Fall 2014

Integrative high-throughput study of arsenic hyper-accumulation in Pteris vittata

Qiong Wu
Purdue University

Graduate School Form 30
(Revised 08/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Qiong Wu

Entitled
INTEGRATIVE HIGH-THROUGHPUT STUDY OF ARSENIC HYPER-ACCUMULATION IN PTERIS VITTATA

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Michael Gribskov

Daisuke Kihara

JoAnne Banks

Ann Rundell

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Michael Gribskov

Approved by: Richard J. Kuhn 12/11/2014

Head of the Department Graduate Program Date

INTEGRATIVE HIGH-THROUGHPUT STUDY OF ARSENIC HYPER-ACCUMULATION IN PTERIS VITTATA

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Qiong Wu

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ROS: reactive oxygen species

GST: glutathione S-transferase

GSH: glutathione

PC: phytochelatin

GO: gene ontology

KOG: Eukaryotic Orthologous Groups

UGT: UDP-glycosyltransferases

CYP: Cytochrome P450

NER: named entity recognition

SVM: support vector machine

TF-IDF: term frequency inverse document frequency

Pi: inorganic phosphate

GAPDH: glyceraldehyde-3-phosphate dehydrogenase

ABSTRACT

Wu, Qiong. Ph.D., Purdue University, December 2014. Integrative High Throughput Study of Arsenic Hyper-accumulation in *Pteris vittata*. Major Professor: Michael Gribskov.

Arsenic is a natural contaminant in the soil and ground water, which raises considerable concerns in food safety and human health worldwide. The fern *Pteris vittata* (Chinese brake fern) is the first identified arsenic hyperaccumulator[1]. It and its close relatives have un-paralleled ability to tolerant arsenic and feature unique arsenic metabolism. The focus of the research presented in this thesis is to elucidate the fundamentals of arsenic tolerance and hyper-accumulation in *Pteris vittata* through high throughput technology and bioinformatics tools. The transcriptome of the *P. vittata* gametophyte under arsenate stress was determined using RNA-Seq technology and Trinity *de novo* assembly. Functional annotation of the transcriptome was performed in terms of blast search, Gene Ontology term assignment, Eukaryotic Orthologous Groups (KOG) classification, and pathway analysis. Differentially expressed genes induced by arsenic stress were identified, which revealed several key players in arsenic hyper-accumulation. As part of the efforts to annotate differentially expressed genes, the literature of plant arsenic tolerance was collected and built into a searchable database using the Textpresso text-mining tool, which greatly facilitates the retrieval of biological facts involving arsenic related genes. In addition, an SVM-based named-entity recognition system was constructed to identify new references to genes in literature. The results provide excellent sequence resources for arsenic tolerance study in *P.vittata,* and establish a platform for integrative study using multiple types of data.

CHAPTER 1. INTRODUCTION

Arsenic is a natural metalloid that is ubiquitously present as an environmental contaminant. It is ranked as the 52nd most abundant element in the earth's crust and 26th in the ocean [2]. Arsenic is often found to associate with sulfur and metals in the forms of MAsS and $MAsS_2$, where M stands for Fe, Ni or Co [3]. Release of arsenic into the environment is through both natural phenomena such as weathering and volcanic emissions, and anthropogenic activities such as ore mining, smelting operations, and burning fossil fuels. Moreover, indiscriminate use of arsenical pesticides, herbicides, food additives and wood preservatives until the mid-90s has led to extensive contamination of agricultural and industrial land worldwide [4].

Arsenic contamination of soil and drinking water affects many regions of the world including the US, which raises global concerns about environmental health and food safety. Arsenic, especially its inorganic forms, is extremely toxic to most organisms, even at very low concentration. As is a Group 1 human carcinogen; human exposure to As is mainly through consumption of contaminated drinking water and plant-based food. Chronic exposure to As is associated with elevated dose-dependent risk of cancers, particularly skin, lung, and bladder cancers [reviewed in [5]]. Other reported long-term

effects of arsenic ingestion include skin lesions, neurotoxicity, cardiovascular diseases, abnormal glucose metabolism, and diabetes.

The increasing awareness of the deleterious effects of arsenic exposure on human health has led to a lowering of the guidelines for the amount of As in drinking water. The U.S. Environmental Protection Agency (EPA) has lowered the maximum As contaminant level in drinking water from 50 μg/L to 10 μg/L beginning January 23, 2006, and proposed a new standard value of 5 μg/L. The Word Health Organization (WHO) has also reviewed As guidelines in drinking water and established a provisional guideline of 10 μg/L to promote worldwide regulatory enforcement of higher standards for safe drinking water [6].

1.1 Arsenic in plants

Besides As contamination in drinking water, dietary arsenic intake is another major contributor to arsenic exposure to humans. Rice is of particular concern regarding the entry of As into food chain due to the massive scale of its consumption and its efficient assimilation from the edible portion. Rice is the staple food for billions of people worldwide, and the intake of As though rice ingestion can be substantial. The issue was first been recognized in regions with geographically-elevated arsenic concentration in groundwater such as Bangladesh, West Bengal (India), China, and Thailand [7, 8], where As-contaminated groundwater is also widely used for crop irrigation. In Bangladesh, one of the worst affected countries, cooked rice accounts for ~56% of the daily arsenic intake [9]. World market surveys have also revealed that rice grains contain considerable higher levels of inorganic As than other sources of food [10], and the total arsenic content was

even higher in samples from the U.S. and France, than those from India [11]. Given the significant consequences of arsenic exposure for human health, there is an urgent need to study the mechanisms of As assimilation and metabolism in plant, in order to develop agricultural and genetic techniques to minimize the uptake and translocation of As to the edible parts.

Moreover, understanding how plants take up and metabolize arsenic has important implications for phytoremediation of arsenic contaminated soils. Phytoremediation is the use of plants to eliminate or mitigate pollutants from the environment, which involves the combination of extraction, filtration, transformation, stabilization, and volatilization of the contaminants by plants. Traditional physical and chemical technologies for As remediation have not been very successful and cannot be applied to large areas [12]. On the other hand, much interest has developed in phytoremediation of arsenic since the discovery of the As-hyperaccumulating Chinese brake fern (*Pteris vittata*) [1]. Cropping As-hyperaccumulators [1, 13, 14] provides an *in situ*, large scale, cost-effective and eco-friendly alternative to chemecally detoxifying contaminated soils. The phytoremediation potential of hyperaccumulators has been tested in hydroponic environments, and the accumulation factor (ratio of arsenic concentrations in plant tissues to arsenic concentrations in the hydroponic solution) can be as high as 138 [15].

1.2 Chemistry and toxicity of Arsenic in Plant

Arsenic (atomic number 33; atomic weight 74.9216) belongs to subgroup Va of the periodic table, and its outer electronic configuration is $4s^2 4p^3$. Due to the transitional properties between nonmetallic and metallic groups, it is often described as a metalloid.

The most common oxidation states of As are -3 (arsine), 0 (arsenic), +3 (arsenite) and +5 (arsenate). Arsenate[As(V)] and arsenite[As(III)] are the predominant species in soil, depending on the surrounding redox state [4]. Arsenate dominates in aerobic conditions, whereas arsenite is the predominant form in less-aerated environments such as flooded rice paddies.

Arsenate closely resembles phosphate in many aspects and can replace phosphate in critical biochemical reactions. For example, pentavalent As(V) interferes with the synthesis of ATP by competing with phosphate during oxidative phosphorylation and forming unstable and short-lived ADP-As, and thus affects the cell energy cycle [16]. Upon absorption, As(V) is rapidly reduced to As(III), which is a more toxic form of As[17], by ACR2 arsenate reductases [18, 19]. The toxicity of As(III) is mainly due to its high sulfhydryl reactivity and the generation of oxidative stress. As(III) can bind up to three sulfhydryl groups [20], and such cross-linking ability can have profound effects on protein folding and potentially inactivate proteins [21, 22]. The Cys-rich binding targets of As(III) include transcription factors, signal transduction proteins, proteolytic enzymes, metabolic enzymes, redox regulatory enzymes, and structural proteins [reviewed in [22]].

Oxidative stress is another major factor contributing to plant arsenic toxicity. Exposure to inorganic arsenic gives rise to reactive oxygen species (ROS) such as superoxide ($O_2\bullet^-$), the hydroxyl radical (•OH), and H_2O_2 [17, 23, 24], which can damage proteins, nucleic acids, carbohydrates (e.g., cell wall polysaccharides), and cause peroxidation of membrane lipids [25]. The defense strategies for oxidative stress in plants involve both enzymatic reactions, such as induced production of catalase, superoxide dismutase (SOD), and increased synthesis of non-enzymatic antioxidants including

glutathione (GSH), phytochelatin (PC), ascorbate, carotenoids, and anthocyanin [reviewed in [26]]. Thiol-containing glutathione is an important cellular antioxidant and a precursor of phytochelatin. Its production is an essential process for detoxifying a range of metals and metalloids. Arsenite has high affinity for thiols such as glutathione (GSH) and phytochelatin (PC). Direct formation of As(III)-GSH or As(III)-PC complexes, and As(III)-induced PC synthesis, deplete the GSH pool in the cytoplasm, further reducing the amount of GSH available for quenching ROS and thereby indirectly increasing oxidative damage to the cell [17].

1.3 Arsenic uptake and transport

Arsenic can be present as either inorganic or organic species in the environment. Of the two inorganic forms, arsenate occurs predominantly as $H_2AsO_4^-$ and $HAsO_4^{2-}$ in aerobic environments, while arsenite (as $H_3AsO_3^0$ and $H_2AsO_3^-$) is more prevalent in anaerobic conditions like submerged soils. Many factors could affect the phyto-availability of As in soil: oxidation state of arsenic, soil properties such as pH and mineral content, the microbial community, the presence of other ions, etc. In terms of binding reactivity, arsenate can bind to most soil minerals and easily precipitate from the soil, while arsenite binding is selective and dependent on specific chemical conditions. For example, arsenate forms strong surface complexes on oxides/hydroxides of Al, Fe and Mn; and aluminosilicate may retain appreciable concentrations of arsenate [27]. In contrast to arsenate, arsenite exhibits a limited affinity for most soil minerals, with the exception of iron (hydr)oxides and magnetite, to which it binds more extensively than arsenate due to the formation of inner-sphere moieties [27]. As a result, the amount of

phytoavailable arsenic is very limited in aerobic soils because of the strong retention of As(V) by soil minerals. However, flooding of paddy soils leads to the reduction of As(V) to As(III) and the reductive dissolution of ferric oxyhydroxides, releasing the adsorbed or co-precipitated As back to the soil solution [28]. Meanwhile, inorganic arsenic species can be converted into organic forms by microbial methylation. Most common organic forms include mono-, di- and tri- methylated derivatives of As(V)/As(III). Methylated arsenic species may also originate from the remainder of As-containing pesticide or herbicides.

1.3.1 Arsenate uptake

Arsenic mainly exists as arsenate in aerobic soils. As an inorganic phosphate (Pi) analog, arsenate is taken up by plant roots via high-affinity phosphate transporters (Pht). Physiological and electrophysiological studies support the idea that arsenate shares and competes with phosphate for the same transport system: 1) suppression of high-affinity phosphate transporters decreases the uptake of arsenate [29, 30]; 2) increasing phosphate supply strongly inhibits the uptake of arsenate [8, 31, 32]; and 3) under low Pi conditions, arsenate may outcompete Pi for entry into the plant by repressing genes involved in the phosphate starvation response while inducing other As(V)-regulated genes [33]. Specific genes that could mediate the uptake of arsenate have been identified. For example, the *A. thaliana* mutant *pht1;1-1* displays enhanced As-resistance and better growth than wildtype [34]. The double mutant *pht1;1Δ4Δ*, which lacks two phosphate transporters expressed in roots, showed even stronger resistance to arsenate without much growth reduction, indicating that Pht1;1 and Pht1;4 mediate a significant portion of arsenate

uptake in *Arabidopsis* [34]. Different phosphate transporters differ in their affinity for arsenate. In a kinetic study of As(V) influx in ferns, given the same arsenate concentration in the growth media, the As-hyperaccumulating species *Pteris vittata* and *Pteris cretica* showed lower Michaelis-Menten kinetic parameters, K_m, than the non-accumulating *Nephrolepis exaltata* (L.), suggesting higher affinity of the transport protein for arsenate in hyper-accumulating ferns [35].

1.3.2 Arsenite uptake

Arsenite is the predominant species in anaerobic, reducing environments. Under normal soil conditions, arsenite remains mostly as neutral arsenous acid (H_3AsO_3, with pK_a=9.2, 12.1 and 13.4). It has been proposed that plants take up As(III) through Nodulin 26-like Intrinsic Proteins (NIPs), which belong to the aquaporin family of major intrinsic proteins (MIPs). Aquaporins are membrane channel proteins that allow the transport of water, small neutral molecules (glycerol, urea, boric acid, silicic acid), hydrogen peroxide, gases (ammonia, carbon dioxide, nitric oxide) [reviewed in [36]] and metalloids [reviewed in [37]]. It is worth mentioning that solute movement via aquaporins can be bidirectional, depending on the concentration gradient [37, 38]. Recent studies have demonstrated that a number of NIPs are permeable to arsenite [38-40], and the capacity of NIPs to transport arsenite is conserved across plant species [16, 38]. Expression of several Arabidopsis NIPs improves yeast growth on As-containing medium, probably due to increased As(III) efflux. In rice roots, the OsNIP2;1/OsLsi1 silicon transporter, which is constitutively expressed at the distal side of exodermal and endodermal cells, acts as the major Si and As(III) uptake protein. Rice mutants defective in Lsi1 show drastically

reduced short-term As(III) uptake at the root [40] as well as reduced arsenic accumulation. Expression of Lsi1 in *Xenopus* oocytes and yeast increases arsenite transport activity by 3-5 folds. However, the long-term impact of Lsi1 deficiency on arsenic accumulation, in shoots and grain of field grown rice, is less prominent than Lsi2, another important arsenite transporter in rice.

Lsi2, a previously known silicon effluxer, which is localized at the proximal side of the same cells as Lsi1, has been shown to function as an efflux carrier of arsenite from the exodermal and endodermal cells into root stele and vascular tissue [40]. Lsi2 is not an aquaglyceroporin but is distantly related to ArsB, the bacterial arsenite efflux protein [40]. Loss of function of Lsi2 significantly affects As accumulation but not short-term uptake. In comparison with wildtype, two independent *lsi2* mutants show markedly reduced arsenite concentration in xylem sap (73% and 91% lower) and grain (63% and 51% lower) [40]. These results indicate that Lsi2 plays a more crucial role than Lsi1 in translocating As to the shoots and ultimately to the grain. As a whole, Lsi1 and Lsi2 work together to facilitate the uptake of silicon and arsenite from the soil into root cells, and efflux it to the stele.

1.3.3 Uptake of methylated As species

A number of methylated arsenic species are also present in small amounts in the soil, and they may originate from either the residue of As-containing pesticides/herbicides, or the transformation of inorganic arsenic through microbial methylation. For example, methylated pentavalent arsenic species such as monomethylarsonic acid (MMA) and dimethylarsinic acid (DMA) are widely used as

herbicides for weed control on cotton, orchards, and lawns, or as a defoliant of cotton (U.S. Environmental Protection Agency, [41]). Li et al. [42] found that MMA and DMA can enter rice roots through the aquaporin channel OsLsi in the protonated, neutral forms, although the uptake efficiency is much lower than that of inorganic species [8]. Surprisingly, the other important player in the silicon pathway, OsLsi2 is not involved in the efflux of DMA or MMA toward the stele [42]. Given that MMA and DMA have relatively low *pKa* (4.2 and 6.1, respectively), they can easily dissociate in alkaline conditions. Increasing the external pH would lead to significant dissociation of MMA and DMA, and thus a smaller portion of uncharged molecules available for transport through aquaporin channels, and less uptake. Li et al. [42] showed that the uptake MMA(V) and DMA(V) in rice seedlings increases with decreasing medium pH due to the increasing portion of undissociated molecules, which confirms that only neutral methylated As species can be taken up by transporters in rice roots. Opposite to the root uptake, upon absorption, methylated arsenic has much greater mobility in plant tissues [42, 43], and the translocation of As species from roots to shoots was in the order TMA(V) > DMA(V) > MMA(V) > As(V) [44]. During grain filling, DMA is transported to the grain through both the phloem and the xylem with substantially greater efficiency than arsenite. When As species were fed directly to the flag leaves, DMA(V) and MMA(V) were efficiently translocated to the grain, whereas arsenate was rapidly reduced within the flag leaves and retained as arsenite [45].

1.3.4 Long distance transport

The mobility of As from roots to shoots is limited in most plants, except for

hyperaccumulators. Wild-type *A.thaliana* only translocates 2.6% of As taken up by roots to the shoots when exposed to arsenate [46]. Raab et al. [43] examined the uptake and translocation of As(V), MMA and DMA from roots to shoots in 46 plants species and reported that the root-to-shoot transfer factor (TF, the ratio of shoot As dry weight, and root As dry weight) in arsenate-treated plants ranged from 0.01 to 0.9, with a median of 0.1. Despite the low uptake rate of DMA, it was translocated more efficiently with TF ranging from 0.02 to 9.8, with a median of 0.8.

In most plant species studied, arsenite is the predominant form of As found in the xylem sap, accounting for approximately 60-90% of total As, regardless of the form of arsenic that is supplied to plant roots [16]. Although As is also present as arsenate in the xylem sap, studies with phosphate transporter mutants of *A. thaliana* suggest that arsenate is not the major form loaded into the xylem. Mutations in AtPho1, a xylem loading phosphate transporter, showed no effect on root-to-shoot As distribution in *A. thaliana* [46]. The pho2 mutant accumulates excessive Pi, but not arsenate, in the shoots [46]. It appears that As is loaded into the xylem mainly as free arsenite, and no As-thiol complexes were detected in the xylem sap [47], which is consistent with the fact that roots have a high capacity for arsenate reduction. After the rapid conversion of arsenate to arsenite, non-hyperaccumulating plants either sequester As in root vacuoles, or efflux As to the environment. In contrast, hyperaccumulating plants have evolved more efficient root to shoot translocation mechanisms that may play an important role in their hypertolerance. Rice loads arsenite into xylem through the highly expressed Lsi2 silicon transporter and shows greater efficiency in the root-to-shoot translocation than other cereal crops [48]. In the As hyperaccumulator *P. vittata*, As is rapidly transported as

As(III) into the fronds, where it is sequestered and accumulated as free As(III) in vacuoles [49, 50]. The exceedingly efficient As translocation may be crucial to its hypertolerance, however, the underlying mechanisms remain to be elucidated.

1.4 Arsenic metabolism

1.4.1 Arsenate reduction

Arsenite is the dominant form of As in plant tissues even when arsenate is supplied [51-53], indicating that arsenate reduction may be the first step of intracellular arsenic metabolism. Arsenate reduction is carried out by specific arsenate reductases. Plant arsenate reductases have been identified in *A. thaliana* [19], *Holcus lanatus* [18], rice [54], and *P. vittata* [55]. These proteins are homologs to CDC25-like (cell division cycle) tyrosine phosphatases, which often exhibit both phosphatase and arsenate reductase activities. The exception is PvACR2 from *P.vittata*, which only has arsenate reductase activity [55]. In vitro experiments demonstrated that plant ACR2s can catalyze arsenate reduction using GSH and glutaredoxin as reductants [18, 55]. Expression of *Arath;CDC25(AtACR2)* in *E.coli* mutant lacking *ArsC*, suppresses the As sensitivity due to the lack of an endogenous arsenate reductase [19]. However, As(III) still dominates As speciation in *Arabidopsis* knockdown lines of *AtACR2*, suggesting there are functional redundancy of arsenate reduction, or alternative non-enzymatic reduction mechanisms in plants. It has been shown that *P. vittata* cytosolic triosephosphate isomerase directly or indirectly functions as an arsenate reductase [56].

1.4.2 Detoxification and sequestration of arsenic

Given that arsenite has high affinity for the sulfhydryl (-SH) groups, complexation of arsenite by thiol compounds such as glutathione (GSH) and phytochelatin (PC) is a major detoxification mechanism for cytoplasmic arsenic in non-hyperaccumulating plants. As-PC complexes have been isolated from arsenate treated plant tissues [57, 58], which are dominated by GS-As(III)-PC_2 and As(III)-PC_3.[59] The biosynthesis and short term accumulation of PCs is significantly induced by arsenate exposure [57, 58]. Gene and enzymes involved in synthesis, metabolism and transport of the PC precursor GSH are up-regulated during arsenate treatment [60]. On the other hand, application of a PC synthase inhibitor increases sensitivity to As [61, 62]. The *Arabidopsis* mutant *cad1-3*, which is impaired in PC synthesis, is 10-20-fold more sensitive to arsenate than wild type [63]. This strongly suggests the essential role of PCs in As detoxification, particularly in As non-hyperaccumulators. Notably, only a small portion (1-3%) of the As in *P.vittata* was found to be chelated with PCs, indicating that PC-based detoxification contributes little to As hyperaccumulation [64].

The As-PC complex is ultimately removed from the cytoplasm by storage within vacuoles. ATP-binding cassette (ABC) family proteins are the major players in transporting metal complexes across membranes. The vacuolar transporter, Ycf1p in yeast, confers arsenite resistance by transporting the As(III)-$(GS)_3$ into the vacuole [65]. Yeast HMT1, another member of the ABC family, transports Cd-PC complexes into the vacuole [66], and may also transport As(III)-PC complexes. In *P. vittata* fronds, As is stored in the vacuoles mainly as free arsenite [49]. An arsenite-specific transporter PvACR3 [67] has recently been isolated from the vacuole membrane, which mediates the

efflux of arsenite into vacuoles, and has been proven to be essential to arsenic tolerance in *P.vittata*.

1.4.3 As hyperaccumulation

Since the first discovery of in *P.vittata* as an As hyperaccumulator [1], more members of the *Pteridaceae* family, especially within the genus *Pteris*, were found to hyperaccumulate As [reviewed in [68]]. In non-hyperaccumulating plants, As tolerance is generally achieved through two mechanisms: 1) reduced As intake by suppression of the high-affinity phosphate transporter [69]; and 2) sequestration of arsenite in root vacuoles by glutathione and PC conjugation [69]. Hyperaccumulators have adapted different strategies to cope with excessive As. *P.vittata* exhibits both higher As uptake rate at roots, and enhanced arsenic translocation to the above-ground portion of the plant. The ratio of the As concentration in the xylem sap of *P. vittata* to that in the nutrient solution was about 2 orders of magnitude higher than that in the nonhyperaccumulators [38]. Energy dispersive X-ray microanalyses (EDXA) also revealed that 96% of total As was located in the pinnae [49], indicating efficient translocation of arsenic from roots to the fronds. Enhanced vacuolar sequestration in fronds is another key mechanism of As detoxification in hyperaccumulators, which is exemplified by the positive linkage between vacuolar transport of As(III) by PvACR3 and the arsenic tolerance ability of *P.vittata* [67].

1.5 Objectives of this study

Understanding the ability of *P. vittata* to hyperaccumulate arsenic has great implications for the design of phytoremediation strategies, and the genetic engineering of

safer food crops. Efforts have been made to elucidate the mechanisms of arsenic hyperaccumulation in *P.vittata* at the molecular level, and important genes have been characterized. However, due to the lack of a genome sequence or mutant library, most genetic studies of *P.vittata* conducted so far were carried out at the single-gene level. The molecular mechanisms underlying arsenic tolerance and hyperaccumulation are still poorly understood. The purpose of the studies presented in this thesis is to investigate the fundamentals of arsenic tolerance and hyper-accumulation in *P. vittata* using high throughput sequencing and bioinformatics tools. We focus on the identification and characterization of As-induced modulation of the *P.vittata* gametophyte transcriptome. Chapter 2 describes the *de novo* assembly of the gametophyte transcriptome from RNA-Seq short read data. Chapter 2 also addresses the identification and characterization of differentially expressed genes under As treatment. Chapter 3 describes the implementation of a text-mining system for arsenic tolerance literature, which could facilitate gene function annotation and discovery of linkage between genes. Chapter 4 proposes a statistical machine learning method that recognizes gene mentions from texts. Identification of additional gene names could improve the indexing of literature in Textpresso, and thus enhance the ability of fact retrieval.

CHAPTER 2. TRANSCRIPTOME ASSEMBLY AND DIFFERENTIAL EXPRESSION ANALYSIS OF PTERIS VITTATA IN RESPONSE TO ARSENIC STRESS

2.1 Introduction

Arsenic is a natural contaminant in the soil and ground water and is the focus of considerable concern in food safety and human health worldwide. Arsenic is extremely toxic to most organisms at very low concentrations (parts per billion), and it is classified as a group 1 human carcinogen. Human exposure to As occurs mainly through consumption of contaminated drinking water and plant-based food. Chronic exposure to As has been associated with dose-dependent elevated risk, particularly skin, lung, and bladder cancers [reviewed in [5]].

The fern *Pteris vittata* (Chinese brake fern) is the first identified arsenic hyperaccumulator [1]. It is highly tolerant to normally toxic concentrations of arsenic and accumulates arsenic up to 2% or more of its dry weight [55]. Several other fern species in the order *Pteridales*, including *Pityrogramma calomelanos* [70], *Pteris cretica*, *Pteris longifolia* and *Pteris umbrosa* [68], have also been reported to have similar abilities to hyperaccumulate As.

More interestingly, the *P.vittata* sporophyte appears to have a unique mechanism for efficiently translocating arsenic from the root to the fronds where it is stored in vacuoles. This distinctive property has raised the possibility of using *P. vitatta* in

phytoremediation of As-contaminated areas [71]. Because of its extraordinary tolerance to arsenic, *P.vittata* has been the focus of extensive study of arsenic uptake, metabolism, and translocation [55, 72-77].

Inorganic arsenic occurs predominantly as two species, arsenate[As(V)], and arsenite[As(III)], depending on the redox environment [4]. As a phosphate analog, arsenate can be taken up via high-affinity phosphate transporters [69], while arsenite is taken up by aquaporin transporters [78]. In non-hyperaccumulating plants, As tolerance is generally achieved through two mechanisms: 1) suppression of the high-affinity phosphate transporter, thereby reducing As uptake [69]; and 2) restricted translocation of arsenate by rapid reduction to arsenite, and subsequent conjugation and sequestration of arsenite in root vacuoles using thiol-containing compounds such as glutathione and phytochelexins (PCs) [69]. On the contrary, *P.vittata* exhibits both a higher As uptake rate, and enhanced arsenic translocation to the above-ground portion of the plant. Energy dispersive X-ray microanalyses (EDXA) revealed that 96% of total As was located in the *pinnae*, probably compartmentalized in the vacuoles of the upper and lower epidermal cells [49], indicating efficient translocation of arsenic from roots to fronds. Only a small portion (1-3%) of the As in *P.vittata* was found to be chelated with PCs, which suggests PCs contribute little to As transport in *P.vittata* [64].

Efforts have been made to elucidate the mechanisms of arsenic resistance in *P.vittata* at the molecular level, and important genes have been characterized. The arsenate reduction to arsenite in *P.vittata* is mainly catalyzed by PvACR2 [55]. Unlike its homologue in *Arabidopsis*, PvACR2 exhibits only arsenate reductase activity and lacks phosphatase activity, which is probably linked to a change in a critical residue that

defines the active site [55]. The *P. vittata* cytosolic triosephosphate isomerase also has shown to either directly or indirectly function as an arsenate reductase [56], indicating functional redundancy of arsenate reduction in *P. vittata*. An arsenite-specific transporter PvACR3 [67] has been isolated from the vacuole membrane where it mediates the transport of arsenite into the vacuole, and was proven to be essential to arsenic tolerance. A glutaredoxin-coding cDNA (Grx) has also been identified in *P.vittata* fronds, and found to be involved in regulating intracellular arsenite levels and thus arsenic resistance [79].

Only a limited number of genes that are linked with arsenic resistance have been identified so far. The details of most of the processes underlying arsenic tolerance and hyperaccumulation still remain to be elucidated. Genome-scale next-generation sequencing technologies such as RNA-Seq, now offer an alternative approach to investigating these mechanisms from a global point of view, providing a powerful approach to studying the transcriptome and enabling the identification of changes in gene expression triggered by arsenic. Here we apply the RNA-Seq technique to characterize arsenic-induced changes in the *P.vittata* gametophyte transcriptome. The gametophyte transcriptome was reconstructed *de novo*, and genes whose expression differs significantly between As(V)-present and As(V)-absent conditions were identified. This study provides a valuable genome resource to the fern community and sheds light on the fundamental basis of arsenic tolerance in *P.vittata*.

2.2 Material and Methods

2.2.1 Plant material preparation and Illumina sequencing

The origin of *Pteris vittata* sporophyte has been previously described [80]. Sporophylls from each plant were placed in glassine bags for 2 weeks; spores released within each bag were collected and stored at room temperature. Collected spores were soaked overnight in sterile double-distilled (dd) H_20, surface sterilized in a solution containing 50% bleach and 50% Tween for 5 min, and rinsed four times in sterile ddH_20. Gametophyte culture medium contained 0.5× MS salts (Sigma M5524, St. Louis MO), pH 6.5. When required, medium was solidified with 0.65% agar (Sigma A9915) prior to autoclaving. Arsenate stock solutions were prepared from monobasic anhydrous potassium arsenate (Sigma A6631) dissolved in 18 MΩ water, sterilized by filtration through a 0.2 μm cellulose acetate filter, and, where necessary, added to previously autoclaved medium. Spores were grown in medium containing petri dishes at 28°C in a growth chamber. One month later, 6 dishes of gametophytes were covered with liquid medium containing ddH_2O, and 6 dishes were covered with liquid medium containing 10 mM KH_2AsO_4. After 24 hours, gametophytes were harvested and frozen in liquid nitrogen. Samples were subsequently stored at -80°C.

In each treatment, gametophytes from two dishes of the same treatment were combined into one sample for mRNA extraction and library preparation, resulting in 3 replicates for each condition. Frozen gametophyte tissue was ground for at least 30 min in liquid nitrogen with a mortar and pestle. RNA extractions were performed using the RNeasy Plant Mini Kit (QIAGEN, Hilden, Germany). The TruSeq kit (Illumina, Inc, San

Diego CA) was used to prepare cDNA libraries for sequencing. Libraries were sequenced in 2 lanes on an Illumina HiSeq2000 platform producing 100 nt paired-end reads.

2.2.2 Reads cleaning and *de novo* assembly of transcriptome

A series of cleaning steps were conducted to prepare the reads for transcriptome assembly. Raw reads from all samples were checked for alignment to bacterial, viral, rRNA, mitochondrial RNA, and chloroplast DNA using DeconSeq 0.4.1 [81]. Suspicious contaminant reads with greater than 75% identity and 50% coverage were removed. Illumina adapter sequences were removed by a custom perl script, and Trimmomatic 0.22 [82] was used to trim bases with quality scores of 10 or lower from the 3′-end of each read, as well as windows with an average quality < 13 in a window of five bases. Cleaned reads with a minimum length of 30 nt after trimming were kept for transcriptome reconstruction. *De novo* assembly of the *P.vittata* transcriptome was carried out using the Trinity package r 2012-06-08 [83] with k-mer size of 25, and a minimum contig length of 150.

2.2.3 Assessment of the completeness and quality of transcriptome

The quality and completeness of the *P.vittata* transcriptome assembly was assessed in three ways: comparison with known plant protein references, searching for the presence of core eukaryote proteins, and a chimerism test. All assembled transcripts were searched against 4 reference protein databases, including *Arabidopsis thaliana, Oryza sativa, Selaginella moellendorffii,* and *Physcomitrella patens* using blastx [84]. Additionally, we searched the assemblies against *A. thaliana* core proteins in Core

Eukaryotic Gene Mapping Approach (CEGMA) dataset [85] using tblastn [84]. The CEGMA set contains 458 highly conserved proteins that are found universally in eukaryotes, and can be used to evaluate the completeness of a transcriptome. The transcriptome assembly was also tested for the presence of chimerism. To identify potential incorrect assemblies, we first identify the set of unique Arabidopsis proteins by comparing the TAIR10 protein database (TAIR; http://www.arabidopsis.org, TAIR 10 release) against itself using blastp. Unique proteins are the ones that only match to themselves. The assemblies were compared to the set of unique proteins using blastx [84], and those that aligned with two or more different unique proteins in disjoint loci were considered as potential chimaeric.

2.2.4 Expression estimation and statistical analysis

Cleaned reads were divided into paired and single reads and separately aligned to the assembled transcriptome. Reads were aligned with Bowtie 0.12.8 [86]. The bowtie mapper requires a valid alignment to have both of the paired reads matched to the transcriptome in correct orientations, within the size range of the insert. Only one mismatch was allowed per 25 nucleotides, and reads with more than 50 matches in different locations of the transcriptome were discarded. The number of reads corresponding to each predicted transcript in each sample was estimated with RSEM 1.1.23 [87]. RSEM estimates the number of aligned reads at both the ‘isoform-level’ and ‘gene-level’, where the Trinity component is considered to correspond to a gene and individual predicted transcript assemblies are considered to be isoforms. For each isoform or gene, counts of aligned reads estimated from paired and single reads were added

together to get the total number of aligned reads. Only predicted transcripts with more than 5 aligned reads in each of three (or more) samples were considered to be reliably determined. Predicted transcripts with fewer counts were not included in future analyses. We applied and compared the results of three statistical packages in terms of identifying genes with differential expression: edgeR [88], DESeq [89] and EBSeq [90]. *P*-values computed by edgeR [88] and DESeq [89] were corrected for multiple comparisons by the Benjamini and Hochberg method [91] to control the overall false discovery rate (FDR). The EBSeq model takes multiple comparisons into account, thus the output posterior probability of being differential expression is equivalent to (1-FDR) and can be directly used for screening.

GO term enrichment tests were performed by topGO [92] Bioconductor package (http://www.bioconductor.org/packages/release/bioc/html/topGO.html). GO terms with p-value less than 0.05 in the one-sided Fisher exact test were deemed overrepresented in the differentially expressed genes.

2.2.5 Quantitative real-time PCR analysis

10 genes were selected to validate the results from differential expression analysis. Total RNA was reverse transcribed into single-stranded cDNA using the Tetro cDNA Synthesis Kit (Bioline, MA). Real-time RT–PCR was performed in the StepOne Plus Real-Time PCR System (Applied Biosystems, NY) using Quickstart with default parameters. Approximately 3ng cDNA was used as template in 20 µl reactions with SYBR green PCR Mater Mix (Applied Biosystems, NY). At least two biological replicates for each template were performed. PCR conditions were: 20 minutes of pre-

denaturation at 95 ℃, 40 cycles of 3 seconds at 95 ℃ and 30 seconds at 60 ℃ followed by generation of melt curves (15 seconds at 95 ℃, 60 seconds at 60 ℃, and 15 seconds at 95 ℃). Relative expression was determined with the $2^{-\Delta\Delta T}$ method [93] by normalizing to the amount of *Actin* (GenBank accession number KC46369.7), which was constitutively expressed regardless addition of arsenate. Reactions without template added served as negative controls. The ΔCt method was used in calculating relative fold changes. Melt curves were generated and evaluated to ensure the absence of multiple peaks. The primers used for real-time RT–PCR are listed in Table 2.1.

2.2.6 Functional annotation and classification of the *P.vittata* transcriptome

The transcript assemblies were compared with sequences in the NCBI non-redundant protein database (http://www.ncbi.nlm.nih.gov), Swiss-Prot protein database (http://www.expasy.ch/sprot), the *Arabidopsis* protein database (TAIR; http://www.arabidopsis.org, TAIR 10 release), and all plants sequences in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database using blastx [84] with a cutoff E-value of 10^{-5}. Likely coding regions were extracted from transcripts using utilities included in the Trinity package [83], out of which the best candidate open reading frames were translated into protein sequences. We then ran HMMER [94], signalP [95], and TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/), which are included in the Trinotate package (http://trinotate.sourceforge.net/), on the translated protein sequences to identify protein domains, potential signal peptides, and likely transmembrane regions respectively. Gene ontology (GO) terms were assigned to each transcripts based on the blastx comparisons to the Swiss-Prot database using b2g4pipe (v2.5.0), a command line

version of the Blast2GO suite (https://www.blast2go.com/blast2gocli). The top 20 Blast hits, with a cutoff E-value of 1e-6 and similarity cut-off of 55% were used for GO annotation, and the annotations were further processed in the ANNEX step [96], where extra annotations can be retrieved by exploring the relationships between GO terms of different categories. Generic GO terms assigned to each transcripts were then mapped onto the plant specific GOSlim set (http://geneontology.org/page/go-slim-and-subset-guide) by GOSlimViewer in AgBase [97] and by custom Perl scripts.

2.2.7 KOG analysis

Open reading frames (ORFs) for each predicted transcript were extracted by the getorf function in the EMBOSS suite [98]. Only ORFs longer than 90 nucleotides were extracted as translated protein sequences. KOG annotations were obtained by submitting the longest ORF protein sequence for each predicted transcript to the WebMGA server [99] (http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/kog/) with blastp E-value cutoff of 10^{-5} . The outputs were downloaded and analyzed by in-house developed Perl script.

2.3 Results

2.3.1 *De novo* assembly and quality assessment

6 samples were sequenced in 2 lanes of an Illumina HiSeq 2000 flowcell. We obtained approximately 373 million 100-bp paired-end reads. The raw reads were subject to a series of cleaning steps including removing sequence contamination by DeconSeq [81], removing adapter sequences, and removing low-quality bases from read ends using

Trimmomatic [82]. After read cleaning, the remaining 287 million reads were *de novo* assembled with the Trinity package [83]. The assembled transcriptome contained 344,048 predicted transcript assemblies in 190,495 component groups, with lengths ranging from 151 to 17,840 bases, and an N50 of 2,125 bases. A summary of the assembly statistics is given in Table 2.2. 342,544 transcript assemblies have at least one open reading frame (ORF) longer than 90 bases, of which 129,150 (37.7%) have ORFs larger than 300 bases. The length distributions of transcript and the longest ORFs are shown in Figure 2.1.

The quality of the assembled transcriptome was first evaluated by comparing to the predicted transcripts to four plant proteomes: *Arabidopsis thaliana*, *Oryza sativa*, *Selaginella moellendorffii*, and *Physcomitrella patens* using blastx. At a significance level of $E<10^{-5}$, 109,099 (31.71%) predicted transcripts have matches in at least one of the references, and 80,352 (23.35%) have matches in all 4 references (Fig. 2.2.). 82.5% of the *S. moellendorffii* proteins have matches in the assembly, followed by 79.7% of *A. thaliana*, 66.2% of *O. sativa* and 61.1% of *P. patens*. We further applied the CEGMA [85] pipeline to assess the completeness and contiguity of the assembly. CEGMA searches the transcriptome assembly for the presence of a collection of highly conserved single-copy genes and also computes the coverage of each conserved gene. 91.05% of the core eukaryotic genes (http://korflab.ucdavis.edu/Datasets/cegma/) from *Arabidopsis thaliana* were mapped to the predicted transcripts with coverage > 70% of the full protein sequences at E-value of 10^{-50}, and at E-value $<10^{-5}$, all core genes were mapped. With the default setting, CEGMA analysis revealed that 99.6% of the core genes were complete and 100% were partially present in the assembly. These results together suggest that the

Trinity transcriptome assembly contains the transcripts of most known protein-coding genes, and has a good coverage to the full length of coding region.

In order to detect potentially chimeric assemblies, a set of 8401 unique *Arabidopsis* proteins were identified by comparing the TAIR10 protein dataset (http://www.arabidopsis.org, TAIR 10 release) to itself using blastp. The assemblies were mapped to the identified unique Arabidopsis proteins using blastx with an Evalue cutoff of 10^{-5}. Only 3347 (0.973%) predicted transcripts aligned to two or more different unique proteins in non-overlapping loci, and thus were considered as potential chimaeric. This low degree of chimerism was considered to be negligible.

2.3.2 Transcriptome profiling and annotation

The predicted transcripts in the Trinity assembly were annotated by comparing the sequences to the NCBI non-redundant protein database (nr), Swiss-Prot protein database, and TAIR 10 protein database. 124,045 (36.05%) of the 344,048 predicted transcripts have at least one hit in the nr database, 101,380 (29.47%) showed matches in Swiss-Prot, and 91,469 (26.59%) have significant similarity to at least one sequence in the TAIR10 proteome at $E<10^{-5}$. A total of 126,329 predicted transcripts presented at least one significant match in the databases mentioned above. Given that Trinity reports multiple predicted isoforms per gene, this number seems reasonable.

To classify the functions of the predicted transcripts, generic gene ontology (GO) terms were assigned to each sequence by the b2g4pipe (v.2.5.0) program, which is a command line version of the blast2go suite [100]. 87,238 (25.36%) out of 344,048 predicted transcripts yielded significant gene ontology (GO) annotation based on blastx

comparison to Swiss-Prot sequences with an Evalue cutoff of 10^{-6}. There were a total of 11,039 gene ontology terms associated with the assembled transcriptome. Of these, 6547 (59.31%) assignments were denoted as biological process, followed by molecular function (3185, 28.85%) and cellular components (1307, 11.84%). These assemblies were further categorized into groups by mapping them to the plant specific GO slim set (http://geneontology.org/ontology/subsets/goslim_plant.obo). GO slims are reduced set of higher-level GO ontologies, which provide a broad overview of the functional distribution of the assigned GO terms. The Trinity transcriptome assembly covers 99 of the 100 plant GO slim terms (Fig. 2.3).

An additional functional annotation of the assembly was performed by searching for putative orthologs and paralogs in the KOG database [101]. A total of 63,254 (18.38%) predicted transcripts were assigned to 26 eukaryotic orthologous groups (Fig. 2.4). The category of signal transduction mechanisms is the most abundant in the annotated transcripts, accounting for 14.63% of the annotated transcripts, followed by general function prediction (13.56%) and posttranslational protein modification (9.36%).

2.3.3 Analysis of differentially expressed genes

Trinity [83] reconstructs the transcripts by starting with a greedy extension from the most abundant *k*-mers to join overlapping kmers into components and a de Bruijn graph is built for each component. Reads are then assigned to their best matching component. The component graphs are trimmed and corrected according to the read-mapping information, and may be broken into several disconnected subcomponents. After graph cleanup, sequences of possible isoforms/contigs are extracted from each

component or subcomponent. All possible isoforms that can be constructed using all possible alternative splice site are reported by Trinity. Here we treated Trinity components as genes, and contigs within the same component as alternative isoforms. Counts for each gene were derived from summing over counts of all transcript assemblies from the same trinity component. These counts serve as the raw input to statistical packages used for detecting differentially expressed genes (DEGs). Differentially expressed genes were defined as those with an adjusted P-value (FDR equivalent) less than 0.2, and at least 2-fold change in average expression level between treatment conditions. Three statistical methods were used to identify DEGs, namely DESeq [89], EdgeR [88] and EBSeq [102]. DESeq is the most conservative one, and identified 10 DEGs, all of which were found by the other two methods. EBSeq identified 15 DEGs. EdgeR identified 163 DEGs, which include all that were found by DESeq and EBSeq. At FDR of 0.2, the three packages together identified 163 differentially expressed Trinity components having absolute fold-change greater than 2. Among these identified DEGs, 57 were up-regulated and 106 were down-regulated by arsenic treatment. The difference in the numbers of detected DEGs likely arises from different model assumptions, FDR controls, and sensitivities to outliers in the tested approaches [103]. At $E<10^{-5}$, out of the 163 DEGs, 331 (50.77%) predicted transcripts from 57 (35.18%) genes have at least one match in the Swiss-Prot database. 9 predicted transcripts from 3 additional genes mapped to *Arabidopsis* protreins in TAIR10. A total of 60 (36.36%) genes were mapped to proteins in Swiss-Prot or TAIR 10 by blastx with $E<10^{-5}$. The list of differentially expressed genes and their annotations is given in Table 2.4.

To further validate the list of statistically identified DEGs, we collected 69 *Arabidopsis* protein sequences and 24 rice protein sequences whose expression has been experimentally determined to react to arsenic stress [104]. Those arsenic-responsive proteins were compared to the DEGs using tblastn, and were found to match 11 DEGs. The function and fold change of those genes in both RNA-Seq and assay experiment are summarized in Table 2.5.

2.3.4 Validation of expression of selected predicted transcript assembly using qRT-PCR

We selected 10 genes with a wide range of fold change between two conditions for validation by qRT-PCR. According to the differential gene expression analysis, 4 of the selected genes were significantly upregulated in arsenate-treated samples, while the other 6 were considered constitutively expressed in both conditions. As shown in Table 2.6 trends of expression observed in qRT-PCR data were consistent with those inferred from the RNA-Seq expression data for all ten genes.

2.3.5 Arsenic-responsive genes

Arsenate stress significantly affects the expression of genes involved in stress response, ion transport, and signaling pathways. Annotations and fold changes of the differentially expressed genes are given in Table 2.4. Glyceraldehyde-3-phosphate dehydrogenase C subunit 1 (GAPC-1) and PHI type glutathione S-transferase (GST) were the most upregulated genes in As-treated gametophytes, and their expression increased by 374 and 234 fold, respectively. Genes of 4 transporters were markedly

upregulated by arsenate, including one putative carnitine transporter 4, two *ACR3*-like arsenite transporters, and a probable heavy metal transporter, while a phosphate transporter gene exhibited lower expression under arsenate exposure. ACR3 is a vacuolar arsenite transporter that plays a key role in arsenic tolerance of *P.vittata* [75]. It has been reported that the expression of *ACR3* increases by approximately 9 fold in arsenite-treated gametophytes [75]. Here RNA-Seq data showed a similar increase of 5.61-fold in *ACR3* level. A sizable portion of the remaining transcripts are related to signaling and stress response. Arsenate induced the expression of genes encoding homolog of an A20/AN1-like zinc finger family protein, UDP-glucosyl transferases (UGTs), phytosylfokine-alpha receptor 2 and cytochrome P450s, while suppressing expression of homolog of genes of glyoxal oxidase-related protein, peroxidase and heat shock protein 70 (HSP70). Despite the central role of arsenate reduction and PC-chelation in arsenic detoxification and tolerance, phytochelatin synthase (PCS) and arsenic reductase (AR) levels were not altered significantly.

A group of genes that are involved in signal mediation have significantly altered expression during arsenate stress. A potential ethylene-responsive transcription factor rap2-4 was upregulated, two putative histidine kinases, a putative ring-h2 finger protein, and a putative phytosylfokine-alpha receptor 2 coding transcript were upregulated. On the contrary, the expression of homolog of transcription factor myb46, a burp domain-containing protein, and an NAC domain-containing protein were significantly down-regulated.

To further characterize the function of the DEGs, a GO term enrichment analysis was performed against the entire transcriptome. 583 and 11260 GO terms were originally

associated with the DEGs and the transcriptome by the Blast2GO suite [100], respectively, which were mapped to 100 plant specific GO slim terms, and then tested for over-representation using one-sided Fisher's exact test as implemented in topGO [105]. The GO terms with adjusted P<= 0.01 were considered significantly enriched in the DEGs. The analysis revealed major categories of biological processes, molecular functions and cellular components that differ in DEGs from the remaining genes. Biological processes such as response to biotic stimulus, response to external stimulus and cell-cell signaling are overrepresented in DEGs. Transporter activity is then only significantly enriched molecular function in the DEGs. For the cellular components, significantly enriched GO terms include plasma membrane and membrane.

2.4 Discussion

The present work seeks to identify arsenic-induced changes in the transcriptome of the *P.vittata* gametophyte at an early stage of arsenic exposure. Substantial progress was made in recent years in understanding As uptake, translocation, toxicity and tolerance in plants. So far, we have learned that As-hyperaccumulating plants take up the metalloid more quickly than non-accumulators, and efficiently translocate As to the above-ground tissues where it is sequestered in the vacuole as free arsenite. Besides the rapid uptake and translocation, As-hyperaccumulators also possess a greater antioxidant capacity to maintain lower ROS levels. It is conceivable that the ability to hyperaccumulate As requires synergistic contribution from numerous physiological processes. Thus, the recent development of global transcript analysis technologies such as

RNA-Seq adds new dimensions to our understanding of the molecular details underpinning arsenic tolerance.

2.4.1 Stress responsive genes

Arsenic can induce severe oxidative stress in plant cells. Exposure to inorganic arsenic generates reactive oxygen species (ROS) such as superoxide ($O_2•^-$), the hydroxyl radical (•OH), and H_2O_2 [17, 23, 24], which can lead to DNA and protein damage, lipid peroxidation, and depleted antioxidant defense levels [25]. Previous studies have shown that a number of enzymes involved in the antioxidant responses are induced by arsenic exposure [23, 106]. A significant portion of the arsenic-responsive genes identified in the present study are involved in combating oxidative stress and the restoration of redox hemostasis, such as one PHI type glutathione S-transferase (GST), two forms of UDP-glucosyl transferases (UGT), and two cytochrome P450s.

GST is ubiquitously present in prokaryotes and eukaryotes, and has an established role in response both biotic and abiotic stress in plants, including heavy metal exposure [107]. It catalyzes the conjugation of electrophilic toxins with the reduced-form γ-Glu-Cys-Gly tripeptide glutathione (GSH) to form non-toxic peptide derivatives. Apart from the their function in GSH-dependent conjugation, plant GSTs can also act as glutathione peroxidases to directly detoxify toxic electrophiles. Studies have confirmed that higher GST activity contributes to better tolerance of herbicides in maize [108] and soybean [109]. Comparison of the arsenic hyperaccumulator *P.vittata* with the non-arsenic hyperaccumulator *P.ensiformis* revealed that *P.vittata* has an inherently greater

antioxidant potential in terms of higher concentrations of ascorbate (AsA) and glutathione (GSH), with or without arsenic exposure [24]. We observed a 234 fold increase in GST expression upon arsenic exposure, indicating that enhanced GST expression and activity maybe essential to maintain the antioxidant level, and to minimize the detrimental effects of ROS in *P.vittata*.

UDP-glycosyltransferases (UGTs) catalyzes the glycosylation of several classes of small molecules to generate secondary metabolites in plants, and have a vital role in the regulation of cellular homeostasis. We found that two homologs of UGTs increased their expression level by 5.8 and 4.6 fold, respectively, in arsenate-treated *P.vittata* gametophytes. It has been reported that several Arabidopsis UGTs are highly inducible under oxidative stress, pathogen invasion, and UV radiation [110-112]. *P.vittata* is likely to utilize UGTs as mediators to initiate the cascade of abiotic stress response.

Upregulated transcripts that are related to stress response also include those encoding homologs of cytochrome P450 (CYP), A20/AN1 zinc-finger containing protein, DNA glycosylase, and pectin methylesterase (PME). CYPs have a well-established role in the oxidation and detoxification of herbicides in plants (reviewed in [113]). Previous studies have demonstrated that multiple forms of CYPs were up-regulated by As(V) and As(III) exposure in rice [23, 60]. Two forms of CYP homologs were found to be differentially expressed in this work, and their expression increased by 2.75 and 3.85 fold, respectively. The *P.vittata* homolog of A20/AN1 zinc-finger containing protein gene *OsiSAP1* is inducible by various abiotic stresses like cold, salt, drought, and heavy metals, and the overexpression of *OsiSAP1* enhances stress tolerance in transgenic plants [114]. Similarly, PMEs are differentially regulated by multiple environmental stresses to modify

the degree of methylesterification of pectins in plant cell wall and thus regulate stress response.

2.4.2 GAPDH in carbon metabolism and as a potential arsenate reductase

One of the most noticeable changes in gene expression is the upregulation of the transcript encoding a homolog glyceraldehyde-3-phosphate dehydrogenase C subunit 1 (GAPC1), whose expression drastically increased by 374 fold under arsenate treatment. GAPDH catalyzes the conversion of glyceraldehyde-3-phosphate and Pi to 1,3-bisphosphoglycerate, where As(V) can replace the substrate Pi and lead to the formation of unstable and short-lived 1-arseno-3-phosphoglycerate. The 3-arseno-phosphoglycerate product rapidly decomposes and thus uncouples ATP synthesis from glycolysis, resulting in reduced energy output. A resent proteomics study of *P.vittata* fronds revealed that enhanced expression of multiple forms of GAPDH and several other proteins of carbon metabolism under arsenate exposure. Ahsan et al. (2010) [115] also reported increased activity of proteins associated with energy metabolism, such as NADP-dependent malic enzyme, NAD-dependent formate dehydrogenase in the leaves of arsenate-treated rice. Those findings together suggest that coping with As stress requires extra energy input, and the upregulation of GAPDH may be a compensatory mechanism to fulfill the increased energy needed for metabolizing arsenic when the ATP yield from glycolysis is jeopardized.

Furthermore, mammalian GAPDHs are able to convert arsenate to arsenite in vitro in the presence of GSH, NAD and glyceraldehyde-3-phosphate [116]. Reduction of

arsenate to arsenite is the initial step in arsenate detoxification, and occurs rapidly in plants. Arsenic-induced differential expression of arsenate reductase *PvACR2* was not observed in this study, which is consistent with previous finding that *PvACR2* is constitutively expressed in the gametophyte, regardless of arsenate exposure [55]. It has been proposed that a functional redundancy of arsenate reduction exists in plants. For example, expression of a cytosolic triosephosphate isomerase (TPI) isolated from *P.vittata* conferred arsenate resistance to the *E.coli* strain lacking arsenate reductase *ArsC*. TPI is also involved in glycolysis. It appears that arsenate reduction is coupled with glycolysis. Besides its key role in role in glycolysis, GADPH may be directly or indirectly involved in arsenate reduction. It is worth investigating whether the *P.vittata* GAPDH is able to reduce arsenate in the presence of an appropriate electron donor.

2.4.3 Transporter activities

Asenic hyperaccumulators and non-accumulators differ distinctively in the distribution of As within the plant. While non-accumulators tend to retain most As in the root, hyperaccumulators efficiently translocate As to the aerial portion and accumulate extremely high concentration of As in the vacuoles. Rapid root-to-shoot translocation and vacuole compartmentalization are both crucial to arsenic-hyperaccumulation in *P.vittata*, but little is known about the mechanisms responsible for the transport of arsenic into vacuole. ACR3 is the first identified arsenite transporter located on the vacuolar membrane and plays a key role in *P.vittata* As-hypertolerance. Its expression was reported to increase by approximately 9 fold in arsenate-grown *P.vittata* gametophytes [75]. We observed a similar result of 5.61-fold up-regulation. There are two copies of this

arsenite transporter gene in *P.vittata*: *ACR3* and *ACR3;1*. Pv ACR3 and ACR3;1 proteins are 84% identical at the amino acid level, and share highly conserved transmembrane domains, but only the *ACR3* gene was inducible by arsenate exposure [75]. However, we observed a 14.48 increase of *ACR3;1* expression in the present study. In addition, a putative organic cation transporter (OCT) and a copper transporter were also upregulated by 49.49 and 2.59 fold respectively in arsenate treatment. Previously, members of Arabidopsis OCT family have been shown to be localized to vacuolar membrane, and to play a role in adaptation to salt, drought, and cold stress [117]. The copper transport protein has a conserved cys-containing heavy-metal-associated (HMA) domain. The two cysteine residues of HMA domain are critical for the binding and transfer of metal ions like As, copper, cadmium, cobalt and zinc. Given that arsenic has strong affinity to sulfhydryl groups and can bind to reduced cysteines in proteins, the newly identified copper transporter homolog could have an important role in mediating the translocation of arsenic into vacuole.

2.4.4 Signaling pathways

ROS generated by arsenic exposure may trigger the production of messenger molecules such as jasmonic acid (JA), S-adenosyl-l-methionine (SAM), and cytokine, which act to activate or inactivate downstream response cascades. A Trinity gene similar to *Arabidopsis* histidine kinase 4 (AHK4) was upregulated by 2.05 fold by As treatment. The Arabidopsis histidine kinase 4 is a cytokinin-binding receptor that transduces cytokinin signals across the membrane [118]. It was demonstrated that AHK4 mediated cytokinin signaling negatively regulates Pi starvation responses in Arabidopsis by

repressing response genes such as phosphate transporter *PHT1;1* [119], and Arabidopsis mutant defective in *PHT1;1* displays enhanced arsenic accumulation [33]. Given that As(V) is a close analog of Pi, AHK4 mediated repression of Pi starvation responsive genes may promote the uptake of As(V).

A phytosylfokine-alpha receptor 2 (PSKR2) homolog encoding transcript, which is involved in tyrosine-sulfated peptide signaling, was found to be up-regulated by 4.40 fold. PSK-alpha acts as a growth factor that regulates root elongation in plants [120]. PSKR2 has recently been implicated a role in microbial resistance in Arabidopsis, and plants lacking PSKR2 function showed higher susceptibility to fungus and bacterial infection [121, 122]. We also observed a 4.65 fold decrease in the expression of a homolog of burp domain-containing protein 16 (BURP 16). The BURP domain has a highly conserved structure with a hydrophobic signal peptide at the N-terminus. Many BURP domain-containing proteins have been reported to be up or down regulated by biotic and abiotic stresses, including ABA [123], auxin [124], salt and cold [125], etc. However, it’s not clear that whether the PSKR2 and BURP16 homologs respond specifically to arsenic exposure or they are involved in the signaling process of generic stress response.

2.5 Conclusions

This study demonstrates that the transcriptome of *P.vittata* can be assembled from the RNA-Seq short reads without a reference genome, which provides both a high-quality sequence resource and an alternative approach to identify arsenic-responsive genes through transcriptome profiling. Our work also provides a guideline for evaluating the

quality of a *de novo* assembly by leveraging genomic resources from model plant species. Differentially expressed genes have been carefully annotated, and new players in arsenic tolerance have been identified from the gene differential expression analysis. The drastically enhanced expression of GAPDH and its ability to reduce arsenate in the presence of electron donors suggest that it could act as an alternative arsenate reductase. We also identified a putative cation transporter and a putative copper transporter as potential arsenic transporter that could facilitate the influx of arsenic into vacuoles. These findings provide insights on arsenic metabolism and tolerance and help to generate testable hypotheses for future study.

Table 2.1 Primers used in the study.

Gene	**Forward Sequence**	**Reverse Sequence**
Actin	5’-GGGTTTACATTCAGCGAAGC-3’	5'-GCTTTCCCTCCAGTGGACTT-3'
comp74286	5’-ATGAAAAGCTTTCGCTTCAC-3’	5’-GGCGAGCATAACTTGATTGC-3’
comp91117	5’-TGTTGCGGACGAGACAATAG-3’	5’-GGCTATGAGCAACAGCAGTA-3’
comp100020	5’-AGCCAGTCCATTGGCTTGGT-3’	5’-CCAGCCTCAGAGGTTTAGCT-3’
comp110556	5’-ATGGAACCTTTTTGCCTGTT-3’	5’-TGTGCATGGATGCATTTCTT-3’
comp85481	5’-AGCAATGGTGGAAGTAGAGTC-3’	5’-GCTTAACCACACCCTCTTCAG-3’
comp98614	5’-AGCAGTGGCTAGAAGTGGAATC-3’	5-AACCACATCTGCATCAGGAG-’3’
comp99511	5’-ATGGCCTCTCCATCATCAAC-3’	5’-CACATGCGACTTCTCCAAAC-3’
comp103624	5’-GCTGCACTTGCCATACTCAA-3’	5’-GCTCTATGGCATGGTCCAAT-3’
comp96855	5’-GCACCATCGACTGTCTTTTG-3’	5’-TTGGCTCCACTTGCTAAGGT-3’
comp97777	5’-AGTACCACCAGCTGCAATGA-3’	5’-TCCAAGCTTTGCAACACATC-3’

Table 2.2 Statistics of the *de novo* transcriptome assembly.

Total Bases	27,824,880,283
Total Reads	287,177,776
Total Predicted Transcripts Assembled	344,048
Total Genes Components Assembled	190,495
N50 (bases)	2,125
Min Length (bases)	151
Max Length (bases)	17,840
Average Length (bases)	916
Average GC%	47.27

Table 2.3 Distribution of KOG annotations of all predicted transcripts.

Description	Count	Percentage
Signal transduction mechanisms	9,254	14.63%
General function prediction only	8,578	13.56%
Posttranslational modification, protein turnover, chaperones	5,923	9.36%
RNA processing and modification	4,330	6.85%
Function unknown	4,217	6.67%
Carbohydrate transport and metabolism	3,887	6.15%
Transcription	3,818	6.04%
Intracellular trafficking, secretion, and vesicular transport	3,724	5.89%
Replication, recombination and repair	3,539	5.59%
Cytoskeleton	3,399	5.37%
Amino acid transport and metabolism	3,066	4.85%
Translation, ribosomal structure and biogenesis	2,927	4.63%
Lipid transport and metabolism	2,772	4.38%
Secondary metabolites biosynthesis, transport and catabolism	2,450	3.87%
Inorganic ion transport and metabolism	2,348	3.71%
Energy production and conversion	2,209	3.49%
Cell cycle control, cell division, chromosome partitioning	2,058	3.25%
Chromatin structure and dynamics	1,236	1.95%
Cell wall/membrane/envelope biogenesis	1,055	1.67%
Coenzyme transport and metabolism	1,043	1.65%
Nucleotide transport and metabolism	906	1.43%
Defense mechanisms	464	0.73%
Extracellular structures	316	0.50%
Nuclear structure	289	0.46%
Cell motility	39	0.06%
multiple functions	26	0.04%

Table 2.4 List of differentially expressed predicted genes.

	Best Match in Arabidopsis	Best E-value	Function Annotation	FDR	FC
comp96855	AT3G04120.1	9.00E-178	glyceraldehyde-3-phosphate dehydrogenase C subunit 1	1.14E-21	374.08
comp98614	AT2G30870.1	1.00E-63	glutathione S-transferase PHI 10	5.26E-74	234.41
comp103624	AT3G20660.1	1.00E-118	organic cation/carnitine transporter4	4.27E-38	49.49
comp108897	AT4G23160.1	9.00E-44	cysteine-rich RLK (RECEPTOR-like protein kinase) 8	1.52E-41	14.48
comp89355	AT4G12040.2	1.00E-25	A20/AN1-like zinc finger family protein	1.46E-01	7.69
comp98470	AT5G11280.1	8.00E-31	unknown protein	3.92E-05	5.90
comp99546	AT1G22380.1	5.00E-78	UDP-glucosyl transferase 85A3	3.79E-14	5.80
comp99730	AT1G80760.1	4.00E-66	NOD26-like intrinsic protein 6;1	7.15E-19	5.56
comp91673	AT5G13250.1	2.00E-07	RING finger protein	3.79E-02	5.16
comp69192	AT1G22370.2	1.00E-21	UDP-glucosyl transferase 85A5	2.49E-03	4.60
comp64296	AT5G53890.1	5.00E-51	phytosylfokine-alpha receptor 2	9.33E-02	4.40
comp62480	AT2G45570.1	2.00E-65	cytochrome P450, family 76, subfamily C, polypeptide 2	1.83E-01	3.86
comp108099	AT1G10800.1	5.00E-11	unknown protein	1.86E-10	3.71
comp105360	AT5G08250.1	2.00E-68	Cytochrome P450 superfamily protein	3.12E-04	2.75
comp98848	AT5G19090.3	4.00E-19	Heavy metal transport/detoxification superfamily protein	1.04E-03	2.59
comp99639	AT3G19540.1	2.00E-121	Protein of unknown function (DUF620)	8.54E-04	2.44
comp105484	AT1G78080.1	5.00E-30	related to AP2 4	9.62E-03	2.41
comp106272	AT1G76490.1	0	hydroxy methylglutaryl CoA reductase 1	2.90E-04	2.34
comp103740	AT5G57970.2	7.00E-74	DNA glycosylase superfamily protein	4.62E-02	2.31
comp104914	AT5G19730.1	4.00E-113	Pectin lyase-like superfamily protein	2.06E-02	2.30
comp110556	AT2G19920.1	7.00E-124	RNA-dependent RNA polymerase family protein	3.40E-03	2.29
comp95798	AT1G55210.2	6.00E-15	Disease resistance-responsive (dirigent-like protein) family protein	1.20E-03	2.27

comp93130	AT2G29390.2	7.00E-148	sterol 4-alpha-methyl-oxidase 2-2	4.86E-04	2.27
comp109449	AT2G01830.1	0	CHASE domain containing histidine kinase protein	9.22E-03	2.05
comp98134	AT1G72200.1	8.00E-21	RING/U-box superfamily protein	1.20E-02	2.04
comp65351	AT3G57620.1	1.00E-173	glyoxal oxidase-related protein	3.58E-02	0.50
comp93517	AT3G52590.1	3.00E-18	ubiquitin extension protein 1	1.57E-01	0.48
comp92136	AT1G49240.1	7.00E-09	actin 8	3.64E-02	0.44
comp100006	AT5G05340.1	6.00E-83	Peroxidase superfamily protein	4.10E-02	0.44
comp104124	AT5G51550.1	2.00E-113	EXORDIUM like 3	7.66E-03	0.42
comp2149	AT5G07720.1	8.00E-96	Galactosyl transferase GMA12/MNN10 family protein	9.02E-02	0.39
comp89401	AT1G01470.1	1.00E-54	Late embryogenesis abundant protein	1.67E-01	0.37
comp92513	AT3G08500.1	1.00E-35	myb domain protein 83	1.60E-01	0.28
comp94448	AT5G61430.1	9.00E-64	NAC domain containing protein 100	6.04E-03	0.27
comp78861	AT1G75630.1	4.00E-13	vacuolar H+pumping ATPase 16 kDa proteolipid subunit 4	6.06E-05	0.22
comp66898	AT1G72730.1	0	DEA(D/H)-box RNA helicase family protein	1.09E-04	0.22
comp93600	AT1G60390.1	2.00E-22	polygalacturonase 1	7.04E-04	0.22
comp289996	AT2G19760.1	4.00E-19	profilin 1	5.70E-04	0.19
comp117534	AT3G12110.1	0	actin-11	7.20E-18	0.17
comp330452	AT2G41190.1	9.00E-07	Transmembrane amino acid transporter family protein	9.78E-08	0.17
comp73822	AT4G14960.1	3.00E-179	Tubulin/FtsZ family protein	2.80E-06	0.17
comp224786	AT3G12580.1	0	heat shock protein 70	1.03E-11	0.16
comp3806	AT2G27030.2	7.00E-58	calmodulin 5	3.95E-05	0.16
comp30460	AT5G55400.1	1.00E-94	Actin binding Calponin homology (CH) domain-containing protein	6.09E-06	0.15
comp82375	AT3G09630.1	2.00E-148	Ribosomal protein L4/L1 family	8.21E-08	0.15
comp53236	AT3G53750.1	1.00E-155	actin 3	2.63E-13	0.15
comp100331	AT1G10130.1	0	endoplasmic reticulum-type calcium-transporting ATPase 3	6.39E-20	0.14
comp69353	AT3G12110.1	1.00E-120	actin-11	2.07E-12	0.14
comp459543	AT5G36940.1	6.00E-14	cationic amino acid transporter 3	1.35E-09	0.13
comp65966	AT3G19940.1	2.00E-49	Major facilitator superfamily protein	3.58E-06	0.12

comp60465	AT1G56070.1	0	Ribosomal protein S5/Elongation factor G/III/V family protein	5.59E-20	0.12
comp25083	AT2G28720.1	2.00E-42	Histone superfamily protein	2.80E-07	0.11
comp68124	AT3G48850.1	1.00E-114	phosphate transporter 3;2	5.20E-09	0.10
comp70445	ATMG01190.1	0	ATP synthase subunit 1	1.91E-11	0.09
comp97905				1.33E-27	557.36
comp79469				1.01E-08	134.70
comp70105				3.27E-08	12.13
comp57362				6.41E-04	8.70
comp93532				7.60E-02	7.40
comp91909				7.60E-02	7.38
comp87721				1.15E-01	7.03
comp92488				3.67E-06	6.85
comp68154				1.67E-01	5.77
comp96450				3.80E-03	5.23
comp111788				8.87E-16	4.99
comp124096				3.13E-12	4.94
comp85918				3.95E-03	4.81
comp63395				4.63E-02	4.77
comp86746				1.84E-04	4.68
comp83600				9.33E-02	4.38
comp60560				6.81E-02	4.31
comp83022				5.25E-02	4.21
comp67418				1.33E-01	4.11
comp70701				6.46E-03	3.97
comp55747				9.70E-05	3.84
comp85534				1.60E-01	3.12
comp89189				2.02E-07	2.92
comp91715				6.06E-03	2.91
comp92919				4.10E-02	2.75
comp123315				9.13E-02	2.73
comp89827				3.67E-06	2.70
comp50887				1.18E-02	2.58
comp94719				9.71E-04	2.33
comp95395				8.68E-04	2.21
comp96249				8.40E-02	2.15
comp88825				4.85E-03	2.05
comp107255				1.26E-02	0.49
comp108896				1.04E-01	0.48
comp108899				1.06E-02	0.47

comp54403	2.70E-03	0.47
comp113699	4.10E-02	0.46
comp96093	4.90E-02	0.44
comp91701	1.83E-01	0.43
comp56500	1.02E-01	0.41
comp64612	2.87E-02	0.40
comp79188	1.53E-05	0.38
comp103117	1.00E-02	0.38
comp93183	3.44E-02	0.35
comp68602	1.61E-01	0.32
comp87152	2.04E-02	0.30
comp53805	1.10E-01	0.27
comp270317	3.19E-03	0.26
comp71256	1.42E-02	0.25
comp3659	1.04E-02	0.24
comp105551	9.20E-09	0.23
comp68956	8.88E-04	0.23
comp95315	2.59E-05	0.23
comp94752	2.49E-03	0.22
comp84679	7.38E-05	0.22
comp81235	6.60E-06	0.21
comp91381	9.22E-06	0.21
comp91485	4.63E-03	0.20
comp89275	1.64E-05	0.20
comp73204	7.82E-08	0.20
comp86598	3.96E-05	0.20
comp67552	2.23E-05	0.19
comp77578	7.56E-05	0.19
comp2123	7.38E-05	0.18
comp94754	1.08E-09	0.18
comp76631	4.79E-07	0.18
comp180100	1.81E-08	0.18
comp60363	5.29E-08	0.17
comp82873	1.23E-13	0.17
comp51704	2.44E-07	0.17
comp77077	7.22E-07	0.17
comp88395	3.00E-10	0.16
comp136586	4.15E-07	0.16
comp60648	6.67E-11	0.16
comp78244	1.73E-02	0.16
comp79827	7.58E-16	0.16
comp73200	7.92E-14	0.16
comp91928	4.08E-05	0.16

comp172745	5.41E-09	0.16
comp84979	6.91E-12	0.15
comp86499	8.89E-17	0.15
comp76181	6.01E-14	0.15
comp94272	1.22E-10	0.15
comp91953	9.29E-09	0.14
comp254714	1.63E-09	0.14
comp97879	4.60E-21	0.14
comp84087	4.91E-12	0.14
comp38889	4.13E-08	0.13
comp385575	6.92E-04	0.13
comp285509	4.23E-09	0.13
comp92726	2.22E-08	0.13
comp13227	2.59E-05	0.13
comp360747	2.32E-03	0.13
comp49437	2.80E-07	0.13
comp95418	5.20E-09	0.12
comp92591	7.11E-11	0.12
comp52050	3.80E-11	0.12
comp48878	5.20E-09	0.12
comp90300	1.67E-07	0.12
comp92332	7.06E-09	0.12
comp92739	5.76E-10	0.12
comp106175	5.85E-17	0.12
comp95128	5.73E-24	0.11
comp85021	4.29E-21	0.11
comp94707	3.80E-11	0.11
comp92564	1.43E-12	0.10
comp92210	2.74E-12	0.09
comp62926	8.23E-05	0.04
comp89868	2.22E-08	0.03

Note: only 54 of the differentially expressed genes are annotatable by blastx searching for matches in known protein databases.

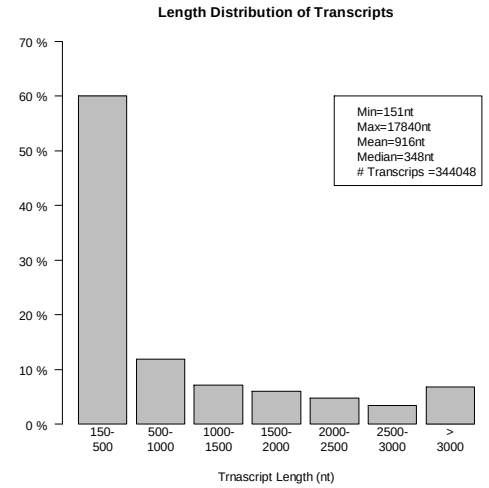
Table 2.5 Set of previously identified As-regulated genes and their matches in Pteris transcriptome

	Previously identified DEG		Pteris DEG	
Description of gene	Locus	Fold change(+/-)	Locus	Fold change(+/-)
cytochrome P450	AT4G31500	(-)1.71	comp100006	(-)2.27
	AT3G48520	(-)1.56	comp102296	(+)1.80
Zinc finger protein	AT5G27420	(-)1.75	comp102778	(+)1.87
glutathione S-transferase	AT3G62760	(+)1.64	comp103643	(-)1.79
	AT1G78370	(+)1.68	comp105360	(+)2.75
peroxidase	AT5G17820	(+)1.68	comp98134	(+)2.04

Table 2.6 Quantitative real time PCR analysis of selected genes.

	qRT-PCR Fold Change	RNA-Seq Fold Change	DEG	Annotation based on blastx results against Swiss-Prot
comp74286	1.296	1.027	N	HASTY 1
comp85481	2.311	1.462	N	glutathione s-transferase f10
comp97777	1.535	1.474	N	histidine-containing phosphotransfer protein 1
comp100020	1.212	1.036	N	*
comp91117	0.7949	0.8894	N	vesicle-associated membrane protein 721
comp99511	0.2731	0.6253	N	transporter ArsB
comp98614	1533	234.4	Y	glutathione S-transferase PHI 10
comp103624	53.79	49.49	Y	organic cation transporter 4
comp96855	15246	374.0	Y	glyceraldehyde-3-phosphate dehydrogenase C subunit 1
comp110556	11.09	2.288	Y	probable RNA-dependent RNA polymerase

*Gene comp100020 has 6 isoforms, none of them has a significant match in the Swiss-Prot database at E-value of 10^{-5}.



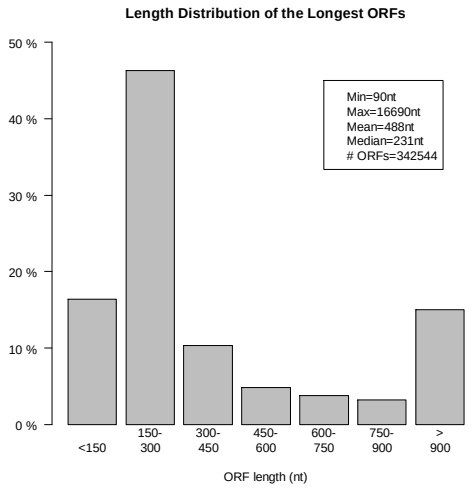


Figure 2.1 Length distributions of the assemblies and the predicted ORFs of the *P.vittata* transcriptome.

The x-axis shows the size category and the y-axis indicates the percentage of assemblies that lie in each bin.

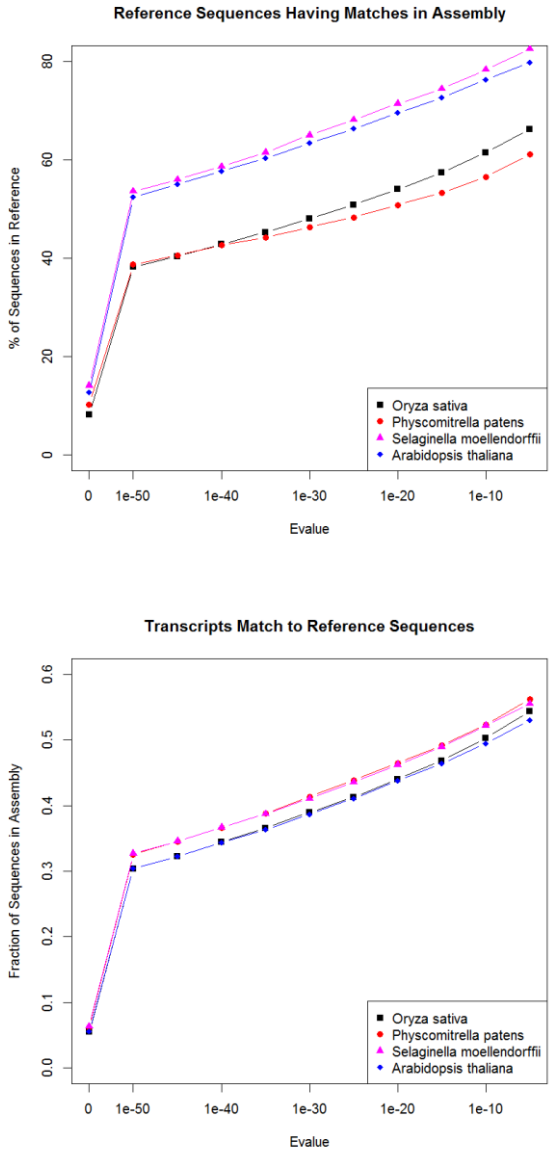


Figure 2.2 Assessment of the completeness of the *P. vitatta* transcriptome by comparing to known plant proteomes.

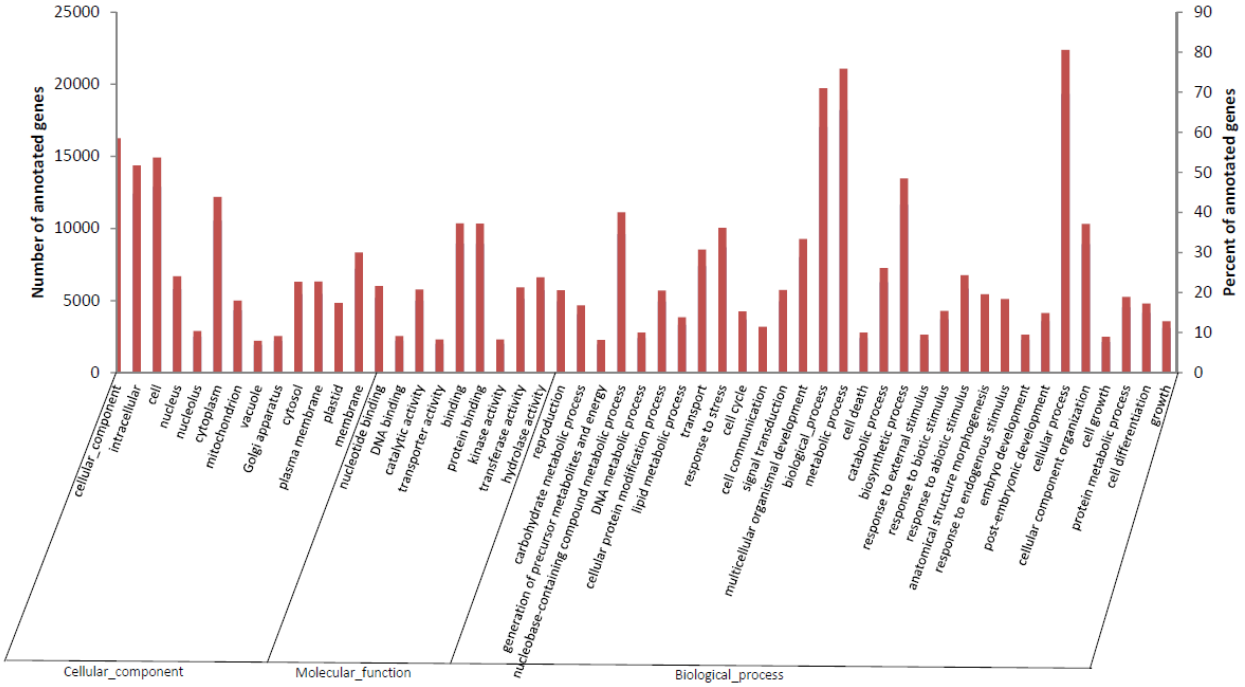


Figure 2.3 Distribution of Plant GO slim terms assigned to *P.vittata* transcriptome assemblies.

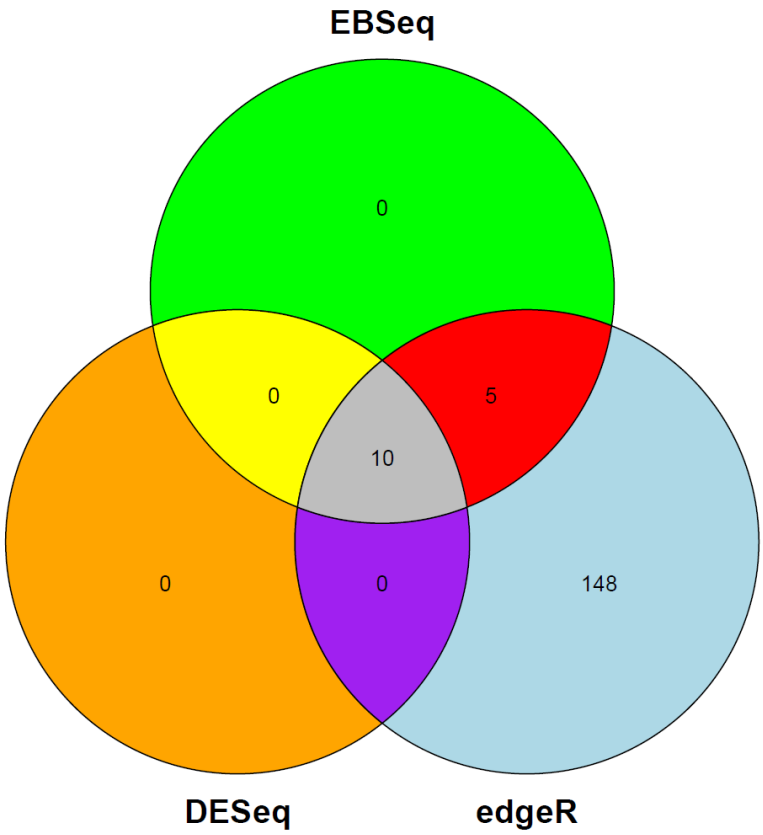


Figure 2.4 Venn diagram comparing differentially expressed genes identified by different statistical packages.

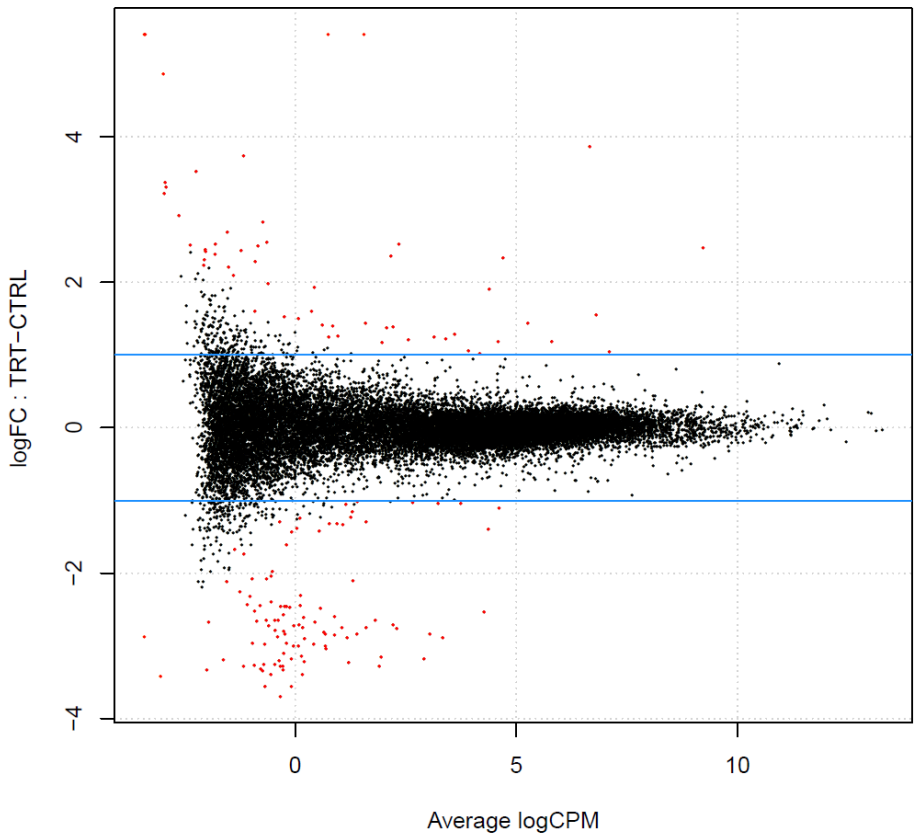


Figure 2.5 MA plot of the differentially expressed genes between As-treated and untreated conditions.

Each point represents a Trinity gene. The mean expression level of each Trinity component is plotted against the fold change. Red points are differentially expressed genes identified by three statistical packages at FDR < 0.2, and black points do not have statistically significant difference in As treatment. Blue lines are levels of 2 fold change.

CHAPTER 3. TEXTPRESSO FOR LITERATURE OF ARSENIC TOLERANCE

3.1 Introduction

Textpresso [126] is one of the most widely used text analysis tools for biological literature [127-130]. Textpresso features two important innovations: first, it performs searches on the full text of the article (full text search), rather than limiting the search to just the title, abstract, and keywords, so that one expects broader information coverage than would be seen in an abstract-only search for a given topic; second, Textpresso can automatically index and cluster literature according to user-defined concepts. In the Textpresso pipeline, a corpus of full texts related to a specific topic are tokenized into sentences, and the sentences into words. Sentences can either be indexed by individual words or semantically. Semantic indexing makes use of a structured lexicon, which contains ontologies of biological entities/concepts, and terms that describe the relationships relating them. Texts of interest are identified by matching to the terms encoded in the lexicon and labeling them with XML tags. A web interface allows users to query a combination of keywords and/or concepts within sentences or the document as a whole. The sentence-level search pinpoints the contexts where query matches appear, which is important for precise retrieval of biological facts. Semantically indexed texts not only allow searches by keyword query, but also enable semantic searches. Thus,

meaningful biological facts can be efficiently recovered. Despite these advantages, Textpresso does have some drawbacks in text processing and database searching. During sentence indexing, Textpresso requires an exact match to lexical terms, thus the lexicon needs to include all possible tenses and forms of verbs and nouns in order to fully cover the potential targets. Such a verbose ontology dramatically slows down the indexing - the most time-consuming step in the Textpresso pipeline. The current Textpresso web interface only provides general keyword searching with "case sensitive" and "exact match" options, which does not take full advantage of its indexing strategy, and will therefore miss a significant number of sentences that match to the query, but use a different tense or form.

We have incorporated a stemming technique into Textpresso in order to decrease the above problems; we call this enhanced version of stemmed-Textpresso. The standard Porter stemmer [131] reduces texts in both the corpus and the lexicon to their common roots. The stemming has been modified to skip words that are potential gene or protein names/symbols. Indexing is performed on the stemmed corpus using a stemmed lexicon, so that sentences are labeled by word roots. In database searches, queries can be stemmed to retrieve more sentences of interest, and all versions of the query keywords are highlighted in the retrieved texts.

3.2 Results

Stemmed-Textpresso has been constructed on a collection of literature studying arsenic tolerance in plants and yeast. The literature database is available at: http://textpresso.genomics.purdue.edu/cgi-bin/cgiwrap/textmine/home. The current

corpus comprises 970 full text articles and 1778 abstracts, together with bibliographic information. The stemmed-textpresso lexicon consists of ontologies describing biological entities or concepts and their relationship together with auxiliary words. Gene Ontology (Gene Ontology Consortium [132]) terms contribute the majority of current Textpresso ontologies. We have added *Arabidopsis* genes, plant structures, cereal plant growth stages, and flowering plant growth stages (Plant Ontology Consortium (Plant Ontology Consortium™, 2002) to the existing Textpresso ontologies. To generate a compact lexicon, entries in ontologies are stemmed and consolidated into their common roots. The stemming procedure reduces the number of lexical entries from 634,748 to 435,923, which not only significantly speeds up the semantic indexing of the corpus, but also finds more matched entities (3,028,991 compared to 2,881,080, an increase of 5.13%). In keyword indexing, each sentence in the text is indexed by each word in the sentence. The number of unique keyword indices decreased from 255,395 to 212,957 (16.6%), while the number of sentences stayed the same. Therefore, the information content of each keyword index is enriched in stemmed-Textpresso.

In addition to the above improvements, stemmed-Textpresso retrieves more related facts than simple keyword searches. By default, stemmed-Textpresso removes the suffixes of query keywords and searches the stemmed database using the word roots. Alternatively, users can skip stemming and choose to search the non-stemmed database. For example, if we search “growth stimulation” in stemmed-Textpresso, the query is first split into “growth” and “stimulation”, which are reduced to their stem forms, “grow” and “stimulat”, respectively. Table 1 shows the difference in the number of sentences retrieved with the stemmed and unstemmmed versions of Textpresso. Stemmed-

Textpresso shows higher recall at the expense of somewhat decreased retrieval precision. Retrieving and formatting the matching documents is a major contributor to the overall search time and accounts for the longer search times reported for stemmed-Textpresso.

3.3 Implementation

The Textpresso 2.5.1 package was downloaded from the Textpresso web site (http://www.textpresso.org/downloads.html). Full text PDF articles are processed using the standard Textpresso procedure up to the point where individual sentences are obtained. The resulting sentences are stemmed by a Perl-implementation of the Porter stemmer (http://search.cpan.org/~creamyg/Lingua-Stem-Snowball-0.952/lib/Lingua/Stem/Snowball.pm). Entries in the lexicon are also stemmed and consolidated. In semantic indexing, words in the stemmed texts are marked up by identifying terms that match those stored in the stemmed lexicon. Each sentence in the stemmed corpus is also indexed by words that constitute the sentence.

To accommodate the changes we made to the literature database, the stemmer is also embedded in the search interface. Query phrases or hyphenated words are parsed, stemmed and reconnected. Unless the “Exact match” or “Case sensitive” option is selected, stemmed keywords or phrase queries are searched in the stemmed corpus. Sentences that contain the query are returned with matches in all syntactical forms highlighted. Users can also choose to search unstemmed queries vs. the corpus without stemming.

3.4 Discussion

The use of the stemmed lexicon greatly simplifies the definition of lexical files by users. Comprehensively including all verb and noun forms in the lexicon is both tedious and error prone. Stemming is most powerful in cases where a number of variations of the query words appear in texts, which can be seen in the example “growth stimulation”. The calculation of the retrieval accuracy is somewhat subjective and may vary; we have used a very strict definition – most of the sentences we assign as negative would, in fact, be of interest given the queries. Despite the lower retrieval precision, stemmed-Textpresso is able to extract significantly more relevant sentences for further examination.

Stemmed-Textpresso uses the Porter stemming algorithm to reduce the number of unique words in the corpus and lexicon, and thus simplifies lexicon construction as well as improves both indexing efficiency and search functionality. The new system not only expedites semantic indexing, but also recognizes more matching sentences in the indexed texts. Stemmed-Textpresso naturally expands the search query and offers better coverage of related biological facts.

Table 3.1 Comparison of search accuracies.

Query	Type	Document Precision	Document Recall	Document F1 score	Sentence Precision	Sentence Recall	Sentence F1 score
growth stimulation	stemmed	0.515 (17/33)	1.0 (17/17)	0.680	0.546 (30/55)	1.0 (30/30)	0.706
	unstemmed	0.833 (5/6)	0.294 (5/17)	0.435	0.857 (6/7)	0.2 (6/30)	0.324
light sensitivity	stemmed	0.661 (37/56)	1.0 (37/37)	0.796	0.692 (54/78)	1.0 (54/54)	0.818
	unstemmed	0.756 (31/41)	0.838 (31/37)	0.795	0.75 (39/52)	0.722 (39/54)	0.736

To compute precision and recall, we strictly define a correct hit as sentence that has query keywords in close proximity with joint meaning relevant to the query.

CHAPTER 4. RECOGNITION OF GENE MENTIONS IN ARSENIC TOLERANCE LITERATURE USING AN SVM CLASSIFIER AND SIMPLE CONTEXT FEATURES

4.1 Introduction

Arsenic toxicity and tolerance has been a focused research area for many years. Search of the simple term ‘arsenic toxicity’ using the PubMed search engine returned 5,248 articles on July, 18, 2014, and this number has been steadily growing in recent years. A large volume of literature makes it challenging for scientists to effectively extract relevant information and identify linkage among different pieces of scientific work. In a previous work, we have constructed the Textpresso-based literature mining tool for arsenic tolerance in plant and yeast. Textpresso provides a platform for organizing literature in a specialized field according to user-defined categories, and enables retrieval of biological facts by both keyword and category. In Textpresso, indices used to tag unstructured text have to be provided by users, and therefore the lack of a complete list of important biological concepts, e.g., all genes and proteins names involved in arsenic tolerance, has limited the retrieval coverage of facts in the literature. Over the past decade, much effort has been devoted to the development of automated extraction of biological entities from free text. Identifying those key concepts of interest is the cornerstone for many downstream text mining tasks, e.g., gene mention normalization, extraction of relationships, ontology construction, etc.

Named entity recognition (NER) in biology is notoriously difficult. First, naming of biological entities is not consistent. There are few well-accepted conventions that researchers consistently apply to present biological entities in text. It is common for a bio-entity to have a number of synonyms, which may be used interchangeably; furthermore, researchers frequently introduce their own names or abbreviations instead of following existing rules. Second, a single biological entity can be represented by multiple concepts depending on the context, which further complicates the task of automated detection and classification.

Recognizing named biomedical entities has been a research focus in biological text analysis, and several systems have been proposed in this field, such as ABNER [133] and BANNER [134]. Most proposed systems apply machine-learning approaches. Machine-learning approaches usually involve training a classifier on a collection of features extracted from a given corpus. Extraction of complex features from the corpus is time-consuming and labor-intense, especially for a new domain where annotated corpora are not always available. The existing NER systems have been trained on specific corpora, and therefore may not be easily transferred to other biological corpus if the representative features significantly differ. Moreover, there are few biomedical NER packages that are currently accessible by the public, and most advances in this field have limited to academic discussions and are not yet been coded into ready-to-use form.

We have developed an innovative biological named entity recognition system using a Support Vector Machine (SVM) and simple contextual features. The idea is based on the observation that gene and protein names frequently co-occur in a restricted set of contexts, therefore the significance of the co-occurrence with contextual words indicates

the likelihood of a word being a gene or protein name. We first extracted contextual words that most often appear in the close neighborhood of pre-identified genes/proteins in the collection of full text papers on arsenic tolerance in plants and yeast. During feature evaluation, we proposed a new measure - supportive ratio, which measures how often a context is used in the description of the target terms. Here context is defined as the words appear in the same sentence and in adjacent to the gene or protein names. During the classification step, SVM-based classifiers have been trained to identify potential biological name mentions. We also evaluated the effect of search scope of context-term co-occurrence on the classification performances. The performance of our system is comparable to ABNER [133] on unseen texts, but requires only contextual features and allows simple adaptation to any corpus.

4.2 Material and methods

4.2.1 Dataset preparation

The experiments were conducted on three types of bio-entities datasets, namely gene symbols, AGI gene codes, and enzyme names. AGI gene codes are generated from a uniform gene nomenclature system for the plant model species, which combines the name of the organism, type of the associated sequence (gene or repeat) and location in the chromosome. Gene symbols on the other hand, do not have a uniform naming system, and typically consist of 3 or 4 letters that define either a single gene or a gene family. The list of gene symbols was downloaded from The Arabidopsis Information Resource ftp (ftp://ftp.arabidopsis.org/home/tair/Genes/, gene_aliases version 2013-08-31). AGI gene codes were retrieved from TAIR10 release of *Arabidopsis* genome, including all locus

names and their variants. The list of enzyme names was downloaded and processed using Perl script through KEGG API service (www.kegg.jp/kegg/docs/keggapi.html).

Each of the retrieved biological names was searched against the full text literature on arsenic tolerance in plants/yeast by Textpresso. The search returned 5,114 sentences that have at least one annotated gene name. 4,114 gene containing sentences which contain a total of 676 annotated gene names were used for contextual feature extraction and training, another 150 sentences were kept to test the predictive performance of proposed name recognition methods. 1,000 AGI names and 1,000 enzymes were randomly selected, and were balanced with an equal number of random words from articles of a different knowledge domain to form training sets. Testing sets for AGI names and enzymes each consists of 500 AGIs/enzymes and 500 random words. Table 4.1 summarizes the counts of different types of entities used in this study.

4.2.2 Context extraction and feature representation

A context is a sequence of $2N+1$ words centered on the target biological entity. We selected $N=3$ and extracted all the contexts *Cg* within the 7-word window surrounding the annotated genes *G* in 4114 gene-containing sentences. Extracted contexts were filtered for stop words and then ranked by their frequencies in the training sentences. The contextual features were manually selected from the top ranking contexts that co-occur most frequently with gene names.

We propose a supportive ratio method that utilizes web evidence to describe the how likely the named entities are accompanied by a specific context in the literature. The supportive ratio is defined as the number of database documents in which a named entity

appears in a specific context, divided by the total number of database documents that have that named entity. All named entities in the training and test sets were queried with and without selected contexts against the PubMed database, and the numbers of documents retrieved were used to calculate the supportive ratio as defined below.

$$\text{supportive ratio} = \frac{\#\ \text{documents containing both named entity and a context}}{\#\ \text{documents containing only the named entity}} \quad (1)$$

The performance of the supportive ratio was also compared to that of 1/0 representation for co-occurrence of named entities and contexts, where 1 stands for occurrence within certain search scope (document level or sentence level). The effects of search scope on classification performance were further tested using the collection of full text articles in the Textpresso database. For this purpose, pairs of named entity and context were searched against Textpresso arsenic articles for co-occurrence at both the document level and the sentence level to generate corresponding word-context vectors for SVM analysis.

Sentences used for labeling were first prepared by tokenization and the removal of stop words and punctuations. Then the word-context co-occurrence vector was created for each token, which is the direct input to classifier.

4.2.3 Support vector machine classification

The support vector machine (SVM) was first introduced by Cortes and Vapnik [135]. SVM is a linear model working in a high dimensional feature space formed by the

nonlinear mapping of the *n*-dimensional input vector *x* into a *K*-dimensional feature space ($K > n$) through the use of a mapping function. The classification is achieved by choosing a separating hyperplane that achieves the maximal margin, where the nearest point to the hyperplane within each class is as far as possible from the hyperplane. In this work, we only consider the NER problem as a binary classification problem. Unlike the multi-classifier approaches that try to distinguish among several types of entities, our approach only attempts to separate single-unit gene/protein mentions from the rest of the text. The SVM classifier was constructed using the LIBSVM [136] software package, which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. An SVM classifier was trained on each type of biological entity and the prediction performance was measured by precision, recall and F_1 score:

$$\text{Precision}=\frac{tp}{tp+fp} \tag{2}$$

$$\text{Recall}=\frac{tp}{tp+fn} \tag{3}$$

$$F_1 = 2\cdot\frac{\text{precision}\cdot\text{recall}}{\text{precision+recall}} \tag{4}$$

where *tp* stands for true positives, *fp* stands for false positive and *fn* stands for false negatives.

4.2.4 TF-IDF filtering

TF-IDF (Term Frequency – Inverse Document Frequency) is defined as the product of Term Frequency (*tf*) and the Inverse of Document Frequency (*idf*). For a specific term t_i in a document d_j, its *tf* is calculated as:

$$tf_{ij} = \frac{frequency_{ij}}{\sum_k frequency_{kj}} \tag{5}$$

where $frequency_{kj}$ is the number of occurrence of the term t_i in document d_j. The denominator is the total number of terms in the document d_j. The *idf* of a term t_i is computed as:

$$idf_i = \log \frac{N}{n_i} \tag{6}$$

where N is the total number of documents in the corpus, and n_i is the number of documents in which the term t_i appears.

The TF-IDF weight of each non-stop word was calculated for every document in the corpus. For each document in the corpus, words in the document were ranked according to the TF-IDF weight from high to low. The TF-IDF threshold was set as the percentage of ranking. During prediction, the SVM-classifier was first applied to label each word in the test set as gene or non-gene. If the TF-IDF filtering is enabled, words must also exceed the threshold on the TF-IDF weight in order to be labeled as gene.

4.3 Results and Discussions

4.3.1 Context extraction and evaluation

54,975 unique words were found in 4,114 gene-containing sentences. After filtering stop words, punctuation and pure numbers, 33 contextual features with different word stems were manually selected from the top-ranking meaningful words that occur most frequently in the 7-word windows centered at gene/protein names. The set of

selected contextual words is listed in Table 4.1. These words describe the most common text environment where a gene or protein name tends to be found.

4.3.2 Feature representation

One of the limiting steps for automated named entities recognition (NER) is the lack of sufficient hand-annotated corpora/training sets, which are time-consuming to obtain. Moreover, most machine-learning-based NER systems are built on corpus of a specific knowledge domain, which are not easily transferable to related but different domains [137]. However, large amounts of unlabeled text are often available for most domains. This fact motivates leveraging web evidence from online literature databases to enrich training data. In a previous study, Brewster et. al [138] exploited web evidence to decide whether a candidate concept belonged to animal behaviors. In their work, a set of semantic patterns containing the candidate term were queried against the entire web via the *Yahoo! BOSS* search engine. If a query phrase was found to have at least one hit, the candidate term was taken as a legitimate term while no hits indicates that this term should be excluded from consideration. In our case, instead of searching the entire web, the PubMed database was queried by a named entity with and without selected contexts through the Entrez programming utilities (E-Utilities) API service [139] (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/). Given a query, the PubMed search engine looks for its presence in the fields of title, abstract, authors, and MESH term tags, etc., of the database articles. Numbers of articles that contain the search query were extracted from the search results, and were used to compute supportive ratio as defined in formula (1). If no article was returned for a specific combination of a named entity and a context,

the supportive ratio of the context for this named entity was set to be 0. Feature vectors using 1/0 representation were also constructed for the same sets of data, where 1 indicates the co-occurrence of a named entity with a certain context at the specified search level (document/sentence).

4.3.3 Comparison of different feature representations

An SVM classifier was trained on each type of named entity with different feature representations. The performance was summarized in Table 4.3. In comparison to 1/0 representation, supportive ratio vector improved the F_1 scores for gene symbols and enzymes by 7.14% and 4.05%, respectively, while the prediction precision on AGI names slightly decreased, leading to a drop of 1.19% in F_1 score. However, even in the worst case, an SVM classifier trained on the supportive ratio still achieved approximately 86.13% in F_1 score, indicating the potential of the proposed approach. Differences in the predictive abilities of SVM classifiers are largely due to the fact that AGI names and enzymes are characterized by unique. For example, the AGI name for a gene always starts with the organism abbreviation followed by the chromosome number, the sequence type (gene or repeats), and the gene id, e.g., AT2G01650. Enzyme names also follow strict naming conventions, and each enzyme is described by a sequence of four numbers preceded by “EC”, where the numbers indicate functional classification, e.g., EC 1.1.1.1. On the other hand, gene symbols are defined by individual researchers, often contain three or four-letter terms resembling words that are much more common in daily use thus less distinguishable from the rest of the text, e.g., the abbreviation of the short

meristemless gene (STM) is the same as the stock name of the STMcroelectronics company.

4.3.4 Comparison of different search levels

The selection of search level could affect the precision of retrieving articles related to a combination of keywords. In articles returned by document-level searches, despite the co-occurrence, the target keywords could appear in different sections of the text, and may not be closely related. The sentence-level search pinpoints the contexts where query matches appear, which is important for precise retrieval of biological facts. Therefore we tested the effect of different search levels on classification performance. A total of 720 gene symbols were annotated in the arsenic tolerance corpus by Textpresso indexing. Annotated gene symbols and 500 words randomly selected from the same corpus were searched against the Textpresso database for co-occurrence with the selected 33 contexts at both full text and sentence levels, and two co-occurrence matrices were constructed accordingly. Table 4.3 shows the summary of training and test data for two search levels.

Classification using sentence level co-occurrence led to improved recall rate at the expense of decreased precision, but the F_1 score was still higher for sentence level vectors. The AUC measures were also comparable between the two levels of co-occurrence. It seems that once enough context features are included in the SVM model, different levels of co-occurrence may have less than expected effects on the overall identification performance. However, the sentence level co-occurrence yields a higher recall rate, which could be important for constructing a complete list of gene names.

4.3.5 Comparison with ABNER on unseen text

ABNER [133] is a state-of-art software tool designed for biological named entities recognition, which is based on a statistical machine learning system using linear-chain conditional random fields (CRFs) [140] with a variety of orthographic and contextual features. It achieved a 69.9% F_1 score for tagging protein/gene names in the BioCreative corpus [133]. The SVM-classifier trained on a sentence-level co-occurrence vector of gene symbols and the context set in previous step was compared with ABNER to tag gene/protein names in 150 new sentences that have 94 gene/protein names in total. The proposed model correctly tagged 68 gene names and achieved 72.34% in recall rate, while ANBER only identified 59 correct names and had a recall rate of 62.77%. SVM-based classifier predicted 262 gene names, which is 95 more than ABNER, but the precision was lower. The SVM classifier and ABNER achieved 25.95% and 35.32% in terms of F_1 score, thus the overall tagging performance of ABNER on unseen text was slightly superior to the proposed model.

To improve the prediction precision of proposed model, the list of predicted gene/protein names were further refined by setting thresholds based on their TF-IDF weight. TF-IDF is high when a term *t* occurs many times within a small number of documents, lower when the term occurs fewer times in one document or occurs in many documents, and the lowest when the term occurs in virtually all documents. Thus TF-IDF weighting scheme can help to identify terms that are highly specific to the topic conveyed by the corpus of literature. By In terms of F1 measure, the best performance of the proposed model was achieved with probability threshold = 0.91, where precision = 37.06%, recall=81.84% and the highest F1-measure is 0.5102.

Neither ABNER nor the proposed model gave completely accurate prediction on the testing sentences. Despite the low precision, the proposed model achieved a high recall rate, which is more important for the purpose of finding the complete list of gene names. After applying TF-IDF screen, the precision of the same model was increased by 14.7%, and the recall rate is still comparable to that of ABNER.

4.4 Conclusions

Overall, the proposed approach is promising in identifying terms such as gene/protein names. Using co-occurrence with selected contexts, the proposed SVM model is able to generate a list that is enriched with gene/protein names. The accuracy can be further improved by filtering the predicted positives with TF-IDF weights. For biologists, an automated system with high recall and even moderate precision (like the current Textpresso) confers a great advantage over skimming text by eye.

Table 4.1 Frequencies of selected contexts in gene-containing sentences.

Context	**Frequency**	**Percentage**	**Rank**
protein	717	17.43	1
gene	525	12.76	2
arsenic	476	11.57	3
expression	458	11.13	4
arsenate	332	8.07	9
activity	306	7.44	11
arsenite	225	5.47	14
cells	218	5.30	15
promoter	211	5.13	17
control	209	5.08	18
tolerance	192	4.67	21
plant	192	4.67	21
levels	189	4.59	22
resistance	172	4.18	27
stress	163	3.96	31
mutant	155	3.77	33
response	154	3.74	34
involved	142	3.45	37
growth	140	3.40	38
increased	139	3.38	39
membrane	134	3.26	42
transport	129	3.14	45
transcription	118	2.87	48
accumulation	116	2.82	49
biosynthesis	111	2.70	52
metal	111	2.70	52
sequence	108	2.63	55
concentration	108	2.63	55
function	90	2.19	69
regulation	83	2.02	75
complex	69	1.68	87
signaling	26	0.63	129
interaction	10	0.24	145

Note: the value of percentage is calculated by dividing the number of gene-containing sentences that have the given context by the total number of gene-containing sentences.

Table 4.2 Counts of different types of entities used for training and testing.

	Gene containing sentences	**Gene Symbols**	**AGI**	**Enzyme**
Training	4114	676(+)/800(-)	1,000(+)/1,000(-)	1,000(+)/1000(-)
Test	150	95(+)/200(-)	500(+)/500(-)	500(+)/500(-)

Note: “+” denotes positive case, and “-” denotes negative case.

Table 4.3 Performance of SVM-based NER system using features evaluated on web evidence.

	1/0 Representation			**Support Ratio**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Gene symbol	0.8878	0.6643	0.7600	0.9008	0.8252	0.8613
AGI name	0.9632	0.994	0.9784	0.9416	1	0.9699
Enzyme name	0.8946	0.9	0.8973	0.9259	0.95	0.9378

Table 4.4 Counts of positive and negative cases in training and testing sets for different levels of co-occurrence.

	Full Text		**Sentence**	
	Train	**Test**	**Train**	**Test**
Gene symbols	575	143	478	119
Random words	373	93	281	70

Table 4.5 Effects of levels of co-occurrence on the performance of SVM classifier.

	Precision	**Recall**	**F1 score**
Full text	0.8558	0.6643	0.7480
Sentence	0.7946	0.7479	0.7705

Table 4.6 Performance of the proposed SVM classifier and ABER on unseen text.

	Precision	**Recall**	**F1 score**
SVM	0.2595	0.7234	0.3820
SVM+TFIDF	0.2792	0.7128	0.4012
ABNER	0.3533	0.6277	0.4521

Note: the best performance of SVM + TF-IDF classifier was achieved when the threshold of TF-IDF was set to top 20%.

REFERENCES

REFERENCES

1. Ma, L.Q., K.M. Komar, C. Tu, W. Zhang, Y. Cai, and E.D. Kennelley, *A fern that hyperaccumulates arsenic.* Nature, 2001. **409**(6820): p. 579-579.
2. Nordstrom, D.K. *Arsenic Geochemistry: Overview of an Underhanded Element*. in *U.S. EPA Workshop on Managing Arsenic Risks to the Environment: Characterization of Waste, Chemistry, and Treatment and Disposal*. 2001. Denver, Colorado.
3. Drahota, P. and M. Filippi, *Secondary arsenic minerals in the environment: A review.* Environment International, 2009. **35**(8): p. 1243-1255.
4. Smith, E., R. Naidu, and A.M. Alston, *Arsenic in the Soil Environment: A Review*, in *Advances in Agronomy*, L.S. Donald, Editor. 1998, Academic Press. p. 149-195.
5. Smith, A.H. and M.M.H. Smith, *Arsenic drinking water regulations in developing countries with extensive exposure.* Toxicology, 2004. **198**(1–3): p. 39-44.
6. Yamamura, S., J. Bartram, M. Csanady, H.G. Gorchev, and A. Redekopp. *Drinking Water Guidelines and Standards*. in *World Health Organization*. 2004. Geneva, Switzerland.
7. Bhattacharjee, Y., *A Sluggish Response to Humanity's Biggest Mass Poisoning.* Science, 2007. **315**(5819): p. 1659-1661.
8. Abedin, M.J., J. Feldmann, and A.A. Meharg, *Uptake Kinetics of Arsenic Species in Rice Plants.* Plant Physiology, 2002. **128**(3): p. 1120-1128.
9. Ohno, K., T. Yanase, Y. Matsuo, T. Kimura, M. Hamidur Rahman, Y. Magara, and Y. Matsui, *Arsenic intake via water and food by a population living in an arsenic-affected area of Bangladesh.* Science of The Total Environment, 2007. **381**(1–3): p. 68-76.
10. Schoof, R.A., L.J. Yost, J. Eickhoff, E.A. Crecelius, D.W. Cragin, D.M. Meacher, and D.B. Menzel, *A Market Basket Survey of Inorganic Arsenic in Food.* Food and Chemical Toxicology, 1999. **37**(8): p. 839-846.
11. Meharg, A.A., P.N. Williams, E. Adomako, Y.Y. Lawgali, C. Deacon, A. Villada, R.C.J. Cambell, G. Sun, Y.-G. Zhu, J. Feldmann, A. Raab, F.-J. Zhao, R. Islam, S. Hossain, and J. Yanai, *Geographical Variation in Total and Inorganic Arsenic Content of Polished (White) Rice.* Environmental Science & Technology, 2009. **43**(5): p. 1612-1617.
12. Mondal, P., C.B. Majumder, and B. Mohanty, *Laboratory based approaches for arsenic remediation from contaminated water: Recent developments.* Journal of Hazardous Materials, 2006. **137**(1): p. 464-479.

13. Srivastava, M., L.Q. Ma, and J.A.G. Santos, *Three new arsenic hyperaccumulating ferns.* Science of The Total Environment, 2006. **364**(1–3): p. 24-31.
14. Wang, H.-S., C.-F. Yung, and F.-R. Chang, *1 Introduction*, in *\mathcal{H}_\infty Control for Nonlinear Descriptor Systems*. 2006, Springer London. p. 1-11.
15. Kalve, S., B. Ketan Sarangi, R.A. Pandey, and T. Chakrabarti, *Arsenic and chromium hyperaccumulation by an ecotype of Pteris vittata -- prospective for phytoextraction from contaminated water and soil.* Current Science (00113891), 2011. **100**(6): p. 888-894.
16. Zhao, F.J., J.F. Ma, A.A. Meharg, and S.P. McGrath, *Arsenic uptake and metabolism in plants.* New Phytologist, 2009. **181**(4): p. 777-794.
17. Hartley-Whitaker, J., G. Ainsworth, and A.A. Meharg, *Copper- and arsenate-induced oxidative stress in Holcus lanatus L. clones with differential sensitivity.* Plant, Cell & Environment, 2001. **24**(7): p. 713-722.
18. Bleeker, P.M., H.W.J. Hakvoort, M. Bliek, E. Souer, and H. Schat, *Enhanced arsenate reduction by a CDC25-like tyrosine phosphatase explains increased phytochelatin accumulation in arsenate-tolerant Holcus lanatus.* The Plant Journal, 2006. **45**(6): p. 917-929.
19. Dhankher, O.P., B.P. Rosen, E.C. McKinney, and R.B. Meagher, *Hyperaccumulation of arsenic in the shoots of Arabidopsis silenced for arsenate reductase (ACR2).* Proceedings of the National Academy of Sciences, 2006. **103**(14): p. 5413-5418.
20. Kitchin, K.T. and K. Wallace, *Arsenite binding to synthetic peptides: The effect of increasing length between two cysteines.* Journal of Biochemical and Molecular Toxicology, 2006. **20**(1): p. 35-38.
21. Cline, D.J., C. Thorpe, and J.P. Schneider, *Effects of As(III) Binding on α-Helical Structure.* Journal of the American Chemical Society, 2003. **125**(10): p. 2923-2929.
22. Ramadan, D., D.J. Cline, S. Bai, C. Thorpe, and J.P. Schneider, *Effects of As(III) Binding on β-Hairpin Structure.* Journal of the American Chemical Society, 2007. **129**(10): p. 2981-2988.
23. Requejo, R. and M. Tena, *Proteome analysis of maize roots reveals that oxidative stress is a main contributing factor to plant arsenic toxicity.* Phytochemistry, 2005. **66**(13): p. 1519-1528.
24. Singh, N., L.Q. Ma, M. Srivastava, and B. Rathinasabapathi, *Metabolic adaptations to arsenic-induced oxidative stress in Pteris vittata L and Pteris ensiformis L.* Plant Science, 2006. **170**(2): p. 274-282.
25. Møller, I.M., P.E. Jensen, and A. Hansson, *Oxidative Modifications to Cellular Components in Plants.* Annual Review of Plant Biology, 2007. **58**(1): p. 459-481.
26. Finnegan, P. and W. Chen, *Arsenic toxicity: the effects on plant metabolism.* Frontiers in Physiology, 2012. **3**.
27. Fendorf, S., M.J. Herbel, K.J. Tufano, and B.D. Kocar, *Biogeochemical Processes Controlling the Cycling of Arsenic in Soils and Sediments*, in *Biophysico-Chemical Processes of Heavy Metals and Metalloids in Soil Environments*. 2007, John Wiley & Sons, Inc. p. 313-338.

28. Bolan, N., S. Mahimairaja, A. Kunhikrishnan, B. Seshadri, and R. Thangarajan, *Bioavailability and ecotoxicity of arsenic species in solution culture and soil system: implications to remediation.* Environmental Science and Pollution Research, 2013: p. 1-10.
29. MEHARG, A.A. and M.R. MACNAIR, *Suppression of the High Affinity Phosphate Uptake System: A Mechanism of Arsenate Tolerance in Holcus lanatus L.* Journal of Experimental Botany, 1992. **43**(4): p. 519-524.
30. Clark, G.T., J. Dunlop, and H.T. Phung, *Phosphate absorption by Arabidopsis thaliana: interactions between phosphorus status and inhibition by arsenate.* Functional Plant Biology, 2000. **27**(10): p. 959-965.
31. Meharg, A.A. and M.R. Macnair, *The mechanisms of arsenate tolerance in Deschampsia cespitosa (L.) Beauv. and Agrostis capillaris L.* New Phytologist, 1991. **119**(2): p. 291-297.
32. Tu, S. and L.Q. Ma, *Interactive effects of pH, arsenic and phosphorus on uptake of As and P and growth of the arsenic hyperaccumulator Pteris vittata L. under hydroponic conditions.* Environmental and Experimental Botany, 2003. **50**(3): p. 243-251.
33. Catarecha, P., M.D. Segura, J.M. Franco-Zorrilla, B. Garc ́a-Ponce, M. Lanza, R. Solano, J. Paz-Ares, and A. Leyva, *A Mutant of the Arabidopsis Phosphate Transporter PHT1;1 Displays Enhanced Arsenic Accumulation.* The Plant Cell Online, 2007. **19**(3): p. 1123-1133.
34. Shin, H., H.-S. Shin, G.R. Dewbre, and M.J. Harrison, *Phosphate transport in Arabidopsis: Pht1;1 and Pht1;4 play a major role in phosphate acquisition from both low- and high-phosphate environments.* The Plant Journal, 2004. **39**(4): p. 629-642.
35. Poynton, C., J. Huang, M. Blaylock, L. Kochian, and M. Elless, *Mechanisms of arsenic hyperaccumulation in Pteris species: root As influx and translocation.* Planta, 2004. **219**(6): p. 1080-1088.
36. Maurel, C., L. Verdoucq, D.-T. Luu, and V. Santoni, *Plant Aquaporins: Membrane Channels with Multiple Integrated Functions.* Annual Review of Plant Biology, 2008. **59**(1): p. 595-624.
37. Mukhopadhyay, R., H. Bhattacharjee, and B.P. Rosen, *Aquaglyceroporins: Generalized metalloid channels.* Biochimica et Biophysica Acta (BBA) - General Subjects, 2014. **1840**(5): p. 1583-1591.
38. Bienert, G., M. Thorsen, M. Schussler, H. Nilsson, A. Wagner, M. Tamas, and T. Jahn, *A subgroup of plant aquaporins facilitate the bi-directional diffusion of As(OH)3 and Sb(OH)3 across membranes.* BMC Biology, 2008. **6**(1): p. 26.
39. Isayenkov, S.V. and F.J.M. Maathuis, *The Arabidopsis thaliana aquaglyceroporin AtNIP7;1 is a pathway for arsenite uptake.* FEBS Letters, 2008. **582**(11): p. 1625-1628.
40. Ma, J.F., N. Yamaji, N. Mitani, X.-Y. Xu, Y.-H. Su, S.P. McGrath, and F.-J. Zhao, *Transporters of arsenite in rice and their role in arsenic accumulation in rice grain.* Proceedings of the National Academy of Sciences, 2008. **105**(29): p. 9931-9935.

41. *Revised reregistration eligibility decision for MSMA, DSMA, CAMA, and cacodylic acid*, U.S.E.P. Agency, Editor. 2006.
42. Li, R.-Y., Y. Ago, W.-J. Liu, N. Mitani, J. Feldmann, S.P. McGrath, J.F. Ma, and F.-J. Zhao, *The Rice Aquaporin Lsi1 Mediates Uptake of Methylated Arsenic Species.* Plant Physiology, 2009. **150**(4): p. 2071-2080.
43. Raab, A., P.N. Williams, A. Meharg, and J. Feldmann, *Uptake and translocation of inorganic and methylated arsenic species by plants.* Environmental Chemistry, 2007. **4**(3): p. 197-203.
44. Jia, Y., H. Huang, G.-X. Sun, F.-J. Zhao, and Y.-G. Zhu, *Pathways and Relative Contributions to Arsenic Volatilization from Rice Plants and Paddy Soil.* Environmental Science & Technology, 2012. **46**(15): p. 8090-8096.
45. Carey, A.-M., G.J. Norton, C. Deacon, K.G. Scheckel, E. Lombi, T. Punshon, M.L. Guerinot, A. Lanzirotti, M. Newville, Y. Choi, A.H. Price, and A.A. Meharg, *Phloem transport of arsenic species from flag leaf to grain during grain filling.* New Phytologist, 2011. **192**(1): p. 87-98.
46. Quaghebeur, M. and Z. Rengel, *Arsenic uptake, translocation and speciation in pho1 and pho2 mutants of Arabidopsis thaliana.* Physiologia Plantarum, 2004. **120**(2): p. 280-286.
47. Raab, A., H. Schat, A.A. Meharg, and J. Feldmann, *Uptake, translocation and transformation of arsenate and arsenite in sunflower (Helianthus annuus): formation of arsenic–phytochelatin complexes during exposure to high arsenic concentrations.* New Phytologist, 2005. **168**(3): p. 551-558.
48. Su, Y.-H., S. McGrath, and F.-J. Zhao, *Rice is more efficient in arsenite uptake and translocation than wheat and barley.* Plant and Soil, 2010. **328**(1-2): p. 27-34.
49. Lombi, E., F.-J. Zhao, M. Fuhrmann, L.Q. Ma, and S.P. McGrath, *Arsenic distribution and speciation in the fronds of the hyperaccumulator Pteris vittata.* New Phytologist, 2002. **156**(2): p. 195-203.
50. Pickering, I.J., L. Gumaelius, H.H. Harris, R.C. Prince, G. Hirsch, J.A. Banks, D.E. Salt, and G.N. George, *Localizing the Biochemical Transformations of Arsenate in a Hyperaccumulating Fern.* Environmental Science & Technology, 2006. **40**(16): p. 5010-5014.
51. Dhankher, O.P., Y. Li, B.P. Rosen, J. Shi, D. Salt, J.F. Senecoff, N.A. Sashti, and R.B. Meagher, *Engineering tolerance and hyperaccumulation of arsenic in plants by combining arsenate reductase and [gamma]-glutamylcysteine synthetase expression.* Nat Biotech, 2002. **20**(11): p. 1140-1145.
52. Pickering, I.J., R.C. Prince, M.J. George, R.D. Smith, G.N. George, and D.E. Salt, *Reduction and Coordination of Arsenic in Indian Mustard.* Plant Physiology, 2000. **122**(4): p. 1171-1178.
53. Xu, X.Y., S.P. McGrath, and F.J. Zhao, *Rapid reduction of arsenate in the medium mediated by plant roots.* New Phytologist, 2007. **176**(3): p. 590-599.
54. Duan, G.-L., Y. Zhou, Y.-P. Tong, R. Mukhopadhyay, B.P. Rosen, and Y.-G. Zhu, *A CDC25 homologue from rice functions as an arsenate reductase.* New Phytologist, 2007. **174**(2): p. 311-321.

55. Ellis, D.R., L. Gumaelius, E. Indriolo, I.J. Pickering, J.A. Banks, and D.E. Salt, *A Novel Arsenate Reductase from the Arsenic Hyperaccumulating Fern Pteris vittata.* Plant Physiology, 2006. **141**(4): p. 1544-1554.
56. Rathinasabapathi, B., S. Wu, S. Sundaram, J. Rivoal, M. Srivastava, and L. Ma, *Arsenic resistance in Pteris vittata L.: identification of a cytosolic triosephosphate isomerase based on cDNA expression cloning in Escherichia coli.* Plant Molecular Biology, 2006. **62**(6): p. 845-857.
57. Schmöger, M.E.V., M. Oven, and E. Grill, *Detoxification of Arsenic by Phytochelatins in Plants.* Plant Physiology, 2000. **122**(3): p. 793-802.
58. Sneller, F.E.C., L.M. Van Heerwaarden, F.J.L. Kraaijeveld-Smit, W.M. Ten Bookum, P.L.M. Koevoets, H. Schat, and J.A.C. Verkleij, *Toxicity of arsenate in Silene vulgaris, accumulation and degradation of arsenate-induced phytochelatins.* New Phytologist, 1999. **144**(2): p. 223-232.
59. Raab, A., J. Feldmann, and A.A. Meharg, *The Nature of Arsenic-Phytochelatin Complexes in Holcus lanatus and Pteris cretica.* Plant Physiology, 2004. **134**(3): p. 1113-1122.
60. Norton, G.J., D.E. Lou-Hing, A.A. Meharg, and A.H. Price, *Rice–arsenate interactions in hydroponics: whole genome transcriptional analysis.* Journal of Experimental Botany, 2008. **59**(8): p. 2267-2276.
61. Schat, H., M. Llugany, R. Vooijs, J. Hartley-Whitaker, and P.M. Bleeker, *The role of phytochelatins in constitutive and adaptive heavy metal tolerances in hyperaccumulator and non-hyperaccumulator metallophytes.* Journal of Experimental Botany, 2002. **53**(379): p. 2381-2392.
62. Hartley-Whitaker, J., C. Woods, and A.A. Meharg, *Is differential phytochelatin production related to decreased arsenate influx in arsenate tolerant Holcus lanatus?* New Phytologist, 2002. **155**(2): p. 219-225.
63. Ha, S.-B., A.P. Smith, R. Howden, W.M. Dietrich, S. Bugg, M.J. O'Connell, P.B. Goldsbrough, and C.S. Cobbett, *Phytochelatin Synthase Genes from Arabidopsis and the Yeast Schizosaccharomyces pombe.* The Plant Cell Online, 1999. **11**(6): p. 1153-1163.
64. Zhao, F.J., J.R. Wang, J.H.A. Barker, H. Schat, P.M. Bleeker, and S.P. McGrath, *The role of phytochelatins in arsenic tolerance in the hyperaccumulator Pteris vittata.* New Phytologist, 2003. **159**(2): p. 403-410.
65. Ghosh, M., J. Shen, and B.P. Rosen, *Pathways of As(III) detoxification in Saccharomyces cerevisiae.* Proceedings of the National Academy of Sciences, 1999. **96**(9): p. 5001-5006.
66. Ortiz, D.F., T. Ruscitti, K.F. McCue, and D.W. Ow, *Transport of Metal-binding Peptides by HMT1, A Fission Yeast ABC-type Vacuolar Membrane Protein.* Journal of Biological Chemistry, 1995. **270**(9): p. 4721-4728.
67. Indriolo, E., G. Na, D. Ellis, D.E. Salt, and J.A. Banks, *A Vacuolar Arsenite Transporter Necessary for Arsenic Tolerance in the Arsenic Hyperaccumulating Fern Pteris vittata Is Missing in Flowering Plants.* The Plant Cell Online, 2010. **22**(6): p. 2045-2057.

68. Wang, J., F.-J. Zhao, A.A. Meharg, A. Raab, J. Feldmann, and S.P. McGrath, *Mechanisms of Arsenic Hyperaccumulation in Pteris vittata. Uptake Kinetics, Interactions with Phosphate, and Arsenic Speciation.* Plant Physiology, 2002. **130**(3): p. 1552-1561.
69. Meharg, A.A. and J. Hartley-Whitaker, *Arsenic uptake and metabolism in arsenic resistant and nonresistant plant species.* New Phytologist, 2002. **154**(1): p. 29-43.
70. Francesconi, K., P. Visoottiviseth, W. Sridokchan, and W. Goessler, *Arsenic species in an arsenic hyperaccumulating fern, Pityrogramma calomelanos: a potential phytoremediator of arsenic-contaminated soils.* Science of The Total Environment, 2002. **284**(1–3): p. 27-35.
71. Shelmerdine, P.A., C.R. Black, S.P. McGrath, and S.D. Young, *Modelling phytoremediation by the hyperaccumulating fern, Pteris vittata, of soils historically contaminated with arsenic.* Environmental Pollution, 2009. **157**(5): p. 1589-1596.
72. Wang, X., L.Q. Ma, B. Rathinasabapathi, Y. Cai, Y.G. Liu, and G.M. Zeng, *Mechanisms of Efficient Arsenite Uptake by Arsenic Hyperaccumulator Pteris vittata.* Environmental Science & Technology, 2011. **45**(22): p. 9719-9725.
73. Lou, L.Q., Z.H. Ye, and M.H. Wong, *A comparison of arsenic tolerance, uptake and accumulation between arsenic hyperaccumulator, Pteris vittata L. and non-accumulator, P. semipinnata L.—A hydroponic study.* Journal of Hazardous Materials, 2009. **171**(1–3): p. 436-442.
74. Vetterlein, D., D. Wesenberg, P. Nathan, A. Bräutigam, A. Schierhorn, J. Mattusch, and R. Jahn, *Pteris vittata – Revisited: Uptake of As and its speciation, impact of P, role of phytochelatins and S.* Environmental Pollution, 2009. **157**(11): p. 3016-3024.
75. Indriolo, E., G. Na, D. Ellis, D.E. Salt, and J.A. Banks, *A Vacuolar Arsenite Transporter Necessary for Arsenic Tolerance in the Arsenic Hyperaccumulating Fern Pteris vittata Is Missing in Flowering Plants.* The Plant Cell Online, 2010.
76. Lei, M., X.-m. Wan, Z.-c. Huang, T.-b. Chen, X.-w. Li, and Y.-r. Liu, *First evidence on different transportation modes of arsenic and phosphorus in arsenic hyperaccumulator Pteris vittata.* Environmental Pollution, 2012. **161**(0): p. 1-7.
77. Mathews, S., B. Rathinasabapathi, and L.Q. Ma, *Uptake and translocation of arsenite by Pteris vittata L.: Effects of glycerol, antimonite and silver.* Environmental Pollution, 2011. **159**(12): p. 3490-3495.
78. Wysocki, R., C.C. Chéry, D. Wawrzycka, M. Van Hulle, R. Cornelis, J.M. Thevelein, and M.J. Tamás, *The glycerol channel Fps1p mediates the uptake of arsenite and antimonite in Saccharomyces cerevisiae.* Molecular Microbiology, 2001. **40**(6): p. 1391-1401.
79. Sundaram, S., B. Rathinasabapathi, L.Q. Ma, and B.P. Rosen, *An Arsenate-activated Glutaredoxin from the Arsenic Hyperaccumulator Fern Pteris vittata L. Regulates Intracellular Arsenite.* Journal of Biological Chemistry, 2008. **283**(10): p. 6095-6101.
80. Gumaelius, L., B. Lahner, D.E. Salt, and J.A. Banks, *Arsenic Hyperaccumulation in Gametophytes of Pteris vittata. A New Model System for Analysis of Arsenic Hyperaccumulation.* Plant Physiology, 2004. **136**(2): p. 3198-3208.

81. Schmieder, R. and R. Edwards, *Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets.* PLoS ONE, 2011. **6**(3): p. e17288.
82. Lohse, M., A.M. Bolger, A. Nagel, A.R. Fernie, J.E. Lunn, M. Stitt, and B. Usadel, *RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics.* Nucleic Acids Research, 2012. **40**(W1): p. W622-W627.
83. Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotech, 2011. **29**(7): p. 644-652.
84. Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. Madden, *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**(1): p. 421.
85. Parra, G., K. Bradnam, and I. Korf, *CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.* Bioinformatics, 2007. **23**(9): p. 1061-1067.
86. Langmead, B., C. Trapnell, M. Pop, and S. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. R25.
87. Li, B. and C. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**(1): p. 323.
88. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
89. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biology, 2010. **11**(10): p. R106.
90. Leng, N., J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould, R.M. Stewart, and C. Kendziorski, *EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.* Bioinformatics, 2013. **29**(8): p. 1035-1043.
91. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Statistics in Medicine, 1990. **9**(7): p. 811-818.
92. Alexa, A., J. Rahnenführer, and T. Lengauer, *Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.* Bioinformatics, 2006. **22**(13): p. 1600-1607.
93. Livak, K.J. and T.D. Schmittgen, *Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method.* Methods, 2001. **25**(4): p. 402-408.
94. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Research, 2011. **39**(suppl 2): p. W29-W37.
95. Petersen, T.N., S. Brunak, G. von Heijne, and H. Nielsen, *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Meth, 2011. **8**(10): p. 785-786.

96. Conesa, A. and S. Götz, *Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics.* International Journal of Plant Genomics, 2008. **2008**: p. 12.
97. McCarthy, F., N. Wang, G.B. Magee, B. Nanduri, M. Lawrence, E. Camon, D. Barrell, D. Hill, M. Dolan, W.P. Williams, D. Luthe, S. Bridges, and S. Burgess, *AgBase: a functional genomics resource for agriculture.* BMC Genomics, 2006. **7**(1): p. 229.
98. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: The European Molecular Biology Open Software Suite.* Trends in Genetics, 2000. **16**(6): p. 276-277.
99. Wu, S., Z. Zhu, L. Fu, B. Niu, and W. Li, *WebMGA: a customizable web server for fast metagenomic sequence analysis.* BMC Genomics, 2011. **12**(1): p. 444.
100. Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón, and M. Robles, *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics, 2005. **21**(18): p. 3674-3676.
101. Koonin, E., N. Fedorova, J. Jackson, A. Jacobs, D. Krylov, K. Makarova, R. Mazumder, S. Mekhedov, A. Nikolskaya, B. Rao, I. Rogozin, S. Smirnov, A. Sorokin, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale, *A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.* Genome Biology, 2004. **5**(2): p. R7.
102. Leng, N., J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould, R.M. Stewart, and C. Kendziorski, *EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments.* Bioinformatics, 2013.
103. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* BMC Bioinformatics, 2013. **14**(1): p. 91.
104. Tripathi, R.D., P. Tripathi, S. Dwivedi, S. Dubey, D. Chakrabarty, and P.k. Trivedi, *Arsenomics: Omics of Arsenic Metabolism in Plants.* Frontiers in Physiology, 2012. **3**.
105. Alexa, A. and J. Rahnenfuhrer, *topGO: Enrichment analysis for Gene Ontology*. 2010.
106. Rai, A., P. Tripathi, S. Dwivedi, S. Dubey, M. Shri, S. Kumar, P.K. Tripathi, R. Dave, A. Kumar, R. Singh, B. Adhikari, M. Bag, R.D. Tripathi, P.K. Trivedi, D. Chakrabarty, and R. Tuli, *Arsenic tolerances in rice (Oryza sativa) have a predominant role in transcriptional regulation of a set of genes including sulphur assimilation pathway and antioxidant system.* Chemosphere, 2011. **82**(7): p. 986-995.
107. Dixon, D.P., I. Cummins, D.J. Cole, and R. Edwards, *Glutathione-mediated detoxification systems in plants.* Current Opinion in Plant Biology, 1998. **1**(3): p. 258-266.
108. Hatton, P.J., D. Dixon, D.J. Cole, and R. Edwards, *Glutathione Transferase Activities and Herbicide Selectivity in Maize and Associated Weed Species.* Pesticide Science, 1996. **46**(3): p. 267-275.
109. Andrews, C.J., M. Skipsey, J.K. Townson, C. Morris, I. Jepson, and R. Edwards, *Glutathione transferase activities toward herbicides used selectively in soybean.* Pesticide Science, 1997. **51**(2): p. 213-222.

110. Mazel, A. and A. Levine, *Induction of glucosyltransferase transcription and activity during superoxide-dependent cell death in Arabidopsis plants.* Plant Physiology and Biochemistry, 2002. **40**(2): p. 133-140.
111. Meißner, D., A. Albert, C. Böttcher, D. Strack, and C. Milkowski, *The role of UDP-glucose:hydroxycinnamate glucosyltransferases in phenylpropanoid metabolism and the response to UV-B radiation in Arabidopsis thaliana.* Planta, 2008. **228**(4): p. 663-674.
112. Langlois-Meurinne, M., C.M.M. Gachon, and P. Saindrenan, *Pathogen-Responsive Expression of Glycosyltransferase Genes UGT73B3 and UGT73B5 Is Necessary for Resistance to Pseudomonas syringae pv tomato in Arabidopsis.* Plant Physiology, 2005. **139**(4): p. 1890-1901.
113. Kreuz, K., R. Tommasini, and E. Martinoia, *Old Enzymes for a New Job (Herbicide Detoxification in Plants).* Plant Physiology, 1996. **111**(2): p. 349-353.
114. Mukhopadhyay, A., S. Vij, and A.K. Tyagi, *Overexpression of a zinc-finger protein gene from rice confers tolerance to cold, dehydration, and salt stress in transgenic tobacco.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(16): p. 6309-6314.
115. Ahsan, N., D.-G. Lee, K.-H. Kim, I. Alam, S.-H. Lee, K.-W. Lee, H. Lee, and B.-H. Lee, *Analysis of arsenic stress-induced differentially expressed proteins in rice leaves by two-dimensional gel electrophoresis coupled with mass spectrometry.* Chemosphere, 2010. **78**(3): p. 224-231.
116. Gregus, Z. and B. Németi, *The Glycolytic Enzyme Glyceraldehyde-3-Phosphate Dehydrogenase Works as an Arsenate Reductase in Human Red Blood Cells and Rat Liver Cytosol.* Toxicological Sciences, 2005. **85**(2): p. 859-869.
117. Kufner, I. and W. Koch, *Stress regulated members of the plant organic cation transporter family are localized to the vacuolar membrane.* BMC Research Notes, 2008. **1**(1): p. 43.
118. Yamada, H., T. Suzuki, K. Terada, K. Takei, K. Ishikawa, K. Miwa, T. Yamashino, and T. Mizuno, *The Arabidopsis AHK4 Histidine Kinase is a Cytokinin-Binding Receptor that Transduces Cytokinin Signals Across the Membrane.* Plant and Cell Physiology, 2001. **42**(9): p. 1017-1023.
119. Sakakibara, H., K. Takei, and N. Hirose, *Interactions between nitrogen and cytokinin in the regulation of metabolism and development.* Trends in Plant Science, 2006. **11**(9): p. 440-448.
120. Matsubayashi, Y., M. Ogawa, H. Kihara, M. Niwa, and Y. Sakagami, *Disruption and Overexpression of Arabidopsis Phytosulfokine Receptor Gene Affects Cellular Longevity and Potential for Growth.* Plant Physiology, 2006. **142**(1): p. 45-53.
121. Loivamäki, M., N. Stührwohldt, R. Deeken, B. Steffens, T. Roitsch, R. Hedrich, and M. Sauter, *A role for PSK signaling in wounding and microbial interactions in Arabidopsis.* Physiologia Plantarum, 2010. **139**(4): p. 348-357.
122. Shen, Y. and A.C. Diener, *<italic>Arabidopsis thaliana RESISTANCE TO FUSARIUM OXYSPORUM 2</italic> Implicates Tyrosine-Sulfated Peptide Signaling in Susceptibility and Resistance to Root Infection.* PLoS Genet, 2013. **9**(5): p. e1003525.

123. Iwasaki, T., K. Yamaguchi-Shinozaki, and K. Shinozaki, *Identification of a cis-regulatory region of a gene in Arabidopsis thaliana whose induction by dehydration is mediated by abscisic acid and requires protein synthesis.* Molecular and General Genetics MGG, 1995. **247**(4): p. 391-398.
124. Datta, N., P. LaFayette, P. Kroner, R. Nagao, and J. Key, *Isolation and characterization of three families of auxin down-regulated cDNA clones.* Plant Molecular Biology, 1993. **21**(5): p. 859-869.
125. Gan, D., H. Jiang, J. Zhang, Y. Zhao, S. Zhu, and B. Cheng, *Genome-wide analysis of BURP domain-containing genes in Maize and Sorghum.* Molecular Biology Reports, 2011. **38**(7): p. 4553-4563.
126. Müller, H.M., E.E. Kenny, and P.W. Sternberg, *Textpresso: an ontology-based information retrieval and extraction system for biological literature.* PLoS Biol, 2004. **2**(11): p. e309.
127. Müller, H.M., A. Rangarajan, T.K. Teal, and P.W. Sternberg, *Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers.* Neuroinformatics, 2008. **6**(3): p. 195-204.
128. Garten, Y. and R.B. Altman, *Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text.* BMC Bioinformatics, 2009. **10 Suppl 2**: p. S6.
129. Urbanski, W.M. and B.G. Condie, *Textpresso site-specific recombinases: A text-mining server for the recombinase literature including Cre mice and conditional alleles.* Genesis, 2009. **47**(12): p. 842-6.
130. Skrzypek, M.S., M.B. Arnaud, M.C. Costanzo, D.O. Inglis, P. Shah, G. Binkley, S.R. Miyasato, and G. Sherlock, *New tools at the Candida Genome Database: biochemical pathways and full-text literature search.* Nucleic Acids Res, 2010. **38**(Database issue): p. D428-32.
131. Porter, M.F., *An algorithm for suffix stripping.* Program, 1980. **14**(3): p. 130-137.
132. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene Ontology: tool for the unification of biology.* Nat Genet, 2000. **25**(1): p. 25-29.
133. Settles, B., *ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.* Bioinformatics, 2005. **21**(14): p. 3191-3192.
134. Leaman, R. and G. Gonzalez, *BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition.* Pac Symp Biocomput, 2008: p. 652-663.
135. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.
136. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**(3): p. 27:1--27:27.
137. Pan, S.J., Z. Toh, and J. Su, *Transfer joint embedding for cross-domain named entity recognition.* ACM Trans. Inf. Syst., 2013. **31**(2): p. 1-27.
138. Brewster, C., S. Jupp, J. Luciano, D. Shotton, R. Stevens, and Z. Zhang, *Issues in learning an ontology from text.* BMC Bioinformatics, 2009. **10**(Suppl 5): p. S1.

139. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Research, 2013. **41**(D1): p. D8-D20.
140. Lafferty, J.D., A. McCallum, and F.C.N. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, in *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, Morgan Kaufmann Publishers Inc. p. 282-289.

VITA

VITA

Qiong Wu
240 S. Martin Jischke Drive
West Lafayette, IN 47907-1971
Phone (765) 494-3574
wu41@purdue.edu

QUALIFICATIONS

- PhD candidate in computational biology and MS in applied statistics.
- In-depth experience in analyzing RNA-Seq data of non-model organisms, including de novo transcriptome assembly, gene expression profiling, functional analysis, etc.
- Experience in statistical consulting, strong interpersonal and communication skills

EDUCATION

- **Purdue University** West Lafayette, Indiana, USA
  Ph.D. in Computational Biology Expected: *12/2014*
  M.S. in Applied Statistics *05/2013*
- **Zhejiang University** Hangzhou, China
  B.S. in Biotechnology 09/2003-07/2007

EXPERIENCES

Purdue University

- **Statistical Consultant at Statistical Consulting Service** 06/2013 – *Present*
  - Assisted clients in all stages of research projects including assess research objectives, experimental design, identify appropriate statistical model, data processing, etc.
  - Provided statistical software services in a drop-in basis.
- **Research Assistant**
  Next-generation sequencing analysis *01/2012 – Present*
  - De novo assembly of the first fern genome: *Pteris Vittata.*
  - Developed evaluation pipeline and new tools for de novo transcriptome assembly.
  - Performed processing, analysis and annotation of differential gene expression experiments using RNA-Seq data.
  Pattern recognition, information retrieval *07/2010 – 10/2011*
  - Develop an SVM-based biomedical named entities recognition system that utilizes contextual features and web retrieved evidences.
  Text mining, database management *06/2009 –06/2010*

- Adapted Textpresso full-text search platform for new corpora of biological literature.
- Constructed domain-specific ontology.
- Incorporated stemmer into Textpresso platform and constructed a prototype of stemmed Textpresso.

- **Teaching Assistant: Bioinformatics, Anatomy and Physiology**
  Teaching assistant for BIOL 478: Intro to Bioinformatics *Fall 2011, Fall 2012*
  Instructor for BIOL 302: Human Design: Anatomy and Physiology *Spring 2010, Spring 2013*
  Teaching assistant for BIOL 44211: Laboratory in Anatomy and Physiology *Spring 2012, Spring 2013*

COMPUTATIONAL SKILLS

Languages/Packages: Perl, R(Bioconductor), Matlab
Statistical software: SAS(macro, SQL)

PROFESSIONAL CERTIFICATES

Completed CFA level I *12/2011*
Certified base programmer for SAS 9 credential *06/2013*

LEADERSHIP EXPERIENCE

Completed Krannert Applied Management Principles (AMP) Program (10 awardees in School of Science) *05/2013*
Vice President of Purdue University Chinese Students and Scholars Association (PUCSSA) *05/2008 -04/2009*

HORNORS

Purdue University Graduate School Summer Research Grant *05/2012-08/2012*

PROFESSIONAL MEMBERSHIPS

International Society for Computational Biology (ISCB), American Statistical Association

PUBLICATIONS

PUBLICATIONS

Wu, Q., Gribskov, M.: **Textpresso with stemming. (In preparation)**

Wu, Q., Gribskov, M., Atallah, N., Banks, J.: **Transcriptome assembly and differential expression analysis of *Pteris vittata* in response to arsenic stress. (In preparation)**