1-1-2012

Genomic and Functional Analysis of Next-Generation Sequencing Data

Philippe Chouvarine

Genomic and functional analysis of next-generation sequencing data

By

Philippe Chouvarine

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Veterinary Medical Sciences
in the College of Veterinary Medicine

Mississippi State, Mississippi

December 2012

Genomic and functional analysis of next-generation sequencing data

By

Philippe Chouvarine

Approved:

Fiona M. McCarthy
Assistant Professor of Basic Sciences
(Co-major Professor)

Bindu Nanduri
Assistant Professor of Basic Sciences
(Co-major Professor)

Daniel G. Peterson
Professor of Plant and Soil Sciences
(Committee Member)

Susan M. Bridges
Professor of Computer Science and Engineering
(Committee Member)

Shane C. Burgess
Professor of Basic Sciences
(Committee Member)

Larry A. Hanson
Professor of Basic Sciences
(Graduate Coordinator)

Kent H. Hoblet
Professor and Dean of Basic Sciences

Name: Philippe Chouvarine

Date of Degree: December 15, 2012

Institution: Mississippi State University

Major Field: Veterinary Medical Sciences

Major Professor: Fiona M. McCarthy, Bindu Nanduri

Title of Study: Genomic and functional analysis of next-generation sequencing data

Pages in Study: 65

Candidate for Degree of Doctor of Philosophy

Advances in next-generation sequencing (NGS) technologies have resulted in significant reduction of cost per sequenced base pair and increase in sequence data volume. On the other hand, most currently used NGS technologies produce relatively short sequence reads (50 - 150 bp) compared to Sanger sequencing (~700 bp). This represents an additional challenge in data analysis, because shorter reads are more difficult to assemble. At this point, production of sequencing data outpaces our capacity to analyze them. Newer NGS technologies capable of producing longer reads are emerging, which should simplify and speed up genome assembly. However, this will only increase the number of sequenced genomes without structural and functional annotation. In addition to multiple scientific initiatives to sequence thousands of genomes, personalized medicine centered on sequencing and analysis of individual human genomes will become more available. This poses a challenge for computer science and emphasizes the importance of developing new computational algorithms, methodology, tools, and pipelines. This dissertation focuses on development of these software tools, methodologies, and resources to help address the need for processing of volumes of data generated by new sequencing technologies. The research concentrated on genome structure analysis, individual variation, and comparative biology. This dissertation presents: (1) the Short Read Classification Pipeline (SRCP) for preliminary genome characterization of unsequenced genomes; (2) a novel methodology for phylogenetic analysis of closely related organisms or strains of the same organism without a sequenced genome; (3) a centralized online resource for standardized gene nomenclature. Utilizing the SRCP and the methodology for initial phylogenetic analysis developed in this dissertation enables positioning the organism in the evolutionary context. This should facilitate identification of orthologs between the species and paralogs within the species even in the initial stage of the analysis when only exome is sequenced and, thus, enable functional annotation by transferring gene nomenclature from well-annotated 1:1 orthologs, as required by the online standardized gene nomenclature resource developed in this dissertation. Thus, the tools, methodology, and resources presented here are tied together in following the initial analysis workflow for structural and functional annotation.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Fiona McCarthy, my mentor, for the opportunity to pursue my Ph.D. study under her guidance. Throughout the entire program, Fiona provided constant support, professional guidance, and intellectual freedom.

In addition, I would like to extend my gratitude to Dr. Shane Burgess for guidance, encouragement, scientific discussions, and career advice. I would also like to thank my committee members Dr. Daniel Peterson and Dr. Susan Bridges, for their generous support and guidance. I am grateful to all members of my graduate committee for their time and effort reviewing the research described in this dissertation and their valuable recommendations and advice.

I would like to thank, Dr. David Ray, Amanda Cooksey, Carole Nail, Dr. Surya Saha, Cathy Gresham, Dr. Brian Baldwin, Dr. Janet Weber, Yachi Spencer, and Swati Kumari for their contribution to the research described in this dissertation. I am also grateful to my friends and family for their support and encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

I. INTRODUCTION 1

Next-generation sequencing 1
Second-generation sequencing 2
Limitations of the second-generation technologies 2
Third-generation sequencing, advantages and limitations 3
Fourth-generation sequencing 3
Effects of next-generation sequencing on biology 4
Big data 5
How this dissertation addresses new challenges in data analysis 5
Genome structure analysis 5
Analysis of individual variation 6
Comparative biology and gene nomenclature 8
References Cited 10

II. AN AUTOMATED, HIGH-THROUGHPUT SEQUENCE READ CLASSIFICATION PIPELINE FOR PRELIMINARY GENOME CHARACTERIZATION[1] 13

Abstract 14
Materials and methods 15
General considerations 15
Technologies 15
Populating the repeat and organellar local databases 16
Populating the "gene sequence" local databases 16
Preparation of query sequences 16
Analysis of random genomic DNA query sequences 17
First-round detection of repeat sequences 18
Classification of query sequences with significant local database BLAST hits 18
Identifying repeats in the "No Hit" query sequences 19

Classification of remaining "No Hit" query sequences 20
Output.. 20
Contig assembly..20
Analysis of reduced representation sequences..20
Results and discussion..20
SRCP analysis of random genomic sorghum query sequences 20
Comparison of Cot analysis and SRCP data..20
The effect of query sequence length on classification....................................21
Analysis of Cot-filtered DNA 22
Conclusions...22
Acknowledgments...22
Appendix A. Supplementary data 23
References 23

III. TRANSCRIPTOME-BASED DIFFERENTIATION OF CLOSELY RELATED MISCANTHUS LINES[1] 24

Abstract...25
Introduction...25
Results and Discussion 26
Plant Materials 26
Transcriptome Sequencing 27
Phylogenetic Analysis 27
Exome Assembly 28
Functional Annotation and Analysis 29
Data..... 30
Methods 30
Transcriptome Sequencing 30
Library Construction 30
Phylogenetic Analysis 30
Exome Assembly and Functional Analysis 31
Amazon EC2 Cloud Computing 32
Author Contributions..32
References 32

IV. MANUAL BIOCURATION TO SUPPORT STANDARDIZED CHICKEN GENE NOMENCLATURE AT CGNC...............................33

Abstract...33
Introduction...34
CGNC Database 36
Implementation and Updates 36
HCOP Orthology Resources...38
BirdBase Resources..39
CGNC Website 40
Searching the Web Interface..40

CGNC Gene Pages 40
Submitting Data to CGNC 41
Downloads 42
Integrating CGNC Data 42
Assigning Nomenclature 43
Automatic Curation: Chicken-Human Orthologs 43
Manual Biocuration Projects 43
Collaborations 46
Future Directions 47
Competing Interests 48
Authors' Contributions 48
Acknowledgements 48
Figures... 50
References Cited 53

V. CONCLUSION 55

References Cited 63

LIST OF TABLES

2.1 Database and Web page addresses 15

3.1 SNPs per aligned bp identified in comparative analysis of cDNA regions common to all samples 27

3.2 Distance matrix...27

3.3 Transcript assembly metrics evaluation using *Arabidopsis thaliana* assemblies 30

LIST OF FIGURES

2.1 General overview of the sequence read classification pipeline (SRCP)................15

2.2 Steps in categorization of random genomic query sequences...............................17

2.3 SRCP-based classification of random sorghum genomic shotgun query sequences 21

2.4 Low-copy and repeat sequence contents of highly repetitive (HR), moderately repetitive (MR), and single/low-copy sorghum DNA libraries as determined by the SRCP and by manual classification 22

3.1 Outline of procedure used to identify SNPs from miscanthus samples.................26

3.2 Phylogenetic tree inferred by SNP analysis in common regions of all seven samples...28

3.3 Impact of k-mer size on characteristics of *Miscanthus x giganteus* exome assembly in Velvet..29

3.4 Distribution of GO annotation for miscanthus sequences compared to *Sorghum bicolor*..31

4.1 CGNC interface for searching chicken gene nomenclature and the corresponding human orthologs 50

4.2 CGNC gene report page...51

4.3 Login and Request Login links on the CGNC front page for gene nomenclature editing by research community..52

CHAPTER I

INTRODUCTION

Computerization of analysis of biological data is key to accelerated research in many areas of biology, including the ones discussed in this dissertation: genome structure analysis, analysis of individual variation, and comparative biology. In this chapter, I review advantages and limitations of next-generation sequencing and the effects it has on biology. One of its major effects is production of large volumes of data that are difficult to manage and analyze due to their size. Further, I will show how the work presented in this dissertation helps tackle analysis of "big data" in biology.

Next-generation sequencing

Rapid advances in sequencing technology revolutionize many areas of biological research. In the last decade the sequencing cost per base has been reduced by more than 100,000 fold [1]. The speed of sequencing has also dramatically increased due to massively parallel approach used in the next-generation sequencers where millions of sequencing reactions with real-time sequence identification can be performed simultaneously [2]. Next-generation sequencing generally refers to sequencing technologies that originated after Sanger capillary sequencing, which required cloning of DNA fragments (digested by enzymes or mechanically sheared) into DNA vectors

(Bacterial Artificial Chromosome (BAC), Yeast Artificial Chromosome (YAC), etc.) for sample amplification.

Second-generation sequencing

The second generation of sequencing technologies (Roche 454, Illumina, SOLiD, and Polonator) overcame the need for vectors by performing *in vitro* cloning amplification. In this process fragmented single-stranded DNA (ssDNA) is ligated to adapters (on both ends), followed by annealing of the adapters to complementary ssDNA on the sequencing media (solid surface or beads). The following polymerase chain reaction (PCR) creates clusters of amplified ssDNA. The sequence identification is based on the polymerase reaction that builds the complementary strand in each of the ssDNA sequences of the amplified cluster. Sequencing, in which such polymerase reaction takes place, is called Sequencing by Synthesis [3].

Limitations of the second-generation technologies

Optical methods, such as pyrosequencing or use of fluorescently labeled deoxynucleotide triphosphates (dNTPs), are utilized to determine incorporated nucleotides. These technologies are using the process of introduction of a single type of labeled nucleotides (either A, T, C, or G) at a time to all sequencing reactions in an amplified cluster of ssDNA to detect the average amplified signal from all reactions. However, with every subsequent incorporation-cycle the signal quality drops due to the lag in incorporation of nucleotides from the previous cycles. This loss of phasing (maintaining synchronous synthesis among identical DNA templates) leads to quality degradation as sequencing progresses toward the 3' end and limits the read lengths produced by these technologies [3]. Other sources of errors in PCR-based sequencing are associated with the PCR process, which introduces editing errors caused by DNA polymerase-catalyzed enzymatic copying and errors due to DNA thermal damage [4].

Third-generation sequencing, advantages and limitations

These problems are addressed by the third generation sequencing technologies (Pacific Biosciences SMRT (http://www.pacificbiosciences.com/) and Helicos (http://www.helicosbio.com/)) allowing single molecule sequencing without PCR amplification. Regardless, when optical identification methods are used in Sequencing by Synthesis, they introduce their own limitations, such as contamination of labeled dNTPs by unlabeled dNTPs (e.g., impurities or hydrolysis products), stray signals from dye molecules that stick to the sequencing surfaces, limitations due to camera read rate capacity, etc. [3]. These issues were addressed in technologies that utilize ionic current for sequence identification. One of the first instruments on the market to utilize this technology was IonTorrent (http://www.iontorrent.com/). While IonTorrent still relies on PCR amplification, which makes it susceptible to problems with the second-generation technologies described above, it identifies the attached nucleotides by change in the pH level associated with the sequencing reaction.

Fourth-generation sequencing

The fourth generation, nanopore-based sequencing technology (Oxford Nanopore (http://www.nanoporetech.com/), Genia (http://geniachip.com/), Nabsys (http://nabsys.com/)) is still in development, but it will potentially address the issues in the previous technologies described above. As described in the introductory materials presented on the web sites of these three companies, nanopore-based sequencing combines single molecule processing with ion current-based sequence identification.

Effects of next-generation sequencing on biology

As shown above, the recent trends in the third- and fourth-generation sequencing technologies are likely to result in increases in read length and sequencing quality, which will make whole genome sequencing faster and less computationally intensive in the future. Availability of long and error-free reads would make it easier to sequence whole genomes of patients in clinical research, regardless of long stretches of DNA varying from the reference genome sequence, thus, making personalized medicine more available. Whole genome sequencing of patients will reveal DNA variations in their personal genomes, which will allow customized healthcare, screening for genetically predisposed risks, and preventive treatment [5]. Of course, a simple knowledge of the entire genome sequence of an individual is not enough to make educated decisions about personalized healthcare for this individual. A systems biology approach, which considers predictive quantitative models for biological systems in a holistic rather than reductionist manner, is necessary to understand how various DNA variations, e.g., gene mutations, present in a given genome can affect gene expression and alter biological pathways [6, 7]. Gene expression profiling of diseased tissues can reveal the stage of the disease and the progress in its treatment. Advances in gene therapy, such as genome editing [8], can be used to directly correct the disease causing DNA mutations in targeted cells.

Big data

The advent of personalized medicine as well as scientific initiatives aiming for sequencing of thousands of new genomes, such as Genome 10K project [9] produce an unprecedented volume of data waiting to be analyzed. This poses a challenge for computer science and emphasizes the importance of developing new computational algorithms, methodology, tools, and pipelines. Note that open source program development, as well as utilizing open source journals for sharing information about availability of new methodology, programs, resources, etc., is extremely important for synergetic scientific research [10]. In the abundance of published research and tools, open source alternatives are likely to be considered first.

How this dissertation addresses new challenges in data analysis

Genome structure analysis

The research presented in this dissertation focuses on development of algorithms, methodologies, tools, and online resources to help make sense of available sequencing data. While easy and accurate whole genome assembly is still out of reach of the current sequencing technologies, genome structure analysis still plays an important role for: (1) identification of repeat elements for their further analysis in regulation, speciation and evolution [11, 12, 13, 14]; (2) identification of the expected percentages of DNA content for validation of future genome assembly [12, 14]; (3) identification of coding sequences for exome assembly [12, 13, 14]. Chapter II of this dissertation will introduce Short Read Classification Pipeline [14] that enables preliminary characterization of organisms without a genome reference by identification of DNA content percentages of various classes of DNA.

Analysis of individual variation

Genetic variation among individual humans plays an important role in personalized medicine research [15]. Individual variation is also essential for finding phylogenetic relationships among closely related organisms or strains of the same organism. Originally, phylogenetic analysis was done by utilizing trait tables with quantified morphological characteristics to identify evolutionary distance among the sampled species. The drawbacks of this approach are that a hypothesis must be made about evolutionary relevance of the traits that should be included [16] and that the same phenotypic trait can be acquired in unrelated lineages [17]. Availability of DNA/RNA sequencing made it possible to perform molecular phylogenetic analysis by identifying orthologous DNA or RNA sequences and performing multiple alignments of such orthologous sequences from all samples. Sequence variations can then be quantified in a distance matrix that is used to construct a phylogenetic tree. The orthologous sequences can be extracted using specific PCR primers followed by PCR amplification for sequencing [18]. The limitation of this method is that phylogeny of a species is determined on the basis of a single gene or a locus [19]. It is possible to combine sequence data from a large number of DNA loci to build a consensus phylogenetic tree [20], however, this would require substantial sequencing effort designing multiple PCR primers or utilizing a fragment polymorphism technique to size separate orthologous sequences [21]. Another approach is to use a genome-wide genetic variation identification by targeting all genes, all microsatellites, etc. These methods employ various genetic markers such as microsatellites, RFLP (Restriction Fragment Length Polymorphisms), AFLP (Amplified Fragment Length Polymorphisms), RAPD (Randomly Amplified Polymorphic DNA), and VNTR (Variable Number Tandem Repeats) [22] to infer phylogeny in organisms without sequenced genomes. However, using SNPs for phylogenetic analysis is more advantageous because they are much more abundant than other markers (1 per 1000 bp in human, 1 per 500 bp in mouse [23], compared to 1 microsatellite in 100,000 bp for human [24], and 1 mutation per 50,000 bp to 1 per 450,000 bp for fragment length polymorphisms [25]) and they can be easily identified using next generation sequencing (NGS). On the other hand, finding SNPs requires knowledge of at least a partial reference genome sequence. This, however, adds additional precision in determining genetic variations among samples because the SNPs can be identified in reference sequences computationally determined to be homologous to all samples in the study. Therefore, for analysis of closely related strains, where finding as many genomic variations as possible in regions known to be homologous in all samples is very important, we will utilize SNP-based analysis. While whole genome sequencing (WGS) can be used for detection of individual variation, due to lower costs current clinical research concentrates on whole exome sequencing (WES), though there is an agreement that WGS will be predominantly used for individual variation research in the future because it provides additional information about genome structure and regulation [26]. The same is true about finding genetic variation between species. In this case, the major problem with using WGS is that many species do not have a reference genome sequenced. Contemporary sequencing efforts largely rely on the second-generation technologies, which are producing large volumes of short-reads (50 - 150 bp), which are difficult to assemble, especially in the repetitive regions. Therefore, sequencing genomes of most eukaryotes is still a challenge and requires incorporation of long reads (Sanger, Pacific Biosciences SMRT , etc.) and/or construction of mate-pair read libraries with variable insert sizes. To avoid the costs and complexity of WGS, complementary DNA (cDNA) sequencing capturing transcript sequences can be used because the most informative genetic variations are located in the coding regions, since they are evolutionary constrained by the function of the proteins they encode [27]. It is likely that not all samples will have the same set of transcripts sequenced and that some of these transcripts will not be sequenced to their full length. To address this issue, partial transcript references homologous to all samples in the study can be assembled. Next-generation sequencing makes it possible to provide significant read alignment coverage and detect coding sequences with very low expression levels, thus increasing the portion of the exome available for genetic variation analysis. Chapter III of this dissertation will cover methodology for utilizing RNA-seq reads for transcript assembly and phylogenetic analysis of closely related organisms without a reference genome.

Comparative biology and gene nomenclature

Advances in next generation sequencing are likely to facilitate generation of thousands of new draft genome sequences in the near future [9]. As genome annotation efforts ensue, the role of comparative biology should become increasingly more important. The existing orthology among the genes of the studied organisms should be identified to transfer gene nomenclature from 1:1 orthologs in well-annotated model organisms to the organisms with less established gene annotation. It is important that such model organisms have standardized and unambiguous gene nomenclature to prevent spreading inconsistencies in gene naming to different organisms used in comparative genomics. Generally, when inconsistent gene nomenclature is used in research, it can cause confusion, duplicated effort, and errors. For example, two research groups may call the same gene two different gene names or use the same name for two unrelated genes and then use each other's results in their research. There should also exist one standard gene naming convention, so that all genes are named following the same rules, e.g., using brief and specific names that convey the character or function of the gene, using American spelling, avoiding tissue specificity or molecular weight designations. Symbols and synonyms should also be standardized. Following such a standardized naming convention (as opposed to calling a gene "smurf" or "pokemon") will ensure that the researchers will get the most meaningful information about the gene from its name, symbol, and synonym. Chapter IV of the dissertation will discuss creation, maintenance, and functionality of a centralized online resource for standardized chicken gene nomenclature. The importance of this resource is that aside from standardizing chicken nomenclature it can also be easily adapted for other model organisms with human orthologs supported by HUGO (Human Genome Organization) Gene Nomenclature Committee (HGNC) (http://www.genenames.org/). According to HGNC, there are currently no Horse, Cow, Chimp, Macaque, Opossum, Platypus or Dog Gene Nomenclature Committees.

References Cited

[1] Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**: 187-197.

[2] Wall P.K., Leebens-Mack J., Chanderbali A.S., Barakat A., Wolcott E., et al. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**: 347.

[3] Fuller C.W., Middendorf L.R., Benner S.A., Church G.M., Harris T., et al. (2009) The challenges of sequencing by synthesis. *Nat Biotechnol*, **27**: 1013-1023.

[4] Pienaar E., Theron M., Nelson M., Viljoen H.J. (2006) A quantitative model of error accumulation during PCR amplification. *Comput Biol Chem*, **30**: 102-111.

[5] Tonellato P.J., Crawford J.M., Boguski M.S., Saffitz J.E. (2011) A national agenda for the future of pathology in personalized medicine: report of the proceedings of a meeting at the Banbury Conference Center on genome-era pathology, precision diagnostics, and preemptive care: a stakeholder summit. *Am J Clin Pathol*, **135**: 668-672.

[6] Snoep J.L., Westerhoff H.V. (2005) From isolation to integration, a systems biology approach for building the Silicon Cell. *Topics in Current Genetics*, **13**: 13-30.

[7] Kan Z., Jaiswal B.S., Stinson J., Janakiraman V., Bhatt D., et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, **466**: 869-873.

[8] Li H., Haurigot V., Doyon Y., Li T., Wong S.Y., et al. (2011) In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature*, **475**: 217-221.

[9] (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered*, **100**: 659-674.

[10] Woelfle M., Olliaro P., Todd M.H. (2011) Open science is a research accelerator. *Nature Chemistr*, **3**: 745-748.

[11] Lorenzi H., Thiagarajan M., Haas B., Wortman J., Hall N., et al. (2008) Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species. *BMC Genomics*, **9**: 595.

[12] Swaminathan K., Varala K., Hudson M.E. (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*, **8**: 132.

[13] Strong W.B., Nelson R.G. (2000) Preliminary profile of the Cryptosporidium parvum genome: an expressed sequence tag and genome survey sequence analysis. *Mol Biochem Parasitol*, **107**: 1-32.

[14] Chouvarine P., Saha S., Peterson D.G. (2008) An automated, high-throughput sequence read classification pipeline for preliminary genome characterization. *Anal Biochem*, **373**: 78-87.

[15] Guttmacher A.E., McGuire A.L., Ponder B., Stefansson K. (2010) Personalized genomic information: preparing for the future of genetic medicine. *Nat Rev Genet*, **11**: 161-165.

[16] Swiderski D., Zelditch M., Fink W. (1998) Why morphometrics is not special: coding quantitative data for phylogenetic analysis. *Systematic Biology*, **47**: 508-519.

[17] Gaubert P., Wozencraft W.C., Cordeiro-Estrela P., Veron G. (2005) Mosaics of convergences and noise in morphological phylogenies: what's in a viverrid-like carnivoran? *Syst Biol*, **54**: 865-894.

[18] Fredslund J., Schauser L., Madsen L., Sandal N., Stougaard J. (2005) PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucleic Acids Res*, **33**: W516-W520.

[19] Tateno Y., Nei M., Tajima F. (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J Mol Evol*, **18**: 387-404.

[20] Pamilo P., Nei M. (1988) Relationships between gene trees and species trees. *Mol Biol Evol*, **5**: 568–583.

[21] Nicod J.C., Largiader C.R. (2003) SNPs by AFLP (SBA): a rapid SNP isolation strategy for non-model organisms. *Nucleic Acids Res*, **31**: e19.

[22] Jones N., Ougham H., Thomas H. (1997) Markers and mapping: we are all geneticists now. *New Phytologist*, **137**: 165-177.

[23] Williams J.L., Dunner S., Valentini A., Mazza R., Amarger V., et al. (2009) Discovery, characterization and validation of single nucleotide polymorphisms within 206 bovine genes that may be considered as candidate genes for beef production and quality. *Anim Genet*, **40**: 486-491.

[24] Wada C., Shionoya S., Fujino Y., Tokuhiro H., Akahoshi T., et al. (1994) Genomic instability of microsatellite repeats and its association with the evolution of chronic myelogenous leukemia. *Blood*, **83**: 3449-3456.

[25] Caldwell D., McCallum N., Mudie S., Hedley P., Ramsay L., et al. (2002) A physical/chemical mutation grid for barley functional genomics. *Annual Report 2001/2002*, Scottish Crop Research Institute, Dundee, Scotland, 157-158 p.

[26] Cirulli E.T., Goldstein D.B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*, **11**: 415-425.

[27] Botstein D., Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, **33** Suppl: 228-237.

CHAPTER II

AN AUTOMATED, HIGH-THROUGHPUT SEQUENCE READ CLASSIFICATION PIPELINE FOR PRELIMINARY GENOME CHARACTERIZATION[1]

[1] Reprint from Chouvarine P., Saha S., Peterson D.G. (2008) An automated, high-throughput sequence read classification pipeline for preliminary genome characterization. Analytical Biochemistry, **373**: 78-87. This article is available from: http://www.ncbi.nlm.nih.gov/pubmed/17868636

Available online at www.sciencedirect.com

ScienceDirect

Analytical Biochemistry 373 (2008) 78–87

ANALYTICAL BIOCHEMISTRY

www.elsevier.com/locate/yabio

An automated, high-throughput sequence read classification pipeline for preliminary genome characterization

Philippe Chouvarine [a,b], Surya Saha [a,c], Daniel G. Peterson [a,b,d,*]

[a] Mississippi Genome Exploration Laboratory, Mississippi State University, Mississippi State, MS 39762, USA
[b] Department of Plant & Soil Sciences, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA
[c] Department of Computer Science & Engineering, Mississippi State University, Mississippi State, MS 39762, USA
[d] Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA

Received 9 July 2007
Available online 10 August 2007

Abstract

In the absence of a complete genome sequence, considerable insight into genome structure can be gained from survey sequencing of genomic DNA. To facilitate high-throughput characterization of genome structure based on shotgun sequence reads, we have developed an automated sequence read classification pipeline (SRCP). The SRCP uses a battery of novel and standard sequence analysis algorithms along with a sophisticated *decision tree* to place reads into "best fit" functional/descriptive categories. Once "primed" with genomic sequence data, the SRCP also permits estimation of gene/repeat enrichment afforded by reduced-representation sequencing techniques. To our knowledge, the SRCP is the only tool that has been designed to provide a description of a genome or a genome component based on sample sequence reads. In an initial test of the SRCP using sequence data from *Sorghum bicolor*, it was shown to provide results similar in quality to results generated by manual classification. Although the SRCP is not a replacement for manual sequence characterization, it can provide a rapid, high-quality overview of genome sequence content and facilitate subsequent annotation. The SRCP presumably can be adapted for analysis of any eukaryotic genome.

Keywords: DNA; Sequence analysis; Transposon; Genome; Bioinformatics; Computational analysis; Genomics; Comparative

Although complete genome sequencing represents an ideal means by which the genomes of organisms can be compared, it is not currently economically feasible for most eukaryotes. This is especially true for the numerous organisms that have large, highly repetitive genomes including many important plants and animals. With this said, sample sequencing of random genomic DNA can be used to gain considerable information about genome structure in lieu of a complete sequence [1,2]. However, it is often difficult for researchers to characterize the sequences they have obtained, especially if they have generated large sequence data sets for organisms for which previous sequencing research has been limited.

At present, numerous automated and semiautomated gene characterization programs are available [3,4]. Likewise, there are a growing number of programs designed to characterize repetitive elements [5–7]. However, to our knowledge, there is no program or pipeline designed to provide an overview of the sequence composition of an entire genome based on shotgun sequence reads. To permit such characterization, we have constructed a sequence read classification pipeline (SRCP)[1] in which a battery of exist-

[*] Corresponding author. Address: Department of Plant & Soil Sciences, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA. Fax: +1 662 325 8742.
E-mail address: dpeterson@pss.msstate.edu (D.G. Peterson).

[1] *Abbreviations used:* SRCP, sequence read classification pipeline; NCBI, National Center for Biotechnology Information; EST, expressed sequence tag; EMC, EST/mRNA/cDNA; BLAST, Basic Local Alignment Search Tool; TIGR, The Institute for Genomic Research; BLAT, BLAST-like alignment tool; XML, Extensible Markup Language; IIS, Internet Information Services; ASP, Active Server Pages; DTS, Data Transformation Services; FTP, File Transfer Protocol; SQL, Structured Query Language; XSLT, Extensible Stylesheet Language Transformations.

0003-2697/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2007.08.008

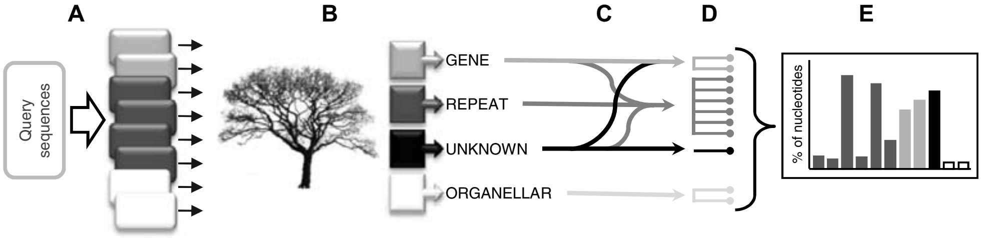


Fig. 1. General overview of the sequence read classification pipeline (SRCP). (A) Query sequences are compared using BLAST (Basic Local Alignment Search Tool) to the contents of two gene (light gray rectangles), four repeat (dark gray rectangles), and two organellar (white rectangles) highly curated, local sequence databases. (B) For each query sequence, data from the BLAST analyses are evaluated with a decision tree algorithm that places that sequence into a "best fit" descriptive gene, repeat, or organellar DNA category; those sequences that do not possess significant homology to sequences in any of the local sequence databases are classified as unknown. (C) Two independent algorithms interrogate those sequences classified as gene or unknown to see if they are possibly repetitive based on their frequency within the data set. Additionally, the unknown sequences are analyzed with *tblastn* to determine if they share significant homology with nontransposon proteins. Based on these secondary analyses, some query sequences are reclassified. (D) Each query sequence is placed into one of 11 final sequence categories. (E) The output of the SRCP is a graph (along with data and statistics) illustrating the composition of the query sequence set.

ing and novel algorithms are used to place random genomic query sequences into descriptive/functional sequence categories. The SRCP calculates the fraction of base pairs in each category, thus providing an overview of genome structure while facilitating initial annotation of query sequences (Fig. 1). In addition, the efficacy of reduced-representation sequencing techniques [8,9] can be assessed by comparing SRCP results for random genomic sequence with SRCP results for gene- or repeat-enriched DNA. With respect to basic configuration, the SRCP uses the program BLAST (Basic Local Alignment Search Tool) to query sequences against highly curated, custom local databases. The BLAST data are filtered, stored in a relational database, and analyzed to derive the final classification of each query sequence. The results of the analysis are available via a Web interface. The system is implemented as a series of Perl scripts, database scripts/queries, and dynamic Web pages.

Table 1
Database and Web page addresses

Database or Web page	Web address
NCBI	www.ncbi.nlm.nih.gov
Core Nucleotide DB	www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nuccore
EST DB	www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucest
Entrez Help Document	www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html
Display Formats	www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Display_Formats
Plastid Organelles	www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plastids.html
Viridiplantae Mitochondria	www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plants.html
The Inst. for Genomic Res.	www.tigr.org/
TIGR Gene Index FTP site	ftp://ftp.tigr.org/pub/data/tgi/
Canad. Bioinf. Help Desk	gchelpdesk.ualberta.ca

Materials and methods

General considerations

1. Because our research is focused primarily on study of seed plants (Phylum Spermatophyta), we developed the SRCP for analysis of sequences from spermatophytes. However, the basic SRCP structure can be adapted for study of any organism or group of organisms.
2. The different sequence categories in the SRCP are based on those used by Peterson and co-workers [10].
3. The addresses of public Web pages and databases not generated as part of our research are given in Table 1.
4. Interested parties can obtain source codes and/or downloads of novel tools and access the contents of our local sequence databases at http://www.mgel.msstate.edu/tools.htm.
5. The version of BLAST (Linux-ia32, Version 2.2.14) used in this pipeline was obtained from the National Center for Biotechnology Information (NCBI).

Technologies

Traditionally, bioinformatics projects have used Linux/Unix platforms. However, there are a number of powerful and often neglected Windows-based software development technologies that afford rich functionality without extensive de novo programming. For this research, we developed a hybrid Linux and Windows system to use the strengths of both operating systems. The power of the Linux operating system lies in its robustness, scalability, and high availability of compatible bioinformatics software. Therefore, we chose to run Linux on the computational server that runs bioinformatics tools. With respect to Windows tools, our database server runs SQL Server 2000 (SQL = Structured Query Language), and we use its built-in Data Transformation Services (DTS) for bulk upload of large XML (Extensible Markup Language) files containing BLAST results. We also use DTS to implement the classification logic of the pipeline (see below). Our Web server runs IIS (Internet Information Services) 6.0, which provides powerful native lock-down mechanisms. The freely available URLScan program (http://www.microsoft.com/technet/security/tools/urlscan.mspx) can be used to secure all versions of IIS. Running IIS allows us to use ASP.NET (ASP = Active Server Pages) for our Web interface. ASP.NET provides a collection of powerful and easily customizable Web controls, most notably the ''data grid'' control, which is ideal for displaying large data sets in a table structure with editable cells.

Populating the repeat and organellar local databases

For all repeat and organellar sequences, we currently download sequence information in the GenBank file format, which includes not only the sequence, its accession number, and its title, but detailed annotation and Internet links.

Spermatophyte transposon, rDNA, and centromere sequences were extracted from the NCBI Core Nucleotide Database by conducting searches using boolean text strings (Supplementary Table 1). Search results were used to create Transposon, rDNA, and Centromere local databases.

Chloroplast genome sequences were downloaded from NCBI's Plastid Organelles page and placed in the Chloroplast local database. Spermatophyte mitochondria sequences were downloaded from NCBI's Viridiplantae Mitochondria page and placed in the Mitochondria local database.

Each local database was assigned a version number containing the date it was populated and a two- or three-letter abbreviation indicating its contents (e.g., the first version of the Mitochondria local database was designated MC_2005-10-01). We update these local databases every 6 months.

Because many repeat sequences are found as annotated sections within larger genomic sequence entries (i.e., are not archived as individual GenBank entries), we developed a Perl script that extracts repeat regions and their annotations from select GenBank files. Extracted repeats were placed in an Annotated Repeat local database. Because of the large number of annotated repeats in plant whole-genome sequences, for this initial test we limited our extraction to manually annotated sequences available for *Sorghum bicolor*.

Populating the ''gene sequence'' local databases

Spermatophyte EST, cDNA, and mRNA (EMC) sequences were originally extracted from the NCBI EST Database and Core Nucleotide Database by conducting searches using a boolean search string (Supplementary Table 1). Because of the relatively large number of retrieved sequences, sequence data were downloaded in FASTA format [11] rather than in GenBank format. Downloaded sequences then were BLASTed (*blastn*) against the Chloroplast, Mitochondria, rDNA, Centromere, and Transposon local databases (see above). Any sequence exhibiting a significant hit (bit score = $S' \geqslant 60$) to one of these local databases was eliminated from the data set by Perl scripts. The remaining sequences were deposited in the EMC local database.

Spermatophyte ''gene'' sequences in FASTA format were downloaded from The Institute for Genome Research (TIGR) Gene Index FTP (File Transfer Protocol) site. Downloaded files were then scanned using a Perl script that eliminates those entries containing the following ''repeat-affiliated'' words in their titles (where asterisks indicate wild-card characters): retrovirus, retroelement, transpos*, gag, pol, polyprotein, env, reverse transcriptase, integrase, stowaway, MITE, miniature, copia, gypsy, RT, helitron, maverick, polinton, mul*, insertional, mitocondri*, chloroplast, capsid, and nucleocapsid. Remaining sequences were then BLASTed against the Annotated Repeats, Chloroplast, Mitochondria, rDNA, Centromere, and Transposon local databases. Sequences exhibiting a significant hit ($S' \geqslant 60$) to one or more of these databases were eliminated using the Perl scripts mentioned above. The remaining sequences were deposited in the Gene Index local database.

Preparation of query sequences

Random *S. bicolor* genomic shotgun sequences (GenBank Accession Nos. CW512190–CW514008) [12] were used as a sample ''unfiltered'' query sequence set. These 1819 sequences, collectively representing 1,088,783 bp, have a mean length of 599 bp (SE ± 38). To study the effect of sequence length on SRCP results, two representations of the sequence data were initially tested. The first representation contained the original GenBank sequences without any size adjustments (i.e., full-length query sequences); the second representation contained the same sequences digitally fragmented into 80- to 179-bp (average 105 bp ± SE 0.14) pieces, that is, short-length query sequences. The level of genome coverage of the short-length query sequence set was the same as that of the full-length query sequence set.

To further explore relationships between query sequence length and classification, a series of sequence subsets were prepared. Each subset contained DNA taken from the random *S. bicolor* genomic sequences used above. Names and details of the subsets are given in Supplementary Table 2.

To examine the ability of the SRCP to estimate gene and/or repeat enrichment afforded by Cot filtration (a reduced representation sequencing technique), Cot-filtered sequences manually classified by Peterson et al. [10] (GenBank Accession Nos. AZ921847–AZ923007) were categorized by the SRCP following analysis of the unfiltered query sequences (see below).

Analysis of random genomic DNA query sequences

The basic steps in analysis of random genomic query sequences are outlined in Fig. 1. Specifics are illustrated in Fig. 2 and further detailed below.

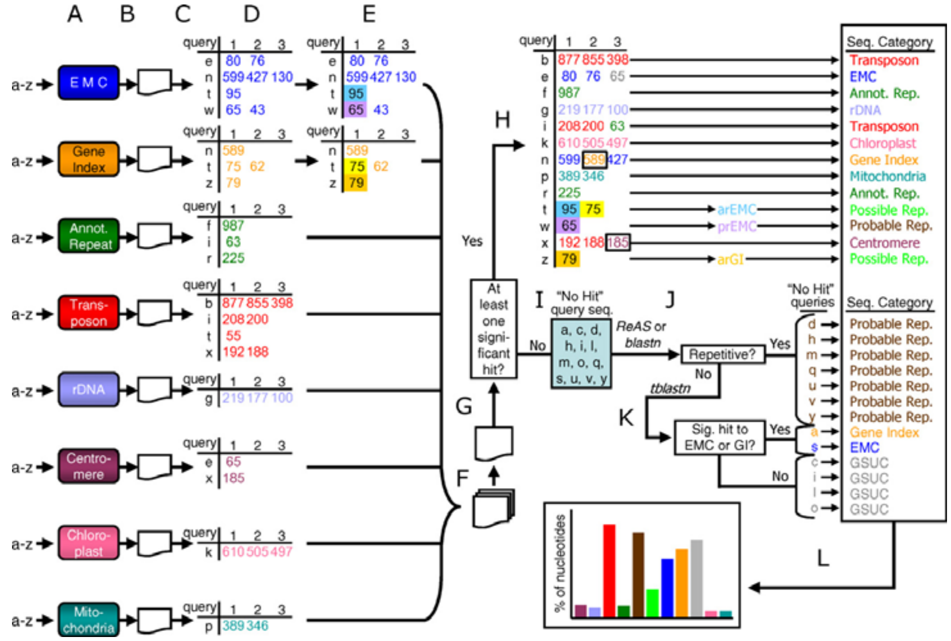


Fig. 2. Steps in categorization of random genomic query sequences. (A) A query sequence set is compared with sequences in the eight local sequence databases. In the diagram, the query sequence set is composed of 26 reads (represented by the lowercase letters a–z). BLAST (Basic Local Alignment Search Tool) parameters are set so that only the three most significant hits (if applicable) for each query sequence are recorded. (B) A Perl script removes unnecessary text and eliminates all hits with bit scores (S') < 45 from the BLAST output files. (C) A script uploads the resulting ‘‘summary’’ files to an SQL Server database. (D) In the SQL database, the BLAST results from each local sequence database are stored in their own data table. In the diagram, each BLAST results table lists only the names of query sequences that produce a hit to a sequence in that local sequence database (left most column) and the bit scores of each query sequence's (up to three) most significant hits. In reality, the data tables contain highly detailed information including each hit's accession number(s), annotation, and alignment information with the query sequence. (E) As a means of detecting repetitive sequences in the EMC (EST/mRNA/cDNA) and Gene Index (GI) local sequence databases, an UPDATE query analyzes the EMC and Gene Index query BLAST data tables to see if multiple query sequences are recognizing the same local database entry, an indication that the entry and the query sequences may represent repetitive elements. On the basis of this analysis, some query sequences are marked as ‘‘Ambiguous Repetitive’’ (ar) or ‘‘Probable Repetitive’’ (pr). In the diagram, arEMCs, prEMCs, arGIs, and prGIs are represented by light blue, violet, gold, and yellow cells, respectively. (F) A UNION query integrates the information from all eight BLAST data tables. (G) A nested SELECT query eliminates hits with bit scores <60 and selects the best three hits from all of the data tables for each query sequence. Each query sequence with at least one $S' \geqslant 60$ hit is included in the query result set. (H) A decision tree assigns each query sequence in the query result set to a descriptive sequence category based on the (up to three) best hits for that sequence. The decision-making process is relatively complex. Rectangles mark instances in which the decision tree assigns a query sequence to a sequence category that differs from the name of the local sequence database to which that sequence shows its most significant hit. For simplicity, all query sequences that are ‘‘called’’ arEMCs or arGIs are assigned to the ‘‘Possible Repeat’’ category, whereas all those ‘‘called’’ as prEMCs or prGIs are assigned to the ‘‘Probable Repeat’’ category. (I) Query sequences that produce no significant hits to any of the local sequence databases are assigned to the temporary ‘‘No Hit’’ group. (J) Depending on the level of genome coverage, either *ReAS* [7] or *blastn* is used to compare ‘‘No Hit’’ sequences to each other. Those query sequences marked as repetitive by *ReAs* or exhibiting significant homology ($S' \geqslant 60$) to a number of other ‘‘No Hit’’ query sequences in excess of a mathematically defined threshold are placed in the ‘‘Probable Repeat’’ category. (K) Remaining ‘‘No Hit’’ query sequences are electronically ‘‘translated’’ by a Perl script into proteins representing each of the six potential reading frames. The program *tblastn* is then used to compare the translated ‘‘No Hit’’ query sequences into translated versions of the EMC and GI local sequence databases. If a translated ‘‘No Hit’’ sequence produces a significant ($S' \geqslant 60$) *tblastn* hit to the EMC and/or GI local sequence databases, it is reclassified based on the highest of its bit scores. If the highest EMC and Gene Index bit scores are equal, the ‘‘Gene Index’’ classification is selected. ‘‘No Hit’’ sequences that are not classified in step J or K are placed in the ‘‘Genome Sequences of Unknown Character’’ (GSUC) category. (L) The query sequence set is displayed in a histogram showing the percentage of base pairs found in each sequence category.

Blast

An entire query sequence set is BLASTed against each of the local databases (Fig. 2A). We set *–b* and *–v blastall* flags to 3 to collect only the top three hits for each sequence, minimizing the sizes of the resulting XML files, which, depending on the number of query sequences, may otherwise become unmanageably large.

The output XML files are processed with a Perl script that creates summary XML files. At this point, hits that do not satisfy a certain minimal bit score threshold may be filtered out using this Perl script. Summary files are then used by DTS scripts to bulk upload the data to an SQL Server database based on the corresponding XML Schema Definition files. Results from each local database BLAST comparison are stored in their own table (Figs. 2B and C).

First-round detection of repeat sequences

A common means used to assess the gene content of a batch of query sequences is comparison of the query sequences with ESTs. However, such an approach requires considerable caution as EST databases often contain numerous repetitive DNA sequences. Some of these repeats are simply organellar, rDNA, or genomic repeat sequences that were not eliminated during the mRNA isolation process. Others are the expressed regions of transposons such as retroelement genes. Because transposons are typically found in numerous copies per genome and contain only genes that promote their own propagation or movement, they are typically classified as repeats. Repeats are eventually "weeded out" of most EST databases, although it may be many years before the culling process is complete.

Recognition of the same EST database entry by multiple genomic query sequences is one means by which query sequence repetitiveness has been estimated and repeat sequence contaminants have been identified in "low-copy-sequence" databases [10,13,14]. In this regard, several SQL queries were used to identify EMC local database entries that were the top significant EMC hit for multiple query sequences. Assuming that query sequences in the EMC BLAST table represent single-copy genes, the average number of times a query sequence would represent a given gene can be predicted by dividing the number of query sequences in the EMC BLAST table by the predicted number of genes for the test organism. For example, in our analysis of the full-length sorghum query sequences, 972 query sequences exhibited their most significant hit ($S' \geqslant 60$) to the EMC local database. If sorghum has roughly 25,000 nonrepetitive gene sequences like *Arabidopsis* [15], the average expected number of hits by an EMC-recognized query sequence to any one of the hypothetical sorghum genes is (972 ÷ 25,000 =) 0.0389. The probability of multiple EMC-recognized query sequences recognizing a particular "single-copy EST" (≈gene) sequence by chance can be roughly estimated using the Poisson probability distribution function,

$$P(X) = \mu^x \div (e^{\mu} X!),$$

where $P =$ probability, $X =$ number of occurrences, and $\mu =$ is the population mean number of occurrences in a unit of space or time [16]. If $\mu = 0.0389$ (see above), the probabilities of two, three, four, and five EMC-recognized query sequences tagging the same single-copy EST by chance are 7.3×10^{-4}, 9.4×10^{-6}, 9.2×10^{-8}, and 7.1×10^{-10}, respectively.

In our implementation, the first value of X to produce a $P(X)$ less than 0.01 can be represented by the variable Y. SQL queries mark a query sequence as an "Ambiguous Repeat EMC" if its most significant hit is to an EMC that is the most significant hit of Y query sequences in the dataset. Any query sequence that has its most significant hit to an EMC that is the most significant hit for $> Y$ query sequences is classified as a "Probable Repeat EMC".

The repeat detection procedure is applied to the Gene Index local database as well with some query sequences being reclassified as "Ambiguous Repeat Gene Index" or "Probable Repeat Gene Index".

Classification of query sequences with significant local database BLAST hits

Local database BLAST results tables are combined in a UNION query. Query sequences with no significant local database hits are not included in the UNION query result set, but, rather, are given the temporary classification of "No Hit" and used to generate a corresponding FASTA file for further analysis (see below). For those query sequences with at least one significant local database hit, an SQL query (see Supplementary Materials, SQL Query) is used to determine the (up to) three best hits with bit scores $\geqslant 60$ for each query sequence from the UNION query result set (Figs. 2F–H).

A DTS script within SQL Server 2000 uses the output of the query above and runs it through a decision tree that places the results in a new table in which each query sequence with at least one hit has three sets of columns for its (up to) three best hits arranged from most significant to least significant (except in instances where two or more bit scores are equal). Generation of this combined results table allows each query sequence to be represented by a single record. Also, the classification calculations are performed only once and stored permanently in the results table precluding the need to run complex SQL SELECT queries over large data tables every time the results are fetched.

Each query sequence with at least one significant hit is classified into one of 11 different categories (see Fig. 2) using the decision tree algorithm mentioned above. The heuristics of this algorithm are presented below:

1. The TIGR Gene Index contains sequences that have been shown to code for protein (and, thus, are likely to actually represent genes), whereas there is no such

prerequisite for a sequence to be included in the EMC Local Database. Consequently, Gene Index is favored over EMC.

2. Because Gene Index and EMC local databases are likely to contain some repeat sequences, significant hits to organellar or repeat local databases are given priority over Gene Index and EMC hits.
3. If the first hit's bit score is at least 20% greater than the next two hits (if any) and the preceding heuristics are not violated, then the query sequence is classified based on the first hit's local database.
4. If the first and second hits or first and third hits are to the same local database, then the query sequence's classification is set to this local database.
5. If a query sequence is not classified in step 1, 2, 3, or 4, it is given the temporary classification of *Flag*. In the case of a *Flag* classification where the two best hits are to different repeat local databases (Ambiguous Repeat EMC, Ambiguous Repeat Gene Index, Probable Repeat EMC, Probable Repeat Gene Index, Annotated Repeats, Transposon, or Probable Repeats), the query sequence is classified by the local database to which it produces the highest bit score. The Probable Repeat local database is used only when analyzing reduced-representation sequences (see below).
6. If the classification is still *Flag* and the two best hits are to EMC and/or Gene Index, EMC is chosen if it has a higher bit score. Otherwise, Gene Index is chosen.
7. If the classification is still *Flag*, at least one of the hits is to Chloroplast, and none are to rDNA, then the classification is set to Chloroplast.
8. If the classification is still *Flag*, at least one of the hits is to rDNA, and none are to Chloroplast, then the classification is set to rDNA.
9. If the classification is still *Flag* and at least one of the hits is to Centromere with a bit score within 20% of the first hit's bit score, the query sequence is given the classification of Centromere.
10. If the classification is still *Flag* and all hits are to EMC, Gene Index, Ambiguous Repeat EMC, Ambiguous Repeat Gene Index, Probable Repeat EMC, or Probable Repeat Gene Index, the classification is set to the repetitive database with the highest bit score.
11. For simplicity, those query sequences classified as Ambiguous Repeat EMC or Ambiguous Repeat Gene Index are placed in the "Possible Repeat" category, whereas those query sequences classified as Probable Repeat EMC and Probable Repeat Gene Index are placed in the "Probable Repeat" category (see Fig. 2).

If *Flag* query sequences remain, they can be manually classified via the SRCP's Web interface or the decision tree algorithm can be modified. Although the decision tree algorithm described above resulted in automated classification of all *Flag* query sequences, other data and/or local database sets may produce unresolved flags indicating that fine tuning of the algorithm may be appropriate.

Identifying repeats in the "No Hit" query sequences

The "No Hit" query sequence group can be further analyzed to identify novel repetitive elements based on their relative iteration in the query sequence set. If the genome coverage is at least 1.58*X*, the "No Hit" query sequence group is analyzed using *ReAS* [7], an ab initio repeat-finding program that has proven especially robust in side-by-side comparisons with other database-independent repeat identification tools (our personal observations). However, the genome coverage in sample sequence-based genome characterization projects is often below the genome coverage levels necessary for most repeat analysis programs. Consequently, we developed a method to calculate which "No Hit" query sequences are probable repeats when genome coverage is below 1.58*X*. First, we determine the *k*-mer length (sequence of length *k*) that will afford one chance in a thousand that two random query sequences will share an identical sequence of length *k* for a genome of size *G*. This determination, based on Batzoglou [17], is made using the following logic:

1. There are four nucleotides in DNA; thus, the total number of potential *k*-mers is 4^k.
2. Because of the double-stranded nature of DNA, a *k*-mer and its exact complement will be considered identical by *blastn*. This means that the number of "unique *k*-mers" is $4^k/2$.
3. Hence, the probability of a given "unique *k*-mer" occurring once in a genome of size *G* is $2G/4^k$.
4. The probability of a specific "unique *k*-mer" occurring twice is $4G^2/4^{2k}$. The probability of any "unique *k*-mer" occurring twice is $2G^2/4^k$ [i.e. $(4G^2/4^{2k}) * 4^k/2$].
5. A 0.001 probability that two reads will share an identical sequence of length *k* by chance is equivalent to $1000 * 2G^2/4^k$. Hence, the length of this unique *k*-mer is $k = \text{ceiling}(\log_4 G^2 + \log_4 2000)$.

The "No Hit" query sequences are BLASTed (*blastn*) against each other with the word size parameter set equal to the *k* calculated as described above. Those query sequences that share a *k*-mer with one or more other "No Hit" query sequences are detected. We then use the Poisson distribution to determine a threshold contig depth *d* [7] that is expected at error rate 0.1% for the level of genome coverage λ as per the equation

$$p = (e^{-\lambda}\lambda^d) \div d!$$

Those query sequences that share a unique *k*-mer to $\geqslant d$ other "No Hit" query sequences (see Supplementary Table 3) are assigned to the "Probable Repeat" sequence category (Fig. 1). When genome coverage is $\leqslant 0.04X$ (and $d + 1 = 2$), the BLAST output file is parsed by a Perl script that classifies query sequences as "Probable Repeats" if they have at least one hit to another query sequence, that is, share a unique k-mer. For data sets with coverage values between $0.05X$ and $1.57X$, we use another Perl script that classifies a query sequence as "Probable Repeat" only if it has at least the minimal number of hits sharing the same k-mer. "Probable Repeat" query sequences are then placed into a consolidated BLASTable local database of the same name. The Probable Repeats local database is used when analyzing sequences that have been generated through reduced-representation sequencing (see below).

Classification of remaining "No Hit" query sequences

As shown in Fig. 2K, all remaining "No Hit" sequences are translated by the Perl script *three_frames.pl*, available from the Canadian Bioinformatics Help Desk, and compared with sequences in the EMC and Gene Index Local Databases using *tblastn* [18]. Such comparison can allow detection of potential gene orthologs that have undergone substantial divergence at the DNA level but have relatively conserved amino acid sequences. Those query sequences producing a significant *tblastn* hit ($S' \geqslant 60$) to an EMC or Gene Index entry are reclassified as described in Fig. 2. "No Hit" query sequences that do not produce a significant *tblastn* hit to EMC and/or Gene Index local databases are placed in the sequence category "Genome Sequence of Unknown Character." This part of the analysis is the most computationally expensive and may be performed using BLAT [19] and/or a computer cluster.

Output

Once classification has been completed, summary statistics are calculated. They can be viewed or saved in an Excel file via a Web interface.

Contig assembly

After classification, all query sequences are collectively analyzed using Phrap (www.phrap.org). An ACE file generated by Phrap is then parsed by Perl scripts that generate two summary XML files::one of the summary XML files contains data grouped by sequences and the other has data grouped by contigs. Both of the XML files include padded sequence data. These data are then bulk uploaded to the SQL Server database. A graphical interface has been designed to permit rapid visualization of contigs and the classification assigned to each query sequence within a contig. Desired outcomes of contig analysis include assembly of genes, characterization of repeat families, correction of potential erroneous classifications, and/or detection of improperly labeled/annotated GenBank/TIGR entries. With respect to error correction, visual inspection of assembly reads aided by color-coded classifications (Supplementary Fig. 1) allows rapid detection of query sequences that appear conspicuously out of place. If deemed appropriate, classifications can be changed and the source of the original classification traced back to the top three hits. It is anticipated that contigs visualized in this manner can potentially limit the *snowballing effect* of incorrect annotations and improve the quality of the local databases.

Analysis of reduced representation sequences

Analysis of reduced-representation query sequences closely follows the scheme used for genomic query sequences (Fig. 2). However, the Probable Repeats local database (see above) generated after analysis of random genomic sequences is used as a ninth local database during the initial classification. Additionally, when analyzing "No Hit" query sequences, the genome size G is replaced by the fraction of the genome in a particular reduced-representation component. For example, according to Peterson et al. [10], the sorghum genome consists of highly repetitive, moderately repetitive, and single-/low-copy components that account for roughly 0.15, 0.41, and 0.24 of the genome, respectively. As the sorghum genome is about 760 Mb [20], the highly repetitive component of sorghum would contain 114 Mb of DNA (i.e., 0.15 * 760 Mb) while moderately repetitive and single/low-copy components would account for 311.6 and 182.4 Mb, respectively. To allow for consistent analysis of all reduced representation-enriched fractions, repetitive query sequences identified in reduced-representation data sets during the "No Hit" repeat analysis are not added to the Probable Repeats local database.

Results and discussion

SRCP analysis of random genomic sorghum query sequences

Initially, two representations of the same *S. bicolor* sequence set were analyzed by the SRCP. The first representation consisted of "full-length" genomic shotgun sequence reads of a size typical of trimmed reads produced via automated Sanger sequencing (mean length = 599 bp). The second representation consisted of the original full-length reads digitally fragmented into pieces between 80 and 179 bp in length (mean length = 105 bp) to simulate short read lengths such as those produced by 454 DNA sequencing [21]. The results of these analyses are summarized in Fig. 3A. As shown, shorter query sequence lengths resulted in an increase in the broadly defined Probable Repeats and Genome Sequence of Unknown Character categories with concomitant decreases in all other classes. This suggests that shortening query sequence length to about 100 bp often disrupts features that permit placement of query sequences into more narrowly defined categories, most notably EMC, Gene Index, and Transposon.

Comparison of Cot analysis and SRCP data

Cot analysis is the study of the kinetics of DNA reassociation in solution. It can be used to learn much about

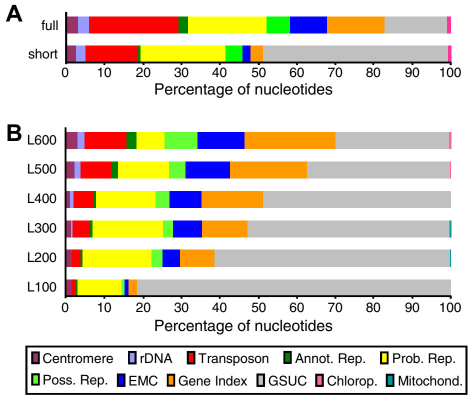


Fig. 3. SRCP-based classification of random sorghum genomic shotgun query sequences. (A) Classification of full-length query sequences (mean = 599 bp) versus short-length query sequences (mean = 105 bp). (B) Effect of query sequence length on classification. Six different query sequence lengths ranging from 100 to 600 bp were tested (see Supplementary Table 2).

the general structure of a genome, including genome size, number and size of kinetic components, amount of repetitive DNA, amount of single-/low-copy DNA, and kinetic complexities of unique and repeat components [22]. To permit comparison with Cot analysis data, percentages of Transposon, Annotated Repeat, Probable Repeat, Centromere, rDNA, and Possible Repeat categories were grouped together and deemed percentage repetitive genomic DNA. Conversely, the EMC and Gene Index categories probably represent single-/low-copy DNA and were grouped as such. The contents of the Genome Sequence of Unknown Character category may represent either low-copy and/or a combination of repetitive and low-copy sequences depending on the depth to which repeat components have been sequenced. With a sufficiently large sequencing depth or with fairly comprehensive repeat local databases, the rigorous repeat search conducted by the SRCP may afford a relatively high probability that sequences that end up in the Genome Sequence of Unknown Character category are also low-copy DNA. For this initial analysis, we conservatively assumed that 50% of the Genome Sequence of Unknown Character bases were low-copy DNA. Half the percentage of the Genome Sequence of Unknown Character category was added to the EMC and Gene Index percentages to yield a rough estimate of genomic single-/low-copy sequences. Based on SRCP analysis of full-length query sequences, the percentage of repetitive DNA in the *S. bicolor* genome is 58.2%, whereas short-length query sequence analysis provides a repeat value of 45.9%. A previous Cot analysis of sorghum [10] suggested that the genome is composed of at least 56% repetitive DNA, a value that falls within the range predicted by full- and short-length SRCP analyses. The percentage of single-/low-copy DNA as detected by SRCP analysis of full-length sorghum query sequences is 32.7%, whereas that of short-length query sequences is 29.4%. The Cot analysis suggested that single-/low-copy DNA makes up at least 24% of the sorghum genome. Considering the various biases inherent in Cot analysis and SRCP classification techniques, the similarity in repeat and low-copy sequence percentages between the two types of results is encouraging.

The effect of query sequence length on classification

The SRCP uses an ‘‘all or nothing’’ approach, assigning every base in a query sequence to a ‘‘best-fit’’ sequence category. Although this is not a perfect classification solution, dissection and annotation of the parts of each query sequence would be a tremendous undertaking. As suggested in Fig. 3A, short query sequence lengths decrease the specificity of classification. Generation of single-read query sequence lengths beyond 600–700 bp is not currently feasible due to limitations of high-throughput capillary electrophoresis, but it is likely that increasing query sequence length much beyond this size would augment the chances that a repeat and a unique sequence occur on the same query sequence.

To further explore the effect of query sequence length on classification, we prepared sequence subsets with different query sequence lengths (Supplementary Table 2) and analyzed the subsets using the SRCP. The results of this analysis are summarized in Fig. 3B. In support of the observations made in analysis of the full-length and short-length query sequences, shorter query sequence lengths limit placement of sequences into gene and repeat classes. The L600 (600-bp sequence length) data set produces the highest levels of bases in the Gene (EMC and Gene Index) and Repeat (Transposon, Annotated Repeat, Probable Repeat, Possible Repeat, Centromere, and rDNA) categories. Compared with the results of the L600 analysis, the L500 set shows similar percentages of bases classified as EMC and Gene Index, but noticeable differences in how sequences are divided among repeat classes. Interestingly, the L600 set (Fig. 3B) shows fewer bases in repeat and low-copy classes compared with the full-length query sequences, which have a mean length of 599 bp (Fig. 3A). The full-length query sequence analysis involved roughly six times as much sequence data as the L600 analysis, and indeed, this may account for the observed differences. Although it is not clear what size query sequence will produce the most accurate description of a genome (and it is likely that optimal query sequence size may differ from genome to genome), our results suggest that 500- to 600-bp fragments provide an adequate compromise between length and classification specificity, while shorter sequences result in disruption of features that permit classification.

Analysis of Cot-filtered DNA

Reduced-representation sequencing techniques are methods that can be used to preferentially isolate and sequence a desired subset of DNA sequences from a larger population of sequences [8,9]. For example, some reduced-representation sequencing techniques are used to isolate and sequence gene-rich regions found within genomic DNA. Others may enrich for repeats or molecular markers. Examples of reduced-representation sequencing techniques include EST sequencing, methylation filtration [14], and Cot filtration [10].

If one is interested in evaluating reduced-representation sequencing-based enrichment using the SRCP, it is best if the SRCP is first used to analyze random genomic DNA from the same organism. This allows establishment of a ‘‘background’’ genome composition and results in generation of a Probable Repeat local database, which can be used to help identify repeats in the reduced-representation sequencing data.

To test the quality of SRCP classification versus manual classification, we first ran sorghum genomic query sequences through the pipeline to generate a Probable Repeat local database for sorghum. Then we used the SRCP to evaluate a set of Cot-filtered highly repetitive, moderately repetitive, and single-/low-copy sequences manually classified and described by Peterson and colleagues [10]. Peterson and colleagues made no attempt was to identify repeats and/or genes in the categories comparable to our ‘‘No Hit’’ group, preventing direct comparisons of repeat and low-copy contents. Consequently, we analyzed the ‘‘No Hit’’ sequences of Peterson et al. [10] with the algorithms depicted in steps J–L in Fig. 2 and made the assumption that 50% of bases given a final classification of Genome Sequence of Unknown Character were low-copy DNA. As with the random genomic DNA, the Cot-filtered sequences were analyzed as ‘‘full-length’’ query sequences (mean ± SE length = 177.5 ± 2.8 bp) and ‘‘short-length’’ query sequences (80–179 bp). The results of the full-length SRCP, short-length SRCP, and manual classification are summarized in Fig. 4. Of note, there is very little difference in the percentages of single-/low-copy and repetitive sequences detected using the three schemes.

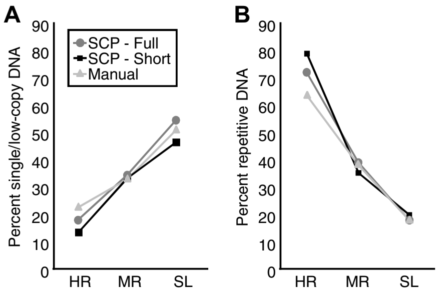


Fig. 4. Low-copy and repeat sequence contents of highly repetitive (HR), moderately repetitive (MR), and single/low-copy sorghum DNA libraries as determined by the SRCP and by manual classification. (A) The increase in low-copy DNA from HR to SL libraries as seen with the manually classified sequences is paralleled by the SRCP classification. (B) The decrease in repetitive DNA from HR to SL libraries as seen with the manually classified sequences is paralleled by the SRCP classification.

Conclusions

The SRCP is an automated means through which genomes can be characterized based on sample shotgun sequencing. To our knowledge, it is the first pipeline designed for this purpose. Moreover, as demonstrated above, it can be used to determine the efficiency of reduced-representation sequencing in a manner that is as accurate as, and certainly much faster than, manual classification. Of note, careful adaptation of the SRCP may advance comparative genomics by affording a rapid means of evaluating divergence that has occurred in ostensibly related species. Although we developed our implementation for the study of higher plant genomes, the SRCP can be easily adapted for study of any group of organisms; the principal adjustment required for use of the SRCP for other subjects is modification of the boolean text strings used in building the local databases (Supplementary Table 1). Alternatively, one can use existing sequence databases, including those developed for model organisms. The implementation of the SRCP described in this article is based on the scale and demands of our current workloads. However, the design is such that it can readily be adapted for larger-scale projects. In such cases, sequence alignment might best be performed on a cluster running a parallelized version of BLAST (at least for alignments performed against the Gene Index and EMC local databases). Techniques such as Extensible Stylesheet Language Transformations (XSLT) may further speed up processing of large XML output files. Once the pipeline is established and performs all steps correctly, it can be further automated via script scheduling and bottleneck elimination in program flow. Additionally, the SRCP is designed to be easily coupled with other scripts that allow further utilization of the sequence data. Indeed, we have begun building a pipeline that will generate consensus sequences for transposons and classify these elements into families based on their sequence structures.

Acknowledgments

This research was supported by Grant DBI-0421717 from the National Science Foundation, Grant 2006-34506-17290 from the U.S. Department of Agriculture Cooperative State Research, Education, and Extension Service, and a grant from the Mississippi Corn Promotion Board.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2007.08.008.

References

[1] W.B. Strong, R.G. Nelson, Preliminary profile of the *Cryptosporidium parvum* genome: an expressed sequence tag and genome survey sequence analysis, Mol. Biochem. Parasitol. 107 (2000) 1–32.

[2] E.F. Kirkness, V. Bafna, A.L. Halpern, S. Levy, K. Remington, D.B. Rusch, A.L. Delcher, M. Pop, W. Wang, C.M. Fraser, J.C. Venter, The dog genome: survey sequencing and comparative analysis, Science 301 (2003) 1898–1903.

[3] A. Lomsadze, V. Ter-Hovhannisyan, Y.O. Chernoff, M. Borodovsky, Gene identification in novel eukaryotic genomes by self-training algorithm, Nucleic Acids Res. 33 (2005) 6494–6506.

[4] V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters, Genome Biol. 7 (Suppl. 1) (2006) S10–S12.

[5] Z. Bao, S.R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes, Genome Res. 12 (2002) 1269–1276.

[6] A.L. Price, N.C. Jones, P.A. Pevzner, De novo identification of repeat families in large genomes, Bioinformatics 21 (Suppl. 1) (2005) i351–i358.

[7] R. Li, J. Ye, S. Li, J. Wang, Y. Han, C. Ye, J. Wang, H. Yang, J. Yu, G.K. Wong, J. Wang, ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun, PLoS Comput. Biol. 1 (2005) e43.

[8] D.G. Peterson, Reduced representation strategies and their application to plant genomes, in: K. Meksem, G. Kahl (Eds.), The Handbook of Genome Mapping: Genetic and Physical Mapping, Wiley-VCH Verlag, Weinheim, 2005, pp. 307–335.

[9] A.H. Paterson, Leafing through the genomes of our major crop plants: strategies for capturing unique information, Nat. Rev. Genet. 7 (2006) 174–184.

[10] D.G. Peterson, S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, A.H. Paterson, Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery, Genome Res. 12 (2002) 795–807.

[11] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444–2448.

[12] J.A. Bedell, M.A. Budiman, A. Nunberg, R.W. Citek, D. Robbins, J. Jones, E. Flick, T. Rholfing, J. Fries, K. Bradford, J. McMenamy, M. Smith, H. Holeman, B.A. Roe, G. Wiley, I.F. Korf, P.D. Rabinowicz, N. Lakey, W.R. McCombie, J.A. Jeddeloh, R.A. Martienssen, Sorghum genome sequencing by methylation filtration, PLoS Biol. 3 (2005) e13.

[13] T.E. Bureau, P.C. Ronald, S.R. Wessler, A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes, Proc. Natl. Acad. Sci. USA 93 (1996) 8524–8529.

[14] P.D. Rabinowicz, K. Schutz, N. Dedhia, C. Yordan, L.D. Parnell, L. Stein, W.R. McCombie, R.A. Martienssen, Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome, Nat. Genet. 23 (1999) 305–308.

[15] The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*, Nature 408 (2000) 796–815.

[16] J.H. Zar, Biostatistical Analysis, Prentice-Hall, Upper Saddle River, NJ, 1996.

[17] S. Batzoglou, Computational Genomics: Mapping, Comparison, and Annotation of Genomes, Dissertation, Massachusetts Institute of Technology, 2000, p. 21.

[18] S.F. Altschul, M.S. Boguski, W. Gish, J.C. Wootton, Issues in searching molecular sequence databases, Nat. Genet. 6 (1994) 119–129.

[19] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.

[20] K. Arumuganathan, E.D. Earle, Nuclear DNA content of some important plant species, Plant Mol. Biol. Rep. 9 (1991) 208–218.

[21] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.

[22] R.J. Britten, D.E. Graham, B.R. Neufeld, Analysis of repeating DNA sequences by reassociation, Methods Enzymol. 29 (1974) 363–405.

CHAPTER III

TRANSCRIPTOME-BASED DIFFERENTIATION OF CLOSELY RELATED MISCANTHUS LINES[1]

[1] Reprint from Chouvarine P., Cooksey A.M., McCarthy F.M., Ray D.A., Baldwin B.S., et al. (2012) Transcriptome-Based Differentiation of Closely Related Miscanthus Lines. PLoS ONE **7(1)**: e29850. doi:10.1371/journal.pone.0029850. This article is available from:
http://www.ncbi.nlm.nih.gov/pubmed/22253803

Transcriptome-Based Differentiation of Closely-Related *Miscanthus* Lines

Philippe Chouvarine[1]*, Amanda M. Cooksey[1], Fiona M. McCarthy[1,2], David A. Ray[1,3], Brian S. Baldwin[4], Shane C. Burgess[1,2]¶, Daniel G. Peterson[1,4]¶

[1] Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, Mississippi, United States of America, [2] College of Veterinary Medicine, Mississippi State University, Mississippi State, Mississippi, United States of America, [3] Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, Mississippi, United States of America, [4] Department of Plant and Soil Sciences, Mississippi State University, Mississippi State, Mississippi, United States of America

Abstract

Background: Distinguishing between individuals is critical to those conducting animal/plant breeding, food safety/quality research, diagnostic and clinical testing, and evolutionary biology studies. Classical genetic identification studies are based on marker polymorphisms, but polymorphism-based techniques are time and labor intensive and often cannot distinguish between closely related individuals. Illumina sequencing technologies provide the detailed sequence data required for rapid and efficient differentiation of related species, lines/cultivars, and individuals in a cost-effective manner. Here we describe the use of Illumina high-throughput exome sequencing, coupled with SNP mapping, as a rapid means of distinguishing between related cultivars of the lignocellulosic bioenergy crop giant miscanthus (*Miscanthus* × *giganteus*). We provide the first exome sequence database for *Miscanthus* species complete with Gene Ontology (GO) functional annotations.

Results: A SNP comparative analysis of rhizome-derived cDNA sequences was successfully utilized to distinguish three *Miscanthus* × *giganteus* cultivars from each other and from other *Miscanthus* species. Moreover, the resulting phylogenetic tree generated from SNP frequency data parallels the known breeding history of the plants examined. Some of the giant miscanthus plants exhibit considerable sequence divergence.

Conclusions: Here we describe an analysis of *Miscanthus* in which high-throughput exome sequencing was utilized to differentiate between closely related genotypes despite the current lack of a reference genome sequence. We functionally annotated the exome sequences and provide resources to support *Miscanthus* systems biology. In addition, we demonstrate the use of the commercial high-performance cloud computing to do computational GO annotation.

Citation: Chouvarine P, Cooksey AM, McCarthy FM, Ray DA, Baldwin BS, et al. (2012) Transcriptome-Based Differentiation of Closely-Related *Miscanthus* Lines. PLoS ONE 7(1): e29850. doi:10.1371/journal.pone.0029850

Editor: Randall P. Niedz, United States Department of Agriculture, United States of America

Received September 19, 2011; **Accepted** December 5, 2011; **Published** January 10, 2012

Funding: This work was supported, in part, by the U.S. Department of Agriculture (2009-34609-20222 and ARS-58-6402-7-241) and the National Science Foundation (DBI-0421717 and MCB-0841821). No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pc79@mafes.msstate.edu

¶ These authors contributed equally to this work.

Introduction

Nucleic acid-based identification techniques are used to improve agronomic species through molecular breeding and/or transgenesis. Moreover, the ability to genetically identify and distinguish between related species, cultivars/strains, and individuals is central to technology commercialization and the protection of intellectual property [1–3]. While a number of restriction site polymorphism-, random amplicon-, and repeat polymorphism-based molecular marker techniques have been developed to compare individuals and construct linkage maps [4], Illumina sequencing makes it affordable to conduct robust assays at the much higher resolution of single nucleotide polymorphisms (SNPs) [5,6]. SNP assays relying on whole genome sequence comparisons are not currently affordable for practical use in commercial settings and for agricultural patents. Moreover, the very large numbers of SNPs in the non-coding regions of genomes, which tend to be under relatively low evolutionary constraint, provide much larger datasets than needed for most mapping and identification/differentiation projects. Exome screening based on high-throughput sequencing, however, is a potential method for comparison of evolutionarily constrained sequences.

Giant miscanthus (*Miscanthus* × *giganteus*), a fast-growing perennial grass that originated in Japan [7], is a hybrid between the diploid *Miscanthus sinensis* ($2n = 2x = 38$) and the tetraploid *M. sacchariflоris* ($2n = 4x = 76$). Its seed sterility (propagation is traditionally via rhizome cuttings), non-invasive nature, efficient C4 metabolism (particularly at cold temperatures), deciduosity, low nutritional requirements, high photosynthetic output, and ability to grow on marginal lands have made it among the most promising dedicated lignocellulosic bioenergy feedstocks [8], especially in areas such as the U.S. and Europe where it has no close wild relatives [9]. Despite the potential of giant miscanthus as a bioenergy crop, very little is known about the molecular mechanisms underlying its basic biology.

Although, giant miscanthus is closely related to sugarcane and sorghum [10], the lack of dedicated functional genomics resources for these three species is a bottleneck for understanding molecular processes underlying the bioenergy qualities of these crops. This lack of molecular genetic data not only hinders strategies aimed at improving giant miscanthus, but it also makes it difficult for plant breeders to prove whether new varieties that they have discovered or developed are genetically different from existing varieties.

Recently, Swaminathan et al. [11] conducted genome survey sequencing and small RNA sequencing in giant miscanthus. Their research revealed that repetitive sequences dominate the giant miscanthus genome. Moreover, the coding regions of the giant miscanthus genome are similar to coding regions in other grasses. Additionally, most small RNAs appear to be the products of transcribed repeats.

Here we describe the use of high-throughput exome sequencing as a means of distinguishing *Miscanthus* × *giganteus* cultivars and *Miscanthus* species. The approach is applicable to technology commercialization, plant improvement, molecular genetic mapping, and phylogenetics. We constructed a first draft of the *Miscanthus* exome from transcript contigs built from cDNA reads of all seven plants utilized in this study. These transcripts were functionally annotated using the Gene Ontology (GO), and the data is publicly available via AgBase [12] (http://www.agbase.msstate.edu).

Results and Discussion

Plant Materials

Seven different plants were utilized in this study. Three of the plants were believed to represent the *Miscanthus* × *giganteus* cultivar 'Freedom'. We designated the 'Freedom' plant first provided to us as FO for 'Freedom', original; the other two 'Freedom' plants were obtained from a field and a nursery, and thus designated FF and FN, respectively. Two plants representing the *Miscanthus* × *giganteus* cultivars 'Illinois' (I) and 'Canada' (C) were also included in the study as was a plant labeled *Miscanthus floridulus* (F). Based upon its physical appearance and growth, the F plant was suspected of actually being *Miscanthus* × *giganteus*. Of note, misidentification and mislabeling of *Miscanthus* species is common [7]. In addition a diploid *Miscanthus sinensis* plant (MS) was used as an outgroup.

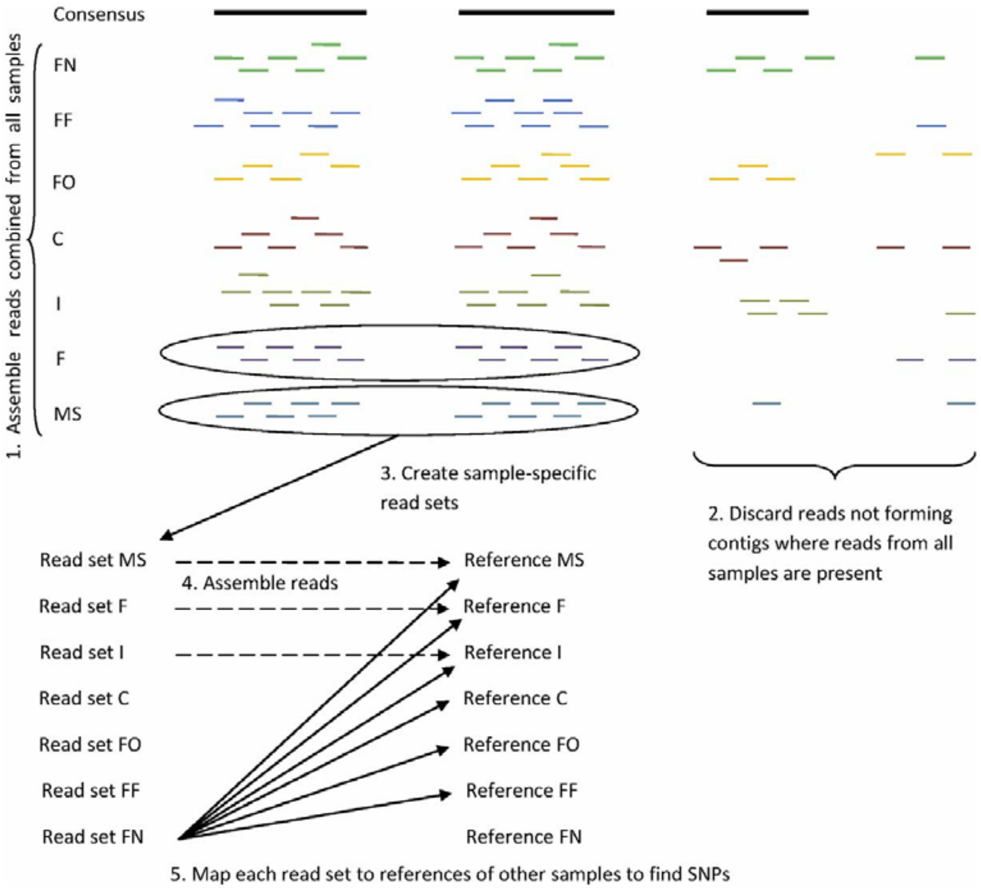

Consensus
1. Assemble reads combined from all samples
FN
FF
FO
C
I
F
MS
3. Create sample-specific read sets
2. Discard reads not forming contigs where reads from all samples are present
Read set MS
Read set F
Read set I
Read set C
Read set FO
Read set FF
Read set FN
4. Assemble reads
Reference MS
Reference F
Reference I
Reference C
Reference FO
Reference FF
Reference FN
5. Map each read set to references of other samples to find SNPs

Figure 1. Outline of procedure used to identify SNPs from miscanthus samples.
doi:10.1371/journal.pone.0029850.g001

Table 1. SNPs per aligned bp identified in comparative analysis of cDNA regions common to all samples.

	FF	FO	FN	I	C	F	MS
FF	-	0.000413390	0.000388363	0.000470852	0.000349889	0.000546697	0.000533935
FO	0.000314511	-	0.000348281	0.000434378	0.000309330	0.000486504	0.000502400
FN	0.000319526	0.000370514	-	0.000472350	0.000359891	0.000531350	0.000557107
I	0.000287344	0.000333024	0.000314453	-	0.000306604	0.000462724	0.000500130
C	0.000356861	0.000409450	0.000387226	0.000479916	-	0.000491909	0.000558566
F	0.000102675	0.000137919	0.000125332	0.000182317	0.000112822	-	0.000236819
MS	0.000187104	0.000244045	0.000230092	0.000334766	0.000212052	0.00060301	-

doi:10.1371/journal.pone.0029850.t001

Transcriptome Sequencing

A rhizome was obtained from each of the seven plants described above; rhizomes were utilized because our research was conducted during the winter, and leaf tissue was not available from all genotypes. mRNA was extracted from each rhizome, reverse-transcribed to produce cDNA, and the cDNA was sequenced using an Illumina Genome Analyzer. We chose to sequence cDNAs because coding sequences are evolutionarily constrained by the function of the proteins they encode [13]. Thus SNPs in coding sequences are likely informative of functional genetic divergence. We generated 8.9 million Illumina reads from cDNA populations obtained from rhizomes of the seven different *Miscanthus* plants described above.

Phylogenetic Analysis

To describe phylogenetic divergence among all seven samples, we used the method shown in Figure 1. We pooled the sequence reads from all samples and assembled the reads into contigs. For this analysis we needed to identify cDNA regions represented in all samples; therefore, we only considered the reads from the contigs where reads from all seven samples were present (14.64% of all reads).

The reads were then compiled into their sample-specific read sets, which ranged from 33,095 to 370,352 reads. The reads within each read set were assembled into contigs. Common regions in the consensus sequences of these sample-specific contigs were used as references for alignment of reads from each of the other read sets. The sums of lengths of the reference sequences in these read sets ranged from 1,315 to 416,163 bp. The resulting alignments for every pair of samples, e.g., alignment of the FF reads to the FO reference and alignment of the FO reads to the FF reference, allowed us to identify two sets of SNPs for each pair of samples (Table 1). In this case, a SNP is a single nucleotide variation between a reference sequence of one sample and consensus of homologous reads of another sample aligned to this reference sequence. To construct a distance matrix we used weighted SNP/bp values. As mentioned above, the number of reads in different sample-specific read sets varied significantly. Thus, SNPs identified by aligning reads from samples with a low number of reads were underrepresented (a smaller subset of them was identified). Therefore, we utilized counts of SNPs per aligned base, which included bases of every aligned read, rather than SNPs per reference base with alignment. This allowed us to add additional weight to SNPs identified by samples with a low number of reads. For each pair of alignments (e.g., FO vs. FF and FF vs. FO) we calculated the mean number of SNPs/bp (SNPs per aligned base) to construct the distance matrix (Table 2). Each of these mean values represents a normalized measure of genetic variation between the compared samples. A neighbor joining tree inferred from the data is presented in Figure 2. To determine nodal support we performed a bootstrap test as described in the Methods section. The resulting support values, calculated using a Majority Rules approach, are provided in the figure.

Our analysis was based on more than 400 million bases of cDNA sequence data from the seven plants. From this data set, we focused on cDNA regions with high quality representation in all seven samples (4.7 million bases total) for SNP analysis. Importantly, the phylogenetic tree constructed from the data exactly represents the known breeding history of the giant miscanthus plants. Of note, a previous AFLP-based approach was unable to demonstrate that sequence differences exist among giant miscanthus cultivars [7] that we differentiated here. Based upon our data, we concluded the following about the seven *Miscanthus* samples:

Table 2. Distance matrix.

	FF	FO	FN	I	C	F	MMS
FF	-						
FO	0.00036395	-					
FN	0.00035394	0.00035940	-				
I	0.00037910	0.00038370	0.00039340	-			
C	0.00035337	0.00035939	0.00037356	0.00039326	-		
F	0.00032469	0.00031221	0.00032834	0.00032252	0.00030237	-	
MS	0.00036052	0.00037322	0.00039360	0.00041745	0.00038531	0.00041991	-

doi:10.1371/journal.pone.0029850.t002

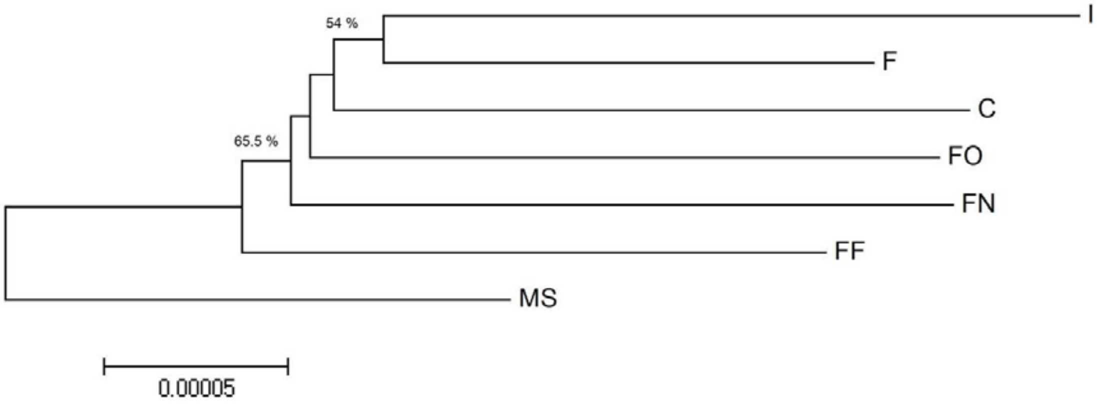


Figure 2. Phylogenetic tree inferred by SNP analysis in common regions of all seven samples. Phylogeny is inferred using weighted SNPs/bp to prepare a distance matrix and generate the neighbor-joining tree for the miscanthus samples.
doi:10.1371/journal.pone.0029850.g002

1. The 'Freedom' plants FO, FF, and FN are more similar to each other than they are to 'Illinois'. On average 'Illinois' is 70% less similar to FO, FF and FN than FO, FF and FN are to each other.
2. The mRNA sequence data from FO, FF, and FN are not sequence identical. This could reflect differences in allele/homolog/paralog expression between the ostensibly genetically identical plants. However, the level of variation is very low, compared with the inter-cultivar or interspecies *Miscanthus* comparisons.
3. 'Canada' is related to 'Illinois' and the three 'Freedom' varieties, but it is more similar to the three 'Freedom' varieties than it is to 'Illinois'. 'Canada' is most similar to FO followed by FN and then FF.
4. F (the plant labeled *M. floridulus*) is related to all other plants in the analysis, but it groups more closely with the giant miscanthus cultivars ('Canada', 'Freedom', and 'Illinois') than it does with MS. Its similarity to giant miscanthus indicates that F is most likely a mislabeled *Miscanthus* × *giganteus* plant.

Our findings strongly suggest that multiple genotypes of giant miscanthus are available. Genetic differences might account for observed differences in optimal growth region, disease resistance/susceptibility, and yield observed between giant miscanthus cultivars. Planting a single genotype over a large geographic area increases susceptibility of the crop to catastrophic loss [14,15]. Our study indicates that the three giant miscanthus cultivars studied (*Freedom*, *Illinois*, and *Canada*) are genetically different and that this diversity can be exploited in future cultivar development.

Exome Assembly

We also produced two miscanthus exome assemblies by separately assembling *Miscanthus sinensis* reads and combined reads from all varieties of *Miscanthus* × *giganteus* using Velvet [16]. Velvet contains a module called Columbus that can be used for assisted transcriptome assembly using transcript sequences of a nearby species. *Sorghum bicolor*, a species with a complete genome sequence and extensive transcript sequence resources [17], is closely related to *Miscanthus* [7], and thus we utilized *Sorghum bicolor* in assisted transcriptome assembly of the *M. sinensis* and *M.* × *giganteus*. Assisted assemblies afforded a significant improvement over non-assisted assemblies as shown in Figure 3. The four graphs represent the effects of varying *k*-mer size on various characteristics of assemblies. For genomic sequence data, the optimal assembly in Velvet is achieved by varying the *k*-mer size to find the maximum N50 and the smallest number of long contigs, while using the expected coverage threshold to minimize misassemblies. This approach is not applicable for transcript assemblies where the number of contigs should ideally be equal to the number of transcripts. For transcript assemblies ideal contig lengths should correspond to actual cDNA lengths and, due to differential gene expression, expected coverage cannot be used. For transcript assemblies, it is more applicable to maximize the contig lengths of longer contigs in the assembly by varying the *k*-mer size. The shorter contig lengths resulting from shorter than optimal *k*-mer length correspond to presence of misassembled transcript fragments. The shorter contig lengths resulting from longer than optimal *k*-mer length correspond to under-assembled contigs due to wasted coverage (unused reads with insufficient overlaps). Velvet outputs only the length of the longest contig (Figure 3, B). However, as shown in this graph, the longest contig in the assisted assemblies of *Miscanthus* × *giganteus* was not affected by varying *k*. Therefore, we calculated the average length of top 100 longest contigs for every assembly (Figure 3, D). We selected the optimal assemblies by finding a peak in this metric – $k = 37$ for the *Miscanthus* × *giganteus* assembly and $k = 23$ for the *Miscanthus sinensis* assembly. To validate this method for selection of optimal transcript assemblies, we assembled *Arabidopsis thaliana* transcripts using Illumina RNA-seq reads from NCBI Short Read Archive (ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/litesra/SRR/SRR018/SRR018346/SRR018346.lite.sra). The reads were assembled using exactly the same assisted assembly pipeline that was applied for the *Miscanthus* transcript assemblies. To estimate quality of each assembly generated by varying the *k*-mer size, we aligned the resulting transcripts to the standard *Arabidopsis thaliana* transcript assemblies downloaded from (ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Arabidopsis_thaliana/) and calculated the number of bases in the regions where our transcript contig sequences aligned without overlapping each other to the standard transcript sequences with 100% identity. The results are shown in Table 3. As we expected, the maximum of the quality metric described above occurred at the

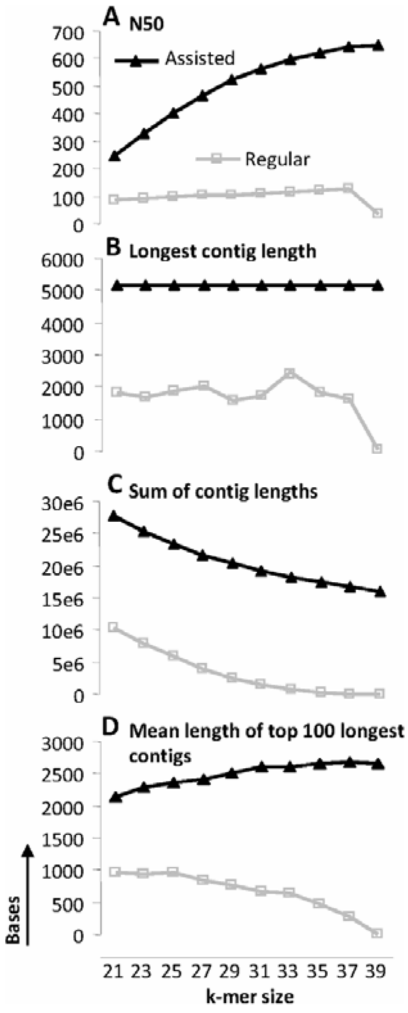


Figure 3. Impact of *k*-mer size on characteristics of *Miscanthus × giganteus* exome assembly in Velvet. Assisted assemblies were assisted with *Sorghum bicolor* transcript references. (A) N50 vs. *k*-mer size. (B) Longest contig length vs. *k*-mer size. (C) Sum of contig lengths, Mb vs. *k*-mer size. (D) Average length of the top 100 longest contigs vs. *k*-mer size.
doi:10.1371/journal.pone.0029850.g003

same *k*-mer size ($k = 19$) as the maximum of the average length of the top 100 longest contigs.

The *Miscanthus* transcript contigs identified using Velvet were processed with the *de novo* transcriptome assembler Oases (http://www.ebi.ac.uk/~zerbino/oases/). This analysis identified 29,795 *Miscanthus × giganteus* transcripts and 14,066 *Miscanthus sinensis* transcripts and generated splicing annotation for these transcripts.

Functional Annotation and Analysis

We did functional annotation of the *Miscanthus* mRNAs using GO. Since these sequences are novel, there is no direct experimental evidence for their function and GO annotation must rely on sequence analysis. The most common type of GO annotation derived from sequence analysis is annotations based on functional motif and domain analysis using InterProScan [18]. Although widely used, InterProScan requires considerable computational power and thus is typically run on clusters. However, a recent trend in bioinformatics is the use of cloud computing for analysis, [19,20] so we tested the use of the publicly available Amazon EC2 cloud to do functional annotation. This approach provided 58,392 GO annotations for 14,098 miscanthus transcripts, 24,874 transcripts were provisionally GO annotated as "ND", (i.e., "No Data"), and the remaining 4,881 transcripts could not be annotated using this procedure (e.g. sequence too short to provide reliable results). When transcripts are grouped into gene models, 32% of *Miscanthus* gene models were annotated with non-"ND" GO terms, indicating a predicted function, and 89% of *Miscanthus* gene models were annotated counting GO terms with the "ND" evidence – these will have to await experimental characterization of function. In comparison, 58% of sorghum genes have GO annotation (based on the current GO Consortium release). Since sorghum gene products are mostly annotated using the same method as we used for *Miscanthus*, we can conclude that our transcript assemblies afforded functional annotation of a comparable percentage of gene products to that of another mostly computationally annotated plant species. Using InterProScan on the Amazon EC2 cloud resulted in the average speed of 3 h 9 min per 1,000 nucleotide sequences (with the average sequence length of 570 bp) at a cost of $21.39 per 1,000 nucleotide sequences. However, mappings from InterPro functional domains to GO are revised on a monthly basis and corresponding GO annotations also need to be updated and this will add to the cost of GO annotation.

We are also providing manually derived GO annotation by transferring annotations from closely related sequences (based on sequence alignments) that have experimentally derived GO annotations [12]. This approach identified 57 GO annotations for eight transcripts. Manual biocuration of plant species within the GO Consortium has focused on the model plant *Arabidopsis thaliana* [21] and, more recently on cereals such as rice and maize [22]. Notably, although *Sorghum bicolor* is closely related to miscanthus, there is currently no experimentally derived GO annotation available for sorghum gene products, so this species was not considered during our manual GO annotation process. This example emphasizes the importance of funded efforts to provide experimentally derived functional annotation for a diverse range of key genes from economically important species.

We compared our functional annotations to those from the closely related *Sorghum bicolor*. The proportion of *Miscanthus* gene products with GO annotation is generally similar to that of *Sorghum bicolor* (Figure 4), indicating that our transcripts are representative of a comprehensive miscanthus model transcriptome. Interestingly, the proportion of miscanthus transcripts annotated to nucleus, plastid and ribosome was twice that of sorghum, while the proportion of miscanthus transcripts annotated to protein modification and transcription was half of that found in sorghum. While caution should be used in interpreting functional annotations from two different and incompletely annotated species, our result is not unexpected in the context of rhizome tissue used in this study. Since rhizomes grow underground, it is expected that chloroplasts would be underrepresented. Moreover, while rhizomes can be very active tissues, the samples used were taken from prolonged cold storage, which may have inhibited transcription and translation (protein modification) in general.

Overall, the total number of GO annotations for *M. sinensis* and *M. × giganteus* is proportional to the number of identified transcripts for these two organisms. Similarly, the larger number

Table 3. Transcript assembly metrics evaluation using *Arabidopsis thaliana* assemblies.

k	Average length of the top 100 longest contigs	Length of the longest contig	N50	Number of megablast hits with 100% identify to the standard transcript sequences produced by the contig sequences	Number of bases in the regions where our transcript contig sequences aligned without overlapping each other to the standard transcript sequences with 100% identity
15	1261	1957	8	661	8571
17	1482	2365	110	73600	1789362
19	**2028**	4616	223	92409	**2189814**
21	1886	4182	165	73506	2124487
23	1732	5050	235	47372	2040209
25	1662	5048	300	31027	1821088
27	1590	5046	346	20384	1493454
29	1457	5044	379	13093	1102776
31	1382	5042	416	7656	750977
33	1253	4260	474	3679	427093
35	1005	4250	510	1362	120707

doi:10.1371/journal.pone.0029850.t003

of sorghum annotations reflects the larger number of known sorghum gene products with GO annotation.

Data

The transcript assemblies, splice annotations, and functional annotations of *Miscanthus* × *giganteus* and *Miscanthus sinensis* are located at http://www.agbase.msstate.edu/cgi-bin/information/Miscanthus.pl. The Illumina reads used in this project can be downloaded from NCBI Short Read Archive using the accession SRA025019.

Methods

Transcriptome Sequencing

Rhizomes were obtained from plants growing in greenhouses or agricultural fields. Individual dormant rhizomes were collected from each of the seven *Miscanthus* clones. Rhizomes were incubated at room temperature on moist paper on a lab bench for 3 days. Small (100 mg) pieces were taken from each rhizome and ground in liquid nitrogen. These pulverized samples were then re-suspended in 1 ml Trizol reagent (Invitrogen) and transferred to ND Pulse tubes (Pressure Biosciences). Samples were processed in a Barocycler (Pressure Biosciences) for 20 cycles of 20 seconds at 35 kpsi followed by 5 seconds at atmospheric pressure. The resulting lysates were passed through QIAshredder columns (Qiagen) according to the manufacturer's protocol. Lysates were phase-separated using the Trizol protocol (Invitrogen). Following addition of isopropanol, RNA was collected on an RNeasy column (Qiagen). Samples were treated with on-column DNase I and washed as per the RNeasy protocol (Qiagen). Each sample was eluted in 30 µl of RNase-free water. Sample quantity and quality were evaluated spectrophotometrically using a Nanodrop (Thermo) and by capillary electrophoresis using a Bioanalyzer (Agilent).

Library Construction

Starting with 10 µg total RNA, library construction was done using the Illumina mRNA-seq sample prep kit. Total mRNA was sampled using polyA beads, chemically fragmented and randomly primed for reverse transcription and second-strand synthesis. The resulting cDNA was end-repaired and an adenosine residue was added to produce single-A overhangs. Illumina paired-end sequence adaptors were ligated to the cDNA fragments. Fragments with lengths of approximately 200 bp were sampled from a 2% w/v agarose gel and amplified by PCR (18 cycles) according to the Illumina protocol. A capillary electrophoresis-based Agilent Bioanalyzer was used to quantify and confirm the fragment size distribution of each library. One microliter of each 10 nM mRNA-seq library sample was diluted 10 fold and denatured. For each denatured library, 6 µl of the 1 nM content was diluted in hybridization buffer to 6 pM for clustering (Illumina Standard Cluster Generation Kit v2) according to the manufacturer's protocol. Single read sequencing (40 bp) of the clustered flow cell was done using Illumina's SBS chemistry (Illumina Sequencing Kits v3) and SCS data analysis pipeline v2.4. Flow-cell image analysis and cluster intensity calculations were carried out by Illumina Real Time Analysis (RTA v1.4.15.0) software. Subsequent base-calling was performed using the Illumina GA Pipeline v1.5.1 software.

Phylogenetic Analysis

To analyze phylogenetic relatedness, we identified SNPs that occur in the regions common to all seven samples. To identify the common regions, Illumina reads from all seven samples were combined and assembled with Velvet. Because SNP identification requires high quality assembly, these Illumina reads were pre-processed prior to assembly. Specifically, we noticed 61% of reads had a single N in the last position; these Ns were removed. Any remaining reads containing Ns were removed. We also set the *-max_gap_count* parameter (the maximum number of gap bases allowed for simplification of two aligned sequences, default: 3) in Velvet to 1, to further improve the assembly quality. Contigs containing at least one read from all seven samples were broken down into sample-specific read sets. Each read set was assembled into a group of sample-specific contigs whose consensus sequences were saved in a reference FASTA file. Each group of sample-specific reads was aligned against each of the other six groups of sample-specific reference sequences using MAQ [23]. All samples except for *Miscanthus sinensis* were from triploid organisms. To account for this we used the *-N 3* option with the *maq assemble* command when aligning reads from such organisms. SNPs were identified using MAQ's *cns2snp* and *SNPfilter* utilities with default parameters. SNP counts were used to calculate the mean of weighted SNPs/bp values for each pair of samples allowing construction of a distance matrix (Table 2). This distance matrix was then analyzed

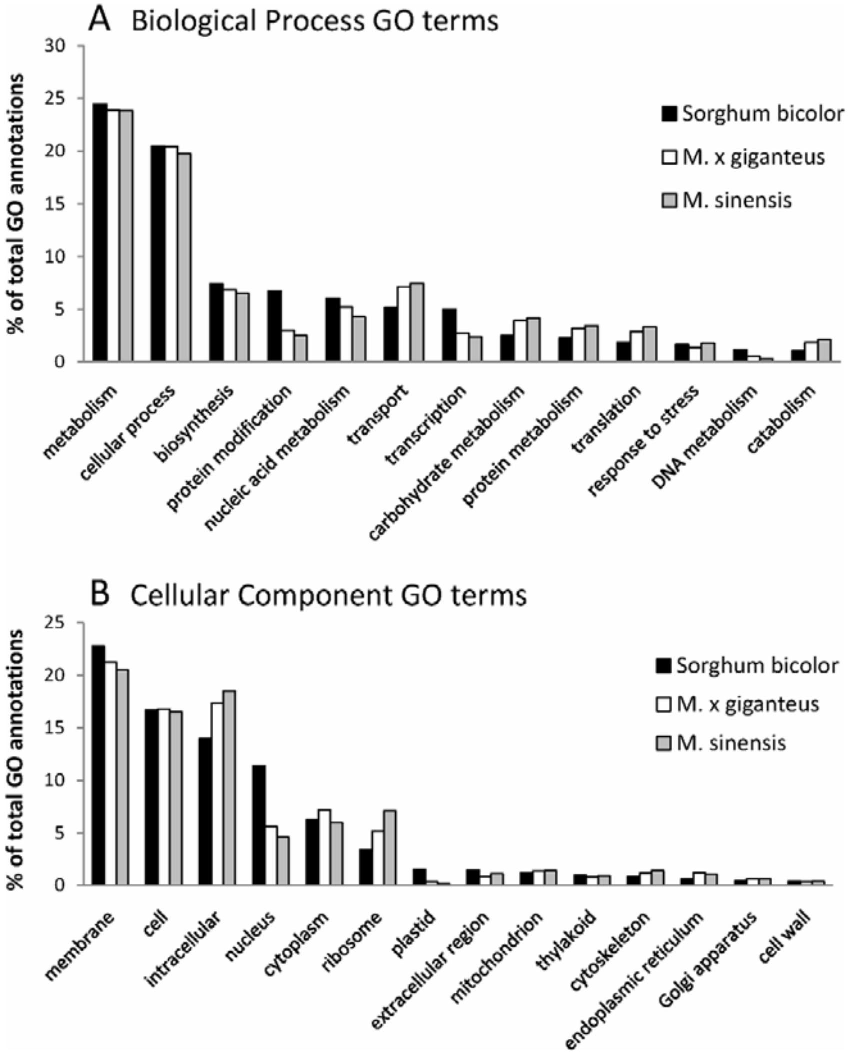


Figure 4. Distribution of GO annotation for miscanthus sequences compared to *Sorghum bicolor*. Sorghum GO annotation was downloaded from AgBase (October 2010) and the Plant GO Slim used to group and compare GO annotations from miscanthus and *Sorghum bicolor*, a closely related species. (A) Biological process GO terms. (B) Cellular component GO terms.
doi:10.1371/journal.pone.0029850.g004

using MEGA 4 [24] to generate the neighbor-joining tree shown in Figure 2. Node support was inferred using a bootstrap test adopted for our method. We created 200 bootstrapped datasets for all 42 alignments that we had, followed by calculation of the mean values of SNPs per aligned base to create 200 distance matrices. These 200 replicates were submitted to the 'neighbor' executable of the PHYLIP 3.67 package. The resulting trees were then submitted to 'consense' to calculate support values.

Exome Assembly and Functional Analysis

We used Bowtie [25] to create alignments (SAM files) to *Sorghum bicolor* transcripts. The transcripts were downloaded from the Gene Index Project (ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Sorghum_bicolor/). The reference sequences, SAM files and unmapped reads were used for cDNA contig assembly in Velvet. We used default parameters without setting coverage cutoff or expected coverage. This was done because expected coverage cannot be assessed for gene expression data. Transcripts were identified by processing the resulting contigs in Oases using default parameters.

The identified transcript sequences were functionally annotated to the GO [26] using standard, GO Consortium compliant biocuration techniques [27]. Since these sequences were not associated with any published functional literature they were GO annotated by manual inspection of BLAST alignments to GO-annotated plant genes using the *GOanna* tool [12] and functional motifs and domains were mapped to the GO using InterProScan. InterproScan IDs were then mapped to GO:IDs and the information formatted as a standard gene association file. We compared these results against GO annotation provided for *Sorghum bicolor* obtained from AgBase (October 2010), as both sorghum and *Miscanthus* have only computationally predicted GO annotations. For each species, GO annotations were summarized into major categories using GOSlimViewer (http://agbase.msstate.edu/cgi-bin/tools/goslimviewer_select.pl) with the Plant GOSlim set. GO annotations were quality checked to meet GO Consortium standards and publicly released via the AgBase database.

Amazon EC2 Cloud Computing

While sequence alignment using MAQ and sequence assembly using Velvet are routinely done using local servers, the InterProScan analysis is extremely CPU-intensive and consequently the program is typically run on a computer cluster. We chose to create an on-demand cluster using 10 high-CPU instances from the Amazon EC2 cloud (http://aws.amazon.com/ec2). InterProScan was installed on an attachable Elastic Block Storage partition. The cluster was started from an instance with the installed StarCluster software (http://web.mit.edu/stardev/cluster/). StarCluster allows specifying an attachable partition available to all cluster nodes via Network File System. We used this feature to make the Elastic Block Storage partition with InterProScan accessible from all cluster nodes. StarCluster also comes with the pre-installed SGE (Sun Grid Engine) queuing system supported by InterProScan. To avoid problems with InterProScan/SGE hanging when processing large files with thousands of nucleotide sequences, we split files into smaller files with up to 1,000 nucleotide sequences, setting the chunk size parameter in InterProScan to 60 and setting the *finished_jobs* parameter in SGE to 20,000. (Increasing the chunk size and the *finished_jobs* parameter allows processing files with longer sequences or a greater number of sequences, but this can decrease the processing speed). For our dataset, this setup resulted in the average speed of 3 h 9 min per 1,000 nucleotide sequences (with the average sequence length of 570 bp) at the cost of $21.39 per 1,000 nucleotide sequences.

Author Contributions

Conceived and designed the experiments: SCB DGP BSB PC DAR FMM. Performed the experiments: AMC PC. Analyzed the data: PC DAR FMM DGP. Contributed reagents/materials/analysis tools: BSB. Wrote the paper: PC AMC FMM DAR BSB SCB DGP.

References

1. Kunihisa M, Fukino N, Matsumoto S (2005) CAPS markers improved by cluster specific amplification for identification of octoploid strawberry (*Fragaria* × *ananassa* Duch.) cultivars, and their disomic inheritance. Theor Appl Genet 110: 1410–1418.
2. Jung J, Park S, Liu W, Kang B (2010) Discovery of single nucleotide polymorphism in *Capsicum* and SNP markers for cultivar identification. Euphytica 175: 91–107.
3. Castro P, Millan T, Gil J, Merida J, Garcia M, et al. (2011) Identification of chickpea cultivars by microsatellite markers. Journal of Agricultural Science 149: 451–460.
4. Semagn K, Bjornstad A, Skinnes H, Maroy AG, Tarkegne Y, et al. (2006) Distribution of DArT, AFLP, and SSR markers in a genetic linkage map of a doubled-haploid hexaploid wheat population. Genome 49: 545–555.
5. Ganal MW, Altmann T, Roder MS (2009) SNP identification in crop plants. Curr Opin Plant Biol 12: 211–217.
6. Rounsley SD, Last RL (2010) Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. Plant J 61: 922–927.
7. Hodkinson TR, Chase MW, Lledo MD, Salamin N, Renvoize SA (2002) Phylogenetics of *Miscanthus*, *Saccharum* and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnLintron and trnL-F intergenic spacers. J Plant Res 115: 381–392.
8. Pyter R, Heaton E, Dohleman F, Voigt T, Long S (2009) Agronomic experiences with *Miscanthus* × *giganteus* in Illinois, USA. Methods Mol Biol 581: 41–52.
9. Heaton EA, Dohleman FG, Long SP (2008) Meeting US biofuel goals with less land: the potential of *Miscanthus*. Global Change Biology 14: 2000–2014.
10. Calvino M, Bruggmann R, Messing J (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. BMC Genomics 12: 356.
11. Swaminathan K, Alabady MS, Varala K, De Paoli E, Ho I, et al. (2010) Genomic and small RNA sequencing of *Miscanthus* × *giganteus* shows the utility of sorghum as a reference genome sequence for Andropogoneae grasses. Genome Biol 11: R12.
12. McCarthy FM, Gresham CR, Buza TJ, Chouvarine P, Pillai LR, et al. (2011) AgBase: supporting functional modeling in agricultural organisms. Nucleic Acids Res 39(Database Issue): D497–D506.
13. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl: 228–237.
14. Oppong-Konadu E, Adu-Dapaah H (1998) The role of genetic diversity in sustainable agriculture. Ghana J Agric Sci 31: 231–240.
15. Cox CM, Garrett KA, Bowden RL, Fritz AK, Dendy SP, et al. (2004) Cultivar mixtures for the simultaneous management of multiple diseases: tan spot and leaf rust of wheat. Phytopathology 94: 961–969.
16. Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18: 821–829.
17. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556.
18. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, et al. (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33: W116–120.
19. Kudtarkar P, Deluca TF, Fusaro VA, Tonellato PJ, Wall DP (2010) Cost-effective cloud computing: a case study using the comparative genomics tool, roundup. Evol Bioinform Online 6: 197–203.
20. Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11: 207.
21. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. (2008) The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36: D1009–1014.
22. Jaiswal P, Ware D, Ni J, Chang K, Zhao W, et al. (2002) Gramene: development and integration of trait and gene ontologies for rice. Comp Funct Genomics 3: 132–136.
23. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851–1858.
24. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
25. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
26. Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 38: D331–D335.
27. Hill DP, Smith B, McAndrews-Hill MS, Blake JA (2008) Gene Ontology annotations: what they mean and where they come from. BMC Bioinformatics 9 Suppl 5: S2.

CHAPTER IV

MANUAL BIOCURATION TO SUPPORT STANDARDIZED CHICKEN GENE NOMENCLATURE AT CGNC

Abstract

Chicken is the *de facto* model bird occupying a key evolutionary niche. However, comparative biology, both within avian species and within amniotes is hampered due to the difficulty of recognizing orthologs and functional equivalents. Standardized gene nomenclature is therefore necessary to facilitate communication between scientists. The international Chicken Gene Nomenclature Consortium (CGNC) provides standardized gene nomenclature for chicken genes. CGNC members initially created a core set of human-chicken orthologs with consistent gene nomenclature as the initial chicken gene nomenclature set. We now report on the development of an interface that allows biocurators and community experts to assign gene nomenclature for chicken and a manual biocuration effort to provide nomenclature for chicken genes without a clear human:chicken 1:1 ortholog. Our current biocuration focus is: (1) manually verifying assigned orthologs; (2) working with domain experts to provide standardized nomenclature for the chicken MHC genes; and (3) assigning nomenclature for genes expressed in hen eggs (that are likely to be bird-specific). We combine manual

biocuration with structural and functional annotation of these genes and gene products. We strongly encourage researchers with domain knowledge to participate in this nomenclature effort. The CGNC website is linked via BirdBase and AgBase or can be accessed directly at http://www.agnc.msstate.edu/.

Introduction

Chicken (*Gallus gallus*) occupies a unique evolutionary niche in vertebrate analyses and is one of the few animals important in both the medicine and agriculture. As the first bird species to have its genome sequenced [1], it is also the best annotated bird genome and serves as the *de facto* model organism for all current and future avian sequencing and annotation projects. Large scale genome sequencing projects such as the Genome 10K Project [2] are already sequencing multiple bird genomes, expanding the number of sequenced avian genomes from three to more than 50. Moreover, advances in sequencing technologies mean that additional, individual bird genome projects are also underway [3,4,5]. As more sequence is obtained from avian species, the need for developing reference genome resources for chicken intensifies. While each bird genome sequence will inform and improve the others, problems caused by propagating poor gene nomenclature will only increase. Lack of standardized gene nomenclature hinders researchers from exploiting the full potential of avian comparative and functional genomic studies.

Although the standardized chicken gene nomenclature was first proposed in 1995 [6], it was not until 2009 that the Chicken Gene Nomenclature Committee (CGNC) formed to provide an international and coordinated effort to provide standardized nomenclature for chicken genes [7]. This initial work focused on developing clear guidelines for assigning chicken gene names and standardizing chicken gene nomenclature with human gene nomenclature where a clear 1:1 ortholog exists.

Here we report an online CGNC resource (http://www.agnc.msstate.edu) that provides the most up-to-date and curated set of chicken gene nomenclature, along with HGNC Comparison of Orthology Predictions (HCOP) [8] verified human orthologs and a detailed gene report containing nomenclature and accession links. Each gene report includes CGNC data, links to external resources, HCOP ortholog data, and maps of neighboring genes for the chicken gene and its human ortholog(s). This database also has a registered user login (available upon request) so that biocurators and community experts can add gene nomenclature information. The website includes guidelines for assigning chicken gene nomenclature, information about ongoing manual biocuration projects, the ability to download gene nomenclature information and a contact address for CGNC biocurators. Interested researchers can help assign nomenclature by registering as CGNC biocurators. The annotation provided by the external experts will be checked for consistency with current guidelines by CGNC biocurators and added to the current CGNC dataset. We also discuss our ongoing projects for manual biocuration of chicken MHC and egg genes and how these projects are informing the development of nomenclature guidelines.

CGNC Database

Implementation and Updates

At the core of the database underlying the CGNC web interface is the dataset of chicken gene nomenclature based on transferring human gene nomenclature to chicken genes in instances where a 1:1 ortholog could be identified [7]; the genes named in this way are classified as "automatic" in the CGNC download statistics and work is ongoing to manually verify these names and collect possible synonyms. The CGNC database brings in NCBI Entrez Gene information (QTL data in Entrez Gene is disregarded). The initial gene nomenclature data stored in the CGNC database comes from NCBI Gallus_gallus.gene_info (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Non-mammalian_vertebrates/) and Ensembl Biomart (http://www.ensembl.org/info/data/biomart.html). The data collected from these sources includes the following fields: CGNC ID, Entrez gene ID, Entrez gene version, Ensembl ID, Ensembl version, gene symbol, gene name, gene synonym, HGNC symbol, human ortholog, HGNC ortholog, curation status, biotype ID, and various tracking fields. Every two months this dataset is automatically updated using a Perl script. This script performs the following steps:

1. It downloads the current Gallus_gallus.gene_info file and adds new non-QTL gene records to our dataset. The gene records selected for insertion have Entrez Gene IDs and gene symbols that are not present in our dataset.
2. Gene names, gene symbols, and gene synonyms in the CGNC gene records without manual annotation are updated using the matching records from NCBI

Gallus_gallus.gene_info if such NCBI records have no CGNC cross-reference (CGNC is not present in the dbXrefs field). This step propagates any non-CGNC initiated nomenclature updates to the CGNC dataset.

3. Automatic CGNC records with obsolete Entrez gene IDs or obsolete Ensembl gene IDs are removed. An error code is added to the manually curated CGNC records with obsolete Entrez gene IDs or obsolete Ensembl gene IDs. Such records will be manually reviewed by our biocurators and most likely deleted, while providing a chance to transfer manual annotation to other records.
4. The current human:chicken ortholog data is obtained from the HCOP [8] as described in the next section. The HGNC nomenclature is transferred to the automatic CGNC records with 1:1 human:chicken orthologs and if the old gene symbol is locus-based (LOC*) then it is moved to gene synonyms. Non-locus-based symbols and names in the automatic CGNC records with 1:1 orthology are overwritten during the nomenclature transfer, because they are not asserted by CGNC and should not be considered established names and symbols as opposed to our manual records. Error flags are added for manual CGNC genes if their human orthology is no longer 1:1 or if now there is a 1:1 human:chicken ortholog with a different gene name or symbol. Error flags are also added to the gene records whose gene symbol became duplicate due to this nomenclature transfer process.
5. Using publically available and locally maintained mappings of Entrez-Ensembl IDs (where local mapping overrides public mapping), Ensembl gene IDs in CGNC records are updated based on the corresponding Entrez gene IDs.

6. An error flag is added to the records missing a symbol and/or name.
7. Error flags are added to records with duplicate symbols or gene IDs.
8. Error flags are added to records in which one NCBI gene maps to more than one Ensembl genes and vice versa. These records will be manually separated by biocurators via adding a NOT-mapping to the locally maintained Entrez-Ensembl ID mapping table.
9. Error flags are added to records with gene symbols starting with LOC* or KIAA* to prioritize them for review by biocurators.

HCOP Orthology Resources

Human:chicken ortholog data is obtained from the HCOP [8]. It is acquired using http://www.genenames.org/cgi-bin/hcop.pl. We retain the following fields from the downloaded dataset: chicken database:ID pairs (e.g., Ensembl=ENSGALG00000004248, Evola(H-InvDB)=HIT000251740, Homologene=55548, Inparanoid=ENSGALP00000006747, OMA=67847, OPTIC=110514, OrthoDB=EOG4H46M4, Treefam=ENSGALG00000004248), chicken chromosome, chicken Entrez Gene ID, Genbank accession, chicken gene name, chicken Genbank or UniProt protein accession, chicken Genbank RNA/mRNA accession, chicken gene symbol, HGNC ID, human assertion IDs, human chromosome, human Entrez Gene ID, human Genbank accession, HGNC assigned gene name, human Genbank and UniProt protein accession, human Genbank RNA/mRNA accession, HGNC assigned gene symbol, CGNC ID, and a list of databases that provided support for orthology prediction. The data fields displayed in the HCOP table of the web interface (Figure 4.1) are: chicken

Entrez Gene ID, chicken gene name, chicken gene symbol, CGNC ID, human Entrez Gene ID, human gene name, human gene symbol, HGNC ID, a list of databases that provided support for orthology prediction, and orthology type (1:1, 1:n, n:1, or n:n). Since the chicken nomenclature data in the HCOP table are taken directly from the HCOP dataset they may vary from the associated recently modified CGNC nomenclature displayed in the main CGNC table. In this case, the HCOP chicken nomenclature only represents a historic record that will be updated as our updates propagate to HCOP and the HCOP dataset is reloaded during the next CGNC update. The downloaded HCOP dataset provides a single record for each orthology relationship without identifying the orthology type (1:1, 1:n, n:1, or n:n). This identification is performed in our database via a series of queries counting the number of orthologs in human for every chicken gene and the number of orthologs in chicken for every human gene. The identified ortholog types are stored in the database and displayed in the HCOP table of the web interface. By displaying human:chicken orthology data in the CGNC web interface we enable biocurators to quickly identify any human orthologs for a particular chicken gene and assign nomenclature accordingly. The HCOP data are updated every two months by reloading the entire table and recalculating the ortholog types.

BirdBase Resources

Every CGNC ID is mapped to the corresponding BirdBase ID (http://birdbase.arizona.edu/birdbase/) and GEISHA (Gallus Expression In Situ Hybridization Analysis) IDs (http://geisha.arizona.edu/geisha/). The IDs are used to link to these resources and mapping tables for these links are updated every two months to reflect any changes.

CGNC Website

Searching the Web Interface

The front page of the CGNC web site (http://www.agnc.msstate.edu/) features two forms for searching CGNC and HCOP datasets. Users may do simple text searches by gene name, gene symbol or by gene name OR synonym (gene name/synonym). They may also search by specifying a public database accession from BirdBase, Entrez Gene, CGNC, or Ensembl. The third type of search is the “Human Chicken Ortholog Predictions Search”, which searches the HCOP dataset for chicken genes and returns information about chicken:human orthology. The HCOP Search can be performed for chicken Entrez Gene ID, chicken gene name, chicken gene symbol, CGNC ID, or the type of orthologous relationship. Search results are displayed in tabulated form with hyperlinks to additional information (Figure 4.1). The last column in the CGNC table contains an HCOP human ortholog link for every gene. When one of the links is selected, all human orthologs are displayed in the HCOP table below. Conversely, selecting an ortholog link in the last column of the HCOP table will display all chicken orthologs in the CGNC table for the corresponding human gene.

CGNC Gene Pages

Searching the CGNC returns results that link to individual gene pages (Figure 4.2). For each gene page, the data are presented grouped as CGNC Data, External Data, Human Orthologs and Avian Orthologs. The CGNC Data group includes: CGNC ID, Last review date, Status (Automatic, Pending. Approved, Entry Withdrawn, or In Review), Species, Gene name, Gene symbol, Synonyms, Chromosome, and Biotype. All these data are stored in the CGNC database. The External Data group contains external IDs formatted as links to the corresponding online resources. The following links are included: NCBI Entrez Gene ID, BirdBase ID, Ensembl gene ID, Chickspress genome browser (http://geneatlas.arl.arizona.edu/), GEISHA ID, and AgBase GO. The Human Orthologs group contains the following HCOP data for the human orthologs: HGNC ID, Entrez Gene ID, Gene name, Gene symbol, and Chromosome. This group also contains the ortholog type (1:1, 1:n, n:1, or n:n) determined by us as described above. The Avian Orthologs group is currently not populated; however, the same types of data as for the human orthologs will be included once the resource expands to other avian species. The Gene Neighbors group contains links to the NCBI gene pages (http://www.ncbi.nlm.nih.gov/gene/) showing the neighboring genes in the Genomic Context section, as well as other relevant gene data. The links are provided for the selected gene and its human orthologs. When avian orthologs are added, avian gene neighbor links will also be included.

Submitting Data to CGNC

Researchers may also contribute their expertise to the CGNC project via a login system for data entry. New users may request a login by contacting CGNC; using this login provides access to the CGNC biocuration database (Figure 4.3). Users can then use the same search strategies to identify chicken genes and their orthologs but the returned results now include an option to edit the nomenclature (name, symbol, synonyms) for any record. A Comments box is used to capture any additional information and the user ID is recorded (as it is for all biocurators). Initially, data provided by new users may also be marked as "in review" until confirmed by CGNC biocurators who check to ensure the names follow CGNC guidelines. All data entered must pass standard quality checks prior to release into the public CGNC database.

Downloads

The Downloads page provides a table with the current annotation statistics for chicken. Nomenclature is grouped into four categories: Automatic, Pending, In Review and Approved. Genes in the Automatic category are assigned their nomenclature based on computational methods: genes in the Pending category have been manually curated, but the biocuration quality has not yet been checked; In Review indicates that the approved gene nomenclature is awaiting further expert review; and genes in the Approved category have manually curated and quality-checked nomenclature. The download menu allows the user to filter and select results based upon curation categories and accession types. The entire unfiltered dataset in the text tab-delimited format can also be accessed from http://www.agnc.msstate.edu/DownloadAll.aspx.

Integrating CGNC Data

The gene page for any CGNC entry can be opened directly by providing a properly formatted URL. For example, to open the gene report for a gene with CGNC ID = 37583 the following URL should be used: http://www.agnc.msstate.edu/GeneReport.aspx?a=37583. Additionally, any gene-specific nomenclature data from the underlying database can also be easily formatted, for example, as an HTTP output with tab-delimited text to be utilized by remote servers. This output can be retrieved using an HTTP GET or POST request or a static URL. To ensure that the chicken gene nomenclature data is widely disseminated, we are happy to collaborate with groups wishing to use these data.

Assigning Nomenclature

Automatic Curation: Chicken-Human Orthologs

There are currently (June 2012) 18,658 chicken genes that have been automatically assigned gene nomenclature based upon 1:1 orthology to human genes that have standardized nomenclature. To confirm these ortholog assertions and capture information about gene synonyms (other names that the gene may also be called in the literature), we are manually checking these records. Students trained in aspects of orthology assertion, synteny and assignment of standardized gene nomenclature check these records by reviewing the HGNC, HCOP and NCBI Entrez gene information. Their data are checked by a trained CGNC biocurator prior to release. This project provides a practical biocuration project for biology undergraduate students to learn key aspects of genomics, comparative biology, database searching and chromosomal structure while contributing to developing fundamental resources for the research community.

Manual Biocuration Projects

Our current manual biocuration is divided into two projects. Our first project is to provide standardized gene nomenclature for chicken MHC genes while our second project focuses on providing nomenclature for genes that are highly expressed in hen eggs. Both of these projects are designed to select gene sets that are important to avian biology but contain genes that are unlikely to be annotated based on orthology with human genes.

The MHC region contains key immune genes involved in disease resistance (or susceptiblilty) and autoimmunity [9], making it a region of interest for immune and disease studies. The chicken major histocompatability complex (MHC) is found on chromosome 16 and this region has been the subject of several studies to fine map these genes [10,11,12,13]. To identify chicken MHC genes we searched for genes annotated by NCBI to occur in this region. This yielded 155 NCBI Entrez Gene IDs, of which 104 had associated gene symbols and or LOC IDs and were associated with the chicken MHC. An additional search using the UCSC Gallus browser (http://genome.ucsc.edu/cgi-bin/hgGateway) supplemented this original list and we also manually included genes from previous chicken MHC studies [10,11]. Discontinued gene annotations and quantitative trait loci (QTL) were excluded to give us a final list of 74 genes. We are now working to provide standardized gene nomenclature for this gene list in conjunction with NCBI, as they review structural annotation of these genes.

As expected, very few genes in this data set have a 1:1 homology ratio with a previously described human gene. Nine genes have strict orthology to named human genes their gene names were changed accordingly to reflect this orthology. Twenty one genes were identified to have a 1:2, 1:n, or n:1 homology ratio with previously described human genes. Chicken MHC genes that have similarity to a human HLA gene are assigned nomenclature that reflects this relationship. The gene name will follow the form: Major histocompatibility complex class # <chain type> <specific name>, (similar to HLA class # <chain type>).

The symbol is retained as the assigned chicken designator. For example, Entrez Gene: 693256 BLB2 becomes:

Gene name: Major histocompatibility complex class II beta chain BLB2, (similar to HLA class II, D beta chain)

Gene symbol: BLB2

This nomenclature is based upon its similarity to HLA class II, D beta chain genes (e.g. HGNC IDs: 4945, 4937, 4953). The remaining 43 chicken MHC genes have no human ortholog. These genes are named systematically based upon their relationship to well-studied chicken MHC genes such as BG2 and based upon their previously published names, while ensuring that gene symbols are unique.

A gene set of chicken egg genes for manual biocuration was determined by combining genes of proteins known to be expressed in egg white [14,15,16], vitelline membrane [17] and yolk [18]. This yielded a list of 201 chicken genes. Of these genes, 105 have strict orthology to named human genes. Their gene names were changed accordingly to reflect this orthology. Forty two genes were identified that had a 1:2, 1:n, or n:1 homology ratio with previously described human genes. The remaining forty seven genes have no human ortholog. Several egg genes are members of large gene families, particularly the SERPIN, SPINK, mucin, and defensin gene families. For gene family members with direct human orthologs gene names were changed accordingly. For gene family members without human orthologs we are working with the HGNC to determine appropriate names that reflect gene family membership. For genes with very well-recognized common names, the appropriate gene family name and symbol were assigned and the common name appended in parentheses in the name field (e.g. MUC6, mucin 6 oligomeric mucus/gel-forming (ovomucin, beta subunint)). In cases where chicken researchers and nomenclature experts have agreed that the common name should be kept, the appropriate gene family names were appended in parentheses in the name field, (e.g.: OVAL, ovalbumin (SERPINB14)). The work to complete manual biocuration is ongoing: we are currently seeking feedback from community experts about the gene nomenclature we have provided for these projects and expect that this data will be revised based upon this feedback and as new information is obtained. The genes to which we are currently assigning nomenclature are available on CGNC as a separate list (and are listed by project) to facilitate community comment. Moreover, as we finish these projects we are seeking community input on future projects that would benefit from manual biocuration. Researchers interested in providing feedback or suggesting future curation projects can contact CGNC biocurators via the website or directly using agabase@hpc.msstate.edu.

Collaborations

The CGNC currently works with the HGNC to ensure that chicken gene nomenclature is consistent where there are clear and strict orthologs and that numbering of family members is sequential. We particularly acknowledge Elspeth Bruford from the HGNC group for her advice and guidance. We owe special thanks to community experts who provide their knowledge and assistance: Marcia M Miller (City of Hope National Medical Center) for her assistance with the MHC project and Janet Fulton (Hy-line International) for her assistance with the egg gene project. CGNC welcomes enquiries from researchers who wish to have gene nomenclature assigned, resources wanting to use the nomenclature, community experts who wish to assist with biocuration or suggest targets for annotation and educators who are interested in incorporating aspects of gene nomenclature in their class work.

Future Directions

In addition to developing additional manual biocuration projects based upon community interest and need, we also expect to develop a core set of chicken gene nomenclature that can be applied to other avian species. Via BirdBase we expect to be able to identify chicken:turkey and chicken:zebra finch orthologs so that we are able to transfer nomenclature to these species. Moreover, we expect that gene annotation, orthology, and literature from these species will also inform chicken gene nomenclature. The CGNC database is designed to encompass chicken:turkey and chicken:zebra finch orthologs data, as it becomes available and this work will link with ongoing efforts to provide a bird comparative genome browser via BirdBase. We expect that this data can next be expanded to inform the additional bird genomes that are or soon will be sequenced.

However, comparative analysis outside of the Aves will also provide valuable information about gene evolution and conservation. We note with interest that the Anole genome project also encompasses a standardized gene nomenclature effort [19] and that sequencing of three Crocodilian genomes [20] will provide valuable information about genes from reptile species more closely related to birds. Comparative genomics amongst these species will only be enhanced by co-operation between resources providing their gene nomenclature and clearly defined guidelines for this biocuration effort.

Competing Interests

The authors declare that they have no competing interests.

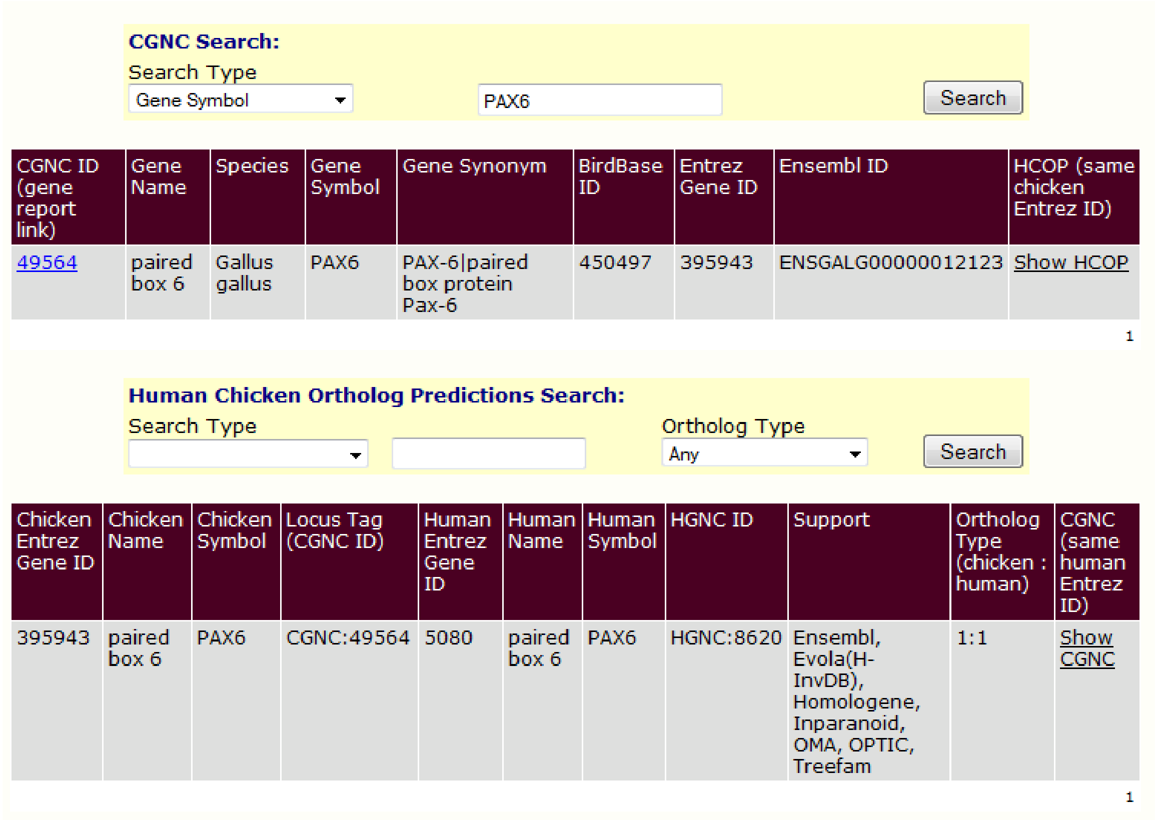
Authors' Contributions

PC and CN both contributed equally to the writing of the first draft; all other authors reviewed the manuscript and were involved in the editorial process. PC developed the CGNC database (including quality control and update procedures), website and biocuration interface. CN provided testing and feedback on the database and biocuration interface, provided manual biocuration and contributed to the development of the nomenclature guidelines. AMC also provided manual biocuration and contributed to the development of the nomenclature guidelines. JW conceived the MHC manual biocuration project and contributed to the development of the nomenclature guidelines. YS and SK verified automatic assignment of chicken:human ortholog nomenclature and provided manual biocuration of chicken genes. FM conceived the project, contributed to the development of the nomenclature guidelines and contributed to the writing of the first draft of this manuscript.

Acknowledgements

This project was supported with funding from the Agriculture and Food Research Initiative Competitive Grant no. 2011-67015-30332 from the USDA National Institute of Food and Agriculture, National Institutes of Health, National Institute of General Medical Sciences Project 07111084 and the US Department of Agriculture Agricultural Research Service NRSP-8 Bioinformatics Program. YS was supported by the supported by the National Science Foundation REU Site grant NSF DBI-1004842. SK was supported in part from funding provided by the Mississippi State University College of Veterinary Medicine.

Figures



CGNC Search:

Search Type

Gene Symbol ▾ PAX6 Search

CGNC ID (gene report link)	Gene Name	Species	Gene Symbol	Gene Synonym	BirdBase ID	Entrez Gene ID	Ensembl ID	HCOP (same chicken Entrez ID)
49564	paired box 6	Gallus gallus	PAX6	PAX-6|paired box protein Pax-6	450497	395943	ENSGALG00000012123	Show HCOP

1

Human Chicken Ortholog Predictions Search:

Search Type ▾

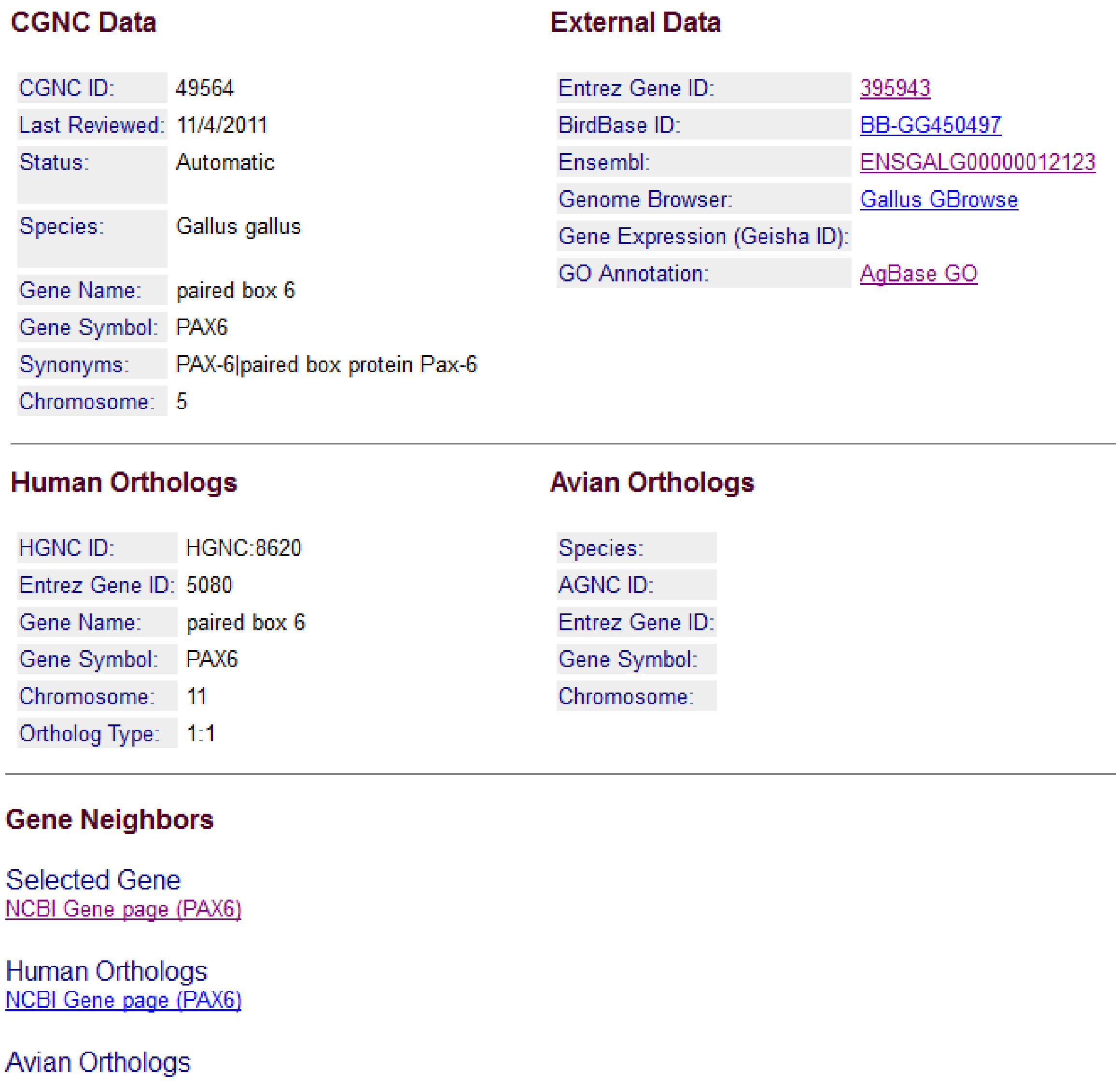
Ortholog Type

Any ▾ Search

Chicken Entrez Gene ID	Chicken Name	Chicken Symbol	Locus Tag (CGNC ID)	Human Entrez Gene ID	Human Name	Human Symbol	HGNC ID	Support	Ortholog Type (chicken : human)	CGNC (same human Entrez ID)
395943	paired box 6	PAX6	CGNC:49564	5080	paired box 6	PAX6	HGNC:8620	Ensembl, Evola(H-InvDB), Homologene, Inparanoid, OMA, OPTIC, Treefam	1:1	Show CGNC

1

Figure 4.1 CGNC interface for searching chicken gene nomenclature and the corresponding human orthologs. The search forms allow to search either CGNC chicken gene nomenclature or HCOP human orthologs. The links in the last column of both tables load the corresponding orthologs for the selected gene. The link in the first column of the CGNC table loads the Gene Report page (Figure 4.2) for the selected gene.



CGNC Data

CGNC ID:	49564
Last Reviewed:	11/4/2011
Status:	Automatic
Species:	Gallus gallus
Gene Name:	paired box 6
Gene Symbol:	PAX6
Synonyms:	PAX-6|paired box protein Pax-6
Chromosome:	5

External Data

Entrez Gene ID:	395943
BirdBase ID:	BB-GG450497
Ensembl:	ENSGALG00000012123
Genome Browser:	Gallus GBrowse
Gene Expression (Geisha ID):	
GO Annotation:	AgBase GO

Human Orthologs

HGNC ID:	HGNC:8620
Entrez Gene ID:	5080
Gene Name:	paired box 6
Gene Symbol:	PAX6
Chromosome:	11
Ortholog Type:	1:1

Avian Orthologs

Species:	
AGNC ID:	
Entrez Gene ID:	
Gene Symbol:	
Chromosome:	

Gene Neighbors

Selected Gene
NCBI Gene page (PAX6)

Human Orthologs
NCBI Gene page (PAX6)

Avian Orthologs

Figure 4.2 CGNC gene report page. The gene report page consolidates locally maintained nomenclature, mappings to external databases, orthology data, and NCBI gene neighbor data for the gene and its orthologs.

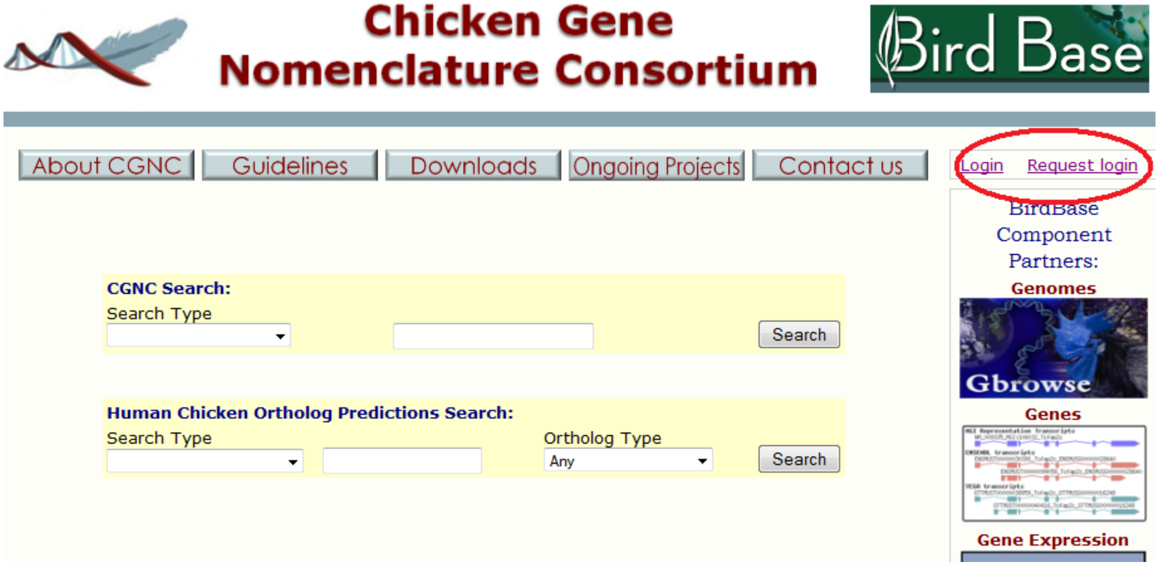


Figure 4.3 Login and Request Login links on the CGNC front page for gene nomenclature editing by research community. Interested researchers are encouraged to request login to the editing interface to submit their gene nomenclature updates.

References Cited

[1] International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**: 695-716.

[2] Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered*, **100**: 659-674.

[3] Anderson A. (2012) CSHL-Led Team Develops De Novo Genome Assembly Method with Error-Corrected PacBio Reads. *GenomeWeb*.

[4] Kraus R.H., Kerstens H.H., Van Hooft P., Crooijmans R.P., Van Der Poel J.J., et al. (2011) Genome wide SNP discovery, analysis and evaluation in mallard (Anas platyrhynchos). *BMC Genomics*, **12**: 150.

[5] Oleksyk T.K. (2011) A Draft Sequence of the Puerto Rican Parrot Genome (Amazona vittata) – a Genome Project funded by a Local Community Effort. *Nature Preceedings*, doi:10.1038/npre.2011.6552.1.

[6] Crittenden L., Bitgood J., Burt D. (1995) Genetic nomenclature guide. Chick. *Trends Genet*: 33-34.

[7] Burt D.W., Carre W., Fell M., Law A.S., Antin P.B., et al. (2009) The Chicken Gene Nomenclature Committee report. *BMC Genomics*, **10 Suppl 2**: S5.

[8] Bruford E.A., Lush M.J., Wright M.W., Sneddon T.P., Povey S., et al. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res*, **36**: D445-448.

[9] Trowsdale J. (2011) The MHC, disease and selection. *Immunol Lett*, **137**: 1-8.

[10] Miller M.M., Bacon L.D., Hala K., Hunt H.D., Ewald S.J., et al. (2004) 2004 Nomenclature for the chicken major histocompatibility (B and Y) complex. *Immunogenetics*, **56**: 261-279.

[11] Shiina T., Briles W.E., Goto R.M., Hosomichi K., Yanagiya K., et al. (2007) Extended gene map reveals tripartite motif, C-type lectin, and Ig superfamily type genes within a subregion of the chicken MHC-B affecting infectious disease. *J Immunol*, **178**: 7162-7172.

[12] Guillemot F., Billault A., Pourquie O., Behar G., Chausse A.M., et al. (1988) A molecular map of the chicken major histocompatibility complex: the class II beta genes are closely linked to the class I genes and the nucleolar organizer. *Embo J*, **7**: 2775-2785.

[13] Ruby T., Bed'Hom B., Wittzell H., Morin V., Oudin A., et al. (2005) Characterisation of a cluster of TRIM-B30.2 genes in the chicken MHC B locus. *Immunogenetics*, **57**: 116-128.

[14] D'Ambrosio C., Arena S., Scaloni A., Guerrier L., Boschetti E., et al. (2008) Exploring the chicken egg white proteome with combinatorial peptide ligand libraries. *J Proteome Res*, **7**: 3461-3474.

[15] Mann K. (2007) The chicken egg white proteome. *Proteomics*, **7**: 3558-3568.

[16] Guerin-Dubiard C., Pasco M., Molle D., Desert C., Croguennec T., et al. (2006) Proteomic analysis of hen egg white. *J Agric Food Chem*, **54**: 3901-3910.

[17] Mann K. (2008) Proteomic analysis of the chicken egg vitelline membrane. *Proteomics*, **8**: 2322-2332.

[18] Mann K., Mann M. (2008) The chicken egg yolk plasma and granule proteomes. *Proteomics*, **8**: 178-191.

[19] Kusumi K., Kulathinal R.J., Abzhanov A., Boissinot S., Crawford N.G., et al. (2011) Developing a community-based genetic nomenclature for anole lizards. *BMC Genomics*, **12**: 554.

[20] St John J.A., Braun E.L., Isberg S.R., Miles L.G., Chong A.Y., et al. (2012) Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol*, **13**: 415.

CHAPTER V

CONCLUSION

Advances in the sequencing technologies are resulting in an increasing amount of sequence data. Currently production of sequencing data outpaces our ability for making sense of this data. This creates a critical need for developing bioinformatics methodologies, tools, and resources that enable researchers to rapidly analyze and gain information from this increasing amount of sequence data. This dissertation focuses on development of tools and resources to support genome structure analysis, individual variation, and comparative biology.

Next-generation sequencing projects are producing either random genomic DNA sequences or transcript (reverse transcribed complimentary DNA) sequences. The former are necessary for genome characterization projects, in which the composition of non-transcribed regions of a genome is included in the analysis. Although complete genome sequencing reveals the entire genome structure, which lends itself for annotation and, thus, represents an ideal means by which the genomes of organisms can be compared, it is not currently economically feasible for most eukaryotes. This is especially true for the numerous organisms that have large, highly repetitive genomes including many important plants and animals. Alternatively, sample sequencing of random genomic DNA can be used to gain considerable information about genome structure in lieu of a complete sequence [1,2]. However, it is often difficult for researchers to characterize the sequences they have obtained, especially if they have generated large sequence data sets for organisms for which previous sequencing research has been limited. There are various programs for gene characterization [3,4] and also a growing number of programs for characterization of repetitive elements [5,6,7]. However, to my knowledge, there is no program or pipeline designed to provide an overview of the sequence composition of an entire genome based on shotgun sequence reads. Such characterization can be possible using a sequence read classification pipeline (SRCP) presented in this dissertation. In this pipeline, a battery of existing and novel algorithms are used to place random genomic query sequences into descriptive/functional sequence categories. Classified reads represent percentages of each descriptive category. Knowledge of these percentages will play an important role in further efforts to sequence and annotate the corresponding genome. Thus, the SRCP addresses the lack of preliminary genome characterization software/methodologies that hindered inference of genome composition in the initial analysis on not-yet-studied organisms. The limitation of the approach used in the SRCP is that it provides a limited view of genome structure of a researched organism. Future development of this pipeline would benefit from inclusion of various comparative analyses with related organisms, such as assigning putative gene orthology, annotation of InterPro domains [8] within assembled putative protein sequences, analysis of conserved synteny between species, identification of ultraconserved elements (UCEs), comparison of gene families and pathways, etc. Future advances in sequencing technologies should provide long reads (several thousand bases long) (http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencingadvantage/), which would change the approach from classifying individual reads as a whole to assembly and identification of the DNA elements present within the assembled sequences. It is likely that such assemblies will represent a fairly accurate reference genome draft, since very long high-quality reads should easily overcome problems with assembly of repeat regions, which would change the focus of the analysis to preliminary annotation of the assembled genome.

Another comparative analysis that can be performed in the initial stages of a study of an organism without a sequenced genome is inferring phylogeny of closely related species. The ability to genetically identify and distinguish between related species, cultivars/strains, and individuals is central to technology commercialization and the protection of intellectual property [9,10,11]. While a number of restriction site polymorphism-, random amplicon-, and repeat polymorphism-based molecular marker techniques have been developed to compare individuals and construct linkage maps [12], next generation sequencing makes it affordable to do genome wide analysis using single nucleotide polymorphisms (SNPs) [13,14]. SNP assays relying on whole genome sequence comparisons are not currently affordable for practical use in commercial settings and for agricultural patents. Moreover, the very large numbers of SNPs in the non-coding regions of genomes, which tend to be under relatively low evolutionary constraint, provide much larger datasets than needed for most mapping and identification/differentiation projects. Therefore, exome screening based on next-generation sequencing can be used for comparison of evolutionarily constrained sequences. This represents a challenge for organisms without sequenced genomes as there are no references to which RNA-seq reads can be aligned. This dissertation presented methodology to overcome this challenge by using the RNA-seq reads for assembly of partial references common to all strains in the analysis, which then can be used for SNP-based phylogeny analysis. The significance of this methodology is that it enables high-resolution phylogeny inference of very closely related strains of the same organism without sequencing individual genomes of each strain. This scenario is common for commercial varieties of many agricultural species. In this dissertation, this methodology was applied to miscanthus, an emerging bio-energy crop [15]. Besides that, the RNA-seq reads sequenced for this study were utilized for creation of transcript assemblies of *Miscanthus x giganteus* and *Miscanthus sinensis*. The assembled transcript sequences can play an important role in further sequencing of exonic regions during the whole genome assembly process. A potential problem with relying exclusively on SNPs for phylogeny inference is that just a small number of mutations, many of which may be non-SNP variations (copy number variations, insertions/deletions, inversions, etc), may separate very closely related strains of the same organism. In this case, overlooking the contribution of the non-SNP variations can skew the analysis results. For these cases, sequencing the entire individual genomes and performing complete comparison of all variations would provide the most accurate result. Advances in sequencing technologies should make this approach feasible in the future.

Upon sequencing of a genome, the next step in the analysis is providing a reliable genome annotation, most typically to identify coding regions. While there are multiple approaches for identifying genes in a DNA sequence [16,17,18,19,20], there are relatively fewer resources for providing gene nomenclature that is the basis of future functional and comparative analyses. Standardized gene nomenclature made available to the researchers as a centralized resource prevents confusion in gene naming, e.g., naming the same gene two different gene names or using the same name for two unrelated genes. It also ensures that orthologs have the same name across species to facilitate comparative genomics. A standard gene naming convention guarantees that all gene names, symbols, and synonyms are designated following the same rules, e.g., using brief and specific names that convey the character or function of the gene, using American spelling, avoiding tissue specificity or molecular weight designations. Following such a standardized naming convention ensures that the researchers will get the most meaningful information about the gene from its name, symbol, and synonym. However, even well researched standard organisms like chicken suffer from lack of reliable gene nomenclature. Problems such as duplicate gene IDs for the same gene name, duplicate gene names for the same gene ID, and inconsistent gene synonyms are common when multiple research groups name genes without having an access to a standardized gene nomenclature resource. Consequently, comparative biology is hampered due to the difficulty of recognizing orthologs and functional equivalents [21]. Therefore, standardized gene nomenclature is necessary to facilitate communication between scientists [22]. The international Chicken Gene Nomenclature Consortium (CGNC) provides standardized gene nomenclature for chicken genes [22]. CGNC members created ChickGeneNames, an annotation tool displaying human-chicken orthologs, to form a core of chicken gene nomenclature.

This dissertation presents a ‘first pass’ gene nomenclature resource created by transferring nomenclature from 1:1 orthologs of a related species and a web-based interface to build on this core set of genes through manual biocuration to assign nomenclature. The nomenclature resource utilizes community expertise via a password-protected web interface allowing interested researchers with domain knowledge to suggest gene nomenclature updates to the biocurators of this resource. As an example, we populated this resource with chicken gene nomenclature. Due to rapidly increasing number of sequencing projects we project that the number of species with sequenced genomes and nascent gene annotation efforts will increase proportionally. At this point orthology prediction can be used to facilitate gene annotation and comparative biology. Hence, there is a need for extensible platform to include related orthology species. For example, in the future, the chicken resource can include other avian species and capture orthology among their genes and human genes, as well as their genes and chicken genes. The general database setup, web interface, and automated update scripts can be easily adapted for gene nomenclature resources dedicated to many other organisms. Thus, the gene nomenclature resource developed in this dissertation addresses the lack of versatile software dedicated for gene nomenclature curation and standardization throughout the scientific community. This software represents a tool that can be utilized by multiple gene nomenclature committees as a standardized way to catalog, curate, and disseminate gene nomenclature data. A limitation of the current version of this resource is that it can only be adapted for organisms with a curated set of human orthologs supported by HUGO (Human Genome Organization) Gene Nomenclature Committee (HGNC) (http://www.genenames.org). Particularly, 1:1 human-chicken orthologs are utilized to transfer human gene nomenclature to chicken genes automatically. Generally, to establish a similar resource for a vertebrate not supported by HGNC, one would have to identify human orthologs first. To establish a similar resource for a plant organism one would have to identify orthologs between the plant of interest and *Arabidopsis thaliana*, which is the *de facto* plant model organism in plant molecular biology [23] with an established reference genome that has been intensively studied and annotated. The results, including gene nomenclature, can be found in The Arabidopsis Information Resource (TAIR) [24]. Regardless of the availability of orthologs from organisms with standardized gene nomenclature this resource would establish a centralized source of gene nomenclature for any organism and enable communication among the researchers. As the orthologs become available and manual biocuration efforts progress the quality of the gene nomenclature will improve.

In conclusion, the research presented in this dissertation has produced new computational algorithms, methodologies, tools, and pipelines that help address the need for processing of volumes of data generated by new sequencing technologies. Sequence data for thousands previously unstudied organisms will become available in the near future, e.g., the Genome 10K project [25]. Utilizing the SRCP (enhanced with comparative biology features) and the methodology for initial phylogenetic analysis developed in this dissertation, researchers will be able to position the organism that they study in the evolutionary context. Knowledge of the genome composition will support hypotheses of evolutionary events, such as genome duplication, that led to genetic variation from the related organisms [26]. Addition of various comparative biology analyses mentioned above to the SRCP will facilitate identifying the profile of genetic variation, which will help inferring more information about these evolutionary events. This comparative biology approach will also facilitate identification of orthologs between the species and paralogs within the species and, thus, enable functional annotation by transferring gene nomenclature from well-annotated 1:1 orthologs, as required by the online standardized gene nomenclature resource developed in this dissertation. Thus, the tools, methodology, and resources presented here are tied together in following the initial analysis workflow for the thousands of organisms slated for sequencing in the near future.

References Cited

[1] Strong W.B., Nelson R.G. (2000) Preliminary profile of the Cryptosporidium parvum genome: an expressed sequence tag and genome survey sequence analysis. *Mol Biochem Parasitol*, **107**: 1-32.

[2] Kirkness E.F., Bafna V., Halpern A.L., Levy S., Remington K., et al. (2003) The dog genome: survey sequencing and comparative analysis. *Science*, **301**: 1898-1903.

[3] Lomsadze A., Ter-Hovhannisyan V., Chernoff Y.O., Borodovsky M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, **33**: 6494-6506.

[4] Solovyev V., Kosarev P., Seledsov I., Vorobyev D. (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol*, **7 (Suppl. 1)**: S10–S12.

[5] Bao Z., Eddy S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, **12**: 1269-1276.

[6] Price A.L., Jones N.C., Pevzner P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21 Suppl 1**: i351-358.

[7] Li R., Ye J., Li S., Wang J., Han Y., et al. (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol*, **1**: e43.

[8] Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., et al. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res,* **33**: W116-120.

[9] Kunihisa M., Fukino N., Matsumoto S. (2005) CAPS markers improved by cluster-specific amplification for identification of octoploid strawberry (Fragaria x ananassa Duch.) cultivars, and their disomic inheritance. *Theor Appl Genet*, **110**: 1410-1418.

[10] Jung J., Park S., Liu W., Kang B. (2010) Discovery of single nucleotide polymorphism in Capsicum and SNP markers for cultivar identification. *Euphytica*, **175**: 91-107.

[11] Castro P., Millan T., Gil J., Merida J., Garcia M., et al. (2011) Identification of chickpea cultivars by microsatellite markers. *Journal of Agricultural Science*, **149**: 451-460.

[12] Semagn K., Bjornstad A., Skinnes H., Maroy A.G., Tarkegne Y., et al. (2006) Distribution of DArT, AFLP, and SSR markers in a genetic linkage map of a doubled-haploid hexaploid wheat population. *Genome*, **49**: 545-555.

[13] Ganal M.W., Altmann T., Roder M.S. (2009) SNP identification in crop plants. *Curr Opin Plant Biol*, **12**: 211-217.

[14] Rounsley S.D., Last R.L. (2010) Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. *Plant J*, **61**: 922-927.

[15] Hodkinson T.R., Chase M.W., Lledo M.D., Salamin N., Renvoize S.A. (2002) Phylogenetics of Miscanthus, Saccharum and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnLintron and trnL-F intergenic spacers. *J Plant Res,* **115**: 381-392.

[16] Math´e C., Sagot M.-F., Schiex T., and Rouz´e P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**, **no. 19**: 4103–4117.

[17] Zhang M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.*, **3**: 698–710.

[18] Stormo G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.*, **10, no. 4**: 394–397.

[19] Claverie J.M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Gene*., **6**: 1735–1744.

[20] Burset M. and Guigo R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**: 353–367.

[21] Kuzniar A., van Ham R.C., Pongor S., Leunissen J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet,* **24**: 539-551.

[22] Burt D.W., Carre W., Fell M., Law A.S., Antin P.B., et al. (2009) The Chicken Gene Nomenclature Committee report. *BMC Genomics*, **10 Suppl 2**: S5.

[23] Meyerowitz E.M., Pruitt R.E. (1985) Arabidopsis thaliana and Plant Molecular Genetics. *Science*, **229**: 1214-1218.

[24] Swarbreck D., Wilks C., Lamesch P., Berardini T.Z., Garcia-Hernandez M., et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, **36**: D1009-1014.

[25] (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered*, **100**: 659-674.

[26] Venkatesh B., Kirkness E.F., Loh Y.H., Halpern A.L., Lee A.P., et al. (2007) Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome. *PLoS Biol*, **5**: e101.