BIOINFORMATICS TOOLS & ANALYSES: SUPPORT FOR BIOFUELS STUDIES IN PLANTS & BACTERIA

by

WEN-CHI CHOU

(Under the Direction of Ying Xu)

ABSTRACT

This dissertation mainly focuses on bioenergy studies from the viewpoint of Bioinformatics. The bioenergy in the study is called second-generation biofuel, which is produced from cellulosic biomass. The difference from a first-generation biofuel is that the cellulosic biomass is not from edible plants. My studies can be classified into two fields, plants and bacteria. In the plant field, a better plant model is required in order to produce cellulosic biomass via genomics modifications. The goals of the genomics projects include increasing the composition of plant cell walls and reducing the difficulties of plant cell wall degradation. In the bacteria field, a better bacterium is necessary to produce a higher yield of ethanol. Through genomics studies, I want to understand the mechanism and functional pathway of ethanol biosynthesis to re-engineer the pathway.

My dissertation consists of three bioinformatics projects, which all utilized bioinformatics tools and analyses to reach the ultimate goal of increasing the amount of ethanol. The first project, in the plant field, is the prediction of plant Golgi resident proteins. I identified novel Golgi proteins because most of the enzymes associated with plant cell wall biosynthesis are located at Golgi. A machine-learning based method was used to identify Golgi proteins. With

those identified Golgi proteins, other scientists can then possibly focus on studying a reasonable number of enzymes. The second project, in the plant field as well, is to determine a complete set of transcribed sequences in switchgrass. The transcribed sequences have been used to design microarray chips for studying transcriptome expression profiles of switchgrass. I applied two-step *de novo* assembly on Sanger and 454 sequencing data to achieve the goal of getting transcribed sequences. The last project focuses on constructing transcriptome structure maps of a thermophilic bacterium, *Clostridium thermocellum* that can degrade plant cell walls and produce ethanol. I used a machine-learning based method together with strand-specific RNA-seq data to identify genome-wide transcription units, which are functional elements of a genome. The transcriptome structure maps will help to understand more about how *Clostridium thermocellum* synthesizes ethanol.

INDEX WORDS: Bioinformatics, Biofuels, Next-generation sequencing, Machine learning, Plant cell wall, Golgi proteins, Switchgrass, Transcribed sequences, *Clostridium thermocellum*, Transcription unit, and Transcriptome.

BIOINFORMATICS TOOLS & ANALYSES: SUPPORT FOR BIOFUELS STUDIES IN PLANTS & BACTERIA

by

WEN-CHI CHOU

BS, Chung Shan Medical University, Taiwan, 2000

MS, Taipei Medical University, Taiwan, 2002

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

BIOINFORMATICS TOOLS & ANALYSES: SUPPORT FOR BIOFUELS STUDIES IN PLANTS & BACTERIA

by

WEN-CHI CHOU

Major Professor: Ying Xu

Committee: Russell L. Malmberg
Chung-Jui Tsai
William York

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

First of all, this dissertation is dedicated to my wife, Hsin-Ping. She deserves the most credits on my dissertation. She always encourages me when I feel down with my research, and she kindly supports me with her statistical knowledge. We share the joy together after all the difficulties. I must also thank my daughter, Jessie, for bringing me limitless love and happiness.

Secondly, I want to dedicate this work to my mother. She made me a better person. I can devote to my work overseas without worries because she takes care of my families in Taiwan.

Finally, I dedicate this dissertation to my late eldest brother. He left this world a few weeks before I completed my PhD. He always concerned about my life and work. I want to share my little achievement with him. I love him, and may he rest in peace.

ACKNOWLEDGEMENTS

This dissertation would not have been completed without the support, the guidance, and the help of advisory committee members.

First of all, I owe my deepest gratitude to Dr. Ying Xu for the support and encouragement over the past five years. I learned passion for science and diligent attitude toward work from him. I believe it will have long-term influence on my career development. I would also like to show my gratitude to my committee of Dr. Russell L. Malmberg, Dr. Chung-Jui Tsai, and Dr. William York for guiding me from rough research ideas to complete studies.

I want to thank my collaborators Maor bar-palad, Jiyi Zheng, Steven Brown, and Shihui Yang for providing experimental data sets and many useful discussions.

I would like to thank Yanbin Yin, who is a good friend and a helpful colleague. He is always willing to discuss everything about the science and profession, and to give me his best suggestions. Many thanks to other lab members, who have given me helpful support, including Joan Yantko, Fenglou Mao, Juan Cui, Sha Cao, Phuongan Dam, and Victor Olman.

I would also like to thank IOB students, Anuj Srivastava and Timothy Shaw. We had worked as a team for a science contest. The experience helped me a lot on my own research projects.

Finally, I would like to thank my families, especially my wife who is always by my side through good and bad times.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Because of the many concerns about fossil fuel depletion, carbon dioxide emissions, and greenhouse gas increases, scientists continue to look for environmentally friendly solutions such as renewable energy. Biofuels are known as one form of renewable energies. Since 2006, there have been many debates over choosing good research directions of biofuels, and these discussions keep moving bioenergy research forward. One practical plan is to use cellulosic biomass and produce cellulosic ethanol (Ragauskas, et al., 2006). Although cellulosic ethanol is not the most efficient bioenergy such as biohygrogen, using ethanol is the most achievable short-term goal without changing the present transport system (Brower, 2006). Unlike the first-generation biofuel, which is generated from edible crops such as corn and sugarcane, the second-generation biofuel comes from crops such as switchgrass and *Populous* so that competition for a popular food and feed supply can be avoided.

In September 2007, the U.S. government funded three DOE bioenergy research centers to achieve higher usage of bioenergy alternatives over the next 15 years (Heaton, et al., 2008). Computational Systems Biology Lab (CSBL), as a member of Bioenergy Science Center (BESC), one of the DOE bioenergy research centers, has focused on the biological challenges of bioethanol production from cellulosic biomass as well as on lowering high production costs (Regalbuto, 2009). Three major research areas of producing bioethanol include: 1) Optimization of plant cell wall for facile degradation and ethanol production, 2) Improvement of the ability of bacteria for deconstruction of cellulosic biomass into sugars, and 3) Engineering of metabolic

pathways in bacteria for more ethanol production. My projects fell under the first and the second research areas.

In Chapter 2, a new method for identifying Golgi proteins in plants is presented. Although Golgi proteins occur in animal and plant cells, this research targets plant cells because Golgi is considered the location where plant cell wall biosynthesis takes place. When more plant Golgi proteins are identified, finding more cell wall synthesis related enzymes becomes easier because the searching range is narrowed. I studied the features of existing Golgi resident proteins and built a machine-learning based classifier to identify novel plant Golgi proteins. The features of Golgi proteins include the number of transmembrane domains, the thickness of the transmembrane domains, and the functional annotations. The classifier, *GolgiP*, is presented as a web service that takes protein sequences and gives results by different user-selected classification modules.

In Chapter 3, a new procedure for constructing a complete set of transcribed sequences of switchgrass is introduced. The switchgrass transcribed sequences can be used to design microarray chips and study gene expressions to understand more about the regulation of plant biomass production. Using novel and published sequencing data sets, the most comprehensive collection was presented. In the study, eleven and a half million 454 reads and 169,079 Sanger reads were used to determine 80,211 non-redundant transcribed sequences. Combining with other published transcribed sequences, 139,200 transcribed sequences were identified. This work also provided analyses to determine the directions of sequences and to check the coverage of our results over the other closely related plant species.

In Chapter 4, a new algorithm of identifying transcription units (TUs) in bacteria is proposed. Transcription units are the functional elements of a genome. In bacteria, a transcription

unit contains one or more than one gene in its coding region. The goal of this algorithm is to identify a complete list of genes in a transcription unit. In other words, the main bioinformatics task includes how to determine whether two continuous genes in the same strand are co-transcribed in one unit. Here, *Clostridium thermocellum* (*C. thermocellum*) is used as a case study as *C. thermocellum* can deconstruct plant cell walls and produce ethanol. I expect to learn more about ethanol synthesis pathways in *C. thermocellum* if I can construct its transcriptome structure, which is an entire list of transcription units. By using patterns of transcription units, such as continuity and stability, I developed machine-learning based classifiers to identify TUs. The proposed method is able to construct transcriptome structure maps based on RNA sequencing data.

Conclusions of my studies and a discussion of future work are provided in Chapter 5.

CHAPTER 2

GOLGIP: PREDICTION OF GOLGI RESIDENT PROTEINS[1]

2.1 ABSTRACT

We present a novel Golgi-prediction server, *GolgiP*, for computational prediction of both membrane-associated and non-membrane-associated Golgi resident proteins in plants. We have employed a support vector machine-based classification method for prediction of such Golgi proteins, based on three types of information: dipeptide composition, transmembrane domain(s), and functional domain(s) of a protein. The functional domain information is generated through searching against the Conserved Domains Database (CDD), and the transmembrane domain (TMD) information includes the number of TMDs, the length of TMD, and the number of TMDs at the N-terminus of a protein. Using *GolgiP*, we have made genome-scale predictions of Golgi resident proteins in 18 plant genomes, and have made the preliminary analysis of the predicted data. The *GolgiP* web service is publically available at http://csbl1.bmb.uga.edu/GolgiP/

[1] Wen-Chi Chou, Yanbin Yin, and Ying Xu. 2010. *Bioinformatics*. 26(19):2464-2465.

2.2 INTRODUCTION

The Golgi apparatus is an essential cellular organelle found in most, if not all, eukaryotic cells, serving as an intermediate station of the secretory pathway that transports proteins out of a cell. In addition, Golgi is also a major site for protein post-translational modifications (e.g. glycosylation (Nilsson, et al., 2009)) and synthesis of various polysaccharides. The plant cell walls are mainly comprised of lignins, glycosylated proteins, and polysaccharides, most of which are synthesized in Golgi (Lerouxel, et al., 2006).

Identification of the Golgi resident proteins represents a very challenging and a highly important problem for the understanding of the biological processes taking place in Golgi. While there are 1183 mouse and human Golgi resident proteins identified (Sprenger, et al., 2007), only a little over 400 plant Golgi proteins have been experimentally identified. A key challenging issue is that plant Golgi proteins do not seem to have known targeting signals while proteins targeted at other cellular compartments, like nucleus or extra-cellular space do. Most of the existing computational tools for subcellular localization predictions are designed for the general subcellular localization prediction, and their predictions for Golgi resident proteins are less than adequate (Sprenger, et al., 2006). Only one program has been specifically designed for prediction of Golgi localized proteins, but it focuses only on transmembrane Golgi proteins (Yuan and Teasdale, 2002). The issue is that only 25% of Golgi proteins of *Arabidopsis thaliana* are estimated to contain transmembrane regions (Schwacke, et al., 2003), indicating the inadequacy of the current programs. Based on this consideration, we have designed a support vector machine (SVM) based classifier, called *GolgiP*, to predict both Golgi localized transmembrane proteins and non-transmembrane proteins. *GolgiP* currently provides multiple models for predicting plant Golgi proteins, based on the specific needs of a user.

2.3 MATERIALS AND METHODS

We have collected a large dataset comprising of 402 known Golgi proteins and 5,703 known non-Golgi proteins of *Arabidopsis thaliana* (91.2%), *Oryza sativa* (8.2%), and other plants (0.7%), from the SUBA (Heazlewood, et al., 2007) and the UniProt (Apweiler, et al., 2004) databases, as well as manually curated from the published literature. The non-Golgi proteins are proteins that have subcellular localization annotations, but not identified as Golgi according to the above databases. The redundant sequences in our dataset were removed by CD-hit using 95% sequence identity as the cut-off (Li and Godzik, 2006). Four-fifths of the data were used to train the classifier and the remaining one fifth was used to test the trained classifier, where the dataset was randomly partitioned into training and test data sets.

To train an SVM-based classifier for Golgi proteins, we have examined three different sets of features, all computed from protein sequences. The first set of features is related to the dipeptide composition (DiAA). For each protein in our training set, we calculated the composition of dipeptides. The second set of features is related to transmembrane domains (TMDs). We used TMHMM (Krogh, et al., 2001) and Phobius (Kall, et al., 2004) to predict the number of TMDs, the average length of TMDs, the number of TMDs within the N terminal region consisting of 70 amino acids, the length of the first TMD within the N terminal region, and the orientation of the N-terminal (i.e. in the cytosol side or in the Golgi lumen side). The third set of features is related to functional domains (FunD). We searched proteins in our datasets against the CDD database using RPS-BLAST (Marchler-Bauer, et al., 2009) with an e-value cutoff < 0.01. We did this because the Golgi apparatus is where proteins get post-translational modifications such as glycosylation (Nilsson, et al., 2009), and where the syntheses of most polysaccharides take place (Nilsson, et al., 2009). In addition, Komatsu et al. found that the

distributions of functional categories of proteins vary in different membranes such as plasma membrane, vascular membrane, and Golgi membrane, respectively (Komatsu, et al., 2007). Hence, it is expected that enzymes for the Golgi-related activities should be located in Golgi. The CDDs found for the Golgi proteins are then collected as the third set of features.

We applied the LIBSVM package (Fan, et al., 2005) to train the classifier. We used the Radial Basis Function kernel, and tuned the cost (c) and gamma (γ) parameters to optimize the classification performance on the training data set.

2.4 RESULTS

We used the aforementioned three sets of sequence features, and trained three SVM classification models. In addition, we combined all three sets of features to train a comprehensive model. The training performances are shown in Table 2.1.

Table 2.1. The performances of the four SVM models

Models	Sensitivity	Specificity	Accuracy	MCC
Transmembrane Domain (TMD)	63.84%	93.73%	91.77%	0.54
Functional Domain (CDD)	47.50%	100.00%	96.54%	0.68
Comprehensive	82.84%	99.45%	98.36%	0.85
Dipeptide Composition (DiAA)	99.63%	99.97%	99.95%	0.99

The performances were measured by 5-fold cross-validation. The measurements are sensitivity $= TP/(TP + FN)$, specificity $= TN/(TN + FP)$, accuracy $= (TP + TN)/(TP + FN + TN + FP)$, and Matthews correlation coefficient (MCC) $= (TP \times TN - FN \times FP)/\sqrt{(TN + FN) \times (TP + FN) \times (TN + FP) \times (TP + FP)}$, where TP, TN, FP and FN denote the numbers of true positive, true negative, false positive and false negative

predictions, respectively. The performances of the models that were trained based on different sets of the selected features indicate that the accuracies of our classifiers vary from 91.77% to 99.95%; the MCC values vary from 0.54 to 0.99.

We have compared the models with the other Golgi protein prediction tools, including PSORT (Nakai and Horton, 1999), WoLF PSORT (Horton, et al., 2007), and Yuan's Golgi predictor (Yuan and Teasdale, 2002) by using the testing data set.

As shown in Table 2.2, Yuan's Golgi predictor has good sensitivity but the lowest specificity and the lowest accuracy. PSORT and WoLF PSORT are two general subcellular localization prediction tools, and have a moderate level of classification performance, which may not be adequate to serve as a plant Golgi protein predictor based on our analysis. Our program, *GolgiP*, exhibits better overall performances with a higher accuracy and MCC value.

Table 2.2. Evaluation of Golgi protein prediction tools

Tools	Sensitivity	Specificity	Accuracy	MCC
Yuan	71.64%	23.18%	26.37%	-0.03
WoLF PSROT	15.92%	92.69%	87.63%	0.08
GolgiP-TMD	61.73%	67.75%	67.75%	0.15
PSORT	43.53%	83.10%	80.49%	0.17
GolgiP-DiAA	71.64%	80.76%	80.16%	0.31
GolgiP-Comprehensive	72.84%	98.42%	96.73%	0.73
GolgiP-FunD	57.50%	100.00%	97.21%	0.75

The performances were sorted by MCC values.

2.5 DISCUSSION

We have applied *GolgiP* with the functional domain model to predict Golgi proteins on 18 selected fully sequenced plant genomes using the same cutoff. The reason we chose the functional domain model is that the model performs the best specificity, and therefore tends to avoid false positive results in this genome-wide prediction. The numbers and percentages of the predicted Golgi proteins in these organisms are shown in Table 2.3. Across algae, moss, monocot and dicot plants, the average percentages of predicted Golgi proteins is 7.25% among all the encoded proteins from these genomes. The stability in the percentage of the predicted Golgi proteins across different genomes indirectly suggests the reliability of our predictions. The trend of distribution of Golgi proteins from lower to higher plant species shows a similar percentage of Golgi proteins. This may suggest that the functionality of the Golgi apparatus has evolved and matured fairly early in the plant evolution.

Table 2.3. Application of the *GolgiP* program on 18 plant genomes

Clade	Species	# predicted Golgi proteins/# Total proteins	%
red algae	*Cyanidioschyzon merolae 10D*	430/5014	8.58%
green algae	*Micromonas pusilla CCMP1545*	716/10475	6.84%
green algae	*Micromonas strain RCC299*	833/9815	8.49%
green algae	*Ostreococcus lucimarinus*	706/7651	9.23%
green algae	*Ostreococcus tauri*	656/7725	8.49%
green algae	*Chlamydomonas reinhardtii*	982/14598	6.73%
green algae	*Volvox carteri f. nagariensis*	1025/15544	6.59%
Moss	*Physcomitrella patens ssp patens*	2344/35938	6.52%

spike moss	*Selaginella moellendorffii*	2912/34697	8.39%
monocot	*Oryza sativa*	4240/67393	6.29%
monocot	*Brachypodium distachyon*	2446/32255	7.58%
monocot	*Sorghum bicolor*	2197/35899	6.12%
monocot	*Zea mays*	4748/75387	6.30%
Dicot	*Vitis vinifera*	2008/30434	6.60%
Dicot	*Arabidopsis thaliana*	2727/33410	8.16%
Dicot	*Medicago truncatula*	1856/30028	6.18%
Dicot	*Glycine max*	5262/75778	6.94%

In conclusion, we developed a Golgi protein prediction tool, *GolgiP*, and demonstrated its superior performance in predicting plant Golgi proteins over existing prediction servers. In addition, our predictions across multiple plant genomes give an estimation of the percentage of plant Golgi proteins across different plant organisms, which is in general agreement with the previous estimations.

2.6 ACKNOWLEDGEMENTS

We would like to thank Dr. Maor Bar-Peled for his helpful discussions. This work is supported in part by the BioEnergy Science Center (BESC) grant from the Office of Biological and Environmental Research in the DOE Office of Science and National Science Foundation (DOE 4000063512).

CHAPTER 3

INTEGRATED DE NOVO TRANSCRIPTOME ASSEMBLY OF SWITCHGRASS (PANICUM VIRGATUM L.)[1]

3.1 ABSTRACT

Switchgrass (*Panicum virgatum* L.), a perennial C4 grass, is one of the important bioenergy crops for producing cellulosic ethanol. In this study, mRNAs of two switchgrass strains, Summer VS16 and Alamo AP13, were sequenced by 454 and Sanger sequencing for designing microarray chips of switchgrass. The chips will be used to study the expression profiles of Switchgrass to understand its regulatory mechanisms. Due to the lack of switchgrass genome sequences, we developed an integrated *de novo* transcriptome assembly pipeline to construct genome-wide transcribed sequences. Eleven and a half million 454 reads and 169,079 Sanger reads were assembled into 77,854 and 30,524 contigs, respectively, which were then trimmed and merged to 128,058 non-redundant transcribed sequences of switchgrass. Further analyses of these sequences were performed with our own developed tools for quality assessment. The identified transcribed sequences had been used to design microarray chips.

[1] Wen-Chi Chou, Yanbin Yin, and Ying Xu.

To be submitted to Bioinformatics.

3.2 INTRODUCTION

Switchgrass (*Panicum virgatum* L.) is a native of North American, is one of the perennial C4 grasses, and is selected for the purpose of bioenergy production. Switchgrass has also been used for soil conservation and forge production on the Great Plains of the U.S. (Bouton, 2007; Keshwani and Cheng, 2009; McLaughlin and Adams Kszos, 2005; Schmer, et al., 2008; Yuan, et al., 2008). Several swichgrass studies aimed to improve the existing model for producing more biomass from switchgrass. The approaches included breeding and genetic engineering based on the current genomic information of switchgrass (Bouton, 2007; Chuck, et al., 2011; Fu, et al., 2011).

In order to study gene expression of switchgrass so that we can understand more about the regulations of plant cell walls, this study provides a complete set of transcribed sequences for designing microarray chips. The genome of switchgrass is either tetraploid or octoploid and contains many repetitive sequences. Thus, constructing the whole genome sequences of switchgrass is a very challenging task because it is difficult to assign those repetitive genes to the correct locations in the genome. Researchers are still working to obtain the whole genome sequences of switchgrass. Without the support of genome sequences, to collect a complete set of transcribed sequences becomes relatively difficult. This study used next-generation sequencing to generate a large amount of reads from various tissues and different cultivars, and the purpose is to produce higher coverage over the switchgrass transcriptome.

This study provides deep analyses on using different assembly tools to determine which tools are efficient and sufficient for reads from both Sanger and 454 sequencing. Sanger sequencing is a traditional method to identify transcribed sequences and produce longer reads to make assembly work easier. However, 454 sequencing provides higher coverage of a transcriptome, and can

produce a more comprehensive set of transcribed sequences. Combining with other published transcribed sequences of switchgrass, we present the most complete collection of transcribed sequences of switchgrass. In addition, the proposed workflow in this study can also be applied to assembly transcriptome sequences of other plant species.

3.3 MATERIALS AND METHODS

In order to prepare RNA samples for seqeuencing, Summer VS16 and Alamo AP13 were first cultured *in vitro* respectively for two and four months, and then the rooted plants were transferred and grown in a growth chamber. Shoot, root and reproductive organs were harvested. The harvested materials were stored at -80ºC before they were used to isolate total mRNAs and produce cDNA libraries for 454 and Sanger sequencing.

With regard to 454 sequencing, the isolated mRNAs were used to generate cDNA libraries by reverse-transcription and emulsion PCR. The cDNAs were sequentially fragmented, size-selected, and sequenced by standard Roche-454 FLX protocols to obtain 454 reads. As requested, 454 reads were submitted to NCBI Sequence Read Archive (SRA). Before the 454 reads were used to perform *de novo* assembly, the reads were pre-processed to remove adaptor sequences, poly A/T repeats, low quality sequences, low complexity sequences, and reads shorter than 100 bp. In addition, to filter the contamination of bacterial and viral RNA, BLAST search was applied to find the bacterial and viral sequences available in the NCBI nucleotide database. SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) was applied to detect the RNA sequences to clean the unwanted RNAs from plant chloroplast and mitochondria. After the removal of the contaminated and unwanted sequences, total reads of 1,507,321 and 10,172,776 were obtained from Summer VS16 and Alamo AP13 mRNAs, respectively. The detailed information of each 454 reads data set, including the accession numbers of NCBI SRA, tissues,

growth stages, number of cleaned reads, and average read length is displayed in Table 3.1 and Table 3.2.

Table 3.1. 454 reads of Summer VS16

Accession numbers of NCBI SRA	**Tissues**	**Growth stages**	**Number of cleaned reads**	**Average length (bp)**
SRX026147	Whole shoot	Leaf development	259,106	201
SRX026148	Whole root	Leaf development	205,466	222
SRX026149	Whole shoot	Stem elongation	194,426	194
SRX026150	Whole root	Stem elongation	174,053	190
SRX026151-2	Whole shoot	Reproductive stage	219,230	189
SRX026153-4	Whole root	Reproductive stage	234,107	205
SRX026155-6	Panicles including seeds	Reproductive stage	220,933	212
			Total: 1,507,321	Average: 202

Table 3.2. 454 reads of Alamo AP13

Accession numbers of NCBI SRA	**Tissues**	**Growth stages**	**Number of cleaned reads**	**Average read length (bp)**
SRX057824	Whole shoot	Stem elongation	733,173	202
SRX057825	Whole root	Stem elongation	667,612	206

SRX057830	Whole shoot	Leaf development	1,236,020	419
SRX057831	Whole root	Leaf development	1,214,630	375
SRX057828	Whole shoot	Stem elongation	1,357,290	223
SRX057829	Whole root	Stem elongation	1,040,192	404
SRX057827	Whole shoot	Reproductive stage	547,278	320
SRX057826	Whole root	Reproductive stage	998,691	388
SRX057834	Panicles including seeds	Reproductive stage	1,096,949	384
SRX057833	Whole shoot	Stem elongation 2 w/drought	362,346	213
SRX057832	Whole root	Stem elongation 2 w/drought	918,585	337
			Total: 10,172,766	Average: 316

For Sanger sequencing, normalized and full-length-enriched cDNA libraries were generated from isolated mRNA of various Alamo AP13 tissues. The cDNA libraries were ligated into pDNR-Lib construct cDNA clones to enrich full-length cDNAs. The enriched cDNAs were then sequenced by a Sanger sequencing instrument, ABI 373 DNA analyzer, to obtain Sanger reads. Unwanted sequences including vector sequences, low quality sequences, low complexity sequences, poly A/T repeats at the ends of reads, and reads shorter than 100 bp were removed. The procedure in 454 reads to remove contaminated and unwanted sequences was also executed. 169,079 Sanger reads were from Alamo AP13 mRNAs. To estimate the percentages of the full-

length cDNA clones in the total cDNA clones, BLAST searches against foxtail millet (*S. italica*), sorghum (*S. bicolor*), and maize (*Z. mays*) were used to identify the full-length cDNA clones. The information of each Sanger reads data set including numbers of cDNA clones, numbers of reads and cleaned reads, average read length, and estimated percentages of full-length cDNA clones is displayed in Table 3.3.

Table 3.3. Normalized full-length cDNA libraries and Sanger reads

Tissues and conditions	Number of cDNA clones	Number of reads	Number of cleaned reads	Average read length (bp)	Estimated percentages of full length cDNA clones (%)
Aerial tissues at multiple stages without specific treatment	19,968	39,936	35,660	669	14.3
Underground tissues at multiple stages without specific treatment	15,744	31,487	28,588	658	14.7
Pooled RNAs from 32 samples including all possible tissues without abiotic stresses	57,600	115,200	104,831	612	9.3

	Total:	Total:	Total:
	93,312	186,623	169,079

In order to identify the genome-wide transcribed sequences, both 454 and Sanger reads were used to perform *de novo* transcriptome assembly. For 454 reads, Newbler v2.3 was applied with "cDNA" option and various setting of assembly parameters. A stringent setting using a minimum overlap of 100 bp and an identity over 99% was chosen after comparing different assembly results. Although 454 sequencing produced more reads and higher coverage than Sanger sequencing, Sanger reads are still valuable because of their longer read lengths, higher quality reads and the directional paired-end information.

To obtain comprehensive Alamo AP13 transcribed sequences, the selected assembly results of 454 reads were then assembled with Sanger reads. Several assembly tools such as CAP3, TGICL, and PAVE were attempted to combine the results of 454 read assembly and Sanger reads. Since PAVE can use the paired-end property of Sanger reads to output longer and more accurate results among others, PAVE was selected as the main tool for final merged assembly. The workflow of the two-step *de novo* assembly is illustrated in Figure 3.1. Comparing to a one-step assembly, the two-step assembly was able to give more reasonable reults because most of the assembly tools were designed to deal with eithor only 454 read or only Sanger reads.

We also included published RNA sequencing reads of switchgrass cultivars, which include Alamo, Kanlow, and sixteen other cultivars, from NCBI. These published reads were then merged with assembly results of 454 reads by PAVE.

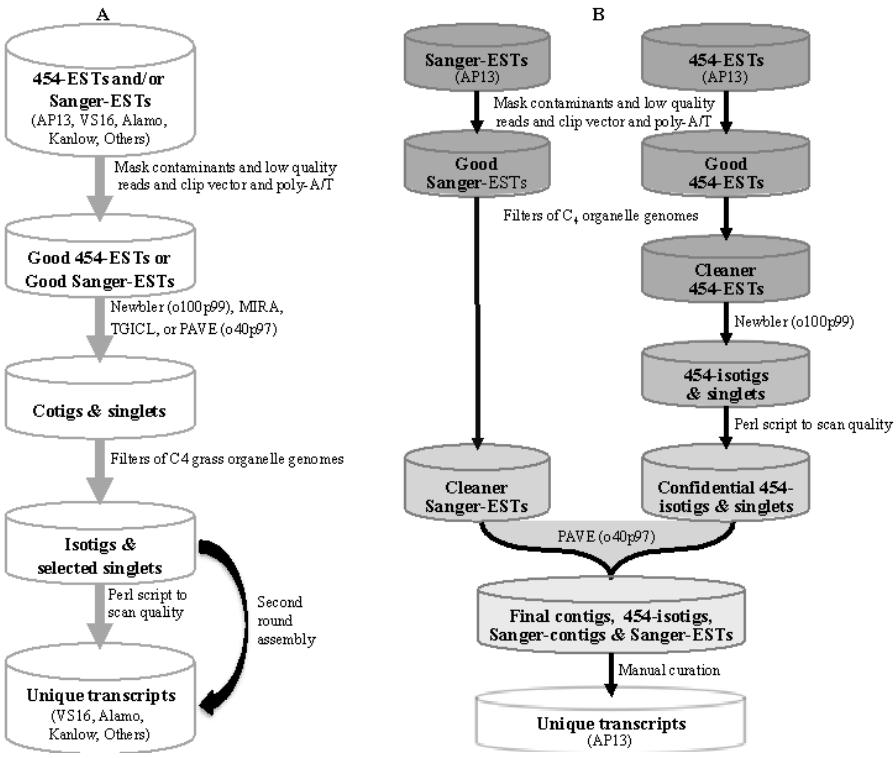


Figure 3.1. Diagrams of the one-step (A) and two-step (B) methods used in the beginning and the final assembly works. An isotig is a cluster of contigs with potential splice variants. o indicates a length of minimum overlap; p indicates a percentage of minimum identity.

After several *de novo* assembly procedures, five sets of results were produced and used to determine a complete set of universal and low-redundant switchgrass transcribed sequences. To illustrate the redundancy, an alignment case of one putative full-length transcribed sequence is displayed in Figure 3.2. To remove redundancy, BLAST search was used to identify redundant sequences, which shared over 90% identity and over 80% overlap in alignments.

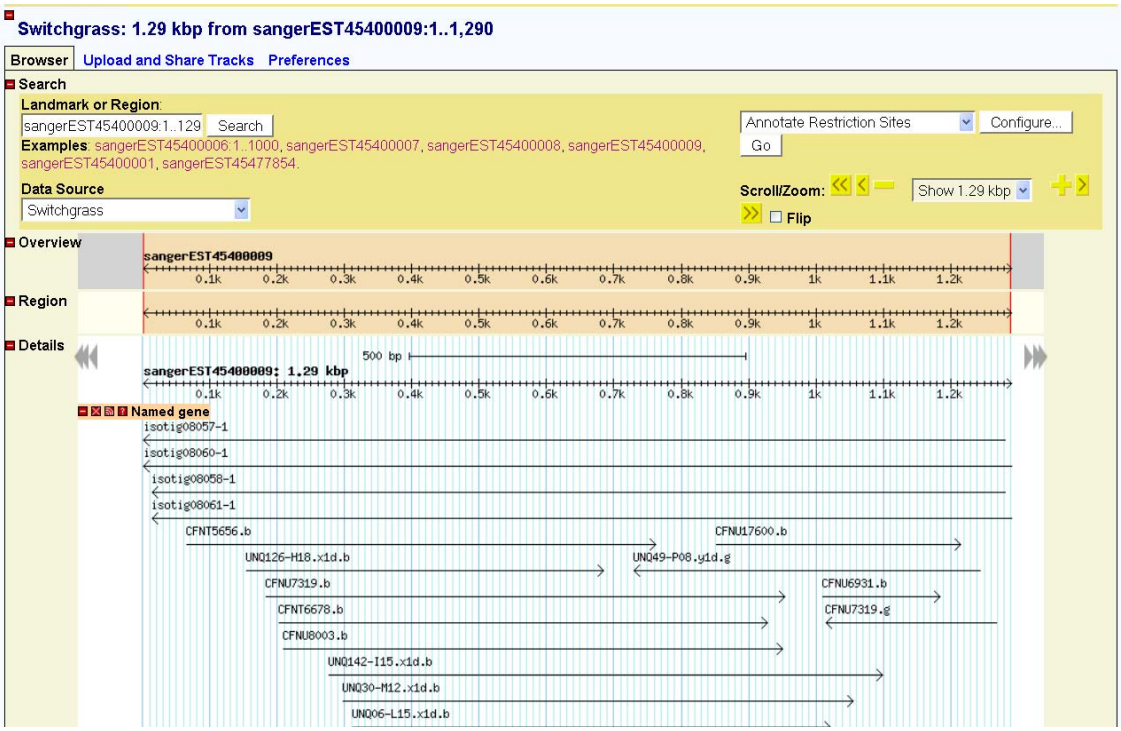


Figure 3.2. One sequence named sangerEST4540009 is a putative full-length transcribed sequence. Sanger reads and assembly results of 454 reads were aligned to the transcribed sequence. Arrows indicate orientations of aligned sequences.

The directions of transcribed sequences were determined by the information of directional Sanger reads, BLAST search, and Pfam search. Approximately eleven thousand transcribed sequences cannot be assigned with a single direction and were stored as two sequences with either forward or reserve direction.

3.4 RESULTS

Before the two-step assembly procedure was decided in this study, different assembly tools were attempted to perform one-step assemblies using both 454 and Sanger reads. The one-step assembly results and comments are shown in Table 3.4. Three terms in Table 3.4 are defined as follows: a contig is a non-redundant sequence, which combines shorter sequences with

similarity; an isotig is a cluster of contigs to present a potential alternative-spliced sequence while isogroup present a group of isotigs; a singlet is a sequence that cannot be aligned to any other sequence.

Table 3.4. One-step *de novo* assembly results by using different programs, such as TGICL, Newbler, PAVE and MIRA.

Assembler	**TGICL**	**Newbler**	**PAVE**	**MIRA without Q-score**	**MIRA with Q-score**
Number of contigs/isotigs	171,589	34,430	85,730	102,861	60,928
Number of singlets	141,586	192,515	159,578	1,747	2,786
Number of putative transcriptional units	313,175	226,945	245,308	104,608	63,714
Number of unused reads	200,379	0	0	190,053	none
comments	poor assembly results	too many redundant transcribed sequences in the results	requiring larger computing resources	too many short contigs	too many short contigs

Before PAVE was selected to perform the second step assembly of two-step process, different assembly tools such as CAP3 and TGICL were evaluated with default and stringent parameters. The second step assembly results are shown in Table 3.5.

Table 3.5. A comparison among the *de novo* assembly results using CAP3 and TGICL with different parameters.

	Number of contigs	**Number of singlets**	**Number of transcribed sequences**
CAP3 with default parameters (OL40 and ID94)	24,657	18,126	42,783
CAP3 with stringent parameters (OL100 and ID99)	31,550	58,810	90,360
TGICL with default parameters (OL40 and ID94)	26,770	7,050	33,820
TGICL with stringent parameters (OL100 and ID99)	26,177	2,399	28,576

OL indicates a length of minimum overlap; ID indicates a percentage of minimum identity.

After the removal of the contaminated and unwanted sequences, 11,680,087 and 169,079 reads of 454 and Sanger sequencing were extracted and used to perform a two-step *de novo* transcriptome assembly. With only 454 reads of Alamo AP13 from various tissues, 36,438 transcribed sequences were determined. With 454 and Sanger reads together, 36,543 transcribed sequences were identified. The detailed assembly statistics are shown in Table 3.6.

Table 3.6. A comparison among the *de novo* assembly results using different data sets of 454 reads from Alamo AP13 ESTs.

	454 reads from shoot libraries	**454 reads from root libraries**	**454 reads from all libraries**	**All 454 reads and Sanger reads**
Number of cleaned 454 reads	4,287,130	4,887,093	10,172,766	10,172,766
Number of Sanger reads	0	0	0	64,248
Number of contigs/isotigs ≥ 50 bp	61,052	64,094	102,144	103,153
Number of contigs/isotigs with length between 500 to 5,000 bp	0	0	92,623	94,136
Number of contigs/isotigs < 50 bp	16	49	38	39
Number of singlets	1,329,657 (31.0%)	396,723 (8.1%)	4,129,017 (40.2%)	4,119,928 (40.1%)
	26,459	27,681	36,438	36,543
Number of isogroups	4,287,130	4,887,093	10,172,766	10,172,766
Max contigs/isotigs length (bp)	0	0	0	64,248

After the transcribed sequences were identified, each sequence will be assigned a direction: forward or reverse. Sanger reads of full-length ESTs and partial ESTs were first used to assign directions. Some transcribed sequences assembled only by 454 reads were assigned directions according to the results of BLAST and Pfam search. Selected assignment cases are shown in Table 3.7. The rest of transcribed sequences, in which directions cannot be determined, were duplicated and labeled with either forward or reverse.

Table 3.7. Direction determination by using BLAST and Pfam searches.

Contigs/isotigs ID	**Number of forward BLAST hits**	**Number of reverse BLAST hits**	**Number of forward Pfam hits**	**Number of reverse Pfam hits**	**Assigned direction**
VS16_Isotig_23270	249	1	2	0	Forward
AP13_Contig_24223	248	22	2	0	Forward
VS16_Isotig_23188	248	3	2	0	Forward
AP13_Isotig_67203	247	3	2	0	Forward
AP13_Isotig_70532	247	21	2	0	Forward
AP13_Isotig_61691	246	4	2	0	Forward
AP13_Isotig_53588	2	248	0	3	Reverse
VS16_Isotig_26348	0	250	0	2	Reverse
AP13_Contig_27699	3	247	0	2	Reverse
AP13_Isotig_64600	4	246	0	2	Reverse
AP13_Isotig_76879	5	245	0	2	Reverse

AP13_Isotig_75608	6	244	0	2	Reverse

Combining with other assembly results of published switchgrass ESTs, all identified transcribed sequences were collected into a final result set called, PviTS, which was consisted of 128,058 non-redundant sequences.

Table 3.8. Transcriptome assembly and EST data resources used for switchgrass transcribed sequences collection, PviTS.

Switchgrass Genotypes	**Tissues and conditions**	**Number of Sanger reads**	**Number of 454 reads**	**Number of contigs/ isotigs**	**Number of Singlets**	**Number of transcribed sequences collected into PviTS**
Alamo AP13	Described in this manuscript	167,901	10,172,766	77,854	0	77,990
Summer VS16	Described in this manuscript	0	1,507,321	30,524	0	2,221
Alamo A4	Callus and vascular bundle	58,251	0	15,186	3,750	10,017
Kanlow population	Multiple tissues and	346,752	0	48,084	8,576	34,412

	conditions					
Other 16 switchgrass	Mainly leaves with multiple abiotic stresses	139,222	0	25,671	14,152	13,058
Alamo AP13	unknown	0	0	1,500	0	1,502

To estimate the completeness of PviTS, transcribed sequences and coding sequences of foxtail millet (*S. italica*), sorghum (*S. bicolor*), and maize (*Z. mays*) from Phytozome v6.0 (http://www.phytozome.net/) were used to calculate the coverage over the downloaded data sets by BLAST search with a E-value cut-off of 0.0001. The results of coverage are shown in Table 3.9.

Table 3.9. Estimated transcriptome coverage of PviTS by comparing with transcribed sequences of other plant species.

	Number of reference sequences	**Number of sequences aligned to PviTS (BLAST e-value ≤ 0.0001)**
Transcribed sequences of *S. italica*	38,038	34,043 (89.5 %)
Transcribed sequences of *S. bicolor*	29,448	27,457 (93.2 %)

Transcribed sequences of *Z. mays*	53,764	50,339 (93.6 %)
Coding sequences of *S. italica*	38,038	33,786 (88.8 %)
Coding sequences of *S. bicolor*	29,448	27,448 (93.2 %)
Coding sequences of *Z. mays*	53,764	49,390 (91.9 %)

3.5 DISCUSSION

The goal of the present study is to generate a complete set of transcribed sequences in switchgrass. Here, three different types of sequencing data sets, including our own 454 reads, our own Sanger reads, and published Sanger reads, were used to perform *de novo* transcriptome assembly. In the beginning of this study, the one-step assembly strategy was used to deal with the sequencing reads and PAVE, among several assemblers, gave the best assembly results based on the number of identified transcribed sequences and the length distribution of the sequences. However, PAVE needs higher computing requirements such as bigger memory size and longer computing time. Thus, Newbler was later chosen to assemble 454 reads in this study, even though Newbler resulted in more redundant sequences. The redundancies were removed according to the results from BLAST search.

In order to find the best assembly tool to merge Sanger reads and assembly results of 454 reads, CAP3 and TGICL were first used to find better assembly parameters. When the

parameters were more stringent, longer contigs were generated. However, more singlets were produced because the stringent parameters made reads hard to find their partners.

We used 454 and Sanger reads of Alamo AP13 to test the two-step assembly strategy. In Table 3.6, the assembly result of using both 454 read and Sanger reads generated the most amount of long isotigs because of the longer length of Sanger reads.

In order to obtain a more complete collection of transcribed sequences, other published switchgrass transcribed sequences were included into PviTS. The newly identified transcribed sequenced from this study is 58% of total collected sequences. We also estimated the coverage of PviTS over the other three closed plant species. The sequences of PviTS were able to cover from 89.5% to 93.6% of transcribed sequences in foxtail millet, sorghum, and maize.

In the near future, more analyses such as functional annotations can be used to examine the sequence quality of PviTS. Recently, the genome of foxtail was published and it will be a very informative reference to verify PviTS because foxtail is very close to switchgrass. Our work had been used to design microarray chips of switchgrass, which were used in several experiments to further study gene expressions of switchgrass.

3.6 ACKNOWLEDGEMENTS

This work was supported by a grant from the U.S. Department of Energy (# DE-PS02-06ER64304). The BioEnergy Science Center (BESC) is supported by the Office of Biological and Environmental Research in the DOE Office of Science.

CHAPTER 4

GENOME-WIDE IDENTIFICATION OF PROKARYOTIC TRANSCRIPTION UNITS USING STRAND-SPECIFIC RNA-SEQ DATA[1]

4.1 ABSTRACT

RNA sequencing (RNA-seq) is an application of next-generation sequencing for accurately measuring whole transcriptome expression profiling. RNA-seq data is used to investigate transcriptome maps in prokaryotes. A transcriptome map consists of the completely annotated transcription units (TUs), which are the structural and functional elements in prokaryotic genomes. This study presents the first computational method to systematically construct a genome-wide transcriptome map of a given prokaryotic genome annotation and its RNA-seq data. The features of continuity and stability of expression signals from RNA-seq data were used to identify TUs. An artificially constructed TU dataset was specially built by splitting each gene into three parts to mimic TU structures. Also, machine-learning based classifiers were built to study the artificially constructed TU dataset and to identify real TUs. Results show that the classifiers are able to correctly identify not only known TUs but also many novel TUs. In addition, the classifiers can identify dynamically arranged TUs cross the different RNA-seq data.

[1] Wen-Chi Chou, Shihui Yang, Steven Brown, and Ying Xu.

To be submitted to Bioinformatics.

4.2 INTRODUCTION

Having a transcriptome structure map composed of all transcription units (TUs) is critical for studying regulations of a prokaryotic genome since TUs are well known as structural and functional elements (Cho, et al., 2009). To obtain a complete TU list by traditional biological experiments such as polymerase chain reaction is labor-intensive and time-consuming. RNA sequencing (RNA-seq) is an application of next-generation sequencing for accurately measuring the whole transcriptome expression profiles. RNA-seq is able to detect all transcripts in one run of the sequencing experiment. In this study, strand-specific RNA-seq was conducted to detect gene expression signals separately from forward strand and reverse strand.

A TU is sequentially composed of one promoter, one transcriptional start site (TSS), one RNA-coding region containing either one open reading frame (ORF) or a cluster of ORFs , and one transcription terminator (Pierce, 2004). A cluster of ORFs is a functional unit that is transcribed into a single RNA molecule and is also called a polycistronic operon in prokaryotes (Jacob, et al., 1960; Wang, et al., 2004). TUs are dynamically arranged and reflect the various products of transcriptions while operons refer to a maximum and static set of continuous genes co-transcribed into a TU (Okuda, et al., 2007). Operons have been identified by different methods. Operon identification mainly relies on computational predictions before the emergence of whole-transcriptome profiling experiments such as tiling array and RNA-seq. Prediction algorithms were commonly built based on 1) DNA sequence structures, such as intergenic distances and phylogenies crossed related species, 2) functional annotations, such as gene ontology (GO) and clusters of orthologous groups of proteins (COG), and 3) limited experimental data, such as DNA microarray data that do not cover most of intergenic regions (Brouwer, et al., 2008; Siqueira, et al., 2011). In addition, some databases, including ODB,

DTBSD, OperonDB, and DOOR, collected predicted and experimentally verified operons (Mao, et al., 2009). However, the current operon collections are not complete due to the lack of appropriate experimental data for verifying the operons (Okuda and Yoshizawa, 2011).

High-throughput transcriptome profiling technologies, such as tiling array and RNA-seq, have been used to provide genome-wide transcriptome maps for several bacteria. Sharma et al. used strand-specific RNA-seq to identify 337 primary operons in *Helicobacter pylori*, and Oliver et al. identified 355 operons in *Listeria monocytogenes* by non-strand-specific RNA-seq. Oliver et al. used continuous expressions over intergenic regions to detect TUs (Oliver, et al., 2009; Sharma, et al., 2010). Güell et al. and Toledo-Arana et al. used dense tiling arrays to identify 139 and 517 polycistronic operons in *Mycoplasma pneumoniae* and *Listeria monocytogenes.* They used stable expressions of two adjacent genes on the same strand to detect TUs (Guell, et al., 2009; Toledo-Arana, et al., 2009).

Although there have been some genome-wide studies to reveal operon or TU maps in bacteria, there is still no analytical method to construct transcriptome structure maps systematically and automatically. This study investigated properties of RNA-seq data such as expression or sequencing biases, and used the properties to identified TUs. Our method can be broadly applied to any bacterial RNA-seq data that generate from strand-specific RNA-seq libraries. Based on the attributes of a TU, both continuous expressions over intergenic regions and stable expressions of two adjacent genes on the same strand were considered in our method. In addition, our method does not rely on either experimentally validated TUs or predicted operons to build a machine-learning based TU classifier.

4.3 MATERIALS AND METHODS

In this study, we used four strand-specific RNA-seq data sets collected on *C. thermocellum.* Sample preparation was discussed by Yang et al. (Yang, et al., 2012). The four data sets are generated to study how ethanol affects *C. thermocellum* growth rate. One sample as the control group was cultured under a normal growth condition, and the other three samples treated with 0.5% ethanol shock. After total RNAs were obtained from samples, ribosomal RNAs were removed. The remaining RNAs were used to prepare strand-specific sequencing libraries for illumina GA platform (Perkins, et al., 2009). Final dilutions of 25, 25, 33, 41 pM of the control and treatment libraries were loaded onto the sequencing machine in different lanes. Table 1 shows the numbers of reads and estimated depths of each RNA-seq data set. The length of each read is 50 bp, and estimated depths are calculated by using the total read length divided by the total length of all *C. thermocellum* annotated genes (3,268,038 bp). Before the reads were mapped to *C. thermocellum* genome, FastQC (Andrews, 2010) was used to assess the quality of the RNA-seq data. The FastQC results showed all RNA-seq data passed the value checks either on per base sequence quality or on per sequence quality.

Table 4.1. Cultured conditions, concentrations of the libraries, numbers of Reads and estimated depths of four RNA-seq data sets

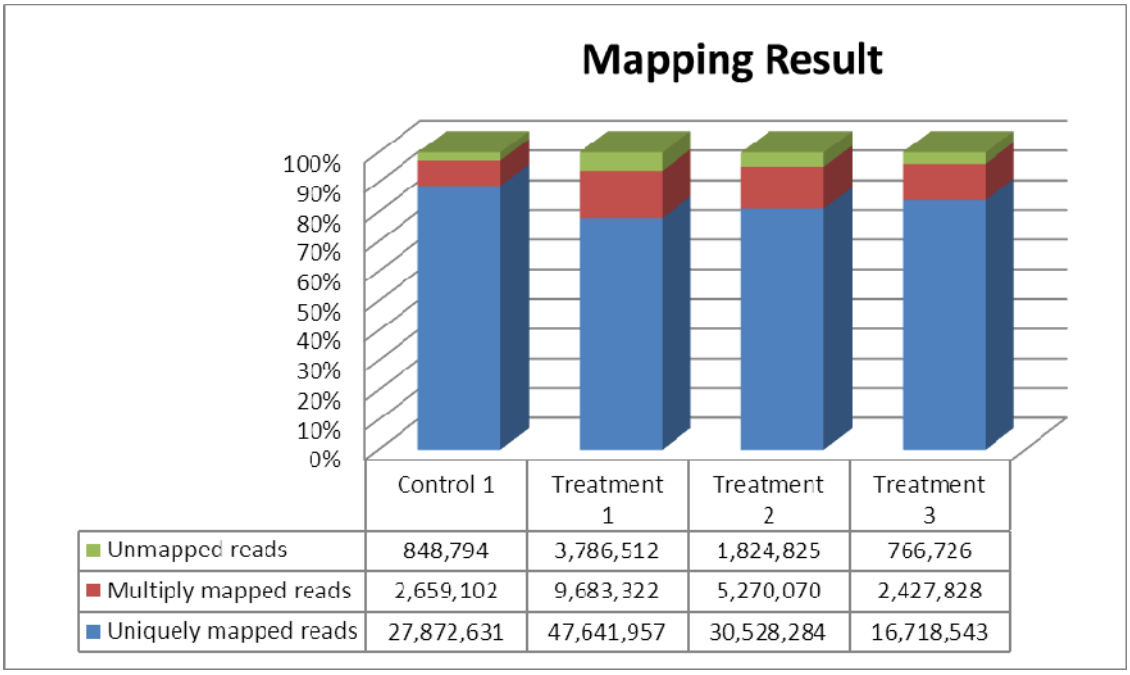
	Culture with additional 0.5% ethanol	**Concentration of RNA library (pM)**	**Number of Reads**	**Estimated depth (X)**
Control 1	No	25	31,380,527	480
Treatment 1	Yes	25	61,111,791	935
Treatment 2	Yes	33	37,623,179	576

Treatment 3	Yes	41	19,913,097	305

The three treatments were used the same sequencing libraries with different dilutions for conducting RNA sequencing.

The RNA-seq reads were mapped to *C. thermocellum* ATCC27405 genome sequence (3,843,301 bp) download from NCBI. The read mapping was used BWA with default parameters, where BWA is a short read aligner using Burrows-Wheeler transform (Li and Durbin, 2009). There were three types of mapping results including uniquely mapped reads, multiply mapped reads, and unmapped reads. Unmapped reads were further mapped by BLAST search with parameters allowing low-complexity regions, e-value below 1e-02, and the minimum length of 40 bps for a consecutive match segment. Final mapping results are shown in Table 2. The mapping results were then used to calculate coverage depths.

Table 4.2. The numbers of uniquely mapped reads, multiply mapped reads, and unmapped reads after mapping all reads to *C. thermocellum* genome.



	Control 1	Treatment 1	Treatment 2	Treatment 3
Unmapped reads	848,794	3,786,512	1,824,825	766,726
Multiply mapped reads	2,659,102	9,683,322	5,270,070	2,427,828
Uniquely mapped reads	27,872,631	47,641,957	30,528,284	16,718,543

Percentage of the three types of results are shown in the bar plot, and are indicated by different colors.

Uniquely and multiply mapped reads were used to get genome-wide coverage depths at each base (EBCDs) which are amounts of reads covering each nucleotide. Uniquely mapped reads were first assigned to *C. thermocellum* genome to get uniquely mapped EBCDs. Then the EBCDs were used to assign multiply mapped reads proportionally to proper positions in the genome. The assignments were undertaken according to the expression levels surrounding each matched position. A final whole genome EBCDs was constructed and was used in the rest of analyses. The whole genome EBCDs were used to check the RNA-seq data coverage on *C. thermocellum* genome. The coverage of each RNA-seq is shown in Table 3. The coverage in Table3 is a proportion of expressed nucleotides (with EBCD larger than zero) in target regions. Also, the genome-wide expression view over *C. thermocellum* genome is presented in Figure 1.

Table 4.3. RNA-seq data coverages over *C. thermocellum* genes and intergenic regions on forward and reverse strands.

	Gene-coding regions on the forward strand	**Intergenic regions on the forward strand**	**Gene-coding regions on the reverse strand**	**Intergenic regions on the reverse strand**
Control 1	92.64%	21.67%	90.90%	22.44%
Treatment 1	96.74%	30.42%	94.96%	31.02%
Treatment 2	95.44%	27.10%	93.27%	27.33%
Treatment 3	93.06%	22.77%	90.60%	23.05%

The four RNA-seq data sets had over 90% coverages over gene-coding regions. The coverages over the intergenic regions are below one third and indicate most of the intergenic regions were not expressed.

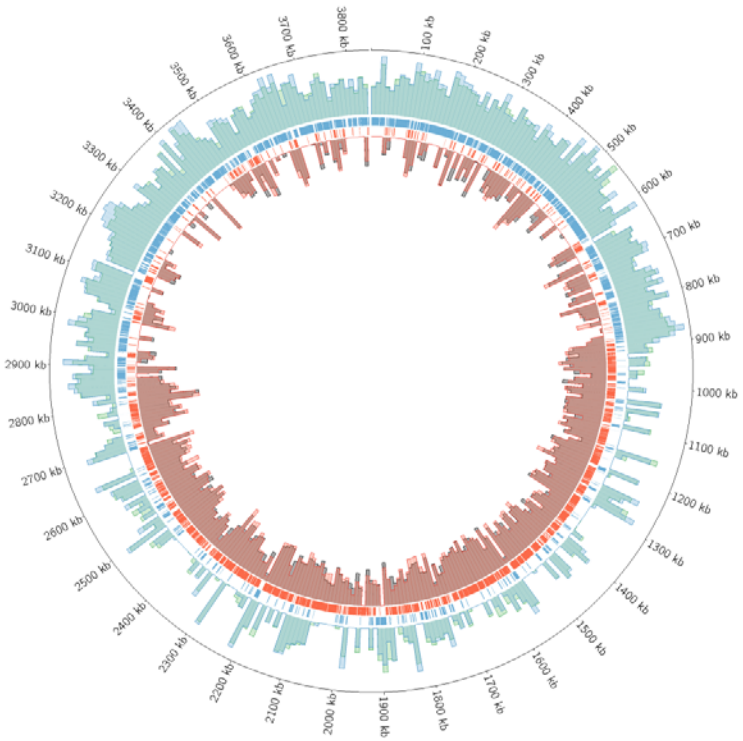


Figure 4.1. Genome-wide expression profile of *C. thermocellum.* Blue strips indicate the genes on the forward strand, and orange strips indicate the genes on the reverse strand. Each bar represents an average EBCD of one kbp. Blue and green bars indicate the average EBCDs of control1 and treatment3 on the forward strand. Red and gray bars indicate the average EBCDs of control1 and treatment3 on the reverse strand.

To ensure RNA-seq data reflect the real expression profile, we used whole genome EBCDs to check the known biases that may affect the accuracy of expression in RNA-seq data. The first one is GC bias: EBCDs decreases when GC content goes higher. The second one is position bias: EBCDs is higher at 5' end of genes, and is lower at 3' end of genes. The plots of

the GC and position biases are shown in Figure 2a and 2b. When we needed to compare expression levels of two genes, 5% EBCDs at 5' end and 3' end of each gene were first removed. The reminding EBCDs of the two genes were binned by GC contents, and the binned EBCDs were used to calculate fold changes.

a. b.

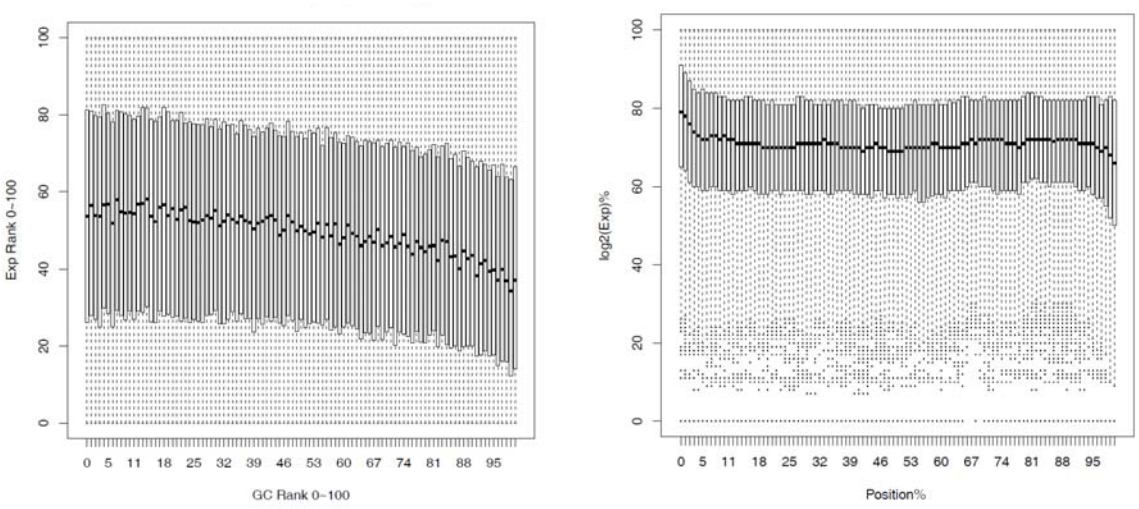


Figure 4.2. GC and position biases of our RNA-seq data. 2a) x axis indicates ranked GC contents in each gene while y axis indicates ranked EBCDs in each gene. The both ranks were normalized from 0 to 100. 2b) x axis indicates position quantiles of each gene while y axis indicates EBCD quantiles of each gene.

After correcting the expression profiles of each RNA-seq data set, a machine-learning based method was used to build TU classifiers. Due to the lack of a certain amount of experimentally validated TUs in *C. thermocellum*, it made machine-learning method difficult. Instead of using a few validated TUs to train classifiers, we generated thousands of artificially constructed TUs (ACTUs) and learned RNA-seq expression patterns of a TU. Each ACTU was a gene and consisted of two regions representing two gene-coding regions and one interval. The lengths of intervals followed a length distribution of real intergenic regions in *C. thermocellum* genome, and each interval was assigned to a region with the least GC content in a selected gene.

The length distributions of two adjacent genes on the same strand and their intergenic region are shown in Figure 3. Because EBCDs of an expressed gene represent either intact or partial RNA-seq expressions of a RNA-coding region in a TU, we used the ACTUs constructed from every expressed gene to learn the RNA-seq expression patterns of a partial RNA-coding region. In other words, the ACTUs can help with examining whether two adjacent genes on the same strand were co-transcribed in the same TU.

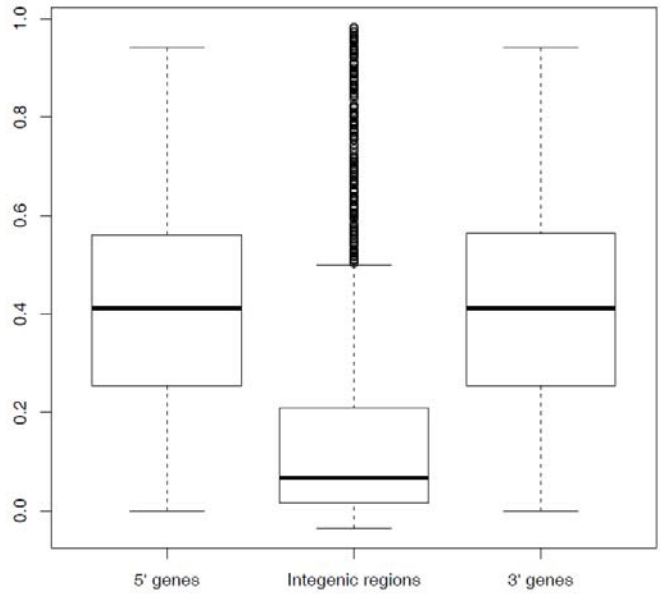


Figure 4.3. Length distributions of two adjacent genes on the same strand and its intergenic region. Y axis indicates the length proportions of the three regions. A total length of two adjacent genes and its intergenic region was normalized to one. The box plot of intergenic regions show the median length of intergenic regions is around 0.9.

In order to know whether two genes are co-transcribed into the same TU, traditionally, biological experiments use PCR and RT-PCR to find continuous expressions of intergenic regions in a target TU. If intergenic regions are continuously expressed, it indicates that genes surrounding the intergenic regions are in the same TU (Newcomb, et al., 2011). We used continuous expressed EBCDs as the first group of features to describe a TU. The features

included number of nucleotide positions with EBCD as zero in two gene-coding regions and intergenic regions.

In addition, a RNA-coding region containing either one or several genes in a TU presumably has a constant expression value. Although RNA sample preparation and sequencing process may produce uncertain biases on RNA-seq results (Trapnell, et al., 2012), we used stable expressed EBCDs as the second group of features to describe a TU. The second group of features tracks the changes of EBCDs through a target TU. The features included 1) fold changes of EBCDs between two gene-coding regions in the target TU, 2) fold changes of EBCDs between both gene-coding regions and the intergenic region, and 3) a variance of EBCDs throughout the target TU. The variance can detect rapidly dropped or raised EBCDs that may indicate unstable expressed EBCDs.

With the two groups of features and the ACTUs, libSVM (Chang, 2011) was used to train two TU classifiers for both forward and reverse strands of each strand-specific RNA-seq data set. Each classifier was trained by a five-fold cross validation with the best parameters for nu and gamma, and sensitivities were used to measure accuracies. The accuracies of the TU classifiers are shown in Figure 4. The TU classifiers were then applied to identify co-transcribed adjacent genes in a TU.

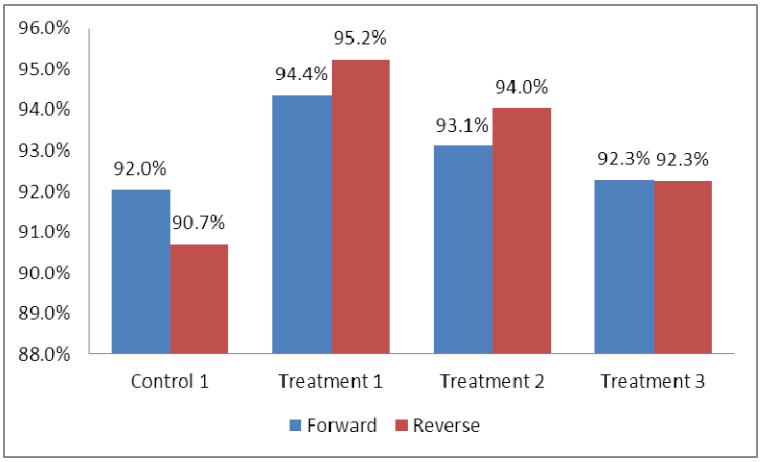


Figure 4.4. Accuracies of one-class SVM training results. Blue and red bars indicate the training accuracies on the forward and reverse strands of each strand-specific RNA-seq data set.

4.4 RESULTS

Using the four TU classifiers, about 1100 TUs in each RNA-seq data were identified. Table 4.4 shows the numbers of TUs identified on forward and reverse strands, and Table 4.5 shows the numbers of TUs composed of different numbers of genes.

Table 4.4. Numbers of identified TUs on forward and reverse strands.

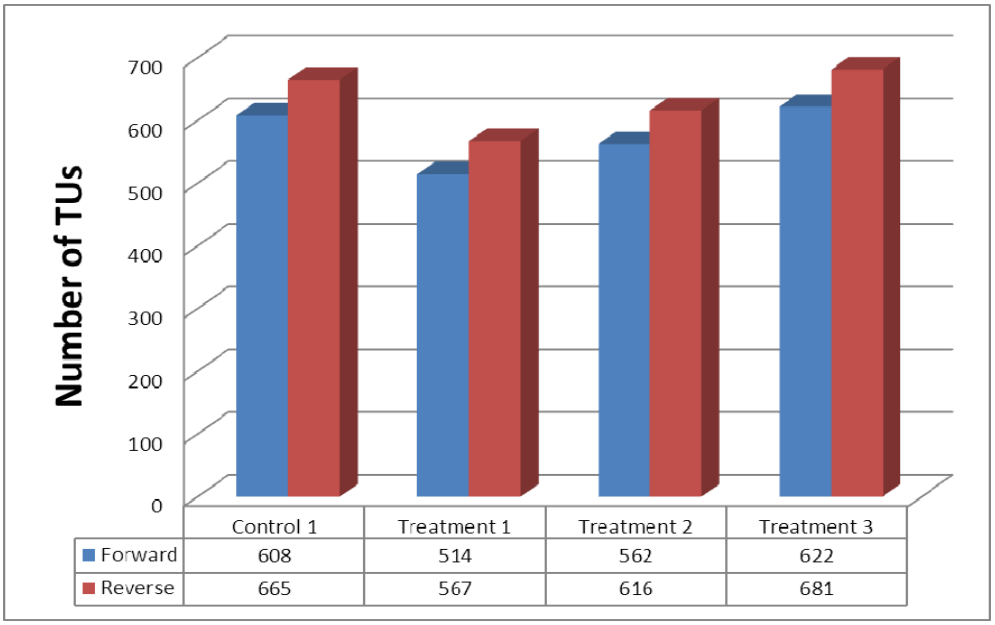


Table 4.5. The numbers of TUs split up and displayed by different numbers of genes.

	1 gene	2 genes	3 genes	4 genes	5/5+ genes
Control 1	705	219	130	61	159
Treatment 1	598	186	110	51	135
Treatment 2	652	202	120	56	147
Treatment 3	721	224	133	62	163

Our TU results were compared with the predicted operon maps provided by DOOR (Mao, et al., 2009). DOOR predicted 56% of genes in *C. thermocellum* are in polycistronic operons. The results showed that from 53% to 59% of genes are expressed and located in polycistronic operons. Figure 4.5 shows a case of different TU results identified by DOOR and our TU classifiers. The four RNA-seq data sets support that Cthe_0001, Cthe_0002, and Cthe_0003 are co-transcribed in a TU while DOOR prediction suggested only Cthe_0002 and Cthe_0003 are co-transcribed.

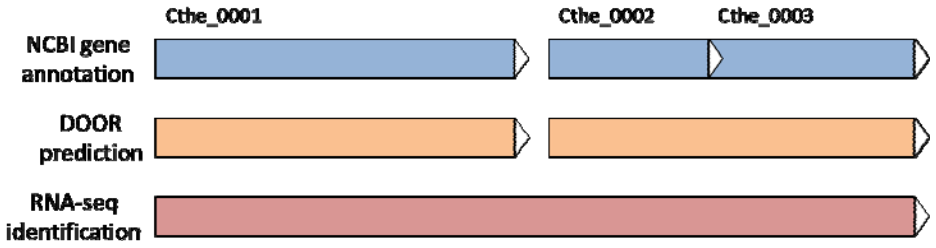


Figure 4.5. TU identified by DOOR (orange) and our classifier (purple).

This method successfully identified several experimental validated operons such as *CelC* operon. *CelC* operon of *C. thermocellum* consists of *celC*, *glyR3*, and *licA* genes (Newcomb, et

al., 2007). The TU identification result is shown in Figure 4.6. This study also identified novel TUs which are different from the literature (Newcomb, et al., 2011). The models together with RNA-seq data identified that *manB* and *CelT* are transcribed separately. The conflicted results may be caused by RNA samples which were collected under different growth conditions. Wet-lab experiments will be needed to verify the conflicting results.

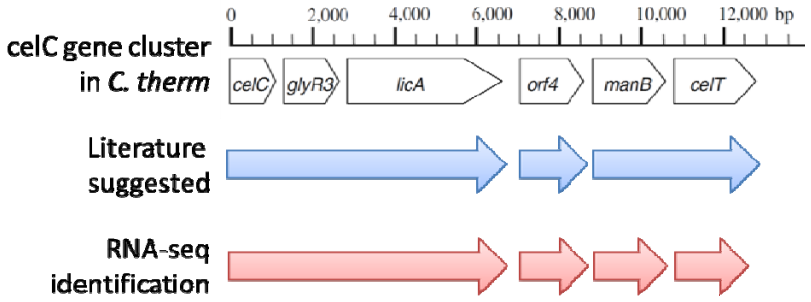


Figure 4.6. *CelC* gene cluster identified by RNA-seq data.

4.5 DISCUSSION

Strand-specific RNA-seq data provide the ability to identify genome-wide TUs. These classifiers use two groups of features to identify TUs. Compared to the predicted operon map, there were more TUs identified. RNA-seq data gave better results than those predicted by only genome sequence information. Many programs such as EDGR (Leek, et al., 2006) use RNA-seq data to detect differential expressed genes. Here, this study provides another level of information to determine how transcriptome structures change over the different experiment conditions. In the results of Control 1 and Treatment 2, there are 235 genes which are not in the same TUs, thus indicating that *C. thermocellum* may use different regulatory systems when growth conditions are changed.

4.6 ACKNOWLEDGEMENTS

This work was supported by a grant from the U.S. Department of Energy (# DE-PS02-06ER64304). The BioEnergy Science Center (BESC) is supported by the Office of Biological and Environmental Research in the DOE Office of Science.

CHAPTER 5

CONCLUSION AND FUTURE WORK

My dissertation consists of three biofuels-related projects: 1) identification of Golgi proteins, 2) de novo transcriptome assembly of switchgrass, and 3) identification of transcription units. The three projects aimed to solve the puzzle of plant cell walls biosynthesis and ethanol production.

GolgiP provides the ability to identify novel plant Golgi proteins, and is still successfully serving researches who want to find Golgi proteins among their interested protein sequences. I had also done some evolutionary studies on proteins involved in ER-to-Golgi transport. The results done by using 18 sequenced plant genomes showed conservation of the transport system. Understanding the ER-to-Golgi transport may provide more insight to identify more Golgi proteins.

Identification of transcribed sequences of switchgrass is a challenging work because of a lack of switchgrass genome sequences. A twofold increase in the amount of transcribed sequences is presented by PviTS. Researches are working on decoding switchgrass genome. I believe PviTS can help to evaluate the quality of genome assembly.

Annotation of transcription units provides the fundamental knowledge to understand the regulations of a prokaryotic cell. My third project delivered a prediction tool to construct transcriptome structure map. The tool will also help to identify dynamically arranged transcription units. There are many different RNA library preparation methods such as 5' cap library and two-terminus paired library developed to accurately identify transcription units. I

hope to make my tool broadly accept RNA sequencing data from different types of RNA libraries.

In sum, I conducted three Bioinformatics projects to study biofuels in plants and bacteria.

REFERENCE

Apweiler, R., *et al.* (2004) UniProt: the Universal Protein knowledgebase, *Nucleic Acids Res*, **32**, D115-119.

Bouton, J.H. (2007) Molecular breeding of switchgrass for use as a biofuel crop, *Current Opinion in Genetics & Development*, **17**, 553-558.

Brouwer, R.W., Kuipers, O.P. and van Hijum, S.A. (2008) The relative value of operon predictions, *Briefings in bioinformatics*, **9**, 367-375.

Cho, B.K., *et al.* (2009) The transcription unit architecture of the Escherichia coli genome, *Nature biotechnology*, **27**, 1043-1049.

Chuck, G.S., *et al.* (2011) Overexpression of the maize Corngrass1 microRNA prevents flowering, improves digestibility, and increases starch content of switchgrass, *Proceedings of the National Academy of Sciences*, **108**, 17550-17555.

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005) Working set selection using second order information for training SVM, *Journal of Machine Learning Research*, **6**.

Fu, C., *et al.* (2011) Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass, *Proceedings of the National Academy of Sciences*, **108**, 3803-3808.

Guell, M., *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium, *Science*, **326**, 1268-1271.

Heazlewood, J.L., *et al.* (2007) SUBA: the Arabidopsis Subcellular Database, *Nucleic Acids Res*, **35**, D213-218.

Horton, P., *et al.* (2007) WoLF PSORT: protein localization predictor, *Nucleic Acids Res*, **35**, W585-587.

Jacob, F., *et al.* (1960) [Operon: a group of genes with the expression coordinated by an operator], *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, **250**, 1727-1729.

Kall, L., Krogh, A. and Sonnhammer, E.L. (2004) A combined transmembrane topology and signal peptide prediction method, *J Mol Biol*, **338**, 1027-1036.

Keshwani, D.R. and Cheng, J.J. (2009) Switchgrass for bioethanol and other value-added applications: A review, *Bioresource Technology*, **100**, 1515-1523.

Komatsu, S., Konishi, H. and Hashimoto, M. (2007) The proteomics of plant cell membranes, *J Exp Bot*, **58**, 103-112.

Krogh, A., *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, **305**, 567-580.

Lemus, R., *et al.* (2002) Biomass yield and quality of 20 switchgrass populations in southern Iowa, USA, *Biomass and Bioenergy*, **23**, 433-442.

Lerouxel, O., *et al.* (2006) Biosynthesis of plant cell wall polysaccharides - a complex process, *Curr Opin Plant Biol*, **9**, 621-630.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics (Oxford, England)*, **22**, 1658-1659.

Mao, F., *et al.* (2009) DOOR: a database for prokaryotic operons, *Nucleic acids research*, **37**, D459-463.

Marchler-Bauer, A., *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database, *Nucleic Acids Res*, **37**, D205-210.

McLaughlin, S.B. and Adams Kszos, L. (2005) Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States, *Biomass and Bioenergy*, **28**, 515-535.

Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem Sci*, **24**, 34-36.

Newcomb, M., Chen, C.Y. and Wu, J.H. (2007) Induction of the celC operon of Clostridium thermocellum by laminaribiose, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 3747-3752.

Newcomb, M., *et al.* (2011) Co-transcription of the celC gene cluster in Clostridium thermocellum, *Applied microbiology and biotechnology*, **90**, 625-634.

Nilsson, T., Au, C.E. and Bergeron, J.J. (2009) Sorting out glycosylation enzymes in the Golgi apparatus, *FEBS Lett*, **583**, 3764-3769.

Okuda, S. and Yoshizawa, A.C. (2011) ODB: a database for operon organizations, 2011 update, *Nucleic acids research*, **39**, D552-555.

Oliver, H.F., *et al.* (2009) Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs, *BMC genomics*, **10**, 641.

Pierce, B. (2004) *Genetics: A Conceptual Approach, 2nd ed.* W. H. Freeman.

Saathoff, A., *et al.* (2011) Switchgrass contains two cinnamyl alcohol dehydrogenases involved in lignin formation, *BioEnergy Research*, **4**, 120-133.

Sarath, G., *et al.* (2011) Ethanol yields and cell wall properties in divergently bred switchgrass genotypes, *Bioresource Technology*, **102**, 9579-9585.

Schmer, M.R., *et al.* (2008) Net energy of cellulosic ethanol from switchgrass, *Proceedings of the National Academy of Sciences*, **105**, 464-469.

Schwacke, R., *et al.* (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins, *Plant Physiol*, **131**, 16-26.

Sharma, C.M., *et al.* (2010) The primary transcriptome of the major human pathogen Helicobacter pylori, *Nature*, **464**, 250-255.

Siqueira, F.M., Schrank, A. and Schrank, I.S. (2011) Mycoplasma hyopneumoniae transcription unit organization: genome survey and prediction, *DNA research : an international journal for rapid publication of reports on genes and genomes*, **18**, 413-422.

Sprenger, J., *et al.* (2007) LOCATE: a mammalian protein subcellular localization database, *Nucleic Acids Research*.

Sprenger, J., Fink, J.L. and Teasdale, R.D. (2006) Evaluation and comparison of mammalian subcellular localization prediction methods, *BMC Bioinformatics*, **7 Suppl 5**, S3.

Toledo-Arana, A., *et al.* (2009) The Listeria transcriptional landscape from saprophytism to virulence, *Nature*, **459**, 950-956.

Wang, L., *et al.* (2004) Genome-wide operon prediction in Staphylococcus aureus, *Nucleic acids research*, **32**, 3689-3702.

Wang, Y., Samuels, T. and Wu, Y. (2011) Development of 1,030 genomic SSR markers in switchgrass, *TAG Theoretical and Applied Genetics*, **122**, 677-686.

Xu, B., *et al.* (2011) Silencing of 4-coumarate:coenzyme A ligase in switchgrass leads to reduced lignin content and improved fermentable sugar yields for biofuel production, *New Phytologist*, no-no.

Yuan, J.S., *et al.* (2008) Plants to power: bioenergy to fuel the future, *Trends in Plant Science*, **13**, 421-429.

Yuan, Z. and Teasdale, R.D. (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains, *Bioinformatics (Oxford, England)*, **18**, 1109-1115.