiu LY, Zhang SM, Ma L, Zhang H: Overexpression of glutathione S-transferase gene increases salt tolerance of arabidopsis. *Russ J Plant Physiol* 2010, 57(2):233-240.

60. VP R, RH S, ER A, RD A: Overexpression of glutathione S-transferase/glutathione peroxidase enhances the growth of transgenic tobacco seedlings during stress. *Nat Biotechnol* 1997, 15:988.

61. Hurkman WJ, Tanaka CK: Effect of Salt Stress on Germin Gene Expression in Barley Roots. *Plant Physiology* 1996, 110(3):971-977.

62. Lane BG, Dunwell JM, Ray JA, Schmitt MR, Cuming AC: Germin, a protein marker of early plant development, is an oxalate oxidase. *Journal of Biological Chemistry* 1993, 268(17):12239-12242.

63. Busov V, Johannes E, Whetten R, Sederoff R, Spiker S, Lanz-Garcia C, Goldfarb B: An auxin-inducible gene from loblolly pine (Pinus taeda L.) is differentially expressed in mature and juvenile-phase shoots and encodes a putative transmembrane protein. *Planta* 2004, 218(6):916-927.

64. Dana MdlM, Pintor-Toro JA, Cubero B: Transgenic Tobacco Plants Overexpressing Chitinases of Fungal Origin Show Enhanced Resistance to Biotic and Abiotic Stress Agents. *Plant Physiology* 2006, 142(2):722-730.

65. Mullen RT, Lisenbee CS, Miernyk JA, Trelease RN: Peroxisomal Membrane Ascorbate Peroxidase Is Sorted to a Membranous Network That Resembles a Subdomain of the Endoplasmic Reticulum. *The Plant Cell Online* 1999, 11(11):2167-2185.

66. Ondrasek G: The Responses of Salt-Affected Plants to Cadmium. In: *Salt Stress in Plants.* Edited by Ahmad P, Azooz MM, Prasad MNV: Springer New York; 2013: 439-463.

67. Nedjimi B, Daoud Y: Cadmium accumulation in *Atriplex halimus subsp. schweinfurthii* and its influence on growth, proline, root hydraulic conductivity and nutrient uptake. *Flora - Morphology, Distribution, Functional Ecology of Plants* 2009, 204(4):316-324.

68. Lefevre I, Marchal G, Meerts P, Correal E, Lutts S: Chloride salinity reduces cadmium accumulation by the Mediterranean halophyte species *Atriplex halimus L. Environmental and Experimental Botany* 2009, 65(1):142-152.

69. Guan C, Rosen ES, Boonsirichai K, Poff KL, Masson PH: The ARG1-LIKE2 Gene of Arabidopsis Functions in a Gravity Signal Transduction Pathway That Is Genetically Distinct from the PGM Pathway. *Plant Physiology* 2003, 133(1):100-112.

70. Kroczyńska B, Coop NE, Miernyk JA: AtJ6, a unique J-domain protein from *Arabidopsis thaliana*. *Plant Science* 2000, 151(1):19-27.

71. Pardo J, Reddy M, Yang S, Maggio A, Huh G-H, Matsumoto T, Coca M, Paino-D'Urzo M, Koiwa H, Yun D-J *et al*: Stress signaling through Ca2+/calmodulin-dependent protein phosphatase calcineurin mediates salt adaptation in plants. *Proceedings of the National Academy of Sciences* 1998, 95(16):9681-9686.

72. Galon Y, Aloni R, Nachmias D, Snir O, Feldmesser E, Scrase-Field S, Boyce J, Bouché N, Knight M, Fromm H: Calmodulin-binding transcription activator 1 mediates auxin signaling and responds to stresses in *Arabidopsis*. *Planta* 2010, 232(1):165-178.

73. Cheong YH, Kim K-N, Pandey GK, Gupta R, Grant JJ, Luan S: CBL1, a Calcium Sensor That Differentially Regulates Salt, Drought, and Cold Responses in *Arabidopsis*. *The Plant Cell Online* 2003, 15(8):1833-1845.

74. Nakajima S, Ito H, Tanaka R, Tanaka A: Chlorophyll b Reductase Plays an Essential Role in Maturation and Storability of *Arabidopsis* Seeds. *Plant Physiology* 2012, 160(1):261-273.

75. Lokhande V, Suprasanna P: Prospects of Halophytes in Understanding and Managing Abiotic Stress Tolerance. In: *Environmental Adaptations and Stress Tolerance of Plants in the Era of Climate Change.* Edited by Ahmad P, Prasad M. Maharashtra, India: Springer; 2012.

76. Martínez-Ballesta MdC, Bastías E, Carvajal M: Combined effect of boron and salinity on water transport: The role of aquaporins. *Plant Signaling & Behavior* 2008, 3(10):844-845.

77. Jarvis DE, Ryu C, Beilstein MA, Schumaker KS: Distinct Roles for SOS1 in the Convergent Evolution of Salt Tolerance in *Eutrema salsugineum* and *Schrenkiella parvula*. *Molecular Biology and Evolution* 2014, 31(8):2094-2107.

78. Hechenberger M, Schwappach B, Fischer WN, Frommer WB, Jentsch TJ, Steinmeyer K: A Family of Putative Chloride Channels from *Arabidopsis* and Functional Complementation of a Yeast Strain with a CLC Gene Disruption. *Journal of Biological Chemistry* 1996, 271(52):33632-33638.

79. Qiu Q, Guo Y, Dietrich M, Schumaker K, Zhu J: Regulation of SOS1, a plasma membrane Na+/H+ exchanger in Arabidopsis thaliana, by SOS2 and SOS3. *Proc Natl Acad Sci USA* 2002, 99:8436 - 8441.

80. Hill CB, Jha D, Bacic A, Tester M, Roessner U: Characterization of Ion Contents and Metabolic Responses to Salt Stress of Different *Arabidopsis* AtHKT1.1 Genotypes and Their Parental Strains. *Molecular Plant* 2012, 6(2):350-368.

81. Senadheera P, Maathuis FJM: Differentially regulated kinases and phosphatases in roots may contribute to inter-cultivar difference in rice salinity tolerance. *Plant Signaling & Behavior* 2009, 4(12):1163-1165.

82. Lee Y, Giorgi F, Lohse M, Kvederaviciute K, Klages S, Usadel B, Meskiene I, Reinhardt R, Hincha D: Transcriptome sequencing and microarray design for functional genomics in the extremophile *Arabidopsis* relative *Thellungiella salsuginea (Eutrema salsugineum)*. *BMC Genomics* 2013, 14(1):793.

83. Hashimoto M, Negi J, Young J, Israelsson M, Schroeder JI, Iba K: Arabidopsis HT1 kinase controls stomatal movements in response to CO2. *Nat Cell Biol* 2006, 8(4):391-397.

84. Ho S-L, Huang L-F, Lu C-A, He S-L, Wang C-C, Yu S-P, Chen J, Yu S-M: Sugar starvation- and GA-inducible calcium-dependent protein kinase 1 feedback regulates GA biosynthesis and activates a 14-3-3 protein to confer drought tolerance in rice seedlings. *Plant Molecular Biology* 2013, 81(4-5):347-361.

85. Cui MH, Yoo KS, Hyoung S, Nguyen HTK, Kim YY, Kim HJ, Ok SH, Yoo SD, Shin JS: An *Arabidopsis* R2R3-MYB transcription factor, AtMYB20, negatively regulates type 2C serine/threonine protein phosphatases to enhance salt tolerance. *FEBS Letters* 2013, 587(12):1773-1778.

86. Hundertmark M, Hincha DK: LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 2008, 9:118.

87. Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Hosokawa D, Shinozaki K: Role of *Arabidopsis* MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *The Plant Cell Online* 1997, 9(10):1859-1868.

88. Swindell W, Huebner M, Weber A: Transcriptional profiling of *Arabidopsis* heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics* 2007, 8(1):125.

89. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26(1):139-140.

90. Haddad F, Baldwin K: Reverse transcription of the ribonucleic acid: the first step in RT-PCR assay. *Methods in molecular biology* 2010, 630:261-270.

91. Livak K, Schmittgen T: Analysis of relative gene expression data using real-time quantitative PCR and the 2-(delta delta C(T) method. *Methods* 2001, 25:402 - 408.

92. Flowers T, Colmer T: Salinity tolerance in halophytes. *New Phytol* 2008, 179:945 - 963.

93. Zhu J: Plant salt tolerance. *Trends Plant Sci* 2001, 6:66 - 71.

94. Glenn EP, Brown JJ, Blumwald E: Salt Tolerance and Crop Potential of Halophytes. *Critical Reviews in Plant Sciences* 1999, 18:227-255.

95. Jiang Y, Zeng B, Zhao H, Zhang M, Xie S, Lai J: Genome-wide transcription factor gene prediction and their expressional tissue-specificities in maize. *J Integr Plant Biol* 2012, 54(9):616-630.

96. You J, Chan Z: ROS Regulation During Abiotic Stress Responses in Crop Plants. *Frontiers in plant science* 2015, 6:1092.

97. Long Y, Scheres B, Blilou I: The logic of communication: roles for mobile transcription factors in plants. *J Exp Bot* 2015, 66(4):1133-1144.

98. Golldack D, Luking I, Yang O: Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. *Plant Cell Rep* 2011, 30(8):1383-1391.

99. Diray-Arce J, Gul B, Khan MA, Nielsen B: 10 - Halophyte Transcriptomics: Understanding Mechanisms of Salinity Tolerance. In: *Halophytes for Food Security in Dry Lands.* San Diego: Academic Press; 2016: 157-175.

100. Ghanekar R, Srinivasasainagendra V, Page G: Cross-Chip Probe Matching Tool: A Web-Based Tool for Linking Microarray Probes within and across Plant Species. *Int J Plant Genomics* 2008, 7.

101. Hameed A, Hussain T, Gulzar S, Aziz I, Gul B, Khan MA: Salt tolerance of a cash crop halophyte Suaeda fruticosa: biochemical responses to salt and exogenous chemical treatments. *Acta Physiologiae Plantarum* 2012, 34:2331-2340.

102. Diray-Arce J, Clement M, Gul B, Ajmal Khan M, Nielsen BL: Transcriptome Assembly, Profiling and Differential Gene Expression Analysis of the halophyte Suaeda fruticosa Provides Insights into Salt Tolerance. *BMC Genomics* 2015, 16(353).

103. Jin J, Zhang H, Kong L, Gao G, Luo J: PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research* 2014, 42(D1):D1182-D1187.

104. Finkelstein RR, Gibson SI: ABA and sugar interactions regulating development: cross-talk or voices in a crowd? *Curr Opin Plant Biol* 2002, 5(1):26-32.

105. Wang W, Tang W, Ma T, Niu D, Jin JB, Wang H, Lin R: A pair of light signaling factors FHY3 and FAR1 regulates plant immunity by modulating chlorophyll biosynthesis. *J Integr Plant Biol* 2016, 58(1):91-103.

106. Babitha KC, Vemanna RS, Nataraja KN, Udayakumar M: Overexpression of EcbHLH57 Transcription Factor from Eleusine coracana L. in Tobacco Confers Tolerance to Salt, Oxidative and Drought Stress. *PLoS One* 2015, 10(9):e0137098.

107. Toda Y, Yoshida M, Hattori T, Takeda S: RICE SALT SENSITIVE3 binding to bHLH and JAZ factors mediates control of cell wall plasticity in the root apex. *Plant Signal Behav* 2013, 8(11):e26256.

108. Zhou J, Li F, Wang JL, Ma Y, Chong K, Xu YY: Basic helix-loop-helix transcription factor from wild rice (OrbHLH2) improves tolerance to salt- and osmotic stress in Arabidopsis. *J Plant Physiol* 2009, 166(12):1296-1306.

109. Sharma R, Mishra M, Gupta B, Parsania C, Singla-Pareek SL, Pareek A: De Novo Assembly and Characterization of Stress Transcriptome in a Salinity-Tolerant Variety CS52 of Brassica juncea. *PLoS One* 2015, 10(5):e0126783.

110. Zhu Q, Zhang JT, Gao XS, Tong JH, Xiao LT, Li WB, Zhang HX: The Arabidopsis AP2/ERF transcription factor RAP2.6 participates in ABA, salt and osmotic stress responses. *Gene* 2010, 457(1-2):1-12.

111. Park JS, Kim JB, Cho KJ, Cheon CI, Sung MK, Choung MG, Roh KH: Arabidopsis R2R3-MYB transcription factor AtMYB60 functions as a transcriptional repressor of anthocyanin biosynthesis in lettuce (Lactuca sativa). *Plant Cell Rep* 2008, 27(6):985-994.

112. Ganesan G, Sankararamasubramanian HM, Harikrishnan M, Ganpudi A, Parida A: A MYB transcription factor from the grey mangrove is induced by stress and confers NaCl tolerance in tobacco. *J Exp Bot* 2012, 63(12):4549-4561.

113. Yang T, Peng H, Whitaker BD, Conway WS: Characterization of a calcium/calmodulin-regulated SR/CAMTA gene family during tomato fruit development and ripening. *BMC Plant Biol* 2012, 12:19.

114. Yue R, Lu C, Sun T, Peng T, Han X, Qi J, Yan S, Tie S: Identification and expression profiling analysis of calmodulin-binding transcription activator genes in maize (Zea mays L.) under abiotic and biotic stresses. *Frontiers in plant science* 2015, 6:576.

115. Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B *et al*: Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. *Plant Cell* 2003, 15(7):1538-1551.

116. Saha G, Park JI, Jung HJ, Ahmed NU, Kayum MA, Chung MY, Hur Y, Cho YG, Watanabe M, Nou IS: Genome-wide identification and characterization of MADS-box family genes related to organ development and stress resistance in Brassica rapa. *BMC Genomics* 2015, 16:178.

117. Kofuji R, Sumikawa N, Yamasaki M, Kondo K, Ueda K, Ito M, Hasebe M: Evolution and divergence of the MADS-box gene family based on genome-wide expression analyses. *Mol Biol Evol* 2003, 20(12):1963-1977.

118. Yang O, Popova OV, Suthoff U, Luking I, Dietz KJ, Golldack D: The Arabidopsis basic leucine zipper transcription factor AtbZIP24 regulates complex transcriptional networks involved in abiotic stress resistance. *Gene* 2009, 436(1-2):45-55.

119. Hsieh TH, Li CW, Su RC, Cheng CP, Sanjaya, Tsai YC, Chan MT: A tomato bZIP transcription factor, SlAREB, is involved in water deficit and salt stress response. *Planta* 2010, 231(6):1459-1473.

120. Lai LB, Nadeau JA, Lucas J, Lee EK, Nakagawa T, Zhao L, Geisler M, Sack FD: The Arabidopsis R2R3 MYB proteins FOUR LIPS and MYB88 restrict divisions late in the stomatal cell lineage. *Plant Cell* 2005, 17(10):2754-2767.

121. Xie Z, Li D, Wang L, Sack FD, Grotewold E: Role of the stomatal development regulators FLP/MYB88 in abiotic stress responses. *The Plant journal : for cell and molecular biology* 2010, 64(5):731-739.

122. Anwer M, Boikoglu E, Herrero E, Hallstein M, Davis A, Velikkakam JG, Nagy F, Davis S: Natural variation reveals that intracellular distribution of ELF3 protein is associated with function in the circadian clock. *Elife* 2014, 3.

123. Boxall SF, Foster JM, Bohnert HJ, Cushman JC, Nimmo HG, Hartwell J: Conservation and divergence of circadian clock operation in a stress-inducible Crassulacean acid metabolism species reveals clock compensation against stress. *Plant Physiol* 2005, 137(3):969-982.

124. Onai K, Ishiura M: PHYTOCLOCK1 encoding a novel GARP protein essential for the Arabidopsis circadian clock. *Genes Cell* 2005, 10:963-972.

125. Benn G, Wang CQ, Hicks DR, Stein J, Guthrie C, Dehesh K: A key general stress response motif is regulated non-uniformly by CAMTA transcription factors. *The Plant journal : for cell and molecular biology* 2014, 80(1):82-92.

126. Rahman H, Yang J, Xu YP, Munyampundu JP, Cai XZ: Phylogeny of Plant CAMTAs and Role of AtCAMTAs in Nonhost Resistance to Xanthomonas oryzae pv. oryzae. *Frontiers in plant science* 2016, 7:177.

127. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y *et al*: The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 2008, 319(5859):64-69.

128. Correa LGG, Riano-Pachon DM, Schrago CG, dos Santos RV, Mueller-Roeber B, Vincentz M: The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes. *Plos One* 2008, 3(8).

129. Smaczniak C, Immink RG, Angenent GC, Kaufmann K: Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* 2012, 139(17):3081-3098.

130. Feller A, Machemer K, Braun EL, Grotewold E: Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *Plant Journal* 2011, 66(1):94-116.

131. Finn RD, Clements J, Eddy SR: HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011, 39(Web Server issue):W29-37.

132. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al*: STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 2015, 43(D1):D447-D452.

133. Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T: PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *Journal of Computational Biology* 2015, 22(5):377-386.

CHAPTER 5: Identification and Evolutionary Characterization of Salt-Responsive Transcription Factors in the Halophyte *Suaeda fruticosa*

Joann Diray-Arce[1*], Anton Suvorov[2*], Collin Hansen[1], Seth M. Bybee[2], Bilquees Gul[3], M. Ajmal Khan[3] and Brent L. Nielsen[1]

[1]Department of Microbiology and Molecular Biology, Brigham Young University,
Provo, Utah USA 84602
[2]Department of Biology, Brigham Young University,
Provo, Utah USA 84602
[3]Institute of Sustainable Halophyte Utilization, University of Karachi,
Karachi, Pakistan 75270

*Equal contributors

E-mail addresses:

Joann Diray-Arce, joann.diray@gmail.com

Anton Suvorov, antony.suvorov@byu.edu

Seth Bybee, seth.bybee@gmail.com

Collin Hansen, collindh@gmail.com

Bilquees Gul, bilqueesgul@uok.edu.pk

M. Ajmal Khan, majmalk@uok.edu.pk

Corresponding Author: Brent L. Nielsen, brentnielsen@byu.edu

ABSTRACT

Transcription factors are key regulatory elements that affect gene expression in response to environmental stress such as salinity. However, specialized plants known as halophytes have the ability to tolerate these harsh environments. Here, we identify and characterize putative transcription factors (TF) in an obligate halophyte *Suaeda fruticosa* that are involved in salt tolerance using RNA-seq data. Specifically, we have analyzed the expression patterns of TF families, protein-protein interactions and evolutionary trajectories to elucidate their roles in salt tolerance. We have detected the top differentially expressed transcription factor (DE TF) families (MYB, CAMTA, MADS-box and bZIP) that appear to be most responsive to salinity. We also found that the majority of DE genes in the four aforementioned TF families cluster together on TF trees, which suggests common evolutionary trajectories. This research represents the first comprehensive transcription factor study of a succulent halophyte. These findings will also provide a foundation for understanding the function of salt-responsive transcription factors to aid target studies of salt tolerance and regulation in plants.

KEYWORDS

Transcription factors, salt tolerance, TF family tree, halophytes, *Suaeda*, profile Hidden Markov model.

INTRODUCTION

Salinity causes significant losses in agricultural production due to the limited capacity of crops to regulate homeostasis (Flowers and Colmer 2008). Halophytes are specialized plants that are known to tolerate high salt concentrations through complex mechanisms of gene expression and protein pathway adaptation (Zhu 2001). In adverse environments, halophytes utilize a variety of physiological and metabolic responses to regulate stress-responsive genes and synthesize functional proteins through a complex signal transduction network to confer tolerance (Flowers and Colmer 2008). Moreover, functional salt tolerance requires integrated adaptations from cellular systems to the whole plant to satisfy energy needs (Glenn, et al. 1999).

Transcription factors (TFs) are proteins that bind to specific DNA sequences to control the rate of transcription of target genes and are essential regulators for gene expression in response to environmental signals including stress (Jiang, et al. 2012). TFs are necessary for controlling cellular processes including the regulation of intercellular mechanisms, cell cycle, growth and reproduction, and stress responses, making TF characterization extremely valuable (Golldack, et al. 2011; Long, et al. 2015). They can alter expression of genes to enhance tolerance to these harsh environments (You and Chan 2015). Despite the wealth of genomic and transcriptomic information on glycophytes and halophytes, there are still many unknown aspects of plant strategies for survival, tolerance and productivity at specific salt concentrations.

New high-throughput technologies allow for the generation of data that address questions of temporal and spatial responses to a variety of stresses and enables more structured gene expression prediction and plant mechanism characterization (Diray-Arce, et al. 2016). Transcriptomic studies have been used to analyze stress-related conditions in crops; however, meta-analysis research on specialized plants including halophytes is very limited (Ghanekar, et al. 2008). Although there have been studies of differentially expressed genes in relation to salt tolerance, studies on plant signaling components and key regulators of salt responses are lacking. Therefore, integration and identification of TFs in adaptive signaling networks are key factors for understanding the adaptations of plants to environmental stress (Golldack, et al. 2011).

Suaeda fruticosa Forssk, 1775 is a perennial leaf succulent halophyte that sequesters NaCl into its vacuoles. Optimal growth of this species occurs at 300 mM NaCl, where plants increase the concentration of leaf Na^+ and Ca^{2+}, creating conditions for enhanced water absorption, while other physiological parameters function normally. Sodium ion buildup begins rapidly at 600 mM NaCl, increasing in ion toxicity leading to a compromised antioxidant system and substantial growth reduction (Hameed, et al. 2012). We utilized RNA-sequencing to assemble the transcriptome and identify differentially expressed genes for this obligate halophyte (Diray-Arce, et al. 2015). In the present study the *S. fruticosa* transcriptome data were analyzed to extract TFs, identify family groups and characterize gene expression patterns in shoots and roots under long-term salinity of low (0 mM NaCl) and optimum (300 mM NaCl) treatment. Hidden Markov model-based domain searches and BLAST-based protein homology searches were used to predict TFs. We reconstructed transcription factor family trees found in PlantTFDBv3.0 to determine the evolutionary relationship of differentially expressed TFs versus non-differentially expressed TFs in *S. fruticosa* and its relationship to TFs of other plant species. We focused on the TF families with highest numbers of differentially expressed genes (MYB, CAMTA, MADS box and bZIP) to determine their characteristics and evolutionary relationships.

RESULTS AND DISCUSSION

Molecular Characterization of Abundant Transcription Factor Families

Transcription factors in different halophytes activate genes involved in cell maintenance, modifications and stress response (Diray-Arce, et al. 2016). To elucidate the roles of and identify to which family each potential *S. fruticosa* TF belongs to, we utilized HMM-based TF domain identification and protein homology search.

Table 5.1 Summary of Transcription Factors Family

TF family	Total	Percentage (%)	TF family	Total	Percentage (%)
FAR1	177	8.18	GRAS	25	1.16
bHLH	142	6.56	HSF	25	1.16
MYB	134	6.19	SBP	24	1.11
RAV	117	5.41	Dof	20	0.92
ARF	86	3.97	LBD	19	0.88
AP2	84	3.88	GRF	16	0.74
ERF	80	3.7	TCP	15	0.69
B3	79	3.65	NF-YB	14	0.65
HB-other	79	3.65	S1Fa-like	14	0.65
ARR-B	76	3.51	NF-YA	12	0.55
bZIP	71	3.28	CPP	11	0.51
NAC	70	3.23	WOX	10	0.46
MIKC	63	2.91	ZF-HD	9	0.42
C3H	57	2.63	NF-YC	8	0.37
M-type	57	2.63	SAP	8	0.37
WRKY	52	2.4	YABBY	8	0.37
C2H2	50	2.31	SRS	7	0.32
G2-like	50	2.31	NF-X1	5	0.23
CO-like	49	2.26	BBR-BPC	4	0.18
HD-ZIP	46	2.13	EIL	4	0.18
GATA	45	2.08	GeBP	4	0.18
CAMTA	44	2.03	LSD	4	0.18
HB-PHD	41	1.89	VOZ	4	0.18
Trihelix	31	1.43	E2F_DP	3	0.14
BES1	26	1.2	NZZ_SPL	3	0.14
Nin-like	26	1.2	STAT	2	0.09
TALE	26	1.2	Whirly	2	0.09
DBB	25	1.16	HRT-like	1	0.05
			Total	2164	100

The assignment of transcription factors per family from PlantTFDBv.3.0 are summarized. This includes the percentage of distribution among the total TF families.

Open reading frame (ORF) annotation of the transcriptome yielded 47,500 protein sequences, that were searched against 57 families (MYB and MYB-related combined) from

PlantTFDBv3.0 containing 129,288 TFs from 83 species of green plants that have been comprehensively annotated with their functional domains, 3D structures, and gene ontology from various databases. In total, our analysis resulted in the identification of 3,110 TFs across the families. The TF assignments are summarized together with the percentage of TF family distribution (Table 5.1).

The results show that the most abundant TF family belongs to FAR1 with 177 identified TFs (8.18%). TF family bHLH is the next highest with 142 members (6.56%), followed by MYB with 134 TF (6.19%) and RAV as the fourth most abundant with 117 TF (5.41%). The smallest family belongs to HRT-like with only one hit. No TFs from the LFY gene family were found. These abundant TFs are likely involved in other functional and structural mechanisms in the plant in addition to salinity stress responses.

Although the FAR1 family has the highest number of identified TFs in *Suaeda*, none are differentially expressed between salt treatments. This suggests that the FAR1 TF family might exhibit another function besides long-term salinity stress regulation. For instance, *Arabidopsis* FAR1 TFs have been reported to bind to promoters of abscisic acid (ABA) genes to activate expression. In particular, under salt and osmotic stress, FAR1 has been shown to trigger the accumulation of ABA (Finkelstein and Gibson 2002). When FAR1 genes lose their functionality (e.g. deletion), sensitivity to ABA-mediated inhibition of seed germination is reduced. Also, FAR1 member fhy3 and far1 mutants exhibit wider stomata, lose water faster, and are more sensitive to drought (Wang, et al. 2016).

The bHLH family is the second highest in abundance with two DE bHLH TFs between long-term no salt and optimum salt treatment. BHLH TFs are involved in salt stress tolerance and developmental processes in tobacco (Babitha, et al. 2015) and rice (Toda, et al. 2013).

Overexpression of some bHLH genes conferred increased tolerance to salt and osmotic stress in *Arabidopsis*. This TF family has been observed to positively regulate salt-stress signals independent of ABA, and have been targets to improve salt tolerance in crops (Zhou, et al. 2009). However, there are limited halophyte studies focusing on the involvement of bHLH TFs in salt, drought and salinity stress (Garg, et al. 2014; Sharma, et al. 2015). RAV is the fourth most abundant TF family identified in this study with two DE genes. The RAV family has been found to modulate drought and salt-stress responses in *Arabidopsis* and is involved in ethylene and brassinosteroid responses (Zhu, et al. 2010).

Identification and Annotation of Differentially Expressed Transcription Factor Genes

We have focused on salt-responsive transcription factors that are differentially expressed (DE) between long-term contrasting laboratory conditions (no salt versus optimum salt concentration). We performed differential expression analysis of the *S. fruticosa* transcriptome using EdgeR (Robinson, et al. 2010). The method compares significant transcript expression levels between specific treatments following a negative binomial model using the Benjamini-Hochberg method for multiple testing correction at a false discovery rate cutoff of 0.05 (Benjamini and Hochberg 1995). We identified 49 DE TFs using a pHMM search against TF family databases from the PlantTFDBv.3.0. The summary of DE TFs among the families highlights that the highest DE TF belongs to the MYB superfamily (MYB and MYB-related) with 8 TF members, CAMTA with 5, MIKC and M-type (both MADS box family) with 4 TFs. bZIP, ARR-B and G2-like all have 3 TF members (Figure 5.1).

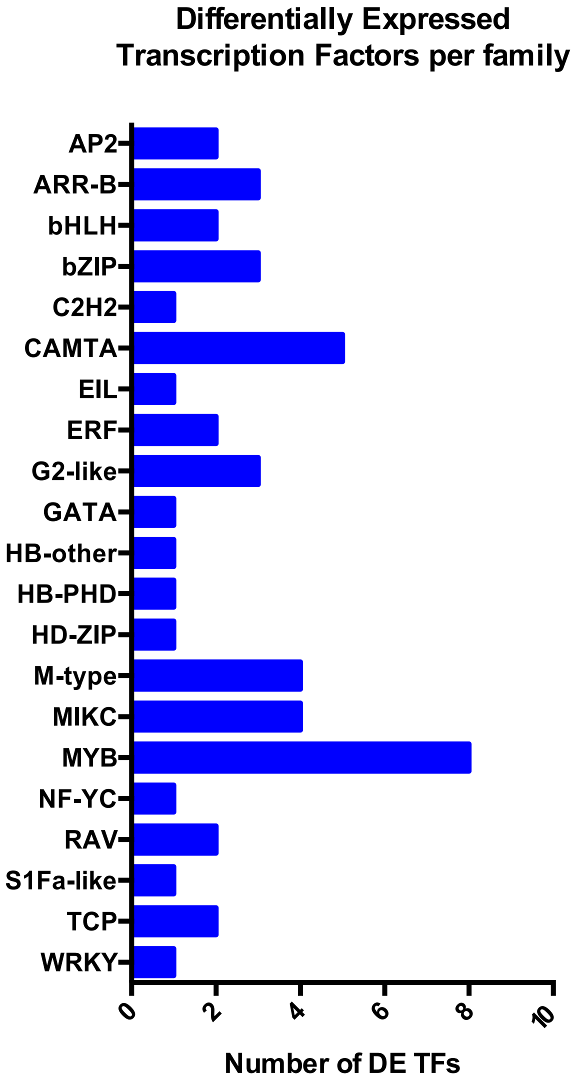


Figure 5.1 Differentially Expressed Transcription Factors Summary
This figure shows the number of differentially expressed transcription factors (DE TF) identified.

We chose the top 4 DE TF families (MYB, CAMTA, MADS-box and bZIP) for expression profiling, phylogenetic tree construction and gene ontology annotation (Figure 5.1). The MYB superfamily contains the highest number of DE TFs between treatments and it is the third most abundant TF family (Figure 5.1 and Table 5.1) found in *S. fruticosa*. Among genome-wide identification and expression analyses related to plant abiotic stress, MYB is one of the most studied TF families in halophytes (Abe, et al. 2003; Garg, et al. 2014). MYB plays diverse physiological and developmental roles that are either induced or repressed under different stress conditions (Golldack, et al. 2011). MYB TFs operate through ABA-dependent or independent pathways. In *Arabidopsis*, MYB2 is induced by salt and drought stress. Rice OsMYB2 encodes a stress-responsive MYB that plays a regulatory role in salt, cold and dehydration (Yang, et al. 2012a). In the halophyte *Avicennia marina* the AmMYB1 gene confers increased salt tolerance with reduced chlorosis and other salt stress symptoms when introduced to tobacco plants (Ganesan, et al. 2012). These findings suggest that the MYB TF family in *S. fruticosa* is the most likely key transcription regulator for salt tolerance regulation.

We identified five calmodulin-binding transcription activators (CAMTA) that are differentially expressed under different salt treatments. At 300 mM NaCl antioxidant enzymes trigger a stress response through the activation of $H_2O_2^-$ mediated Ca^{2+} uptake for Na^+ homeostasis in cells and tissues (Hameed, et al. 2012). Calcium, responsible for the signaling network of growth and development of the plant, are accumulated in the cytosol as Na^+ increases (Anil, et al. 2008). Calmodulin, a major calcium ion sensor, can bind to certain TFs as part of stress response mechanisms. CAMTA TFs are signal proteins that respond to hormonal stimuli such as auxin, ethylene, ABA, salicylic acid and other environmental stresses (Yang, et al. 2012b). *Arabidopsis* AtCAMTA1 is involved in regulation of a broad spectrum of membrane integrity response genes through ABA response to drought stress. Maize CAMTA genes are important regulators of tolerance to environmental stresses (Yue, et al. 2015). Based on these observations, these suggest that CAMTA TFs are also involved in resistance to elevated salinity in *S. fruticosa.*

Four MADS-box DE genes were identified in *Suaeda* upon salt treatment. MIKC and M-type show similar gene hits since both belong to the same MADS-box TF family. MIKC type contains a keratin-like coiled-coil (K) domain while M-type lacks this domain. MADS-box family genes are involved in fruit development, seed pigmentation, floral organ identity determination, and stress response in several species (Parenicova, et al. 2003). In *Brassica rapa*, several MADS-box family TFs were shown to be induced by cold, drought and salt stresses (Saha, et al. 2015). In rice, three genes (OsMADS2, 30 and 55) showed more than 2-fold downregulation in response to dehydration and salt stress (Arora, et al. 2007).

Three DE bZIP TFs were identified upon salt treatment. Group F bZIP family from Arabidopsis and its relative halophyte species was identified to be a key regulator of salt stress adaptation (Yang, et al. 2009). Arabidopsis AREB1, AREB2 and ABF3 are also important genes for signaling under drought stress. Group A bZIP in rice and tomato confers increased tolerance to water deficit and salt stress (Hsieh, et al. 2010). Overall, MADS-box and bZIP families are also potential candidates for salt regulation in *S. fruticosa.*

To validate the results from the transcriptome analysis, we selected three DE genes in each of the top four families (MYB, CAMTA, bZIP and MADS-box) for quantitative reverse transcriptase PCR (qRTPCR) analysis to measure gene expression among different treatments and tissue types (roots and shoots). Specific primers were optimized for the twelve selected TF genes using alpha tubulin as an endogenous control (Supplementary File 2). We amplified cDNA libraries from three biological replicates of roots and shoots for 0 mM and 300 mM treated plants. Based on the qRTPCR results, all gene targets selected correspond with the expression levels observed in the transcriptome analysis (Figure 5.2).

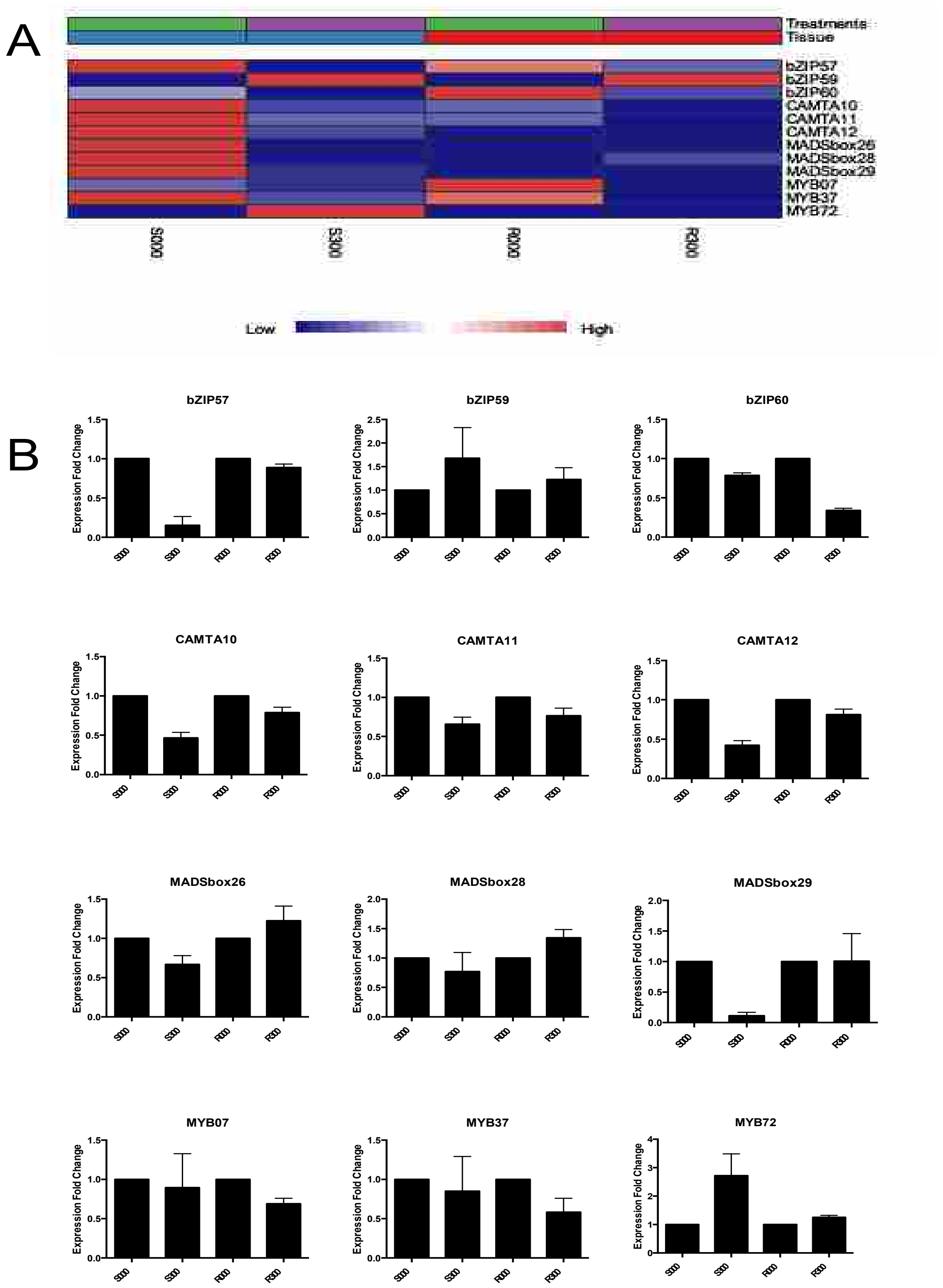


Figure 5.2 RNA-Seq and qRTPCR Analysis of Differentially Expressed Transcription Factors
Heatmap representation of shoots and roots for 0 mM and 300 mM treatments are illustrated from RNA-seq analysis of selected DE TFs from the top DE families (3A). The same DE TFs target are validated using quantitative reverse transcriptase PCR. Expression fold changes are calculated using ΔΔCT and $2^{-\Delta\Delta CT}$ against alpha tubulin as the endogenous control. Standard error of the mean is calculated using data analysis package in Prism Graph Pad.(3B). R000 (roots at 0 mM NaCl), R300 (roots at 300 mM NaCl), S000 (shoots at 0 mM NaCl), S300 (shoots at 300 mM NaCl).

Significant decreases in expression of bZIP57 are observed in the 300 mM treated shoots, which correspond closely with the RNA-sequencing results. Similarly, there is a decrease of CAMTA12 expression in the 300mM shoots. MADSbox29 shows a significant decrease of expression in shoot optimal growth, while MYB72 shows upregulation on the same tissue type and treatment.

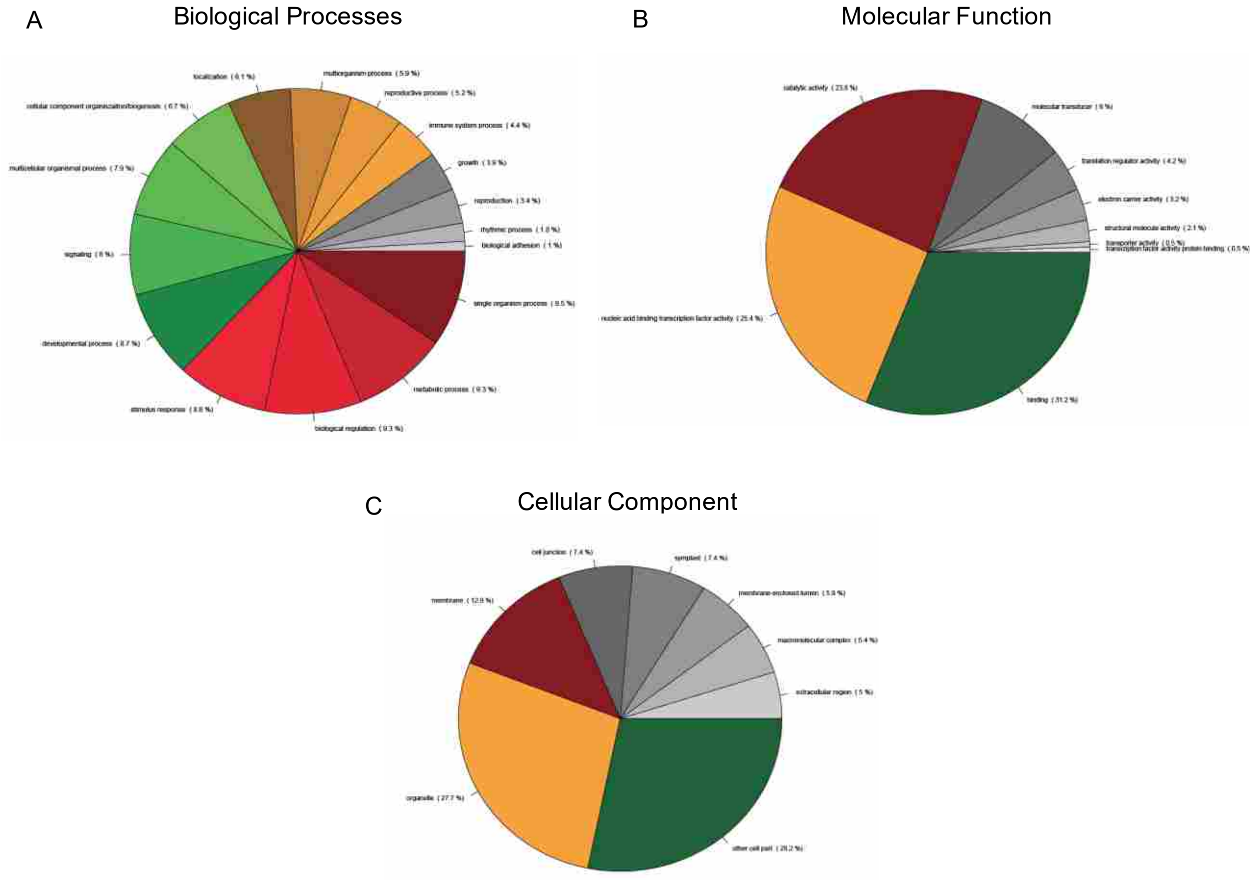


Figure 5.3 Gene Ontology Annotation of DE TFs
Differentially expressed transcripts were classified into 3 main GO annotations: Biological Processes (A), Molecular Function (B) and Cellular Component (C).

To perform functional and process homology-based annotation, we ran BLAST with the DE TFs against SwissProt and NCBI non-redundant protein databases using Blast2GO. Sequences were mapped with GO terms associated with BLAST search hits, and assigned functional terms based on the gene ontology vocabulary (Figure 5.3). The TFs are assigned into three main categories: Biological process refers to the biological objective of the genes or gene products, molecular function as the biochemical activity of the genes, and cellular components as the place where the interaction of the gene product actively functions. Dominant categories include metabolic, developmental and single organism process and stimulus response (each comprising 9%) for biological processes (Figure 5.3A). There are 59 hits (31%) for general binding for the molecular function category (Figure 5.3B), and cellular component category shows 28% of hits for cell part and organelle where the interaction of the genes is happening (Figure 5.3C). This annotation of *S. fruticosa* DE TFs suggests that they are involved in salt regulation but likely perform diverse functions in other regulatory, metabolic and stress response mechanisms.

Protein Interaction Network of Differentially Expressed Transcription Factors

The sequences of DE TFs were analyzed and examined using STRING v10 software to retrieve physical and functional interactions among proteins. The summary network of all identified DE TFs suggests involvement in flowering, stomatal development and stress regulation (Supplementary Figure 3). Importantly, the protein relationships predicted in *S.*

fruticosa using Arabidopsis homologs MYB (FLP1, MYB13, LHY), ARR-B PCL1, and MADS-box AGL24 are involved in one interaction network.

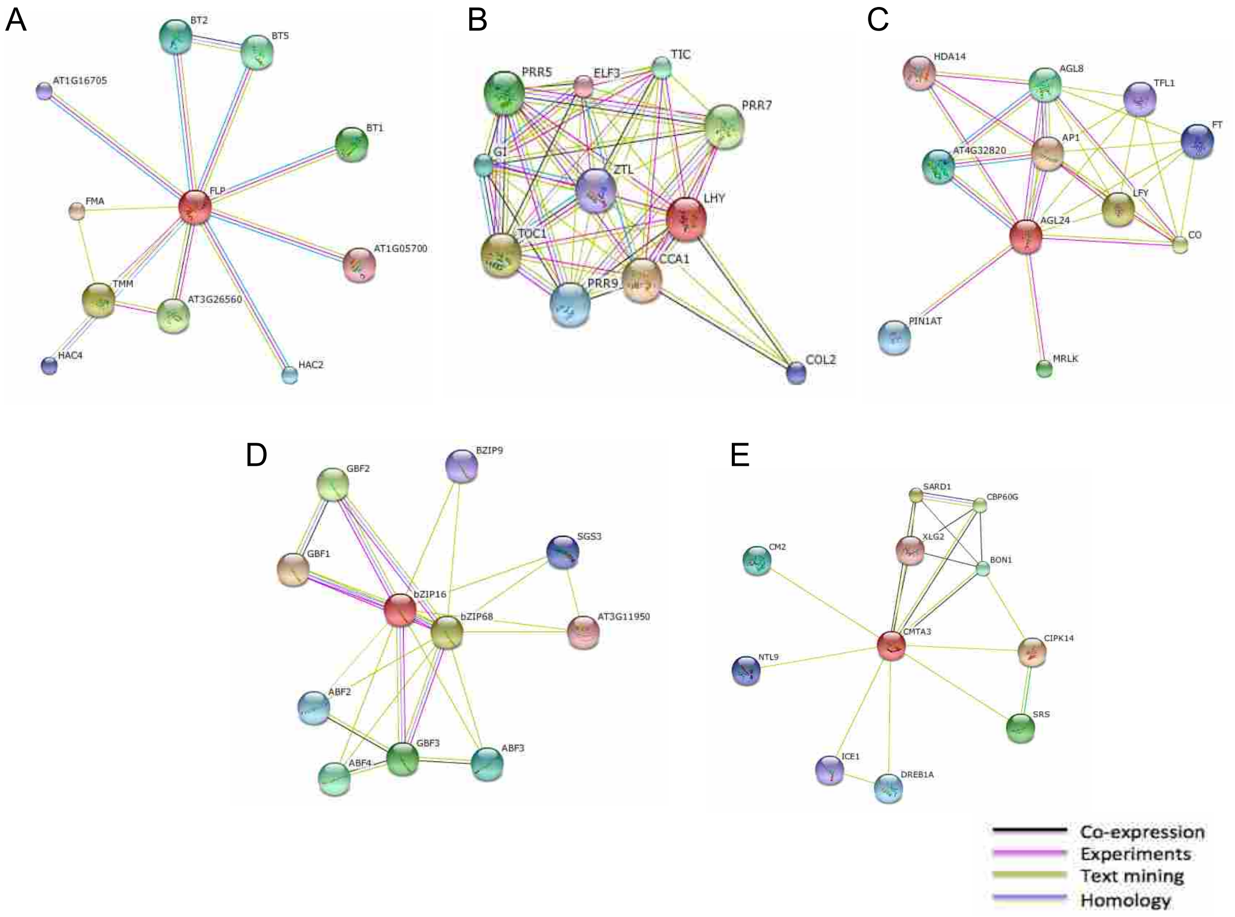


Figure 5.4 Protein-protein Interaction Network Predicted by STRING
Interactions of selected DE TFs from top DE families are illustrated: MYB TF FLP (A), MYB TF LHY and CCA1 (B), MADS-box AGL24 and LFY (C), bZIP family bZIP16 and bZIP 68 (D), CAMTA family CMTA3 (E). Colored lines represented different interactions: black (co-expression), pink (experimental), green (text mining), blue (homology).

Genes that belong to the top DE TF families were also examined for their interactions and functions with other genes (Figure 5.4). From the identified interactions between DE TFs, two *S. fruticosa* genes (Locus_17372_Transcripts_9,12) encoding similar identity with FLP (88% identity) and MYB88 (79% identity) contain a putative MYB transcription factor involved in stomata development (Figure 5.4A). The loss of FLP activity results in failure of guard mother cells to adopt the guard cell fate (Lai, et al. 2005). FLP and MYB88 negatively controls the expression of genes associated with stomatal development but positively regulates gene expression related to stress conditions. Double mutants of FLP and MYB88 are more susceptible to drought and salt stress and lose water significantly faster than wild-type (Xie, et al. 2010). This suggests that these individuals TFs play notable roles in salt regulation.

Four DE genes (Locus_36812_Transcript_1,2,5,6) related to LHY or CCA1 interact with other MYB TFs (Figure 5.4B). CCA1 regulates ELF4 and ELF3 that are involved in circadian control and phytochrome regulation in C3 and CAM leaves (Anwer, et al. 2014). These clock-associated genes in *Mesembryanthemum* are unaffected by salt stress, suggesting compensation of the central circadian clock against development and abiotic stress in specialized plants (Boxall, et al. 2005)

The DE homologue MADS-box AGL24 (Locus_82944_Transcripts_1,3,4,6) also interacts with these MYB homologs (Figure 5.4C). The AGL24 transcriptional activator is predicted to mediate effects of gibberellins on flowering and regulates the expression of LFY genes for floral induction and development. A homologue of MYB13 (Locus_37251_Transcript_2) is involved in response to salt stress, jasmonic acid and gibberellin (Boxall, et al. 2005) and interacts with homologue PCL1 (Locus_119717_Transcript_1,2). PCL1 works as a transcriptional activator involved in circadian rhythm and regulation of flower development in Arabidopsis (Onai and Ishiura 2005).

Other families including three *S. fruticosa* bZIP16 homologs (Locus_50829_Transcript_4,7,8) interact with ABF genes and other bZIP genes (Figure 5.4D). Arabidopsis bZIP16 promotes seed germination and hypocotyl elongation during early stages of seedling development. CAMTA3 homologues (Locus_5187_Transcript_1/9984, 1/9985, 1/9988, 2,9) show interactions with DREB dehydration response elements, regulators of cell death and defense, and other genes important to regulation of plant immunity (Figure 5.4E). Studies of CAMTA3 in other plants reveal that it negatively regulates plant defense and suppresses salicylic acid accumulation and disease resistance. Calcium ion/calmodulin binding through CAMTA3 is critical for wound response. Overexpression of AtSR1/CAMTA3 effectively confers plant resistance to herbivore attack through salicylic acid/jasmonic acid crosstalk regulation (Benn, et al. 2014; Yang, et al. 2012b)

Evolution of Transcription Factor-Encoding Genes in *Suaeda fruticosa*

We reconstructed 57 ML TF family trees using the iterative alignment-tree searching algorithm in PASTA (Supplementary Figure 4). The CAMTA TF family tree shows that the majority of DE and non-DE TF genes formed single monophyletic clades (Figure 5.5A). Whole genome/large-scale chromosomal duplications play a crucial role in increasing copy number of CAMTA TF genes (Rahman, et al. 2016). The close-relatedness of DE TF paralogs found most likely indicates that these genes duplicated separately from other non-DE TFs and subsequently their expression patterns/regulatory mutations were preserved by species-specific environmental constraints related to increased salt concentration (Rensing, et al. 2008). Based on observed patterns, we hypothesize that such large CAMTA family expansions can be explained by small-scale gene duplication events (e.g. via unequal crossing over).

In the bZIP TF family most of the DE and non-DE TF genes were scattered uniformly across the tree; however, all four DE genes formed a single monophyletic cluster (Figure 5.5B). Such distribution of bZIP genes suggests that gene duplications happened before speciation of *S. fruticosa*. One of the major bZIP family expansions was observed on the branch that leads to seed plants (Correa, et al. 2008). Moreover, its evolution-by-gene duplication patterns fit to a random birth-death-model, suggesting that new gene copies occurred as a result of small-scale duplication events rather than whole genome/chromosome duplications (Correa, et al. 2008).

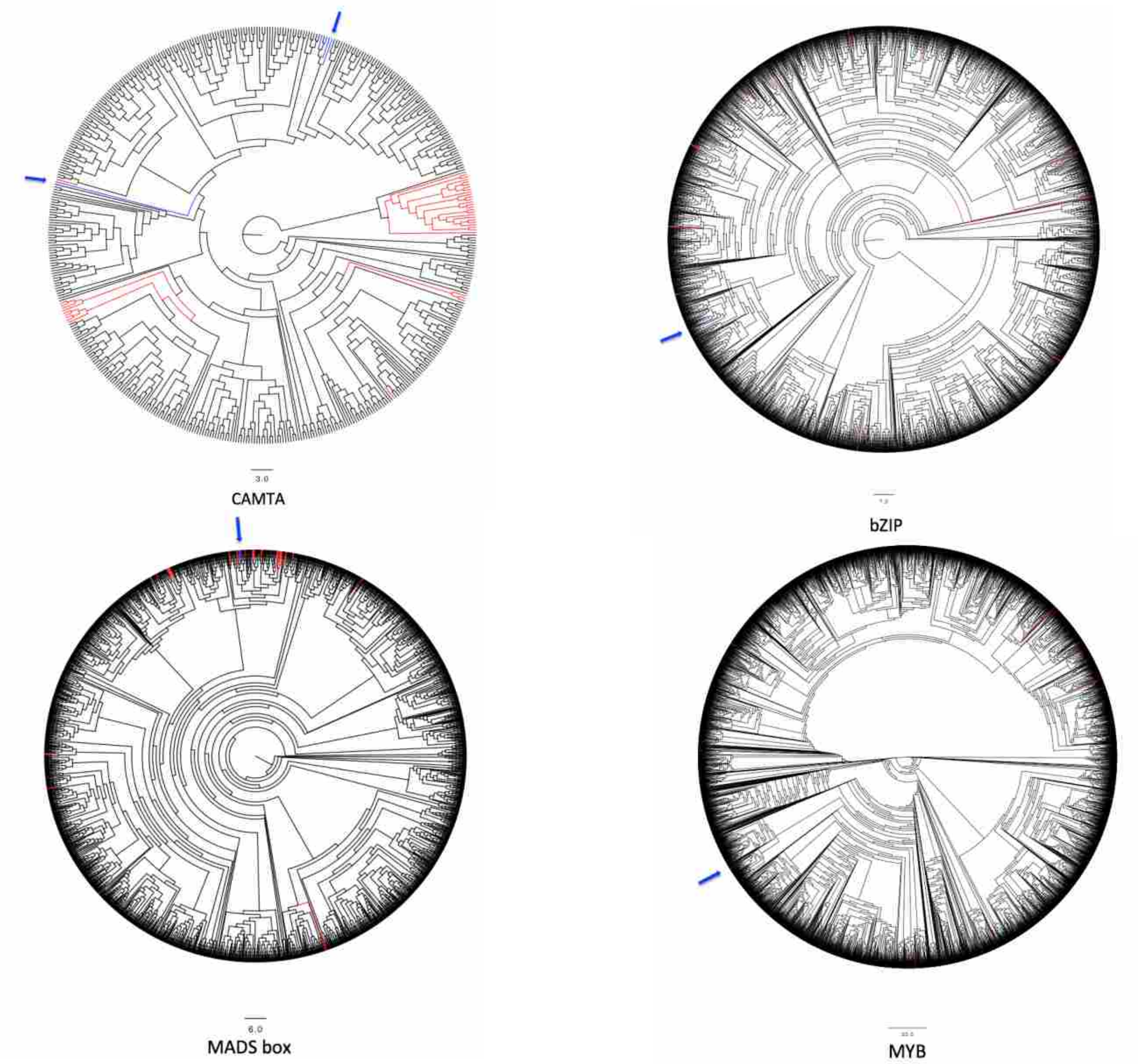


Figure 5.5 Ilustration of Cladogram Trees from Top 4 DE TFs
Evolutionary trees include TFs of green plants identified from PlantTFDBv.3.0 belonging to the respective TF family and identified *S. fruticosa* TFs of that family. Red highlighted lines represent the total *S. fruticosa* TFs while blue lines represent those *S. fruticosa* TFs that are differentially expressed. Arrow indicates the DE TFs locations.

The M-type tree (a subset of MADS-box) also exhibits similar relationships between DE and non-DE TF genes (Figure 5.5C). Nevertheless, all four M-type DE genes cluster with non-DE genes, suggesting their recent adaptive radiation as a response to salt. Most likely these copies appeared via whole genome duplications and intraspecific gene duplications (Smaczniak, et al. 2012). The ancestral functions of MADS-box genes are currently unknown. Some MADS-box genes in Arabidopsis showed that they are polyphyletic with significantly longer branch lengths than for other genes, suggesting that they could be pseudogenized as a result of neutral evolution (Kofuji, et al. 2003).

For the MYB TF family we observed similar patterns where four DE genes formed a monophyletic group whereas non-DE genes were uniformly distributed across a tree (Figure 5.5D). It has been suggested that following duplication events MYB TFs usually undergo sub-functionalization (Feller, et al. 2011).

In conclusion, we have identified transcription factors expressed in *S. fruticosa,* provided phylogenetic trees for top DE TFs, performed expression pattern analysis and annotated individual TFs involved in interaction networks. The results provide basic information on key regulator TFs of *S. fruticosa* and contribute to an increased understanding of salt tolerance mechanisms of a succulent halophyte that may be utilized for the improvement of halophytes as non-conventional crops. Future analyses should include individual examination of the transcription factors identified in relation to salt tolerance between halophytes and salt-sensitive glycophytes.

MATERIALS AND METHODS

Plant Material and Harvest

Suaeda fruticosa seeds were grown at Brigham Young University, Provo, Utah, USA according to the optimized protocol under long term salinity treatment (Hameed, et al. 2012). Plant samples of three biological replicates from roots and shoots were treated at low (0 mM NaCl) and optimal (300 mM NaCl) salt conditions and used for transcriptome sequencing and isolate of RNA for qRTPCR analysis.

Data Acquisition and Analysis

A description of plant samples processed for RNA-Seq and methods for bioinformatics analysis including *de novo* assembly and differential expression analysis are found in our recent paper (Diray-Arce, et al. 2015). Illumina sequences are available at the NCBI Sequence Read Archive under *Suaeda fruticosa* accession SRX973396. Transcriptome sequence information is deposited in the Transcriptome Shotgun Assembly Sequence Database: BioProject ID: PRJNA279962 and PRJNA279890. The following supplementary information files will be publicly available at Dryad upon acceptance. Differentially expressed (DE) genes and the entire assembled transcriptome were translated using Transdecoder software and the protein sequences clustered using CDHIT (Fu, et al. 2012).

Transcription Factor Identification

Transcription factors were identified and searched against the Plant Transcription Factor Database 3.0. HMM profiles of the 57 families were obtained and used to search against the *S. fruticosa* proteome using profile hidden Markov search in HMMER with an E-value cutoff of

10^{-10}. Codes for TF prediction, DE TF identification and phylogenetic tree construction are available in Supplementary File 5.

Differential Expression Analysis of Transcription Factors-Encoding Genes

Analysis of differential expression between treatments of 0 mM and 300 mM NaCl from *S. fruticosa* was performed using the EdgeR package from R. We used the generalized linear models for data analysis for different salt concentrations of treatment and biological replicates. This differentiates the number of expressed transcripts across experimental conditions. We then searched and identified TFs from the differentially expressed list using a profile hidden Markov search in HMMER (Finn, et al. 2011) using an E-value of 10^{-10} against the database from PlantTFDBv3.0. These TFs were then annotated based on gene ontology, their functional domains and structures using BLAST2GO against the nr and Swiss protein databases with a similar E-value cutoff of 10^{-10}. Enrichment analysis for specific gene ontology for biological process, molecular function and cellular components were determined using default parameters. Functional interactions between DE TFs were performed using STRING software version 10. STRING is a widely used database and web interface to explore protein-protein interactions, including physical and functional interactions (Szklarczyk, et al. 2015).

Validation of Differentially Expressed Transcription Factors Through qRTPCR

Transcription factors identified were selected for validation of differential expression using qRTPCR. For each qRTPCR reaction, 1 µg of RNA of 0 mM and 300 mM NaCl treated samples were reverse transcribed into cDNA using oligodT primers, and the cDNA libraries produced were used for qRTPCR using this method (Haddad and Baldwin 2010). Primer sequences are available as supplementary information (Supplementary File 2). We ran second strand synthesis using an ABI Plus One thermocycler with annealing temperature of 58°C. To assess validation for each gene, qRTPCR data were analyzed based on $\Delta\Delta$CT and $2^{-\Delta\Delta CT}$ method. The ΔCT value of each gene was calculated by subtracting the CT value of the endogenous control from the CT value of the target gene.

We selected the alpha tubulin gene as an endogenous control. Primers were designed from the top DE TFs from *S. fruticosa* transcriptome sequences and optimized for RTPCR. We chose to sample 3 gene targets per family. Expression analysis using $\Delta\Delta$CT, $2^{-\Delta\Delta CT}$ and standard error of the mean were calculated using the data analysis package in Microsoft Excel. Data were plotted as mean fold change ($2^{-\Delta\Delta CT}$). Significant differences ($p < 0.05$) were determined using a one-tailed two-sample t-test assuming equal variances for comparison of the fold change values between biological replicates using GraphPad Prism software.

Molecular Genetic Analysis of Gene Structure and Motif Composition of Selected TF Families

In order to generate multiple sequence alignment of an entire TF family and construct a corresponding Maximum-Likelihood (ML) gene tree we used an alignment-tree co-estimation algorithm implemented in PASTA (Mirarab, et al. 2015). PASTA has been shown to produce accurate alignments and generate trees on large datasets. First, we ran PASTA for two iterations to generate TF family alignments and masked sites with <5% data. Second, we used that masked alignment to extract homologous genes from the *S. fruticosa* transcriptome using profile hidden Markov search in HMMER (Finn, et al. 2011) with the E-value cutoff of 10^{-10}. These gene hits were then combined with the original TF family sequences and the alignment and tree was co-estimated again in PASTA (Supplementary Figure 4). Constructed trees from all plant TF families are uploaded and can be viewed using FigTree (Dryad, Diray-Arce, 2016).

List of Abbreviations Used

TF- transcription factors; DE- differential expression/differentially expressed; GO- gene ontology; ML- maximum likelihood

Author's Contributions

JDA and AS conceived, performed bioinformatics, wet lab analysis, and wrote the manuscript. CH performed qRTPCR and assisted in writing the manuscript. SB provided guidance for evolution analysis. BG provided information and seeds for growing the plants. MAK provided expertise in halophyte growth and analysis, obtained funding, and assisted in preparation of the manuscript. BLN provided guidance on the project, obtained funding, and helped format and edit the manuscript. All authors read and approved the final manuscript.

Conflict of Interest Statement

The authors declare no competing interests.

ACKNOWLEDGEMENTS

We thank BYU Fulton Supercomputing Facility and Mark Clement of the BYU CS department for bioinformatics assistance. This research has been supported by a grant from the Pakistan-US Science and Technology Cooperation Program, US Department of State and Higher Education

Commission of Pakistan and by the Department of Microbiology and Molecular Biology at Brigham Young University.

Supplementary Files

Source codes:

```
### PASTA run 1
for i in *.dir; do  cd $i; ~/Scripts/producePASTA.sh *.cdhit ; cd ..; done

### Alignment masking
for i in *.dir; do cd $i; declare -i nn; nn=$( grep ">" *.fasta.cdhit.aln | wc -l)*5/100; python2.7 ~/Soft/PASTA/pasta/run_seqtools.py -infile
*.fasta.cdhit.aln* -outfile *.fasta.cdhit.aln.masked -masksites $nn -filterfragments 1   ; cd ..; done

### HMMER Building and Compression
for i in *.dir; do cd $i; ~/Soft/hmmer-3.1b1-linux-intel-x86_64/binaries/hmmbuild pastarun1.marker001.fasta.cdhit.aln.masked.hmm
pastarun1.marker001.fasta.cdhit.aln.masked ; ~/Soft/hmmer-3.1b1-linux-intel-x86_64/binaries/hmmpress
pastarun1.marker001.fasta.cdhit.aln.masked.hmm; cd ..; done

### HMMER de genes search
for i in *.dir; do cd $i; ~/Soft/hmmer-3.1b1-linux-intel-x86_64/binaries/hmmscan -o unusedde.out --tblout degenes.hits -E 1e-10
pastarun1.marker001.fasta.cdhit.aln.masked.hmm ../data/degenes.newname.fasta.transdecoder.pep; cd ..; done

### HMMER total transcriptome search
for i in *.dir; do  cd $i; ~/Scripts/produceHMMSCAN.sh pastarun1.marker001.fasta.cdhit.aln.masked.hmm ../data/Sfruticosaprotein.fasta; cd ..;
done

### Extract DE gene IDs
for i in *.dir; do  cd $i; grep -v "#" degenes.hits  | awk '{print$3}' > degenes.hits.list; cd ..; done

### Extract Total  gene IDs
for i in *.dir; do  cd $i; grep -v "#" total.hits  | awk '{print$3}' > total.hits.list; cd ..; done

### Extract DE gene seqs
for i in *.dir; do  cd $i; python3.4 ~/Scripts/extract.py ../data/degenes.newname.fasta.transdecoder.cds degenes.hits.list > degenes.hits.dna.fasta;
cd ..; done

### Extract DE gene seqs peps
for i in *.dir; do  cd $i; python3.4 ~/Scripts/extract.py ../data/degenes.newname.fasta.transdecoder.pep degenes.hits.list > degenes.hits.prot.fasta;
cd ..; done

### Extract total gene seqs peps
for i in *.dir; do  cd $i; python3.4 ~/Scripts/extract.py ../data/Sfruticosafinal.cdhit.fasta.transdecoder.pep.uniq total.hits.list > total.hits.prot.fasta;
cd ..; done


###Remove redundancy using CDHIT 100% identity threshold
###for total protein
cd-hit -i Sfruticosaprotein.fasta -o Sfruticosaprotein.cdhit.fasta -c 1.00
###for degenes cds
cd-hit-est -i degenes.fasta.transdecoder.cds -o degenes.fasta.transdecoder.cdhit.cds -c 1.00
###for degenes protein
cd-hit -i degenes.fasta.transdecoder.pep -o degenes.fasta.transdecoder.cdhit.pep -c 1.00
####for whole transcriptome
cd-hit-est -i Sfruticosafinal.fsa -o Sfruticosafinal.cdhit.fsa -c 1.00
####for degenes nucleotide seqs
cd-hit-est -i degenes.fasta -o degenes.cdhit.fasta -c 1.00

#Concatenate TF sequences with Suaeda hits
for i in *.dir; do  cd $i; cat *.fasta.cdhit total.hits.prot.fasta > tf.family.suaeda.prot.fasta; cd ..; done

#Remove * signs (stop codons)
```

```
 for i in *.dir; do cd $i; sed 's/*//g'  tf.family.suaeda.prot.fasta > tf.family.suaeda.prot.nostopcodon.fasta; cd ..; done

#PASTA run 2
for i in *.dir; do  cd $i; ~/Scripts/producePASTA.sh tf.family.suaeda.prot.nostopcodon.fasta; cd ..; done

#Concatenate all total protein sequence into one file
for i in *.dir; do cat *.dir/total.hits.prot.fasta > ../total.combined.prot.fasta; done

###HMMER search degenes 1-15-2016#####
$ for i in *.fasta
> do
> cd $i.dir
> ~/Soft/hmmer-3.1b1-linux-intel-x86_64/binaries/hmmscan -o unused.txt --tblout $i.degenes.hits -E 1e-10 *.hmm
../data/degenes.cdhit.fasta.transdecoder.pep.uniq
> cd ..
> done
```

Supplementary File 2

Primers designed for qRTPCR

MYB37 FWD
CAT GAG GAT GTC GGA GCA TTA T

MYB37 REV
GTT GCA CAG GAC AGG AAT TTG

MYB72 FWD
AGG AAC CTG ATG CTG ATG ATG

MYB72 REV
CAG TGG AGG ATG GTG TTT CTT

MYB07 FWD
GAG GTG TTG TCC GTT GAA GA

MYB07 REV
GAA CGT CGT CCG ACA TAT ACA C

CAMTA10 FWD
GAA AGG CCA GGA ACT TCT CTA C

CAMTA10 REV
TGG CTC CAT GTC TCC TAA CT

CAMTA11 FWD
CCA TTA TCC AGA AGC GAG AGA G

CAMTA11 REV
CAT CAA TTG CGC CAC TAC AC

CAMTA12 FWD
CAA TCT GAG GGC GCT TCT T

CAMTA12 REV
GCT CTC TCG CTT CTG GAT AAT G

MADSbox26 FWD
CTT CTG GCA AAC TCC ATG ATT TC

MADSbox26 REV
GGA TCA AGC TGT TGA GGA AGA

MADSbox28 FWD
TTA AGC CGA ATG CTA GGA GAA G

MADSbox28 REV
GCT TGA GGT CTA CGA TCA CTT T

MADSbox29 FWD
CTT CTG GCA AAC TCC ATG ATT TC

MADSbox29 REV
GGA TCA AGC TGT TGA GGA AGA

bZIP57 FWD
GGA TGA CTA TGG TGC CAA TGA

bZIP57 REV
CGT ATA GCC TGG ATT GGA GAT G

bZIP59 FWD
CGT AGA TCC AGA CTG CGT AAA C

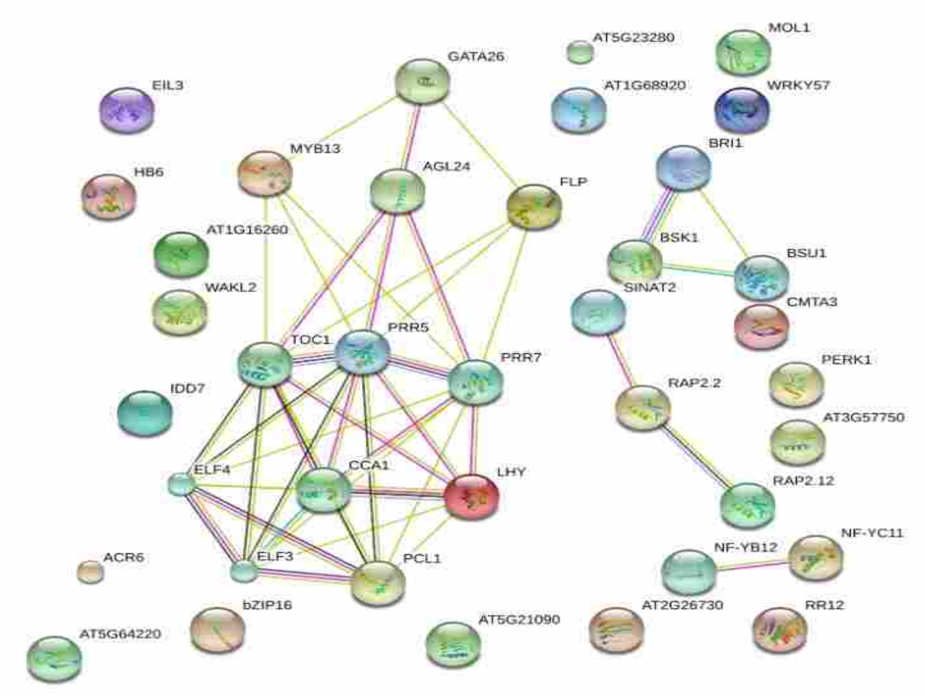
bZIP59 REV
GCC CTA AGC TGC TCG TAA TC

bZIP60 FWD
GGA TGA CTA TGG TGC CAA TGA

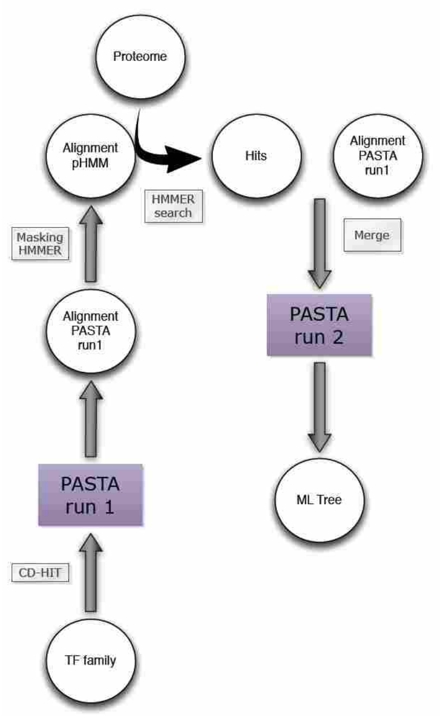
bZIP60 REV
CGT ATA GCC TGG ATT GGA GAT G

A tubulin FWD
CAC GCG CTG TAT TCG TAG AT

A tubulin REV
TGA CCA CGA GCG AAG TTA TTA G



Supplementary Figure 1 Protein Interaction Network of Differentially Expressed Transcription Factors in *Suaeda fruticosa*



Supplementary Figure 2 A Diagram of the Tree Inference Workflow

REFERENCES

Abe H, et al. 2003. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. Plant Cell 15: 63-78.

Anil VS, Rahjkumar P, Kumar P, Mathew MK 2008. A plant Ca2+ pump, ACA2, relieves salt hypersensitivity in yeast. Modulation of cytosolic calcium signature and activation of adaptive Na+ homeostasis. J Biol Chem 283: 3497-3506.

Anwer M, et al. 2014. Natural variation reveals that intracellular distribution of ELF3 protein is associated with function in the circadian clock. Elife 3. doi: 10.7554/eLife.02206.001

Arora R, et al. 2007. MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. BMC Genomics 8: 242. doi: 10.1186/1471-2164-8-242

Babitha KC, Vemanna RS, Nataraja KN, Udayakumar M 2015. Overexpression of EcbHLH57 Transcription Factor from Eleusine coracana L. in Tobacco Confers Tolerance to Salt, Oxidative and Drought Stress. PLoS One 10: e0137098. doi: 10.1371/journal.pone.0137098

Benjamini Y, Hochberg Y 1995. Controlling the false discovery rate: a practival and powerful approach to multiple testing. J Royal Statistical Soc Series 57: 289-300.

Benn G, et al. 2014. A key general stress response motif is regulated non-uniformly by CAMTA transcription factors. Plant J 80: 82-92. doi: 10.1111/tpj.12620

Boxall SF, et al. 2005. Conservation and divergence of circadian clock operation in a stress-inducible Crassulacean acid metabolism species reveals clock compensation against stress. Plant Physiol 137: 969-982. doi: 10.1104/pp.104.054577

Correa LGG, et al. 2008. The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes. PLoS One 3. doi: ARTN e2944 10.1371/journal.pone.0002944

Diray-Arce J, Clement M, Gul B, Ajmal Khan M, Nielsen BL 2015. Transcriptome Assembly, Profiling and Differential Gene Expression Analysis of the halophyte Suaeda fruticosa Provides Insights into Salt Tolerance. BMC Genomics 16. doi: 10.1186/s12864-015-1553-x

Diray-Arce J, Gul B, Khan MA, Nielsen B. 2016. 10 - Halophyte Transcriptomics: Understanding Mechanisms of Salinity Tolerance. In. Halophytes for Food Security in Dry Lands. San Diego: Academic Press. p. 157-175.

Feller A, Machemer K, Braun EL, Grotewold E 2011. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant Journal 66: 94-116. doi: 10.1111/j.1365-313X.2010.04459.x

Finkelstein RR, Gibson SI 2002. ABA and sugar interactions regulating development: cross-talk or voices in a crowd? Curr Opin Plant Biol 5: 26-32.

Finn RD, Clements J, Eddy SR 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-37. doi: 10.1093/nar/gkr367

Flowers T, Colmer T 2008. Salinity tolerance in halophytes. New Phytol 179: 945 - 963.

Fu L, Niu B, Zhu Z, Wu S, Li W 2012. CDHIT-accelerated for clustering the next generation sequencing data. Bioinformatics 28: 3150-3152.

Ganesan G, Sankararamasubramanian HM, Harikrishnan M, Ganpudi A, Parida A 2012. A MYB transcription factor from the grey mangrove is induced by stress and confers NaCl tolerance in tobacco. J Exp Bot 63: 4549-4561. doi: 10.1093/jxb/ers135

Garg R, et al. 2014. Deep Transcriptome Sequencing of Wild Halophyte Rice, Porteresia coarctata, Provides Novel Insights into the Salinity and Submergence Tolerance Factors. DNA Research 21: 69-84. doi: 10.1093/dnares/dst042

Ghanekar R, Srinivasasainagendra V, Page G 2008. Cross-Chip Probe Matching Tool: A Web-Based Tool for Linking Microarray Probes within and across Plant Species. Int. J. Plant Genomics 7. doi: 10.1155/2008/451327

Glenn EP, Brown JJ, Blumwald E 1999. Salt Tolerance and Crop Potential of Halophytes. Critical reviews in plant sciences 18: 227-255. doi: 10.1080/07352689991309207

Golldack D, Luking I, Yang O 2011. Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. Plant Cell Reports 30: 1383-1391. doi: 10.1007/s00299-011-1068-0

Haddad F, Baldwin K 2010. Reverse transcription of the ribonucleic acid: the first step in RT-PCR assay. Methods Mol Biol 630: 261-270.

Hameed A, et al. 2012. Salt tolerance of a cash crop halophyte Suaeda fruticosa: biochemical responses to salt and exogenous chemical treatments. Acta Physiologiae Plantarum 34: 2331-2340. doi: 10.1007/s11738-012-1035-6

Hsieh TH, et al. 2010. A tomato bZIP transcription factor, SlAREB, is involved in water deficit and salt stress response. Planta 231: 1459-1473. doi: 10.1007/s00425-010-1147-4

Jiang Y, et al. 2012. Genome-wide transcription factor gene prediction and their expressional tissue-specificities in maize. J Integr Plant Biol 54: 616-630. doi: 10.1111/j.1744-7909.2012.01149.x

Jin J, Zhang H, Kong L, Gao G, Luo J 2014. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research 42: D1182-D1187. doi: 10.1093/nar/gkt1016

Kofuji R, et al. 2003. Evolution and divergence of the MADS-box gene family based on genome-wide expression analyses. Mol Biol Evol 20: 1963-1977. doi: 10.1093/molbev/msg216

Lai LB, et al. 2005. The Arabidopsis R2R3 MYB proteins FOUR LIPS and MYB88 restrict divisions late in the stomatal cell lineage. Plant Cell 17: 2754-2767. doi: 10.1105/tpc.105.034116

Long Y, Scheres B, Blilou I 2015. The logic of communication: roles for mobile transcription factors in plants. J Exp Bot 66: 1133-1144. doi: 10.1093/jxb/eru548

Mirarab S, et al. 2015. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. Journal of Computational Biology 22: 377-386. doi: 10.1089/cmb.2014.0156

Onai K, Ishiura M 2005. PHYTOCLOCK1 encoding a novel GARP protein essential for the Arabidopsis circadian clock. Genes Cell 10: 963-972.

Parenicova L, et al. 2003. Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell 15: 1538-1551.

Rahman H, Yang J, Xu YP, Munyampundu JP, Cai XZ 2016. Phylogeny of Plant CAMTAs and Role of AtCAMTAs in Nonhost Resistance to Xanthomonas oryzae pv. oryzae. Front Plant Sci 7: 177. doi: 10.3389/fpls.2016.00177

Rensing SA, et al. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319: 64-69. doi: 10.1126/science.1150646

Robinson MD, McCarthy DJ, Smyth GK 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140. doi: 10.1093/bioinformatics/btp616

Saha G, et al. 2015. Genome-wide identification and characterization of MADS-box family genes related to organ development and stress resistance in Brassica rapa. BMC Genomics 16: 178. doi: 10.1186/s12864-015-1349-z

Sharma R, et al. 2015. De Novo Assembly and Characterization of Stress Transcriptome in a Salinity-Tolerant Variety CS52 of Brassica juncea. PLoS One 10: e0126783. doi: 10.1371/journal.pone.0126783

Smaczniak C, Immink RG, Angenent GC, Kaufmann K 2012. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. Development 139: 3081-3098. doi: 10.1242/dev.074674

Szklarczyk D, et al. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research 43: D447-D452. doi: 10.1093/nar/gku1003

Toda Y, Yoshida M, Hattori T, Takeda S 2013. RICE SALT SENSITIVE3 binding to bHLH and JAZ factors mediates control of cell wall plasticity in the root apex. Plant Signal Behav 8: e26256. doi: 10.4161/psb.26256

Wang W, et al. 2016. A pair of light signaling factors FHY3 and FAR1 regulates plant immunity by modulating chlorophyll biosynthesis. J Integr Plant Biol 58: 91-103. doi: 10.1111/jipb.12369

Xie Z, Li D, Wang L, Sack FD, Grotewold E 2010. Role of the stomatal development regulators FLP/MYB88 in abiotic stress responses. Plant J 64: 731-739. doi: 10.1111/j.1365-313X.2010.04364.x

Yang A, Dai X, Zhang WH 2012a. A R2R3-type MYB gene, OsMYB2, is involved in salt, cold, and dehydration tolerance in rice. J Exp Bot 63: 2541-2556. doi: 10.1093/jxb/err431

Yang O, et al. 2009. The Arabidopsis basic leucine zipper transcription factor AtbZIP24 regulates complex transcriptional networks involved in abiotic stress resistance. Gene 436: 45-55. doi: 10.1016/j.gene.2009.02.010

Yang T, Peng H, Whitaker BD, Conway WS 2012b. Characterization of a calcium/calmodulin-regulated SR/CAMTA gene family during tomato fruit development and ripening. BMC Plant Biol 12: 19. doi: 10.1186/1471-2229-12-19

You J, Chan Z 2015. ROS Regulation During Abiotic Stress Responses in Crop Plants. Front Plant Sci 6: 1092. doi: 10.3389/fpls.2015.01092

Yue R, et al. 2015. Identification and expression profiling analysis of calmodulin-binding transcription activator genes in maize (Zea mays L.) under abiotic and biotic stresses. Front Plant Sci 6: 576. doi: 10.3389/fpls.2015.00576

Zhou J, et al. 2009. Basic helix-loop-helix transcription factor from wild rice (OrbHLH2) improves tolerance to salt- and osmotic stress in Arabidopsis. J Plant Physiol 166: 1296-1306. doi: 10.1016/j.jplph.2009.02.007

Zhu J 2001. Plant salt tolerance. Trends Plant Sci 6: 66 - 71.

Zhu Q, et al. 2010. The Arabidopsis AP2/ERF transcription factor RAP2.6 participates in ABA, salt and osmotic stress responses. Gene 457: 1-12. doi: 10.1016/j.gene.2010.02.011

CHAPTER 6: Conclusions and Future Directions

Salinity stress, often interconnected with osmotic and ionic stress, involves signaling processes and transcription controls that activate stress response mechanisms. These signals are important to reestablish homeostasis and protect and repair damaged proteins and membranes. When one or more steps in this process are inadequate to restore cellular homeostasis, this may result in destruction of functional and structural proteins and membranes that can lead to cell death. If all of these processes are regained, this can lead to salinity tolerance or adaptation (Vinocur and Altman 2005). These response mechanisms are found in naturally occurring salt-tolerant plants called halophytes.

In this study we have worked with a previously characterized halophyte using next-generation sequencing to identify groups of genes that are important to salt regulation in halophytes (Chapter 1) (Diray-Arce, et al. 2016). The halophyte we are studying, *Suaeda fruticosa Forssk*, is a member of a large halophytic family Chenopodiaceae and belongs to a potential model genus for studying salt tolerance because of its ability to take up salt to a high concentration and because its physiological and physical characteristics are similar to most halophytes (Chapter 2). Since *Suaeda fruticosa* does not have a reference genome or transcriptome, we have assembled the de novo transcriptome using a genome-independent reconstruction approach and clustering algorithms from RNA-sequencing data. Since typical next generation sequencing results are comprised of very large (gigabase to terabases) data, which requires a very large amount of computing system memory to run algorithm analysis, we utilized various methods of analysis to provide the most preferred assembly with the highest coverage and least redundancy (Chapter 3). We also have compared methods for assembly of the

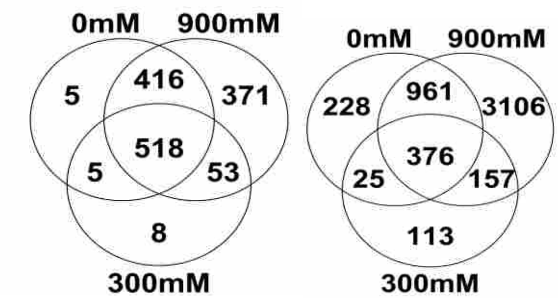
Suaeda fruticosa transcriptome using different bioinformatics algorithms and have optimized the assemblies using clustering methods. In Chapter 4, we reported the first transcriptome analysis of *Suaeda fruticosa* focusing on the identification and annotation of transcripts for this halophyte (Diray-Arce, et al. 2015). We also analyzed differential expression and tissue-specific patterns of the transcriptomic response and identified genes that are induced or repressed in plants grown in optimal salt concentration in comparison with those grown in the absence of salt. There are 519 differentially expressed transcripts; 44 of them are found to be upregulated upon salt treatment and 475 of them are downregulated. These genes have been annotated based on their biological process, molecular functions and cellular component categories. We have identified and analyzed putative salt-tolerance related genes and performed qRTPCR for selected genes for confirmation of relative expression. This study will contribute to comprehensive information about the transcriptome of *S. fruticosa* and will provide a basis for further study of the mechanisms of salt tolerance in succulent halophytes. In Chapter 5, we have identified and characterized putative transcription factors (TF) expressed in *S. fruticosa.* We also have analyzed TF expression patterns and predicted protein-protein interactions and evolutionary trajectories using evolutionary family trees for the top differentially expressed transcription factor (DE TF) families. We have identified the top DE TFs (MYB, CAMTA, MADS-box and bZIP) to understand their roles as the most responsive families in salinity tolerance. The results provide basic information on key regulator TFs of *S. fruticosa* to aid studies on regulation of salt tolerance in plants.

FUTURE DIRECTIONS

Next generation technologies such as transcriptomics and microarray studies allow investigation of changes in gene expression under experimental conditions. However, because of post-transcriptional modifications such as splicing variations, differential degradation of mRNAs and proteins and other modifications, we cannot always imply that the amount of protein made is exactly correlated to the number of transcripts produced for any gene. Since proteins are responsible for a variety of functions, any changes in the internal and external cellular environment can affect the response network, signaling, growth and development of the organism. Direct measurement of protein expression is deemed necessary to provide details on the physiological state and subcellular localization (Ngara and Ndimba 2014). Upon salt stress, plants respond by sensory mechanisms that alter gene and protein expression patterns. Since the plant proteome is highly dynamic, it may show both qualitative and quantitative expression changes after exposure to treatments (Hossain, et al. 2011). As a future direction, proteomics studies are strongly suggested to capture the spatial and temporal changes in expression by comparing different plant species, types and levels of stress, harvesting times, different tissues or subcellular compartments (Salekdeh and Komatsu 2007).

We have conducted preliminary quantitative proteomics analysis on proteins isolated from roots of the same set of plants used for transcriptomics analysis (treated at 0, 300 and 900 mM of NaCl). We performed protein extraction and cleanup, FASP protocol and trypsin digestion, along with high pH fractionation for analysis of proteins using an LC-MS/MS Thermo-Fisher Orbitrap (BYU Chemistry & Biochemistry Department) (Figure 6.1). The results were analyzed using Protein Prospector (UCSF) and Scaffold software (UC Davis). We identified 518 total proteins from all three treatments. In addition, there are 376 shared peptides among the three treatments. We have performed Fisher's exact tests to analyze how many differentially expressed proteins are identified between treatments with a p-value of less than 0.05 and false discovery rate of 1%. Upon examination of all conditions 309 proteins are differentially expressed among them. When comparing 0 and 300 mM NaCl treatment, there are 90 differentially expressed proteins while there are 210 differentially expressed proteins when comparing 0 and 900 mM NaCl treated plants. There are 112 proteins that are differentially expressed when comparing 300 and 900 mM NaCl treated *Suaeda fruticosa* roots.

A: Total unique proteins B: Total unique peptides



C. Statistical design:

Fisher's exact test (p-value < 0.05)
0 and 300 mM- 90 DE proteins
0 and 900 mM- 210 DE proteins
300 and 900 mM- 112 DE proteins
All treatments-309 DE proteins

Figure 6.1 Summary details from LC-MS/MS Identification Technology
Figure 6.1A shows the number of total unique proteins and those shared among the conditions and 6.1B shows the number of total unique peptides and the number of shared peptides among them. 6.1C shows the number of differentially expressed proteins among the treatments.

Table 6.1 Summary of Identified Proteins and Peptides from LC-MS/MS MUDPIT Analysis

Suaeda treatment	Protein groups identified	Proteins	Peptides	Protein groups identified from *Suaeda fruticosa* database
0 mM NaCl	1122	6104	2147	1008=90%
300 mM NaCl	989	5426	1678	883=89.2%
900 mM NaCl	2857	13911	7004	2630=92%
Multireports (combination of all treatments)	3074	14672	7662	

This table shows the total number of identified proteins and peptides from LC-MS/MS MUDPIT analysis. All peptides are identified within 1% false discovery rate. Database used include the translated unigenes from Suaeda fruticosa transcriptome BioProject Accession: PRJNA279890 and the SwissProt Green plant database.

From the de novo transcriptome analysis, we translated the unigenes of *Suaeda fruticosa* using ESTScan to obtain amino acid sequences, generating a total of 52,018 proteins. We have scanned and used this database for searching matches of peptides and proteins from the Orbitrap results. We have identified the total number of matches from the database and determined the number of proteins and peptides from this preliminary analysis (Table 6.1). This data allows us to further characterize protein-coding genes that show the same pattern of expression in protein levels. We plan to correlate and compare this data to see if drastic changes are seen from transcriptomic to proteomic data of different plant treatments for any genes.

Although transcriptome information can provide information on the gene activity of the cell, posttranscriptional gene regulation and mRNA stability affects the correlation between the mRNA levels and the protein produced, therefore requiring the confirmation of results in protein levels. The challenge might be more difficult because of cellular variability and dynamics over time; however, differential level measurements can be accurately measured using proteomics. Quantitative proteomics analysis will be beneficial to identify and quantify expression of key protein markers that change upon introduction of salt or identify proteins that are differentially expressed over a longer time period. Other future directions may include exploration of identified salt-tolerant proteins through targeted proteomics to characterize regulatory proteins involving transcription factor-DNA interactions and to analyze the fine dynamics of protein systems such as protein networks and specific signaling pathway.

REFERENCES

Diray-Arce J, Clement M, Gul B, Ajmal Khan M, Nielsen BL 2015. Transcriptome Assembly, Profiling and Differential Gene Expression Analysis of the halophyte *Suaeda fruticosa* Provides Insights into Salt Tolerance. BMC Genomics.

Diray-Arce J, Gul B, Khan MA, Nielsen B. 2016. 10 - Halophyte Transcriptomics: Understanding Mechanisms of Salinity Tolerance. In. Halophytes for Food Security in Dry Lands. San Diego: Academic Press. p. 157-175.

Hossain Z, Nouri M, Komatsu S 2011 Plant cell organelle proteomics in response to abiotic stress. Journal of Proteome Research 11: 37-48.

Ngara R, Ndimba B 2014. Model plant systems in salinity and drought stress proteomics studies: A perspective on Arabidopsis and Sorghum. Plant Biology 16: 1029-1032.

Salekdeh GH, Komatsu S 2007. Crop proteomics: aim at sustainable agriculture of tomorrow. Proteomics 2007: 2976-2996.

Vinocur B, Altman A 2005. Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Current Opinion in Biotechnology 16: 123-132. doi: http://dx.doi.org/10.1016/j.copbio.2005.02.001

APPENDIX 1: The *Arabidopsis* At1g30680 Gene Encodes a Homologue to the Phage T7 gp4 Protein that has both DNA Primase and DNA Helicase Activities

Joann Diray-Arce[1], Bin Liu[1], John D. Cupp, Travis Hunt and Brent L. Nielsen*

Dept. of Microbiology & Molecular Biology
775 WIDB
Brigham Young University
Provo, Utah 84602 U.S.A.

[1]These authors contributed equally to this work

*Author for correspondence: brentnielsen@byu.edu
Tel. 1-801-422-1102; Fax 1-801-422-0519

Author e-mail addresses:
Joann Diray-Arce: joann.diray@gmail.com
Bin Liu: nbinliu@yahoo.com
John Cupp: john.d.cupp@gmail.com
Travis Hunt: tjcalcio@aol.com
Brent Nielsen: brentnielsen@byu.edu

© 2013 BMC Plant Biology, Volume 13:36, doi: 10.1186/1471-2229-13-36. 4 Mar 2013

ABSTRACT

Background

The *Arabidopsis thaliana* genome encodes a homologue of the full-length bacteriophage T7 gp4 protein, which is also homologous to the eukaryotic Twinkle protein. While the phage protein has both DNA primase and DNA helicase activities, in animal cells Twinkle is localized to mitochondria and has only DNA helicase activity due to sequence changes in the DNA primase domain. However, *Arabidopsis* and other plant Twinkle homologues retain sequence homology for both functional domains of the phage protein. The *Arabidopsis* Twinkle homologue has been shown by others to be dual targeted to mitochondria and chloroplasts.

Results

To determine the functional activity of the *Arabidopsis* protein we obtained the gene for the full-length *Arabidopsis* protein and expressed it in bacteria. The purified protein was shown to have both DNA primase and DNA helicase activities. Western blot and qRT-PCR analysis indicated that the *Arabidopsis* gene is expressed most abundantly in young leaves and shoot apex tissue, as expected if this protein plays a role in organelle DNA replication. This expression is closely correlated with the expression of organelle-localized DNA polymerase in the same tissues. Homologues from other plant species show close similarity by phylogenetic analysis.

Conclusions

The results presented here indicate that the *Arabidopsis* phage T7 gp4/Twinkle homologue has both DNA primase and DNA helicase activities and may provide these functions for organelle DNA replication.

Keywords

DNA primase, DNA helicase, bacteriophage T7 gp4, Twinkle, Organelle DNA replication

BACKGROUND

DNA replication involves the coordinated activity of several enzymes and proteins. These enzymes assist with the unwinding, separation, and copying of double stranded DNA to produce new identical DNA copies [1]. DNA helicase translocates unidirectionally along one strand of the nucleic acid to facilitate replication initiation. The helicase utilizes ATP hydrolysis to separate the DNA double helix into individual strands [2,3]. DNA primase catalyzes the formation of short RNA oligonucleotides used as primers to begin DNA synthesis [4]. DNA polymerase uses the primers and extends the 3' end of the nucleotide chain by adding nucleotides matched to the template strand [1].

Individual genes are usually responsible for encoding each replication enzyme activity. However, bacteriophage T7 gene 4 protein (T7 gp4) and similar proteins from T3, P4 and other phages [4] encode a single protein with both DNA helicase and DNA primase domains. T7 phage has two forms of gp4 protein that are both required for phage genome replication. The longer form has two zinc motifs and has both DNA primase and helicase activity while the shorter one retains only DNA helicase activity [5].

Most eukaryotic organisms have a homologue of the T7 gp4 protein that has been named Twinkle (T7 gp4-like protein with intramitochondrial nucleoid localization). This protein shares close sequence similarity with the bacteriophage T7 gp4 primase-helicase protein [6,7]. Twinkle is a hexameric DNA helicase at the mitochondrial DNA replication fork which unwinds sections of double-stranded DNA [8,9]. The Twinkle homologue lacks DNA primase activity in higher eukaryotes but is suggested to have this activity in *Plasmodium* species [6,10] and *Arabidopsis* *thaliana* and other plants [11,12]. This protein is assumed to play a key role in mitochondrial DNA (mtDNA) replication as it localizes in the mitochondrial nucleoid and matrix. In maize, Twinkle has also been found associated with the chloroplast nucleoid [13], suggesting that this protein may function in both mitochondria and chloroplasts.

Mutations in Twinkle result in mitochondrial-associated diseases in humans [6,14] and mice [15,16]. In humans, coding region mutations in this gene have been linked with autosomal dominant progressive external ophthalmoplegia (adPEO) and are often associated with multiple mtDNA deletions, suggesting a role in mtDNA replication [6]. In mice, Twinkle expression reduction by RNAi resulted in a rapid drop in mtDNA copy number [6,17] while overexpression of the protein led to increases in mtDNA copy number in muscle and heart tissue [15,18].

When the amino acid sequences of Twinkle homologues from a wide variety of eukaryotic species are compared, high homology in the conserved Walker motifs for the DNA helicase domain of the protein has been observed, as summarized in two review papers [4,5]. Critical differences were observed in the primase domain of Twinkle in some model organisms when compared to the N-terminal end of the T7 gp4 protein [19]. The location of the (nonfunctional) primase domain in human Twinkle is at the N-terminal portion of the protein, the same as in phage T7 gp4 and in DNAG-like primases in bacteria and phage [4,11]. But unlike T7 gp4, the N-terminal domain of human Twinkle lacks several motifs required for primer synthesis in T7 gp4, thus leading to the prediction that the Twinkle N-terminal region is generally inactive in humans and metazoa in general [5]. The T7 gp4 protein contains a beta sheet structure and cysteine residues forming two zinc fingers [7] in Motif 1. The N-terminal end of the primase domain of T7 gp4 contains a zinc finger motif but Twinkle in most metazoan species lacks the zinc-binding domain necessary for DNA and amino acid binding for polymerization [5]. Also, human Twinkle does not contain the conserved cysteine residues of a zinc-finger motif critical for DNA binding and primase activity [20]. The zinc finger motif in the primase domain synthesizes pppAC oligonucleotide primers important for the initial step of sequence-specific primer synthesis at the sequence 5′-GTC-3′ [21]. The Twinkle protein from *Arabidopsis thaliana* contains the conserved sequence elements and is predicted to have both DNA primase and DNA helicase activities.

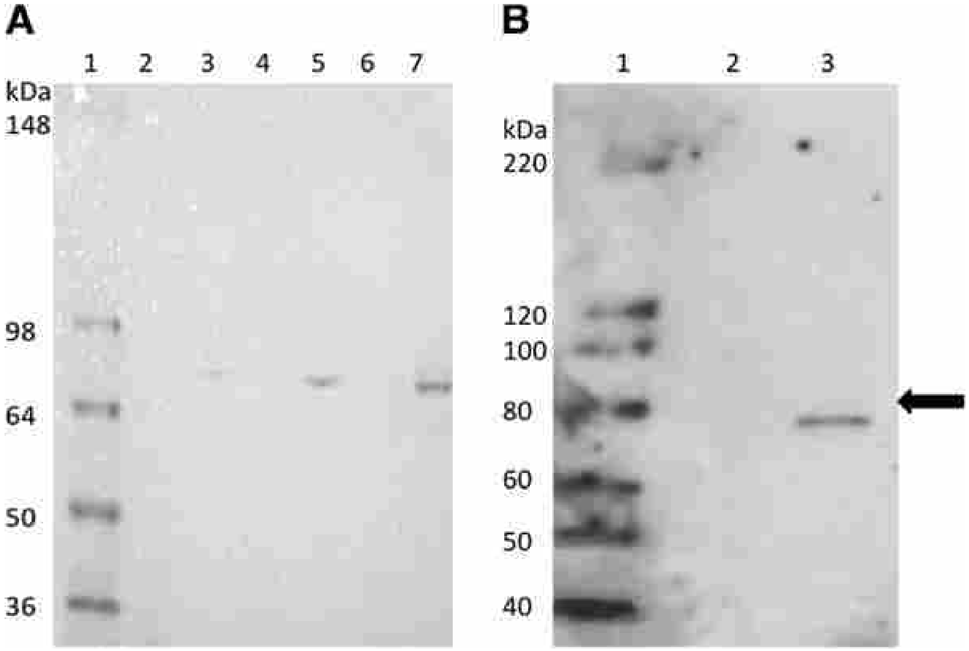
The *Arabidopsis* genome contains two homologues of the bacteriophage T7 gp4 protein. The first (At1g30680) shares homology with the conserved motifs of the DNA primase and DNA helicase domains [5]. The coding sequence predicts a protein of about 80 kDa, which is larger than the full-length 63,000 kDa T7 gp4 protein but similar to the sizes of Twinkle homologues reported in eukaryotes. The second *Arabidopsis* homologue is truncated, sharing the N-terminal primase domain but entirely lacking the C-terminal helicase domain, with a predicted size of ~38 kDa (At1g30660). Since this gene is truncated, it will be designated as a primase homologue, while the full-length gene will be designated as a Twinkle homologue in this paper.

We show here that the *Arabidopsis* T7 gp4 homologue has both DNA primase and DNA helicase activities, the first such report from a higher eukaryote. The gene for this protein is highly expressed in rapidly growing plant tissues and is correlated with organelle DNA polymerase gene expression.

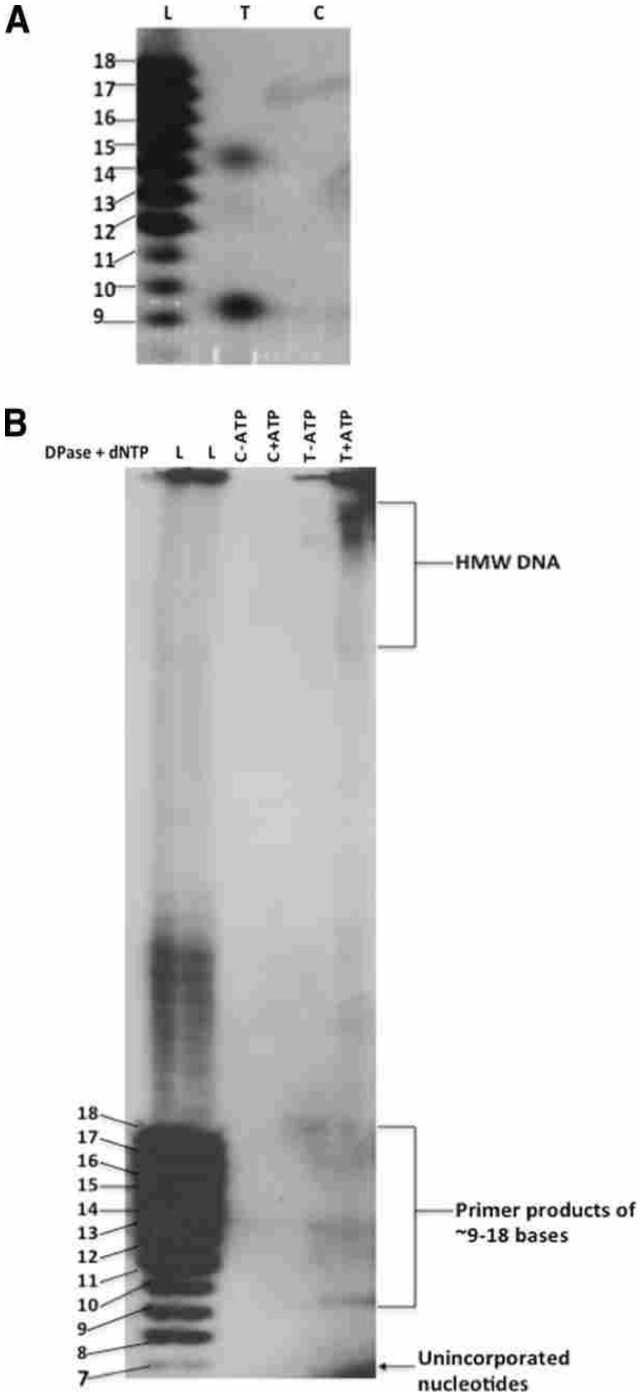
RESULTS

Expression of the *Arabidopsis* Protein in *E. coli* and Demonstration of DNA Primase Activity

The full-length cDNA for the *Arabidopsis* Twinkle gene was obtained and cloned into a bacterial expression vector to produce protein for enzymatic activity assays. The purified protein showed a predominant band of the proper size by gel staining (Figure 1A). Its identity as the expressed protein was confirmed by western blot analysis using an antibody against a synthetic peptide from the *Arabidopsis* protein sequence (Figure 1B). The recombinant protein product is smaller (~74 kDa) than the full-length coding region of the Twinkle homologue since it lacks the N-terminal organelle targeting sequence. The purified protein was used for an *in vitro* assay for DNA primase activity. Gel analysis of the reaction products indicates that the protein is capable of producing RNA primers of ~ 9–18 bases from a single-stranded DNA template (Figure 2).



Appendix Figure 1 Purification of the Recombinant Protein
Panel A shows the Coomassie Blue-stained gel, with increasing amounts of the purified recombinant (lanes 3, 5 and 7) and control (lanes 2, 4 and 6) protein, from left to right. Lane 1, protein molecular weight markers (Invitrogen SeeBlue 2 markers). Lanes 2 and 3, 0.195 ng; lanes 4 and 5, 0.39 ng; lanes 6 and 7, 0.585 ng. Panel B shows a western blot of the purified protein using antibody against the *Arabidopsis* Twinkle homologue. Lane 1 contains molecular weight markers (Invitrogen Magic Markers). Lane 2, control protein; lane 3, 0.5 ng purified recombinant protein. The arrow at the right indicates 80 kDa, the length of the full-length *Arabidopsis* gene product. The recombinant protein is slightly smaller (~74 kDa) as it lacks the N-terminal localization sequence.

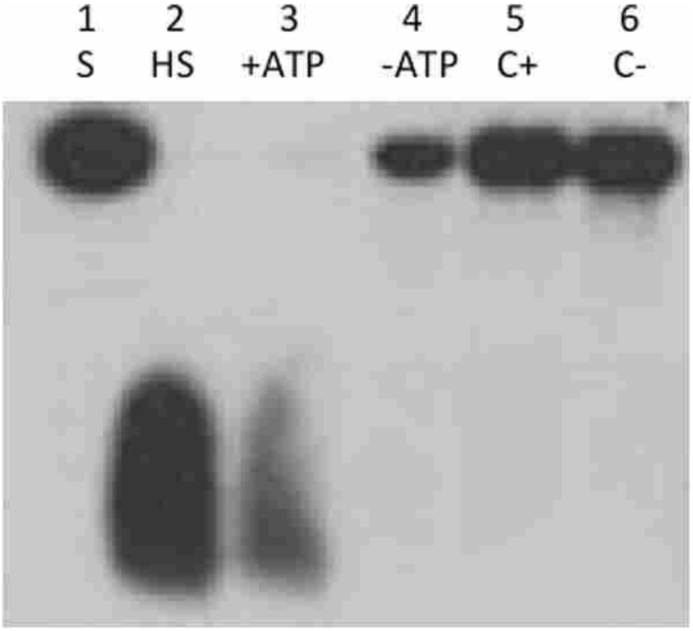


Appendix Figure 2 DNA Primase Assay
The recombinant Twinkle homologue purified from *E. coli* cells was tested for DNA primase activity. Panel A, lane L (DNA single-base ladder), oligo dT_{9-18} included as size markers (same for panel B). Lane T, reaction products with the recombinant protein. Lane C, reaction products using a bacterial fraction with the empty vector as control. Panel B shows incorporation of primers into high molecular weight DNA in the presence (lane 4) but not the absence (lanes 3) of *E. coli* DNA polymerase I and dNTPs. Lanes 1 and 2 are the control protein fraction in the absence (lane 1) and presence (lane 2) of DNA polymerase I and dNTPs.

Stronger intensity of the primers of 9 and 14 bases was consistently observed (close-up shown in Figure 2A), and are similar in size to products reported for other DNA primases [22].
The primers were capable of being extended by DNA polymerase into high molecular weight DNA (Figure 2B), which is a fundamental property of a DNA primase that generates primers for DNA replication. The primer bands are absent in the control lanes (protein from bacteria with the empty vector lacking the *Arabidopsis* gene), indicating that this activity is not due to bacterial DNA primase contamination of the purified recombinant protein. This provides clear evidence for the function of the *Arabidopsis* Twinkle homologue as an active DNA primase, the first such report in a higher eukaryote.

DNA Helicase Activity of the *Arabidopsis* Twinkle Homologue Protein

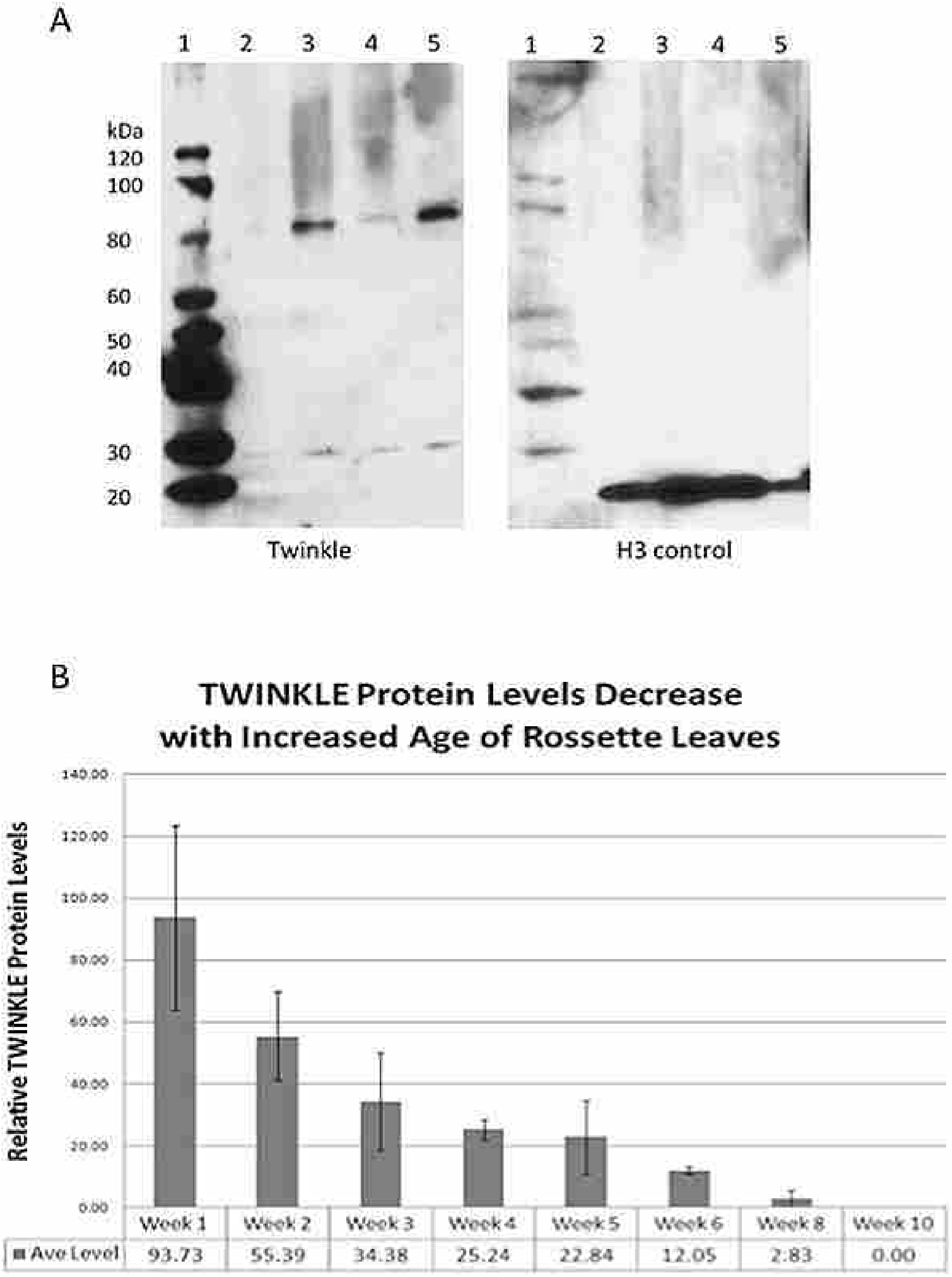
The purified recombinant protein was also assayed for DNA helicase activity. The results indicate that the protein indeed has ATP-dependent DNA helicase activity as predicted (Figure 3). The control protein preparation (vector with no insert) lacked DNA helicase activity in the presence or absence of ATP (Figure 3 lanes 5 and 6). The activity is similar to the DNA unwinding activity we previously detected in soybean mitochondrial extracts [23]. The results from the biochemical assays indicate that the *Arabidopsis* Twinkle homologue has both DNA primase and helicase activities, similar to the phage T7 gp4 protein.



Appendix Figure 3 DNA Helicase Assay
The recombinant Twinkle homologue purified from *E. coli* cells was tested for DNA helicase activity as described in the text. Lane 1 is the control substrate (S). Lane 2 is the heated control (HS), showing separation of the short labeled oligo from the substrate, which runs in this gel as a leading band with a diffuse smear; lane 3 (T+ATP), reaction using the purified recombinant protein with ATP; lane 4 (T-ATP), same reaction without ATP, lane 5 (C+ATP), control protein from *E. coli* cells lacking the expression construct with ATP, lane 6 (C–ATP), same reaction but without ATP.

Western Blot Analysis of *Arabidopsis* Twinkle Homologue Expression

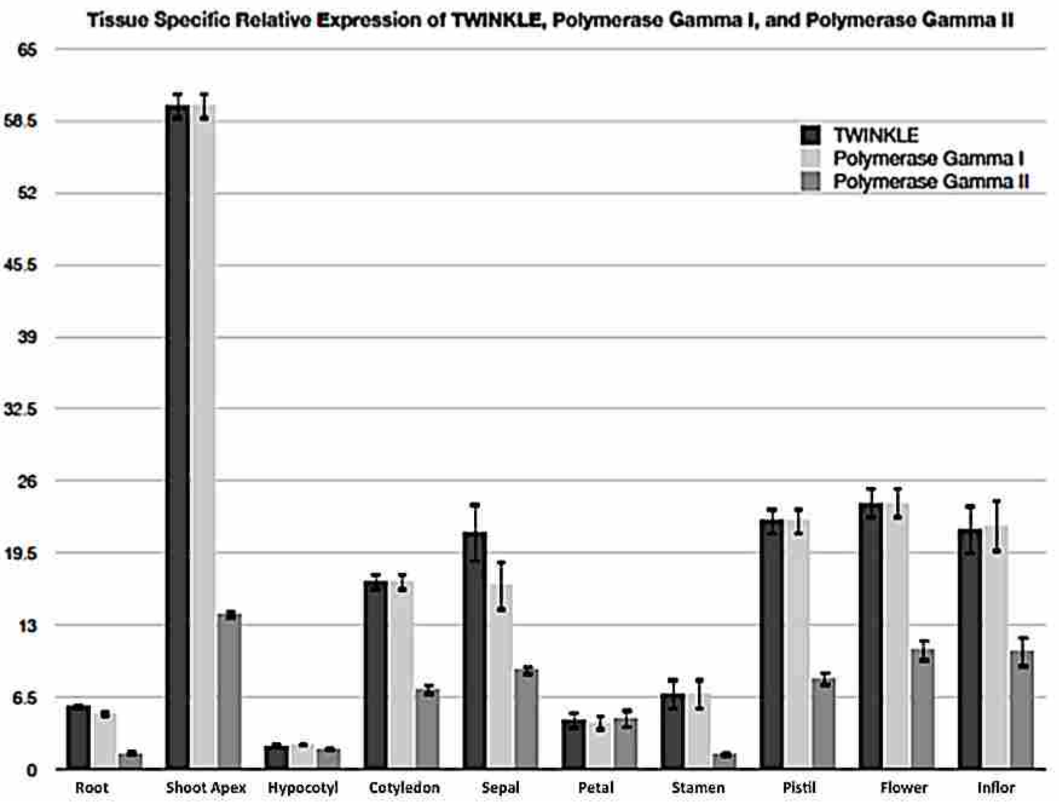
Western blot analysis of Twinkle protein expression levels in different *Arabidopsis* tissues shows that the protein is most abundant in meristem and young leaf tissue and nearly undetectable in mature leaves (Figure 4A). Total rosette leaf tissue from plants was collected at weekly intervals and total protein was recovered from each sample for western blot analysis. The results show relatively high levels of the Twinkle protein in weeks 1–3 of growth, with a subsequent rapid drop in levels until the protein is nearly undetectable after week 5 (Figure 4B). These results are compatible with those reported from the different tissues (Figure 4A) and provide support for the involvement of Twinkle in organelle DNA replication in developing tissues. We conducted western blot analysis which indicated the presence of Twinkle in isolated mitochondria and chloroplasts of *Arabidopsis* (data not shown).



Appendix Figure 4 Western Blot Analysis of Arabidopsis Twinkle Homologue Expression
A. Lane 1, Molecular weight markers, Lane 2, leaf tissue from 6-week plants; lane 3, shoot apex tissue; lane 4, total plant tissue protein; lane 5, cotyledon protein. The panel on the left was incubated with antibody against the Twinkle protein. The panel on the right was incubated with histone H3 antibody as a loading control. B. Relative levels of Twinkle protein relative to a nuclear tubulin protein control in *Arabidopsis* seedlings harvested at the times indicated. The average of three independent western blots is shown for each time point (weeks 1–5 and 10). Error bars indicate the SEM (standard error of the mean).

Analysis of *Arabidopsis* Twinkle Homologue Expression in Different Tissues by qRT-PCR

Quantitative reverse transcriptase PCR analysis of cDNA generated from different tissues indicate that the *Arabidopsis* Twinkle gene is expressed at the highest level in the shoot apex (Figure 5), as expected if the Twinkle protein plays a role in organelle DNA replication in rapidly growing tissues. Twinkle is also expressed at relatively high levels in other developing tissues, especially cotyledons and different parts of flowers including sepals, pistils and the inflorescence (Figure 5). Interestingly, expression levels of Twinkle are very similar to expression levels of DNA Pol gamma I (Figure 5), a dual-targeted DNA polymerase that has been shown to play a role in plant organelle DNA replication and repair [24].



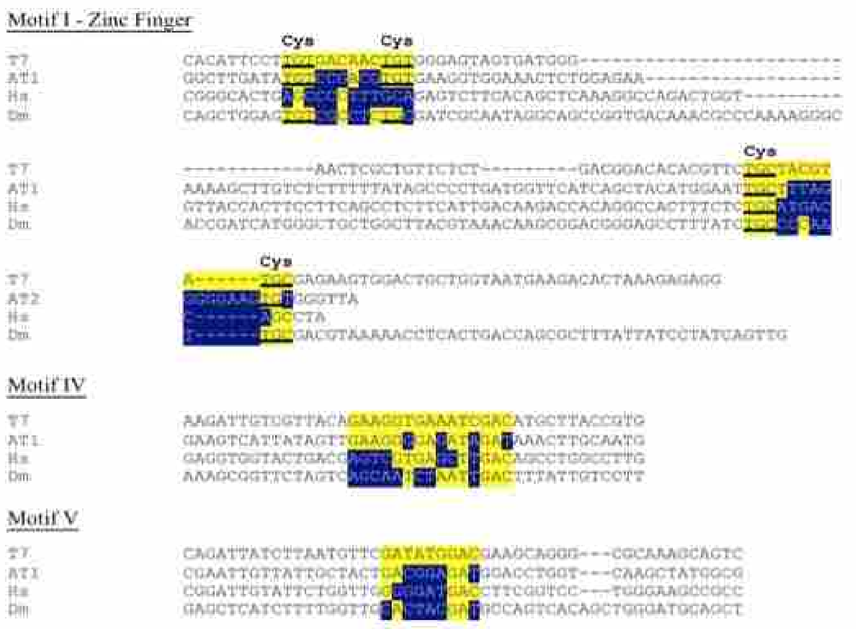
Appendix Figure 5 RT-qPCR Analysis of the Arabidopsis Twinkle Homologue Gene Expression Relative to Organellar Localized DNA Polymerases in Various Tissues
The relative abundance of Twinkle and the two organellar DNA polymerases (Polymerase gamma I and Polymerase gamma II) is shown, and varied among selected organs with highest expression in the shoot apex. The relative expression of Twinkle follows the expression levels of DNA polymerase gamma I. Error bars indicate SEM of three replicates. The Y axis indicates relative expression ($\log_2$) normalized to nuclear actin gene expression. Inflor, inflorescence.

The expression of DNA Pol gamma II is also generally highest in the same tissues that have high Twinkle expression (Figure 5). The similar levels of expression of Twinkle and the organelle-localized DNA polymerases [25] suggest that Twinkle may play a role in replication of organelle DNA.

Analysis of Twinkle DNA and Protein Sequences

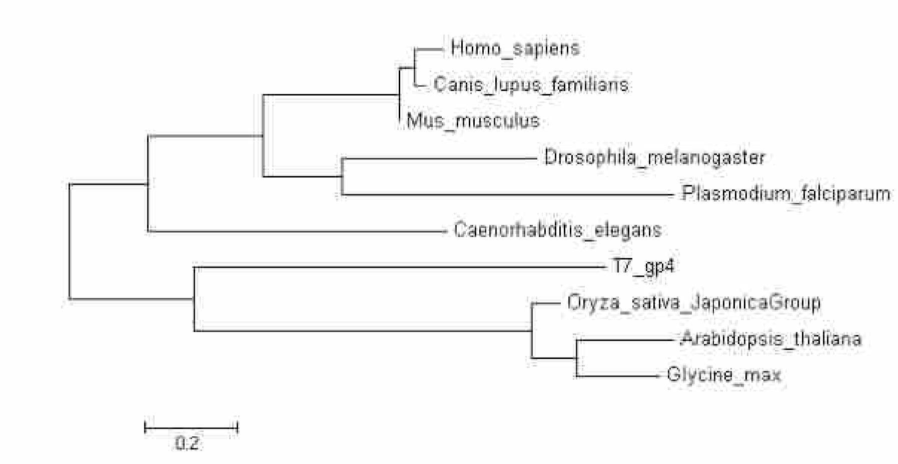
Two separate research groups have reported on the comparison of the amino acid sequences of Twinkle homologues from a wide variety of eukaryotic species, and have shown that there is high homology in the conserved Walker motifs for the C-terminal DNA helicase domain of the *Arabidopsis* protein [4,5]. The human, *Drosophila* and *C. elegans* Twinkle homologues have DNA helicase activity but lack DNA primase activity [4,5]. Upon close examination of the amino acid sequence encoding the primase domain at the N-terminal end of the protein in the plant and animal proteins, some critical differences are apparent. Two zinc fingers formed by cysteine residues in Motif 1 are present in the T7 gp4 protein [7] and in homologues from most eukaryotes, but the four cysteines that form the zinc fingers are absent in metazoans, including humans [4,5]. Analysis of the amino acid sequence alignment of the Twinkle homologues against the T7 gp4 protein shows that only the *Arabidopsis* and other plant Twinkle homologues share all highly conserved elements with the T7 gp4 protein [5]. Additional important differences are observed in other conserved motifs within the primase region of the protein in humans, *Drosophila* and *C. elegans*, while the sequences from a number of lower eukaryotes share the conserved elements with T7 gp4 protein [5]. In particular, the human homologue lacks both zinc finger domains in Motif 1, and the human and *Drosophila* sequences lack the highly conserved residues found in Motif IV and Motif V.

The *Arabidopsis thaliana* Twinkle protein contains the conserved sequence elements and is predicted to have both DNA primase and DNA helicase activities [4]. While the previous analysis of the amino acid sequences of these proteins identified critical differences at some conserved sites in the primase domain region of the protein in metazoa, including the absence of the cysteine residues needed to form the zinc fingers [4,5], we wanted to know if these changes were due to minor mutations in the sequence. However, DNA sequence analysis indicates that the differences in amino acid sequence of the homologues in human and *Drosophila* are not due to single base changes but are due to more significant alterations in the DNA sequence (Figure 6). The base sequence differences that are present in the *Arabidopsis* Twinkle primase domain as compared to the T7 gp4 protein mostly occur in the third position of the codons and do not alter the amino acid sequence.



Appendix Figure 6 DNA Sequence Alignments of Some Twinkle Primase Domain Conserved Regions To Show the Extent of Changes Between Different Organisms
The DNA sequences for the Twinkle protein from T7, *Arabidopsis* (At1g30680), human (Hs) and *Drosophila* (Dm) are shown for the conserved motifs I, IV and V. The locations of the cysteine residues in Motif I are indicated above the sequence while the corresponding codon sequence is underlined in the DNA sequence. The central conserved elements of each motif are shaded yellow. Base differences from the T7 gp4 sequence are shaded dark blue with white lettering.

Phylogenetic analysis of amino acid sequences of Twinkle homologues from several plants and other species shows that the *Arabidopsis* and plant homologues are closely clustered and are most similar to the bacteriophage T7 gp4 protein (Figure 7). The relationship between Twinkle proteins is supported by maximum likelihood phylogenetic analysis of taxonomic samples of Twinkle homologues. This suggests that the Twinkle homologues from humans and other animals are most distantly related to the T7 gp4 protein, supporting the observations from direct DNA and amino acid sequence alignments.



Appendix Figure 7 Phylogenetic Analysis of the T7 gp4 Protein, Plant Homologues, and Selected Eukaryotic Twinkle Protein Homologues
Molecular phylogenetic analysis was performed using the maximum likelihood method. The scale bar indicates the number of substitutions per site.

DISCUSSION

Twinkle has been shown to be the replicative DNA helicase in mitochondria of eukaryotic cells, and mutations that abolish expression of this gene are lethal in animal cells [6,14,15,26]. Twinkle is a homologue of the bacteriophage T7 gp4 protein, which has both DNA primase and DNA helicase activities and contains the highly characterized TOPRIM domain that is conserved in DNA primases, topoisomerases and OLD family nucleases [4]. However, until the present work no Twinkle homologue from a higher eukaryote has been shown to have DNA primase activity. Shutt and Gray have analyzed the sequence of Twinkle homologues from several eukaryote species and have proposed that in addition to being the DNA helicase, Twinkle may also serve as the mitochondrial DNA primase in most eukaryotes except metazoa [5]. As far as we know our present report is the first to show that the Twinkle homologue in a plant species (*Arabidopsis*) has both DNA primase and DNA helicase activities. Other than the truncated primase homologue already mentioned (At1g30660; but there is no information available about whether this protein is functional) no other bacterial or phage-type DNA primase homologues have been found in the *Arabidopsis* genome sequence.

Sequence analysis provides an explanation of why the plant homologue has both activities while the animal homologues lack DNA primase activity (Figure 6). The absence of primase activity in human Twinkle is likely due to the lack of the zinc finger motifs formed by 4 cysteine residues near the N-terminal end of the protein, as well as other amino acid sequence differences at conserved sequences in the primase domain of the protein which have been shown to be responsible for the primase activity (Figure 6) [4]. Sequence variation occurs in other metazoan species, and while some have the zinc fingers, they have differences at other conserved motifs. The *Arabidopsis* homologue, in contrast, retains all conserved motifs [4]. Phylogenetic analysis further supports these findings, indicating that the plant Twinkle homologues are most closely related to the T7 gp4 protein, while the animal homologues are quite distantly related. These results suggest that the bifunctional T7 gp4 homologue may be conserved in higher plants.

The *Arabidopsis* Twinkle protein may function both in mitochondria and chloroplasts, as this protein has been shown to be dual-targeted to both organelles [27]. These reports are based on the analysis of predicted N-terminal targeting sequences of a number of nuclear-encoded

Arabidopsis proteins fused with the GFP coding region. However, it has been shown that targeting of fusion proteins can be affected by the context of the N-terminal sequence with the GFP sequence [28,29]. A recent report on the maize plastid proteome has shown the presence of Twinkle in the chloroplast nucleoid [13].

Mitochondrial genomes range widely in size, from about 16.5 kbp in vertebrates and invertebrates, to 70–100 kbp in yeast and 200–2000 kbp in plants. The replication of animal mtDNA has been characterized in great detail, and in the original model each strand of the duplex DNA replicates at a different time, with the initial replication primed by a short transcript synthesized by the mitochondrial RNA polymerase [30]. The second strand replicates only when it becomes single stranded by progression of the first strand, allowing formation of a characteristic structure to facilitate replication initiation of this strand. In yeast and plants, mtDNA replication appears to be more complex, and may involve a recombination-dependent replication mechanism [23,31-34]. In this case DNA priming may not be required if invading strands provide the priming function for DNA synthesis. However, even in phage systems that replicate by a recombination mechanism a DNA primase is still required for priming synthesis at lagging strands during some phases of DNA replication [4].

A distinct mtDNA primase activity has been reported in some animal and protist cells and mtDNA primase activity has been reported in human cells, but no distinct human protein with this activity has yet been identified. It has been suggested that the DNA primase in animal cells is tightly associated with the mtDNA (γ) polymerase, and is thus difficult to isolate separately [35]. In a trypanosome a mtDNA primase of 70 kDa has been reported [36], while in yeast a mtDNA primase of 67 kDa has been characterized [37], which are both close to the size of T7 gp4 and Twinkle. Our understanding of animal mtDNA replication is complicated by reports of strand-coupled bidirectional replication from a single replication origin, which by its nature should require a DNA primase to synthesize primers for the lagging strand [30,38,39]. It is unclear whether a separate mtDNA primase is present or required in species (including human) with highly compact mitochondrial genomes [40]. Recently it has been shown that *in vitro*, human mitochondrial RNA polymerase is responsible for priming lagging strand mtDNA synthesis. It may be possible that priming of replication of the small animal mitochondrial genome is provided by short transcripts synthesized by the mitochondrial RNA polymerase [40,41].

A DNA primase has been purified and characterized from pea chloroplasts [42], and primers synthesized by that preparation are similar to primers synthesized by the purified *Arabidopsis* Twinkle homologue. The pea enzyme is larger (~90 kDa) than the *Arabidopsis* Twinkle homologue, but it was not characterized for DNA helicase activity. CtDNA replication involves multiple replication origins and bidirectional DNA synthesis [42,43], which would require DNA primase activity for lagging strand synthesis.

Organelle DNA replication appears to be different in plants (as compared to animals), which have very large and complex mitochondrial genomes and likely require multiple sites of lagging strand DNA synthesis. The role of recombination-mediated replication [33,34] may reduce the need for primase-synthesized primers for organelle DNA replication, as an invading DNA strand could provide the 3′ ends for DNA synthesis. However, even in this case it is likely that organelle DNA primase(s) is (are) required in plants. Bacteriophage T4 replicates by multiple mechanisms, including recombination-dependent replication, and requires a DNA primase. The observations that the *Arabidopsis* Twinkle protein is expressed at highest levels in the shoot apex and other developing tissues including young leaves provides strong support for a role of the Twinkle homologue in plant organelle DNA replication, similar to its role in other species [4,5].

Mutations in human Twinkle have been shown to lead to a drastic reduction in mtDNA copy number and disease [17]. RNAi-mediated reduction of Twinkle expression in cultured human cells was found to lead to a rapid drop in mtDNA copy number, while overexpression of Twinkle in mouse tissue was associated with an increase in mtDNA copy number [15,26]. In each of these cases the effect has been associated with the DNA helicase activity of the protein. We showed that this single protein from *Arabidopsis* has both DNA primase and DNA helicase activities in vitro, the same activity as the bacteriophage T7 gp4 protein.

CONCLUSION

The *Arabidopsis* homologue of the bacteriophage T7 gp4 protein has been shown to have both DNA primase and DNA helicase activities similar to the phage protein. It is expressed at highest levels in actively growing tissues, suggesting that it could play a role in organelle DNA replication. Two DNA polymerases have been identified in plants, and both have been reported to be dual targeted to mitochondria and chloroplasts [28,44]. It is likely that this *Arabidopsis* phage T7 gp4 homologue functions along with one or both of these DNA polymerases to accomplish organelle DNA replication. Even if the mtDNA replicates by a recombination-dependent mechanism as suggested by some [23,33,34], DNA priming may be required for lagging-strand DNA replication. This *Arabidopsis* protein may also play a role in control of plant mtDNA (and possibly also ctDNA) copy number as observed in animals [5,17], but this determination will require additional experiments, which will be the subject of future work in our lab.

METHOD

Identification of an *Arabidopsis* Twinkle Homologue

A full-length Twinkle homologue was identified in the *Arabidopsis thaliana* genome (At1g30680, protein molecular weight of 80,401.9 Da). A second, truncated homologue is also present (At1g30660, molecular weight of 37,806.9 Da) near the first gene, but contains only the primase domain of the protein and ends near the linker region [45] joining the primase and helicase domains. Only the full-length gene (At1g30680) was examined in this study.

Recombinant Expression of the *Arabidopsis* Twinkle Homologue

The full-length cDNA for At1g30680 was obtained from Riken (Japan). The full-length coding region for this gene predicts a polypeptide of 709 amino acids, and the MitoProt program [46] predicts the cleavage site after amino acid 91, which is prior to the conserved elements including the zinc fingers in the DNA primase domain of the protein. We generated a construct of the entire conserved coding region of the gene but lacking the DNA sequence for the N-terminal 91 amino acids in the pEXP5-NT/TOPO expression vector (Invitrogen). The construct was then transformed into the *E. coli* BL21 strain (Invitrogen). A total volume of 500 ml of LB was used to grow the bacteria. After it reached O.D.$_{600}$ 0.4-0.6, IPTG was added to the medium to a final concentration of 0.5 mM to induce the expression of the targeted protein. The cells were grown at 30^{0}C for an additional 4 hr and harvested by centrifugation. A control strain containing an empty vector lacking the gene insert was grown under identical conditions. The recombinant protein and control sample were purified under identical conditions using ProBond Nickel-chelating resin (Invitrogen). Native conditions were used and the purification was performed as described in the manual. Protein purity was analyzed by gel electrophoresis and western blot analysis.

DNA Primase Activity Assay

DNA primase activity of the recombinant protein was detected using a previously published procedure [42] using single-stranded M13 DNA as template. A control bacterial fraction was included to eliminate the possibility that bacterial DNA primase was present in the recombinant protein fraction. Single-stranded M13 DNA was incubated with 0.5 ng of the ProBond-purified recombinant or control protein fraction in the presence of rNTPs including α^{32}P-ATP (MP Biomedical). The reaction products were separated in a 20% denaturing polyacrylamide gel (6% urea in 1X TBE). End-labeled oligo(dT)$_{12–18}$ was used as size markers. After electrophoresis the gel was dried and exposed to X-ray film.

DNA Helicase Activity Assay

DNA helicase activity of the ProBond-purified recombinant protein was assayed according to the procedure of Song [47]. The substrate was prepared by annealing (heating for 5 min to 65^0C in 40 mM Tris–HCl, pH 7.8, 50 mM NaCl and slowly cooling to room temperature for 20–30 min) single-stranded M13 circular DNA with a complementary oligonucleotide (5′ GTAAAACGACGGCCAGT 3′) labeled at the 5′ end using T4 polynucleotide kinase (New England Biolabs) and γ^{32}P-ATP (MP Biomedical). The substrate was incubated with 0.5 ng of the recombinant protein in reaction buffer (10 mM Tris–HCl, pH 8.0, 8 mM $MgCl_2$, 1 mM dithiothreitol, 5 mM ATP, 1 ng ^{32}P-labeled helicase substrate) for 30 min, after which the reaction was terminated by adding EDTA to 2 mM, and the reaction products were separated by electrophoresis through a native TBE 6% polyacrylamide gel. The same bacterial protein control was included. The gel was then dried and exposed to X-ray film.

Western blot Analysis of Twinkle Homologue Expression in Different Tissues

Protein fractions were prepared from different tissues of *Arabidopsis thaliana* by grinding in liquid nitrogen and suspending in 1X SDS-loading buffer. The proteins were heated to 95^0C for 5 min and separated by electrophoresis in 8-20% SDS-PAGE gels. Proteins were transferred to PVDF membrane and after blocking in 5% skim milk the membrane was incubated with antibody that had been raised in rabbit (by Sigma-Genosys) against a synthetic peptide from a unique region of the Twinkle protein (KASRIVIATDGDGPG). This sequence is shared in both the full-length and truncated *Arabidopsis* genes (At1g30680 and At1g30660). The sequence of the peptide antigen was compared to the entire *Arabidopsis* proteome to ensure it does not share homology with any other protein besides the Twinkle homologues (NCBI-Blast). A control blot against the histone H3 protein was performed for normalization of signal strength. Bound antibody was detected using the Pierce Supersignal Western Chemiluminescence kit followed by exposure to X-ray film.

For time course analysis, total leaf tissue was extracted from *Arabidopsis* plants at weekly intervals starting at 1 week of age. The tissue was flash frozen in liquid nitrogen and stored at -80^0C. Total protein was extracted from 50 mg of crushed and homogenized tissue with 1X SDS-loading buffer [48]. Samples were quantified (BioRad RC DC protein assay kit) and normalized prior to electrophoresis by SDS-PAGE. Western blots were conducted as described above. Protein levels were determined by averaging mean pixel intensities measured with Un-Scan-It software (Silk Scientific, Orem, Utah) from three independent western blots.

Gene Expression Analysis by qRT-PCR

RNA was isolated using the PureLink RNA Mini Kit (Invitrogen) from young *Arabidopsis* seedlings. For very small tissues more than 200 young plants were used to obtain enough sample. Shoot apex tissues were taken as the very tip of the young shoots and include the apical meristem. The RNA was quantified and 1 μg was added to a reverse transcription reaction with SuperScriptIII (Invitrogen). The cDNAs from these reactions were diluted with 100 μl of sterile water and added to qPCR reactions as described by the manufacturer (Roche). qPCR reactions consisted of 1X SYBR Green PCR Master Mix (Roche), and 50 nM of each primer. Primers for the *Arabidopsis* Twinkle gene were 5′-TCCCCAGAGTCCCAACTCCTGTTGA-3′ and 5′-TCCCTGTTCCGCCAATTTACGCC-3′; for DNA polymerase gamma 1 (At3g20540) were 5′-CCTGAATACCGTTCACGTGCCCA-3′ and 5′-AGCCGCACTTCCCTGAACAGGA-3′, and for DNA polymerase gamma 2 (At1g50840) were 5′-TTCCGGCGTCAAAGTCACGTGC-3′ and 5′-TGCACTTCCCTGGACTGGAGTGT-3′. Reactions were carried out in a LightCycler 480 System (Roche) for 45 cycles (95°C for 10 secs, 58°C for 10 secs, 72°C for 20 sec) after initial 5 min incubation at 95°C. The fold changes in gene expression were calculated using the ΔΔCt method [49], with the Tub 4 tubulin gene (At5g44340) as an internal control.

Phylogenetic Analysis

Protein sequences for Twinkle homologues were downloaded from Gen Bank with the following accession numbers: *Homo sapiens* (NP_068602.2), *Caenorhabditis elegans* (F46G11.1), *Drosophila melanogaster* (NP_609318.1), *Plasmodium falciparum* (NP_702000.1), T7 gp4 (P03692.1), *Mus musculus* (AAL27647.1), *Canis lupus familiaris* (XP_543974.1), *Arabidopsis thaliana* (ACI49800.1), *Glycine max* (XP_003546288.1), and *Oryza sativa* Japonica group (BAD46002.1). Multiple sequence alignment was performed using MUSCLE [50] and the evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [51]. The tree with the highest log likelihood (−3556.6701) is shown. Initial trees for the heuristic search were obtained automatically as follows. When the number of common sites was < 100 or less than one fourth of the total number of sites, the maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 10 amino acid sequences. The coding data was translated assuming a standard genetic code table. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 199 positions in the final dataset. Evolutionary analyses were conducted in MEGA 5 [50].

ABBREVIATIONS

qRT-PCR, quantitative reverse-transcriptase PCR; mtDNA, mitochondrial DNA; ctDNA, chloroplast DNA; Twinkle, T7 gp4-like protein with intramitochondrial nucleoid localization

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR'S CONTRIBUTION

JDA performed the tissue-specific western blot and phylogenetic analyses and helped write the manuscript. BL helped conceive the project and made the recombinant protein construct and purified the protein, and performed some of the preliminary assays. JDC performed the qRT-PCR and western blot time course analyses. TH performed the DNA and amino acid sequence analyses. BLN helped conceive the project, performed the primase and helicase assays, and wrote the manuscript with JDA. All authors read and approved the manuscript.

ACKNOWLEDGEMENT

We thank Elisabeth Haws, Tatiana Alarcon, Cynthia Perry and Kali Tibbitts for assistance with some of the experiments. This research was supported in part by a grant from the NIH and by the BYU Mentoring Environments Grant Program.

REFERENCES

1. Kornberg A, Baker T: *DNA Replication, vol 2.* New York, New York: W.H. Freeman and Co.; 1991.

2. Lionnet T, Spiering MM, Benkovic SJ, Bensimon D, Croquette V: Real-time observation of bacteriophage T4 gp41 helicase reveals an unwinding mechanism. *Proc Natl Acad Sci USA* 2007, 104(50):19790–19795.

3. Tougu K, Peng H, Marians KJ: Identification of a domain of Escherichia coli primase required for functional interaction with the DnaB helicase at the replication fork. *J Biol Chem* 1994, 269(6):4675–4682.

4. Ilyina TV, Gorbalenya AE, Koonin EV: Organization and evolution of bacterial and bacteriophage primase-helicase systems. *J Mol Evol* 1992, 34(4):351–357.

5. Shutt TE, Gray MW: Twinkle, the mitochondrial replicative DNA helicase, is widespread in the eukaryotic radiation and may also be the mitochondrial DNA primase in most eukaryotes. *J Mol Evol* 2006, 62(5):588–599.

6. Spelbrink JN, Li FY, Tiranti V, Nikali K, Yuan QP, Tariq M, Wanrooij S, Garrido N, Comi G, Morandi L, Santoro L, Toscano A, Fabrizi GM, Somer, H, Croxen, R, Beeson D, Poulton J, Suomalainen A, Jacobs HT, Zeviani M, Larsson, C: Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nat Genet* 2001, 28(3):223.

7. Bernstein JA, Richardson CC: A 7-kDa region of the bacteriophage T7 gene 4 protein is required for primase but not for helicase activity. *Proc Natl Acad Sci USA* 1988, 85(2):396–400.

8. Patel SS, Picha KM: Structure and function of hexameric helicases 1. *Ann Rev Bioc* 2000, 69(1):651–697.

9. Korhonen JA, Gaspari M, Falkenberg M: TWINKLE Has 5' -> 3' DNA helicase activity and is specifically stimulated by mitochondrial single-stranded DNA-binding protein. *J Biol Chem* 2003, 278(49):48627–48632.

10. Seow F, Sato S, Janssen CS, Riehle MO, Mukhopadhyay A, Phillips RS, Wilson RJ, Barrett MP: The plastidic DNA replication enzyme complex of Plasmodium falciparum. *Mol Bioc Para* 2005, 141(2):145–153.

11. Leipe DD, Aravind L, Grishin NV, Koonin EV: The bacterial replicative helicase DnaB evolved from a RecA duplication. *Gen Res* 2000, 10(1):5–16.

12. Emanuelsson O, Nielsen H, von Heijne G: ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Prot Sci* 1999, 8(5):978–984.

13. Majeran W, Friso G, Asakura Y, Qu X, Huang M, Ponnala L, Watkins KP, Barkan A, van Wijk KJ: Nucleoid-Enriched Proteomes in Developing Plastids and Chloroplasts from Maize Leaves: A New Conceptual Framework for Nucleoid Functions. *Plant Physiology* 2012, 158(1):156-189.

14. Wanrooij S, Luoma P, van Goethem G, van Broeckhoven C, Suomalainen A, Spelbrink JN: Twinkle and POLG defects enhance age-dependent accumulation of mutations in the control region of mtDNA. *Nucleic Acids Res* 2004, 32(10):3053–3064.

15. Tyynismaa H, Sembongi H, Bokori-Brown M, Granycome C, Ashley N, Poulton J, Jalanko A, Spelbrink JN, Holt IJ, Suomalainen A: Twinkle helicase is essential for mtDNA maintenance and regulates mtDNA copy number. *Human Mol Gen* 2004, 13(24):3219–3227.

16. Tyynismaa H, Mjosund KP, Wanrooij S, Lappalainen I, Ylikallio E, Jalanko A, Spelbrink JN, Paetau A, Suomalainen A: Mutant mitochondrial helicase Twinkle causes multiple mtDNA deletions and a late-onset mitochondrial disease in mice. *Proc Natl Acad Sci USA* 2005, 102:17687–17692.

17. Sarzi E, Goffart S, Serre V, Chretien D, Slama A, Munnich A, Spelbrink JN, Rotig A: Twinkle helicase (PEO1) gene mutation causes mitochondrial DNA depletion. *Ann Neurol* 2007, 62(6):579–587.

18. Inoue T, Ide T, Tyynismaa H, Yoshida M, Ando M, Tanaka A, Todaka K, Kang D, Suomalainen A, Sunagawa K: Overexpression of mitochondria DNA helicase, Twinkle, ameliorates cardiac remodeling and failure in mice. *Circulation Res* 2008, 118(18):S314–S315.

19. Kusakabe T, Richardson CC: Gene 4 DNA primase of bacteriophage T7 mediates the annealing and extension of ribo-oligonucleotides at primase recognition sites. *J Biol Chem* 1997, 272(19):12446–12453.

20. Kusakabe T, Richardson CC: The role of the zinc motif in sequence recognition by DNA primases. *J Biol Chem* 1996, 271(32):19563–19570.

21. Kusakabe T, Hine AV, Hyberts SG, Richardson CC: The Cys4 zinc finger of bacteriophage T7 primase in sequence-specific single-stranded DNA recognition. *Proc Natl Acad Sci USA* 1999, 96(8):4295–4300.

22. Laquel P, Litvak S, Castroviejo M: Wheat DNA primase (RNA primer synthesis in vitro, structural studies by photochemical cross-linking, and modulation of primase activity by DNA polymerases). *Plant Physiol* 1994, 105:69–79.

23. Manchekar M, Scissum-Gunn K, Song D, Khazi F, McLean SL, Nielsen BL: DNA recombination activity in soybean mitochondria. *J Mol Biol* 2006, 356(2):288–299.

24. Parent JS, Lepage E, Brisson N: Divergent Roles for the Two PolI-Like Organelle DNA

Polymerases of Arabidopsis. *Plant Physiology* 2011, 156(1):254-262.

25. Cupp J, Nielsen B: *Arabidopsis thaliana* organellar DNA polymerase IB mutants exhibit reduced mtDNA levels with a decrease in mitochondrial area density. *Physiol Plantarum* 2012.

26. Tyynismaa H, Suomalainen A: Mouse models of mitochondrial DNA defects and their relevance for human disease. *EMBO Rep* 2009, 10(2):137–143.

27. Carrie C, Kühn K, Murcha MW, Duncan O, Small ID, O'Toole N, Whelan J: Approaches to defining dual-targeted proteins in Arabidopsis. *Plant J* 2009, 57(6):1128–1139.

28. Christensen A, Lyznik A, Mohammed S, Elowsky CG, Elo A, Yule R, Mackenzie SA: Dual-domain, dual-targeting organellar protein presequences in Arabidopsis can use non-AUG start codons. *Plant Cell* 2005, 17:2805–2816.

29. Mackenzie SA: Plant organellar protein targeting: a traffic plan still under construction. *Trends Cell Biol* 2005, 15:548–554.

30. Clayton D: Mitochondrial DNA Replication: What We Know. *IUBMB Life* 2003, 55(4–5):213–217.

31. Manchekar M, Scissum-Gunn KD, Hammett LA, Backert S, Nielsen BL: Mitochondrial DNA recombination in Brassica campestris. *Plant Sci* 2009, 177(6):629–635.

32. Backert S, Nielsen BL, Borner T: The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. *Trends Plant Sci* 1997, 2(12):477–483.

33. Backert S, Borner T: Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant *Chenopodium album (L.)*. *Current Gen* 2000, 37:304–314.

34. Oldenburg DJ, Bendich AJ: Mitochondrial DNA from the liverwort Marchantia polymorpha: circularly permuted linear molecules, head-to-tail concatemers, and a 5' protein. *J Mol Biol* 2001, 310(3):549–562.

35. Wong TW, Clayton DA: Isolation and characterization of a DNA primase from human mitochondria. *J Biol Chem* 1985, 260(21):11530–11535.

36. Hines J, Ray DS: A mitochondrial DNA primase is essential for cell growth and kinetoplast DNA replication in *Trypanosoma brucei*. *Mol Cell Biol* 2010, 30:1319–1328.

37. Murthy V, Pasupathy K: Characterization of mitochondrial DNA primase from *Saccharomyces cerevisiae*. *J Biosciences* 1994, 19:1–8.

38. Holt IJ, Lorimer HE, Jacobs HT: Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell* 2000, 100(5):515–524.

39. Bogenhagen DF, Clayton DA: Concluding remarks: The mitochondrial DNA replication bubble has not burst. *Trends Bioc Sci* 2003, 28(8):404–405.

40. Wanrooij S, Fuste JM, Farge G, Shi Y, Gustafsson CM, Falkenberg M: Human mitochondrial RNA polymerase primes lagging-strand DNA synthesis in vitro. *Proc Natl Acad Sci USA* 2008, 105(32):11122–11127.

41. Fuste JM, Wanrooij S, Jemt E, Granycome CE, Cluett TJ, Shi Y, Atanassova N, Holt IJ, Gustafsson CM, Falkenberg M: Mitochondrial RNA polymerase is needed for activation of the origin of light-strand DNA replication. *Mol Cell* 2010, 37(1):67–78.

42. Nielsen BL, Rajasekhar VK, Tewari KK: Pea chloroplast DNA primase: characterization and role in initiation of replication. *Plant Mol Biol* 1991, 16:1019–1034.

43. Tuteja N, Phan TN, Tewari KK: Purification and characterization of a DNA helicase from Pea chloroplast that translocates in the 3′-to-5′ direction. *Eur J Bioc* 1996, 238(1):54–63.

44. Ono Y, Sakai A, Takechi K, Takio S, Takusagawa M, Takano H: NtPolI-like1 and NtPolI-like2, bacterial DNA polymerase I homologues isolated from BY-2 cultured tobacco cells, encode DNA polymerases engaged in DNA replication in both plastids and mitochondria. *Plant Cell Physiol* 2007, 48:1679–1692.

45. Guo S, Tabor S, Richardson CC: The linker region between the helicase and primase domains of the bacteriophage T7 gene 4 protein is critical for hexamer formation. *J Biol Chem* 1999, 274(42):30303–30309.

46. Claros MG: MitoProt, a Macintosh Application for studying mitochondrial proteins. *Comput Appl Biosci* 1995, 11:441–447.

47. Song D: *Homologous strand exchange and DNA helicase activities in plant mitochondria.* Provo: Brigham Young University; 2005.

48. Weigel D, Glazebook J: *Arabidopsis: a laboratory manual.* Cold Spring Harbor, New York, USA: Cold Spring Harbor Laboratory Press; 2002.

49. Livak K, Schmittgen T: Analysis of relative gene expression data using real-time quantitative PCR and the 2-(−delta delta C(T) method. *Methods* 2001, 25:402–408.

50. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: *MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood.* Mol Biol Evol: Evolutionary Distance and Maximum Parsimony Methods; 2011.

51. Jones D, Taylor R, Thornton J: The rapid generation of mutation data matrices from protein sequences. *Comp App Bios* 1992, 8:275–282.