AN ABSTRACT OF THE THESIS OF

Steven T. Hill for the degree of Master of Science in Computer Science presented on June 10th, 2016.

Title: The Pursuit of Hoppiness: Propelling Hop into the Genomic Era

Abstract approved: ____

David A. Hendrix John A. Henning

Hop (*Humulus lupulus L. var lupulus*) is a plant of great cultural significance, used as a medicinal herb for thousands of years, and for flavor and as a preservative in brewing beer. Studies of the medicinal effects of the unique compounds produced by hop have led to interest from the pharmacy and healthcare industries. Although many industries have interest in the plant itself and scientists are interested in the evolution of the *Cannabaceae* sex chromosomes, little effort has gone into developing genomic resources. *H. lupulus* is a highly heterozygous, repeat-rich, plant genome with a size of about 2.8 gigabases. This combination presents an immense challenge to studying the genomics of hop. Here we present, a web portal for studying hop genomics, a novel hop genome assembly, gene annotations for both draft genome assemblies, an evolutionary biology study regarding the hop sex chromosomes, and a novel way of modeling transcripts using deep learning. The combination of these manuscripts provides a framework for the future of hop genomics.

The Pursuit of Hoppiness: Propelling Hop into the Genomic Era

by
Steven T. Hill

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

June 10, 2016
Commencement June 2017

Master of Science thesis of Steven T. Hill presented on June 10th, 2016

APPROVED:

Major Professor, representing Computer Science

Co-Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Steven T. Hill, Author

ACKNOWLEDGEMENTS

I would like to thank all the scientists that built the foundations for this work to build upon.

CONTRIBUTION OF AUTHORS

Dr. Aaron Liston provided expert background knowledge and contributed to writing. Jamie Coggins developed all of the sequencing libraries for genotyping by sequencing. Ramcharan Sudarsanam assisted in quality control of the genome annotations. Rachael Kuintzle provided expert advice, thorough analysis, contributed to writing, developed figures, and conceived experiments. Erich Merrill assisted with package development, experimental design, guidance in implementation, and expert advice. David Hendrix contributed manuscript writing, experimental design, extensive review, and assisted in analysis. John Henning assisted with library preparation, manuscript writing, analysis conclusions, data analysis, and advising

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER 1: INTRODUCTION

In this thesis, we focus on the process of annotating a new and complex genome. Genome annotation requires a number of steps. Typically researchers begin with sequencing. DNA-sequencing has traditionally been the most powerful method of annotating a genome. It is possible to identify genes from sequence content alone. More recently, it also has been possible to partially annotate a genome using RNA-sequencing. This method lacks the ability to sequence untranslated regions. In the chapters to come, we address the issues of genome annotation, not only applied to hop, but also theoretical work using newer statistical models for annotating long intergenic non-coding RNAs (lincRNAs).

1.1 History of Hop Breeding and genetics

Hop (*Humulus lupulus* L.) is a traditional herb that has been used for thousands of years as a medicinal agent (Chopra et al. 1986a). In India, it was used as a sleep aid as a method of reducing anxiety. However, the use of hop as a bittering agent in beer is relatively new. Although the first use of hop in beer is disputed, the first written report of hop being used in beer was in 877 AD by a French monk, Abbot Adalhard (Delyser and Kasper 1994). Adalhard left a recipe for making a porter that included using hop cultivated in the monasteries garden. However, it was not until the 20th century that humans became interested in breeding of hop.

The first hop breeder was Professor Ernest Salmon of Wye College, England. Wye College established a breeding program in 1906, and Professor Salmon quickly began working on releasing hop varieties (Salmon 1917). Salmon has a collection of publications discussing the genetics of hop, breeding of new cultivars, and about the challenge of disease faced by growers. Particularly of interest was the cultivar Brewer's Gold. Brewer's Gold was a cross with a European hop and a wild Canadian hop; it is notable for not only being a great hop in brewing, but also it is the ancestor of most high-bittering potential hop varieties used in brewing today.

While hop breeding was beginning to take off in Europe, World War I marks a massive growth of the hop industry in the states. Much of the agriculture in Europe was destroyed, opening the door for hop production in the United States (Tomlan 2013). There was a significant industry for hop growing in the Willamette valley throughout the late 19th century and pre-prohibition 20th century. Prohibition affected hop growers hard, and many were forced to switch to growing a different crop or shut down entirely (Landis 1939).

The USDA-ARS hop-breeding program was initiated at the end of prohibition. Stanley Brooks was the first USDA hop breeder, with the objective of breeding downy mildew resistant hop. Although the breeding program was initiated in 1933, the first hop cultivar released was Cascade in 1974 (Brooks et al. 1972). Cascade is one of the most cultivated hop varieties today (USDA 2016), and marked the beginning of the "hop heavy" beer styles.

Shortly after the release of Cascade, triploid (3X=30) varieties Willamette and Columbia were released. These two varieties have also had a large impact on the brewing industry (Haunold et al. 1976a; Haunold et al. 1976b). Willamette and Columbia were released out of the polyploid hop breeding project (Haunold 1972). Triploid hop have not been used much outside of the USDA breeding program, although the New Zealand breeding program has released and continues to release triploid hop varieties.

1.2 Brief history of molecular breeding in Hop

Hop has been studied for cytogenetic research for more than half a century (Ono, 1962). The unusual diversity of chromosome counts made it a popular target for using cytological techniques to study sex chromosomes (Ono 1962a; Karlov et al. 2003a). Although these methods were very successful, the limited capabilities of early cytogenetics did not answer questions about the genotype of different hop cultivars.

Just recently, molecular breeding has become popular among hop breeders. The first reported identification of molecular markers in hop was in 1995 – as a simple study of genetic diversity and heritability of different traits (Brady et al. 1996; Pillay and Kenny 1996). With the ability to quickly genotype different hop cultivars, a natural progression was to develop genetic maps. Developing a genetic map is the process of using recombination as a measure of genetic distance among individual molecular markers. Using these genetic distances, which correlate closely to physical distances, it is possible to determine which traits are genetically linked. However, constructing a genetic map is very similar to a well-known NP-complete problem, the minimal Hamiltonian-path problem (Crescenzi and Kann 1997).

There have been several genetic maps published, beginning as early as 2000, and most recently 2015 (Seefelder et al. 2000; Koie et al. 2004; Cerenak et al. 2006; Henning et al. 2011; Henning et al. 2015b). A challenge in developing linkage maps is the lack of guidance for arranging markers along a physical map. Without a reference genome, it is impossible to know the accuracy of a given genetic map. Due to the non-linearity and NP-hardness of the problem, it is unlikely that any of the published maps are 100% accuracy with respect to the physical distances of each marker.

Most recently, genome wide association studies (GWAS), quantitative trait loci (QTL), and RNA-sequencing studies have gained popularity within the hop breeding community. QTL studies for many different traits have been performed, and as the genomic resources increase, the quality of these studies should improve as well (Patzak et al. 2012; Clark et al. 2013a; Jakse et al. 2013; McAdam et al. 2013; Henning et al. 2015a). An obvious extension of these works is to develop QTLs using only markers that fall within coding regions of genes. Although no work has been published on this, there is data available and manuscripts in preparation regarding the subject.

Finally, in 2014 the release of a draft hop genome assembly was released for the cultivar Shinsuwase (Natsume et al. 2014a). The following chapters in this thesis not only use this resource, but they also extend it to include a gene annotation and provide an alternative genome assembly. The work in this thesis is meant as a resource and tool to continue the development of hop genomic resources. Although the resources are not perfect and lack a finished reference assembly, the number of genomic resources available for hop has significantly increased in just the past two years.

1.3 Machine learning in plant genetics

Machine learning has a rich history in plant genetics. Some of the most common uses of regression in plant breeding are: heritability estimation, genotype by environment interaction, genome wide association studies, quantitative trait loci discovery, and genomic selection (Smith and Kinman 1965; Henderson 1975; Wright 1976; Seaton et al. 2002; Listgarten et al. 2012). It is easy to understand why that is true; plant breeders are involved in an intense multi-objective optimization problem. In fact, it is very easy to model the problem of plant breeding as a Markov decision process (MDP) -- a known class of problems which involves sequential decision making. Many statistical and computational frameworks already exist for MDP problems (Anantharam et al. 1987).

Linear regression provides some major advantages for plant breeders. It’s stable, online, and has a closed form solution that is guaranteed to be optimal if the problem is convex.
If the optimization problem is non-convex, linear regression is only guaranteed to converge on a local minimum. This is particularly troubling considering biochemical pathways are often a combination of genes producing different compounds required for the production of a certain chemical. Similarly, traits related to size and yield are far too complex to model with a linear method. An example of these phenomena in hop is the prenylated-flavonoid Xanthohumol. Although the biochemical pathway is well understood, regulation is unknown and there is some evidence suggesting that the Humulone pathway plays a role in the regulation of Xanthohumol (Stevens et al. 1997a). Linear regression would almost certainly fail on a trait like Xanthohumol due to the multiplicity of pathways and control involved in its production. Recently, there has been some work to incorporate newer non-linear statistical models to plant breeding, Although they have been consistently more successful than linear models, the majority of quantitative genetics software packages are still centered around linear regression (Spindel et al. 2015).

1.4 Machine learning in genomics

While machine learning in plant breeding is often focused around regression, genomics typically relies on generative models to represent some unknown distribution. A generative model will try to model the underlying distribution directly, contrary to discriminative models, which are only interested in modeling $p(y = 1|x)$. These models have the advantage of interpretability, and a natural fit within genomic frameworks.

Hidden Markov Models (HMMs) find their place in a diverse array of genomic problems, including both base-calling during sequencing, and gene finding while annotating a genome (Stanke et al. 2006a; Timp et al. 2012). Bayesian nets and other probabilistic models are the dominating method of discovering regulatory pathways (Hecker et al. 2009). The speed and interpretability of the models makes them easy to debug and adapt to new problems.

Discriminative models also have their place in genomics. There are many different instances where classification is an important task. Typically discriminative models problems are sub-problems of things modeled using generative models (Abrusán et al. 2009a). A generative model can be defined as a model trying to directly model the joint distribution, $P(x, y)$. By doing this, you implicitly also learn $P(y=1| x)$, given that one has the prior $P(x)$. While discriminative models may not be as interesting as generative models, they still are vital in the process of genome annotation. Perhaps most importantly, they are much easier to train, requiring less data and learn a more simple representation of the data.

1.5 Deep Learning in Bioinformatics

Deep learning refers to “stacked”, or multi-layered neural networks. A neural network can simply be seen as a linear model with a non-linear transformation as the last step. In the simplest case, linear regression can be transformed into a single-layer neural network by adding a sigmoid function to the linear output of the model. Functions applied to the output of the regression model are called activation functions. The sigmoid function transforms the regressor into a classifier by forcing the values into the range of [0, 1]. The simplest way to think about a multi-layer neural network is by taking the output of a linear model, and applying it as input to another linear model. However the non-linear activation functions transform the models into non-linear models. Training these models is done with a first-order technique known as stochastic gradient descent.

Deep learning initially gained popularity in the 1980s with the invention of multi-layer perceptron (MLP). A MLP is like the model described above, the output of each linear model is passed as input to the next, and the final activation is a sigmoid function that transforms the output into a probability. This probability can be used for classification, however by using the identity matrix as an activation function, you construct a multi-layer regressor.

In 2012, Alex Krizhevsky brought deep learning to the front of scientist's attention by winning the ImageNet competition. Krizhevsky and his team beat competition by having an error rate 10% lower than the second place team (Krizhevsky et al. 2012). This kind of margin of error is rare for classification competitions; deep neural networks dominated almost every subsequent computer vision competition. Other fields in machine learning and computer science quickly began adopting neural networks. Convolutional neural networks, as used by Krizhevsky in the ImageNet competition, began gaining large amounts of momentum in every field interested in classification. Shortly after Krizhevsky's win in 2012, recurrent neural networks also began to gain popularity. In 2008, Alex Graves published his PhD dissertation titled "Recurrent neural networks for sequence labeling." Although this paper did not immediately make a large impact, it did lead to the now-common use of recurrent neural networks (RNN).

Recurrent neural networks can be thought of as feed forward neural networks (described previously), except that each "layer" is represented by a step in time. There is a special "hidden layer," which is computed by combining the previous hidden layer and the input at the current step in time. This kind of model lends itself favorably to any sort of temporal model, but also towards models that want a sequence of outputs rather than a single output. Long short-term memory is the dominant model used for recurrent neural networks. Long short-term memory applies two separate set of gates which exist simply to aid the optimization process through stochastic gradient descent (Hochreiter and Schmidhuber 1997).

In bioinformatics, deep learning is a relatively new concept. It is starting to gain traction as a tool for feature extraction. That is, use a deep neural network to understand cellular biology. The most successful use of deep learning is a program called DeepBind (Alipanahi et al. 2015). Although this model used a convolutional neural network (CNN), it only used a single layer and the primary finding in the paper was the use of perturbation techniques to understand the features that the CNN learned.

Similarly, a program called DeepSEA used a single-layer convolutional neural network to understand non-coding variants. DeepSEA is an excellent example of the strengths of deep learning. The authors trained the model to predict the chromatin state at each nucleotide. While this is not directly useful, observing changes in the chromatin level predictions when changing nucleotides provides information about the effect of each mutation. Not only did they abuse the interaction between chromatin state and functionality, they also achieved extremely high accuracy on a dataset not included during training (Zhou and Troyanskaya 2015).

Recurrent neural networks have yet to gain much traction in bioinformatics. An extension of the DeepSEA model was published as a program named DANQ. The models have the same goal, same underlying model, same unique training method, however, DANQ uses a bi-directional Long short-term memory (LSTM) model after the convolutional layer (Quang and Xie 2015).

Recurrent neural networks fit naturally as a replacement for the generative models currently being used. They are robust to long-term state dependencies, easy to train (fewer parameters), naturally fit the sequential nature of DNA, and can be parallelized using modern graphics processing-unit (GPU) technologies. However, no such publication has incorporated these models into bioinformatics pipelines. There is a natural fit in gene finding, base calling, regulatory network identification, sequence classification, and sequence modeling or labeling. The last chapter of this thesis is devoted to recurrent neural network models for transcript classification; the first work of it’s kind.

Chapter 2: HopBase: A unified resource for *Humulus* Genomics

Steven T. Hill, John Henning, Ramcharan Sudarsanam, and David Hendrix

2.1 ABSTRACT

Hop is a plant of great cultural significance, used as a medicinal herb for thousands of years, as well as flavoring and preserving beer. Studies on the medicinal effects of the unique compounds produced by hop has led to interest from pharmaceutical, healthcare and animal livestock industries. Recent developments in hop genomics research include published draft genome and transcriptome assemblies. Although research into the genomics of hop have gained interest, there is a critical need for centralized online genomic resources. To support the growing research community, we report the development of an online resource "HopBase.org." In addition to providing a gene annotation to the existing 'Shinsuwase' draft genome, HopBase makes available genome assemblies and annotations for both the cultivar 'Teamaker' and male hop accession 'USDA 21422M'. These genome assemblies, gene annotations, along with other common data, coupled with a genome browser and BLAST database enable the hop community to enter the genomic age.

2.2 INTRODUCTION

Hop is a large, climbing, dioecious plant in the Rosid class that has been used for medicinal, brewing, and preservative activities for millennia (Hamel and Chiltoskey 1975; Chopra et al. 1986b; Neve 2012). The *Humulus* genus contains three species, *Humulus japonicus, Humulus lupulus,* and *Humulus yunnanensis*, two of which, *Humulus japonicus* and *Humulus lupulus,* are known to produce compounds with beneficial pharmaceutical properties (Sung et al. 2015).Little is known about *Humulus yunnanensis* and it may be extinct, even though there has been effort to find a living plant (Boutain 2014). *Humulus* also has three typical sex chromosome configurations: *Humulus lupulus* (2n = 18 + XY), *Humulus lupulus var. cordifolius* (2n = 16 + X1X2 Y1Y2), and *Humulus japonicus* (2n = 14 + XY1Y2) (Ono 1962b). The simplicity of *H. lupulus var. lupulus* makes possibly the more tractable of these configurations for genome assembly. These configurations provide an interesting platform for studying sex chromosome evolution in plants and several research projects have been focused around this already (Karlov et al. 2003b; Hill et al.).

Cytogenetic research and genome assembly suggest that the hop genome is approximately 2.8Gb and highly repetitive (Danilova et al. 2003; Natsume et al. 2014b). Large amounts of repetitive DNA cause difficulties in short-read genome assembly due to the inability to assemble through repetitive regions. As a result, the repeat regions are larger than current mate-pair technology and require expensive long-read sequencing methods to assemble. Efforts using short-read sequencing techniques have been extensive and exhaustive and the resulting assemblies, while incomplete, are now available (Natsume et al. 2014b).

Currently, there exists published work on a high-marker-density genetic map (Henning et al., 2015), several RNA sequencing datasets (Clark et al. 2013b; Natsume et al. 2014b), a draft genome assembly, a plethora of research surrounding the essential oils (Stevens et al. 1997b; Miranda et al. 1999; Aron and Shellhammer 2010), and many other secondary resources. Furthermore, we have deep-sequenced, assembled, and annotated another female hop variety, ‘Teamaker’, The assembly was used to guide the assembly of the first male hop genome (USDA

21422M) coupled with the identification of male specific DNA and pseudo-autosomal regions of the sex chromosomes (chapter 3). None of these resources include a public annotation, and no attempt has been made to consolidate this information into a single resource. Standardizing data and providing a unified access has been a challenge in genome annotation and bioinformatics for some time. The consolidation of information allows for a much cleaner and easier flow of information among hop researchers. The objective of our work was to assemble both a male and female hop genome and to couple the information from these assemblies along with all other online hop genome information into a single resource available to hops researchers and breeders alike.

2.3 MATERIALS AND METHODS

2.3.1 Teamaker Genome assembly.

The Teamaker genome assembly used libraries selected in accord to the ALLPATHS-LG recipe (Gnerre et al. 2011). Reads were adapter trimmed and filtered for a mean quality of at least 30 using the program Skewer (Jiang et al. 2014). Duplicated reads were removed using a custom C++ program (https://github.com/hillst/dedup_paired_reads). This resulted in an estimated coverage of 109x (Table 2.6.1). Assembly was performed using the ALLPATHS-LG assembly with ploidy as 2 and using a minimum contig size of 500. Lower values resulted in unfeasible computation time and memory usage. The resulting assembly was gap filled using GapCloser 1.0 (Luo et al. 2012).

2.3.2 Transcript guided assembly.

In addition to the above mentioned assembly, we used SOAPdenovo-trans to perform a *de novo* transcriptome assembly (Xie et al. 2014). RNA-seq reads were acquired from Shinsuwase assembly (Natsume et al. 2014b). All libraries corresponding to the cultivar Shinsuwase were downloaded from the DNA Databank of Japan ID: DRA002630. The resulting assembly was 1,102,071 scaffolds with an N50 of 431, indicating many broken transcripts. The contigs were filtered to a minimum length of 1000-bp in order to remove most of the fragmented transcripts. Remaining contigs were then filtered for contaminants using BLAST against the NR database (Pruitt et al. 2005). Hits that were not plants were removed. This resulted in 43926 scaffolds with 2765 N50. Transcripts that did not align to the genome assembly were noted for future use. Genes were then assembled using a transcript guided gene space assembly method. The method is similar to a previously noted method, by Aluome et al (2016) with the addition of contig ordering and gap closing. Transcript guided assembly revolves around the idea that transcripts contain order information about the genome, similar to a mate-pair read. To make use of this information, the assembled mRNA sequences needed to be assembled in the context of the genomic gene space. This results in an assembly of the transcripts, which contains partial or complete 5’ promoter regions, 3’ flanking sequence, and introns.

Whole genome reads were aligned to these transcripts using BLASTN. These reads were then assembled using Velvet with a K of 51 and no other parameters (Zerbino and Birney 2008). The resulting contigs were aligned back to the original transcript using Exonerate (Slater and Birney 2005). The result was considered to be the "order" of the assembled contigs. Contigs were ordered and scaffolded together with N's separating each contig. The gaps were then filled using GapCloser 1.0. This process was repeated five times for each transcript. The result was a final assembly of 1,766,890,029-bp with an NG50 of 41,006-bp.

2.3.4 Repeat library construction.

Novel repeats were constructed according to a process whereby k-mers that have a high copy number selected to assemble a repeat and used library (Li and Waterman 2003). Jellyfish was used with k=31 to identify high copy k-mers (Kurtz et al. 2008). The 173 bp library was used for this method. The k-mers that had more than 120 copies were called repetitive, as this was roughly 6 times the expected coverage. These k-mers were then assembled using velvet to give an initial set of repeat sequences. Sequences of less than 64-bp were removed. The remaining sequences were blasted against the MIPS eudicot repeat database and NR (Nussbaumer et al. 2013). Chloroplast, mitochondria, and rRNA were placed into their own categories. Sequences with a functional annotation to plants that were not repeats were also removed from the library and marked for future analysis. The final set of repeats had an N50 of 212 and contained 9615 repeats. These repeats were then annotated using pretrained models of TEclass (Abrusán et al. 2009b). TEclass uses hierarchical classification; it classified 98% of repeats, 85.8% of the retrotransposon class and 14.2% of the DNA transposon class (Table 2.6.4). This is in accordance with other angiosperms. This library was combined with the MIPS Eudicot repeat database to create the final repeat library for use in masking.

2.3.5 Shinsuwase and Teamaker assembly annotation.

The genome annotation was performed in a multi-step fashion. First, the genome was masked using RepeatMasker along with the previously described repeat database. The remaining genome was then used (Smit et al. 1996). The RNA-seq reads described previously in the transcriptome assembly were aligned to the genome using HISAT and were assembled using StringTie (Kim et al. 2015; Pertea et al. 2015). The result had far too many genes, likely due to the unusually high volume of RNA-seq. Most of the genes were single exon with low read coverage and were thus called noise. Genes were filtered using outlier detection via one-class SVM trained using scikit-learn (Pedregosa et al. 2011). Outliers were then called genes and used as the first set of genes. MAKER-P was then run on the masked genome with the StringTie transcripts used as external information. Augustus and SNAP were used as gene finders with the provided *Arabidopsis* models (Korf 2004; Stanke et al. 2006b). Finally, the peptide sequences of the remaining genes were extracted and aligned to the TAIR10 *Arabidopsis* mitochondria and chloroplast protein sequences using BLASTP (Rhee et al. 2003; Johnson et al. 2008). Genes that matched and had an E-value > 0.0001 were removed and called pseudo-genes. After reviewing the annotation, it became clear the masking of the genome assemblies was not sufficient. Thus, another pass was made, removing all genes that contained the keywords “gag”, “pol”, “Retrotransposon”, and “Retroelement.”

The remaining genes were then scanned for functional annotations using BLASTP against a database of known hop genes, TAIR 10, and against Uniprot (Rhee et al. 2003; UniProt Consortium 2008). The annotation with the highest e-value was chosen. This gave a set of 22,201 and 16,161 annotated genes in the Shinsuwase and Teamaker annotations respectively (Table 2.6.3). The difference in total genes and annotated genes can be characterized by the difference in assembly methods. ALLPATHS-LG is known to be a conservative assembler, possibly excluding highly heterozygous genes or broken genes. Similarly, an aggressive assembler may include these genes as two separate scaffolds. In any case, it is clear much work needs to be done before the hop draft genome can be called complete.

2.3.6 21422M Annotation.

The genome assembly of 21422M was the same used in the next chapter. The genome was annotated in a simpler fashion, as the identification of complete genes is not likely in a genome with low sequencing coverage (approximately 18X). The RNA sequencing reads from the previously published transcriptome assembly were quality filtered with a mean quality of 30 and adapters were removed using Skewer (Clark et al. 2013b; Jiang et al. 2014). Reads were then aligned using HISAT to the 21422m assembly (Kim et al. 2015; Hill et al.). Reads were assembled using StringTie v1.0. The result was annotated using alignments to TAIR10 protein coding genes using BLASTX. Alignments with an e-value < 0.0001 were included. This resulted in 22,754 genes.

2.4 RESULTS

Genome assemblies for both a male and female hop accession were developed and fully annotated to the degree possible given the repetitive nature of the hop genome and the difficulties associated with said assembly. Overall sequencing depth for Teamaker was 209X prior to read processing (Table 2.6.1). Fragment library's (101-bp) had insert sizes of 143-bp, 173-bp and 250-bp. This resulted in 63.1X coverage after removal of duplicates and quality control. In addition, mate-pair, paired-end reads (101 bp) with insert sizes ranging from 3000 – 9000 bp were sequenced for an additional coverage of 46X. Sequencing libraries with insert sizes outside normal library preparation of approximately 250-bp insert size proved difficult to develop and losses due to quality control reflected this. Ultimately, the total coverage for sequencing Teamaker after removal of duplicated reads and quality control was approximately 109X.

The Teamaker genome assembly compares with that published for the Shinsuwase genome with each having their respective strengths and weaknesses (Table 2.6.2) (Natsume et al. 2014b). The Teamaker assembly has slightly higher alignment to transcriptome assembly while alignments to Public EST data is slightly higher with the Shinsuwase genome. The Shinsuwase genome also has a slightly higher alignment to CEGMA core genes than Teamaker. It is likely that the higher alignment of Teamaker with public transcriptome data is due to the use of transcriptome-guided genome assembly as an aid to assembling the genome. Finally, the Teamaker genome (with N's) is closer to actual size than that observed for Shinsuwase. Gene annotation was more successful using the Shinsuwase genome assembly with the exception of Stringtie Transcripts (Table 2.6.3).

One feature common to both assemblies is the presence of large numbers of DNA repeats (Table 2.6.4). These repeats varied in size from 100-bp to greater than 300-bp. The vast majority of repeats consisted of long terminal repeats (LTR) and retro-transposons. The next group, long-interspersed-nuclear-elements (LINE's) made up the majority of repeat sequences that were greater than 300-bp in length. Finally, large numbers of DNA transposons were observed with most ranging in size from 100 – 200-bp. It is likely that much of the missing portion of both genomes are repetitive elements. It is observed that regions on the boundary of scaffolds had a much higher copy number than portions in the center of scaffolds. This is not surprising as assemblers have a very hard time with repeat regions.

In-house development of genetic linkage maps demonstrated Teamaker as superior for use in identifying SNP markers that are map-able to linkage groups. One means to ascertain the quality of marker placement in a genetic map is to calculate the nearest neighbor (NN) fit value for all markers in a map. There was no statistical difference between average NN fit values for markers between the map developed using Teamaker versus the map developed using Shinsuwase (NN_Teamaker average fit = 0.107; NN Shinsuwase average fit = 0.109). However, the linkage map obtained using SNPs identified using the Teamaker genome covered a larger portion of the genome (761.05 cM) with a shorter average distance between markers (0.5 cM) than that obtained from the Shinsuwase genome (470.76 cM with average distance between markers of 0.7 cM). Genetic maps for a population segregating for short stature hops were made using SNPs identified using reference-guided TASSEL v 3.0 pipeline (Bradbury et al. 2007). In the case of SNP markers identified using the Shinsuwase genome, only 677 markers mapped to 10 different linkage groups (data not shown). Use of Teamaker genome assembly for SNP identification under the same default conditions as used for Shinsuwase resulted in a genetic map with 1530 markers mapped to 10 different linkage groups (data not shown). The same phenomenon was observed in the development of a genetic map for a population segregating for downy mildew resistance (data not shown). These observations are reported not as a means of accessing assembly quality but as a suggestion for use in identifying markers for linkage or association mapping studies.

2.5 DISCUSSION

It is likely that much of the missing portions of both genomes are repetitive elements. It is observed that regions on the boundary of scaffolds had a much higher copy number than portions in the center of scaffolds. This is not surprising as assemblers have a very hard time with repeat regions. The creation and unification of hop genomic resources paves the way for a complete genome assembly. The accessibility and centrality of the software is vital for the application of 3rd generation sequencing and assembly. Furthermore, it is possible to compare and contrast the different draft genomes and even ultimately repair and clean them when a complete genome assembly is available.

There are stark differences between the two assemblies. The Shinsuwase assembly ultimately was annotated to have a higher number of genes. In addition, the RNA-seq dataset had a higher percentage of alignments. Perhaps the simplest explanation is the format of the genome assemblies themselves. The Shinsuwase assembly was published with all gaps reduced to a single "N", which could cause spurious gene isoforms called from the different gene finders.

Another explanation for the discrepancy between the two cultivars is lineage. Shinsuwase was an offspring of open pollenated Saazer grown in Japan. It is possible that the male plant contains pedigree from the *Humulus lupulus var cordifolius* subspecies. Genetic distances computed from SNPs within the deep sequencing of Teamaker, USDA 21422M, Shinsuwase, and *Humulus lupulus var cordifolius* suggest that this is the case. Shinsuwase is by far the cultivar most closely related to the wild Japanese hop (Supplemental data).

We propose the discrepancy between the two assemblies as a result of the different assembly methods. The Shinsuwase assembly was performed using CLC assembly cell and the SSPACE scaffolder. In contrast, the Teamaker assembly was performed using ALLPATHS-LG. It is well known that ALLPATHS-LG is a more conservative assembler and scaffolder than the combination of CLC assembly cell and SSPACE. Groups who used CLC or SSPACE (no group used both) and participated in Assemblathon 2 performed worse in quality metrics on average than groups which used ALLPATHS-LG (Bradnam et al. 2013). In contrast, these groups performed as well or better than ALLPATHS-LG groups when measured on continuity (N50). In other words, ALLPATHS-LG will produce higher quality, yet smaller and shorter genome assemblies (conservative), while alternative methods will result in lower quality yet longer and larger assemblies (greedy).

This can explain the discrepancy in the number of genes. Perhaps the simplest explanation is the format of the genome assemblies themselves. If an assembler is more conservative about separating different haplotypes – especially large insertions or deletions – it would be less likely to duplicate genes which appear only once within the genome. On the contrary, a more conservative assembler would be less likely to correctly separate genes, which have a copy number higher than one.

While both approaches could be argued as "better," it is more useful and constructive to consider the cases in which each is useful. The greedy approach is more useful when researchers are interested in knowing *what* exists within the true hop genome. An example could be RNA-seq quantification. The more conservative method is when you need high resolution of the hop genome. An example would be researchers who are interested in the genotypes of different hop cultivars.

The final difference between assembly methods is related to the transcriptome guided genome assembly of missing genes from the Teamaker assembly. Since the target genes were directly taken from the transcriptome (which as filtered for contaminants), it is expected that the Teamaker assembly would contain a higher number of EST and transcriptome alignments. On the contrary, a less stringent filtering of contaminants from the transcriptome assembly could have provided higher scores for the Shinsuwase assembly. Since the data was generated by the same experiment, it would not be surprising for the whole genome sequencing reads to contain scaffolds from non-plant organisms.

SUPPLEMENTAL MATERIALS AND METHODS

SYSTEM IMPLEMENTATION

The server itself is a 32 AMD-x64 CPU machine with 32 Gigabytes of RAM and a 10 Gigabit/second pipe to the Oregon State University ISP. The HopBase stack consists of, Linux CentOS 6.6 final, Apache2, PHP5, Symfony2, Bootstrap3, and AngularJS. The liberal use of modern front-end libraries, specifically AngularJS 1.0 and Bootstrap provides a modern look-and-feel for HopBase while Symfony provides maintainable backend architecture using a mature MVC framework. The three assemblies available are USDA 21422M, Shinsuwase, and Teamaker (Henning et al. 2008; Natsume et al. 2014; Hill et al. 2016). Each assembly includes an annotation using the RNA sequence data provided by Natusme et al.

The BLAST web tool is implemented using SequenceServer (Priyam et al. 2015). SequenceServer is a standalone tool for interfacing with the command line NCBI BLAST. The databases included on the website correspond to each genome assembly, coding sequences, predicted protein sequences, and other specialty databases. In particular, the male specific region is a standalone BLAST database. Access to an easy-to-use BLAST interface specific to hop will greatly help the hop research community.

The resources page host’s raw data for bulk download: that is, files for genome assemblies, various annotation formats, and other processed resources (VCF, BAM, gene expression). It also includes the standardized ID format for submission from users. Downloading and accessing the raw files for bioinformatics can be a challenge, especially when there are multiple resources present as well as locations for these resources. A central location containing each of the abovementioned files provides scientists an easy starting point for working on *Humulus* genomics.

The JBrowse server is hosted on the same machine at jbrowse-hopbase.cgrb.oregonstate.edu (Skinner et al. 2009). Each genome assembly is provided as a separate tab within the front-end framework. This allows for quickly switching between contexts and allowing for the data to be loaded in the background. Each JBrowse includes the final annotations, the StringTie annotations, repeat annotations, gene expression for each available tissue type, as well as predicted motifs for known plant transcription factor binding sites. In addition, JBrowse includes

RNA-seq experimental expression data for genes and known transcripts across several different hop varieties.

The HopBase mailing list provides for rapid information regarding updates when pushed to production. If a new annotation is produced, or a new draft of the genome is available, it is easy to notify users of this information. This provides a convenient alternative to frequently checking the website for updates.

SNPs were called from 15x of whole genome sequencing reads for the cultivars Teamaker, Shinsuwase, USDA 21422M, and Cordifolius. SNPs were all called using GATK and the corresponding best practices pipeline. Co-ancestry was computed using the relatedness phi as implemented in vcftools; Large negative values indicate individuals from different populations, where as positive values within a population is an approximation of the kinship coefficient (Li et al. 2009; Manichaikul et al. 2010; Danecek et al. 2011). From these statistics, the fact that Shinsuwase, Teamaker, and USDA 21422M are from the same population is a given, and is widely accepted among hop breeders. In addition, Teamaker and USDA 21422M are clearly from a different population than Cordifolius, which again is accepted among Hop breeders. However, Shinsuwase and Cordifolius have a relatedness score of nearly 0, which indicates unrelated individuals within a population. While the sample number is low, the genotype data suggests a relationship between Cordifolius and Shinsuwase that is not shared among other cultivated hops.

2.6 TABLES

Table 2.6.1 Sequencing libraries.

Mate Pair insert Size (bp)	Number of raw reads	Number of Dedup + QC Reads	Portion lost from dedup + QC	Estimated Coverage
9000	796503434	164452668	0.793531753	6.090839556
6000	363664930	96117630	0.73569728	3.559912222
5000	830281020	611993950	0.262907455	22.66644259
3000	618181114	379821668	0.385581896	14.06746919
Mate pair Total	2608630498	1252385916	0.519906742	46.38466356
Fragment library insert size (bp)				
143	1655421082	708994796	0.571713322	26.25906652
173	1176857672	606418512	0.484713805	22.45994489
250	419621690	388494910	0.074178196	14.38870037
Fragment total	3251900444	1703908218	0.476026943	63.10771178

Table 2.6.2: Comparison of Shinsuwase assembly and Teamaker assembly.

	Shinsuwase v1 (Natsume et al. 2015)	Hopbase Teamaker v1 (current)
Transcriptome Assembly	70%	76%

alignments		
Public ESTs alignments	94%	96%
CEGMA genes	89%	85%
NG50 (without N's)	5,050	9,231
NG50 (with N's)	N/A	41,006
Assembly size (with N's)	2,049,209,000	2,770,850,934
Assembly size (without N's)	1,775,776,000	1,766,890,029

Table 2.6.3: Gene annotation for Shinsuwase and Teamaker assemblies.

	Shinsuwase	Teamaker
Stringtie Transcripts	1120693	1137597
StringTie w/ SVM Transcripts	97288	77118
MAKER genes	46735	39831
MAKER after pseudogene removal	39672	28434
MAKER after repeat removal	35482	24919
Unknown protein	13281	8758
Annotated genes	22201	16161
Total remaining genes	35482	24919

Table 2.6.4: Distribution of repeats in Teamaker assembly by length.

	100-200	201-300	301+	Total
LTR	2094	780	353	3227
Unclear	155	51	52	258
DNA Transposon	1024	240	60	1324
Retro	1533	545	212	2290
LINE	303	601	618	1522
SINE	621	62	3	686
nonLTR	235	67	6	308
LTR + Retro	3627	1325	565	5517

Table 2.6.5: Relatedness of different individuals with WGS reads.

INDV1	INDV2	N_AaAa	N_AAaa	N1_Aa	N2_Aa	PHI
USDA21422M	USDA21422M	316418	0	316418	316418	0.5
USDA21422M	Cordifolius	26110	79853	316418	53545	-0.361106
USDA21422M	Shinsuwase	237926	1821	316418	389779	0.331754
USDA21422	Teamaker	245479	18564	316418	324193	0.325238

M						
Cordifolius	USDA21422M	26110	79853	53545	316418	-0.361106
Cordifolius	Cordifolius	53545	0	53545	53545	0.5
Cordifolius	Shinsuwase	20020	20108	53545	389779	-0.0455558
Cordifolius	Teamaker	23928	55859	53545	324193	-0.23241
Shinsuwase	USDA21422M	237926	1821	389779	316418	0.331754
Shinsuwase	Cordifolius	20020	20108	389779	53545	-0.0455558
Shinsuwase	Shinsuwase	389779	0	389779	389779	0.5
Shinsuwase	Teamaker	247013	963	389779	324193	0.343273
Teamaker	USDA21422M	245479	18564	324193	316418	0.325238
Teamaker	Cordifolius	23928	55859	324193	53545	-0.23241
Teamaker	Shinsuwase	247013	963	324193	389779	0.343273
Teamaker	Teamaker	324193	0	324193	324193	0.5

Chapter 3: Genomics of the Hop Pseudo-Autosomal Regions

S. T. Hill, J. Coggins, A. Liston, D. Hendrix and J. A. Henning

Euphytica

The Netherlands

May 2016, Volume 209, Issue 1, pp 171-179

3.1 ABSTRACT

Hop is one of the few dioecious plants with dimorphic sex chromosomes. Because the entire Cannabaceae family is dioecious, hop and other members of this family are thought to have a relatively older sex chromosomal system than other plant species. Hop cones are only produced in female hops with or without fertilization. This has lead to most genomic research being directed toward female plants. The work we present provides genomic resources surrounding male plants. We have produced a draft genome for the male hop line USDA 21422M using a novel genome assembly method. In addition, we identified a 1.3Mb set of scaffolds, which appear to be the male specific region based upon specificity with male hop accessions. This set includes a smaller high confidence total length18Kb set of scaffolds, which are supported by over 500 individuals, including the USDA world collection of hop varieties and two mapping populations, with genotyping-by-sequencing. We also have identified a portion of the Teamaker x 21422M linkage map to be associated with the pseudo-autosomal region (PAR). Within the genomic scaffolds, we identified a set of genes that are sex-linked and likely located in the PAR.

3.2 INTRODUCTION

Humulus lupulus L. var. lupulus (European hop) is a dioecious (2n = 2X = 18A + XX /XY), perennial, climbing plant that is harvested for its female flowers. Its primary use is as flavoring and bittering additive in beer. Females Hops produce lupulin glands, which in-turn produce more than 1000 known essential oils (Eri et al. 2000) as well as the bittering acids responsible for beer bittering. Although males do not produce cones, they do produce lupulin glands (with much lower production) in both flowers and on leaves (Figure 1). This suggests female versus male differences in fitness and evolutionary function of these compounds. Because of the importance of the female flower, breeding and genomic work has been almost entirely focused on females. Less than 6% of all flowering plants are dioecious, and only a few of these are documented as heterogametic like hop_ENREF_6 (Ming et al. 2011). In the family *Cannabaceae*, *Cannabis sativa* (2n = 18 + XY)*, Humulus japonicus* (2n = 14 + XY1Y2)*,* and *Humulus lupulus* (2n = 18 + XY) all have heterogametic sex chromosomes.

Although hop typically has two sex chromosomes, there are six systems known to exist, spanning from one to three pairs of sex chromosomes with various sizes of the Y-chromosome. Differences primarily occur within *var. cordifolious* (Table 3.6.1) (Ono 1961). *Humulus lupulus* is furthermore one of the only plants to have flowers that morphologically diverge early in development (Shephard et al. 2000). These characteristics contribute to the hypothesis that Cannabaceae possesses a relatively well-established and presumably older sex chromosome system (Charlesworth 2015).

Hop is sometimes known to exhibit pseudo-monoecious flowering, however male flowers on pseudo-monoecious hop plants have never been reported to produce viable pollen. Hop is known to have an X:A ratio for sex determination, suggesting the structural genes are located on the autosomes, while the genes responsible for completing pollen development are in the sex determining region (SDR), that is the region specifically responsible for determining sex (Shephard et al. 2000).

Thus far, there have been several cytogenetic experiments involving the sex chromosomes of Hop, *Humulus japonicus*, and Hemp (Divashuk et al. 2014), (Grabowska-Joachimiak et al. 2011), (Divashuk et al. 2011). Additionally, a SSR marker was developed for screening male hops at a young age (Jakse et al. 2008). However no work has been done to unravel the control of sex determination in hop using genomics.

Traditionally, sex chromosomes have been studied in animals. More recent studies have focused upon the evolution of sex chromosomes and floral development in plant species. The primary difference between sex determination in plants and mammals is that morphological differences between sexes appear very late in the life cycle of plants. Most research on sexual differentiation and sex chromosomes has been done on *Silene,* papaya and *Rumex* (Liu et al. 2004); (Filatov 2005); (Hough et al. 2014). Unlike *Rumex*, the Y-chromosome is essential in hop for development of pollen in male plants (Shephard et al. 2000). Nevertheless, *Humulus* remains largely unstudied, even though plants in this family show evolutionary advanced stages of sex chromosomes, particularly, between stages 5 and 6 (Divashuk et al. 2011). Stage five is recognized by a small, degenerating Y-chromosome undergoing heavy recombination suppression that is enriched with repetitive elements. Stage 6 occurs with the loss of the Y-chromosome and an X:Autosome sex determination ratio (Ming et al. 2011).

The pseudo-autosomal region (PAR) is defined as the recombining region of sex chromosomes. Recombination within PAR does not follow normal segregation patterns as in autosomal chromosomes – portions of the region may be genetically linked to the sexual determining region causing recombination suppression of the alleles near the SDR. This leads to major differences in allelic frequency between sexes for loci in the PAR near the SDR boundry. This can cause an aggregation of genes with different fitness levels for both sexes. Genes with different allele frequencies in each sex may then gain tighter linkage to the SDR causing a cascade effect until the loci is ultimately subsumed by the SDR. Loci in the SDR undergo recombination suppression due to a lack of pairing during meiosis and are considered completely sex linked.

While several cytogenetic studies on hop sex chromosomes exist there are no studies on the molecular basis for sex determination. The objectives of this study were to identify the pseudo-autosomal region in hop sex chromosomes as well as identify male specific regions of the Y-chromosome along with putative genes located on these regions

3.3 MATERIALS AND METHODS

3.3.1 Plant Material, DNA Extraction and Library Preparation.

All accessions used in the study were maintained at the USDA-ARS Hop Breeding and Genomics program located outside Corvallis, OR. Rhizome cuttings were obtained from each accession and grown out under clean conditions in a glasshouse at Oregon State University (Corvallis, OR) with disease and insect infestations controlled with regular chemical applications. Young leaves of approximately 4 cm^2 were collected and placed under ice until prepared for DNA extraction in the lab. DNA extraction was performed immediately after leaf tissue samples were collected. Qiagen Plant DNAeasy Kits (Qiagen Inc, USA) were used with modifications to the protocol as outlined by Henning et al (2015). These modifications resulted in samples possessing high quality DNA samples with large fragment sizes of approximately 25 kb. Library preparation for genotyping by sequencing (GBS) was performed as reported by (Elshire et al. 2011). GBS sequencing was performed on the Illumina HiSeq 2000 platform (Illumina Inc) with 48 genotypes per lane. A total of **511** accessions were GBS-sequenced to a depth on average of 5X (Table 3.6.2).

3.3.2 SNP Identification.

All SNPs utilized in the study were identified using TASSEL 3 (Glaubitz et al. 2014) GBS pipeline and two different hop genome assemblies. SNP identification was performed twice, once against the variety ‘Shinsuwase’ assembly (Natsume et al. 2015) and another against the 21422M MSR (see below). Default settings for TASSEL GBS pipeline were used for SNP ID. This provided two sets of SNPs (male specific and autosomal) for further analysis. The resulting data sets provided initial SNP sets of 1,098,285 SNPs for Shinsuwase-based and 80,168 SNPs for 21422M MSR. Further filtration of the raw SNPs for both data sets was utilized so that only SNPs present in 80% of all accessions were obtained: 260,318 for Shinsuwase and 23,943 for 21422M MSR.

3.3.3 21422M genome assembly.

The genome for 21422M was assembled using a novel method called “transcriptome guided genome assembly”, which uses transcript sequences as a guide for local gene-space genome assembly. The transcripts used for this process were taken from the transcriptome assembly present on HopBase.org. Reads for the HopBase.org transcriptome assembly were acquired from the DNA Databank of Japan (DDJP) id: DRP002426. RNAseq reads corresponding to cultivar ‘Shinsuwase’ tissue types leaf, flower, immature cone, intermediate cone, mature cone, and lupulin glands. Reads were cleaned and QCed using Skewer v0.1.120 with a mean quality score required of 30 (Jiang et al. 2014). The HopBase transcriptome was assembled using SOAPdenovo-trans v1.03 with a K-value of 23 and default settings (Xie et al. 2014). Contigs smaller than 1000 were removed, as they were most likely fragmented transcripts. This resulted in a set of 37,324 contigs. Contigs were then filtered for contaminants using BLAST against the NCBI non-redundant database (NR) (Johnson et al. 2008). After removing all non-plant hits, this resulted in a remaining set of 36,808 contigs.

Our implementation of transcript guided assembly, called Cantina (Hill et al unpublished) is available at (https://github.com/hillst/Cantina). The assembly resulted in .081 Gb out of the estimate 2.8 for *Humulus lupulus cv 21422M* with an N50 of 3654. The small size is due to the focus of assembly around the known gene space. A total of 25,185 out of 36,808 transcripts were successfully assembled. This is likely due to the low coverage of genomic sequencing, however it is still a valuable resource for investigating the male hop plant and is included in this publication.

3.3.4 Male Specific Region identification.

The male specific region (MSR) is defined as regions of the male genome that do not contain alignments from any female cultivars, yet contain alignments from many or most male cultivars.

The whole genome sequencing reads from 21422M, a single lane of paired-end HiSeq 2000 with a 250bp insert size, were assembled using velvet v1.2.10 (Zerbino and Birney 2008). A K-value of 51 was used with exp_cov set to auto and scaffolding enabled. The resulting scaffolds were filtered for contigs > 200bp.

Contigs that contained alignments from sequencing reads from female whole-genome sequencing were used to filter out regions shared by both sexes. Paired-end whole genome sequencing reads from Teamaker, 21422M, and from Shinsuwase were down-sampled to 10x to match 21422M and decrease computation time. These reads were then aligned to the scaffolds from Teamaker, 21422M and Shinsuwase. Loci with no reads from the female libraries (Teamaker and Shinsuwase) were called male specific. This resulted in an assembly of 20,202,198 bp. This resulting set was used for calling SNPs as described above. These regions were then further filtered using GBS reads to identify high confidence loci and to remove loci in which GBS reads from female samples aligned. Alignments were performed using BWA v 0.7.12 with default settings (Li and Durbin 2009). The difference in number of male and female samples is due to the focus of sequencing on females and the uneven distribution of males and females within a population (Table 3.6.2). Contigs containing any female GBS alignments were removed, resulting in a set of 1.3Mb. This set is denoted our putative SDR, although it is severely limited by the fragmented MSR assembly for 21422M and has much room for improvement. Contigs containing alignments present in at least 80% of the male accessions were called male specific with high confidence due to the large number of samples. Due to limited GBS cut sites within the MSR, these contigs resulted in a small total length of 18 Kb, these loci are the best candidates for molecular male markers.

SNPs used for identifying the linkage group from Henning et al. (2015) containing the PAR region were selected from the data set consisting of 35,922 SNPs, each SNP belonged to one of the 10 linkage groups. These were ultimately chosen for use in mixed linear models (MLM) analysis in TASSEL v5.21. Kinship and Q-matrices were not utilized for MLM as the population had a clearly defined genetic make-up consisting of a full-sib family from the mapping population between ‘Teamaker x 21422M’. The statistical threshold for marker significance of 5.85 on the –log10 scale was determined by Bonferroni correction (Dunn 1961). See Supplementary Figure 1 for a workflow diagram of the above procedures.

3.4 RESULTS AND DISCUSSION

3.4.1 Male Specific Region (MSR) identification.

Male specific regions are typified by large stretches of repeat DNA and retrotransposons (Zhang et al. 2008); (Oyama et al. 2010); (Divashuk et al. 2014). These regions do not undergo recombination with the X-chromosome and therefore genes located in this region will be fixed. It is presumed that borders between MSR and PAR are regions where sexually antagonistic genes are located and are undergoing evolution (Oyama et al. 2010); (Charlesworth 2015); (Hough et al. 2014). Nonetheless, little is known about the function (if any) of the repeat DNA and retrotransposons in hop.

Our study identified a 1.3 Mb set of DNA scaffolds that appear to be unique to male hop accessions. This DNA set contains a subset totaling18 Kb in length of DNA that were validated by lack of alignment from 385 female lines present in our GBS data set as well as alignments from 80% of the 117 males making up the GBS set. The MSR (and putative SDR) identified herein provides a set of DNA useful for both the development of male markers for selection, and the exploration of markers related to sex that are shared among males and females. It may be possible to use this region as a basis for identifying molecular mechanisms for sex determination as proposed by Zhang et al. (2008). In addition, while several publications have cited the identification of "male markers" (Polley et al. 1997; Seefelder et al. 2000; Danilova and Karlov 2006; Jakse et al. 2008; McAdam et al. 2013), most have been identified by means of segregating loci—meaning recombination with the X-chromosome. A preferable marker system would be one utilizing a male marker located on the MSR of the Y-chromosome where no recombination occurs and marker evaluation could be a simple inexpensive PCR "presence/absence" of the marker.

Divashuk et al. (2011) identified the long arms of both the X and Y-chromosomes as the PAR for hop sex chromosome. It follows that the MSR we've identified would reside upon the short-arms and potentially covers the centromere. The cytogenetic research by Divushuck et. al (2011) identified the regions showing X-Y pairing to be external to the centromere. Thus, linkage maps in *Humulus* species would not show markers from the MSR as one of the linkage groups but would only show markers present in the PAR. Linkage maps are developed through genetic marker data for loci segregating in the population. Linkage distance between loci is calculated based upon the recombination rate in the population. If X- and Y-chromosomes do not pair and undergo chiasma, no recombination will be possible. Thus, only a portion of the long arms of the X- and Y-chromosome pair and undergo chiasma. Those regions not pairing would be considered the MSR on the Y-chromosome while those regions that do pair, show recombination and are thus the PAR.

3.4.2 Pseudo-autosomal Region (PAR) identification.

The natural follow-up to identifying the MSR was to attempt to find sex-linked SNPs in the female genome assembly. The only linkage group in Henning et al (2015) that contained sex-linked SNPs was also the clearly sex enriched linkage group, linkage group 4 (LG4) (Figure 2). Only LG4 contained GBS markers that were significantly associated with sex. To explore this relationship further, SNPs hypothesized to be present within the PAR were identified by performing a mixed linear model (MLM) by using the TASSEL v5.21 GUI on the sexual phenotypes of all the GBS individuals previously mentioned. The SNPs were also tested against the Teamaker x 21422M linkage map (Henning et al. 2015). LG4 was statistically enriched for sex-associated SNPs and thus concluded to be the pseudo-autosomal region—presumably carrying alleles from the X and Y-chromosome (Figure 3.7.2).

The identification of the PAR opens the door for further sex chromosome studies in *Humulus*. *Humulus lupulus* is in an advanced stage of sex chromosome evolution showing relatively small estimated PAR sizes compared to other dioecious plants with heterogametic sex chromosomes (Divashuk et al. 2011Divashuk et al. 2011). The size of Linkage group 4 after including genomic scaffolds is only 5Mb, however this number is much lower than expected due to the fragmented genome assembly. The nature of the genes within the PAR (Supplementary Table 3.6.1), in addition to the identification of cytogenetic markers, may pave the way for understanding the unusual distribution of sex determination in the *Humulus* genus.

3.4.3 Sexually antagonistic selective genes.

In addition to observing an increase in pairwise diversity across the PAR, albeit missing the SDR genes, we also expect to observe genes acting in a sexually antagonistic fashion as we near the SDR boundary (Hough et al. 2014; Otto et al. 2011). To identify sexually antagonistic genes we first identified markers from the overall pool of all GBS markers that were > 95% homozygous in females and at least 50% heterozygous in males (Figure 3.7.3). These markers were then scanned for flanking genes to identify genes located nearby with the presumption that they act as sexually antagonistic genes (Supplementary Table 3.6.2).

Not every gene contains the ApeKi cutsites used by GBS; many genes did not even have the possibility of being identified in the previous analysis. To explore potentially excluded genes, the PAR (LG4) was also scanned for genes. These genes were then added to our list and are putatively sex-linked, but more specifically noted as PAR genes (Supplementary Table 1). Although some of the genes identified in this step showed homology to other plant species, most of the genes had unknown function, likely due to the lack of quality annotation for hop. Specialization occurring within this region could be particularly interesting to plant breeders. If there are genes associated with any of the flavoring components, favorable alleles should be fixed on males (on the haploid X chromosome). This sort of information would allow for a nearly guaranteed inheritance of a desirable allele by selection and utilization of male parental lines possessing the desirable alleles. These regions may also be of interest for genetic engineering. If a locus is tightly linked to sex, the trait will recombine less frequently and show little change from parent to offspring. In particular, the WRKY1 transcription factor, known to be responsible for the last step of prenylation in the Xanthohumol pathway and involved in disease resistance (Majer et al. 2014), is located on LG4 (HL.SW.v1.0.G043711). Additionally a WRKY domain binding protein also exists on the PAR (HL.SW.v1.0.G020812). This further suggests female-specific specification occurring within the PAR.

One of the sex-linked genes identified within the PAR region is annotated as Acetyl-CoA carboxylase 1 (Supplementary Table 1). This gene codes for a protein that helps catalyze the first step of the humulone biosynthesis pathway. Interestingly, humulone is a compound produced predominantly in females, with trace amounts being found in male flower. This suggests that there is some specialization occurring in the PAR involving the bitter acid biosynthesis. We then used the whole genome sequence alignments described earlier to try and identify a copy number variation occurring within the genome. However, the results showed evidence of only one copy (Figure 3.7.4). Interestingly, there was a sharp spike near the middle of the gene, showing 7 copies in females, and 14 copies in males. We then took this gene and looked for conserved protein coding domains through InterProScan5.

The peak at about 2000bp in Figure 3.7.4 corresponded perfectly with the biotin-lipoyl attachment domain, which is known to be critical for the function of Acetyl-CoA carboxylase 1 (Russell and Guest 1991). This suggests that the biotin-lipoyl coding domain is either extremely important for male hops, or that the region is duplicated many times on the Y-chromosome—potentially in the MSR. The latter has been observed in humans (Skaletsky et al. 2003) where there is a set of genes that are palindromic and high copy number. Although there is no direct evidence, the tight linkage of these genes with the male sex suggests it is near the MSR boundary. By further analyzing the gene family containing the biotin-lipoyl attachment domain, it may be possible to phylogenetically unravel the evolution of sex chromosomes in the *Cannabaceae* family.

3.5 CONCLUSION

The work described in this study offer a beginning for the understanding of dioecy mechanisms in *Humulus*. Unfortunately, the hop genome assembly is quite rudimentary and much of it is not assembled with only 1.8 Gb out of 2.7 Gb assembled and annotated (http://hopbase.org/). Most of the assembly resides around gene space with little or no information covering large repetitive regions that could potentially be responsible for gene regulation (Hill, unpublished data). With this in mind, it follows that the MSR for the Y-chromosome would not be included in the current hop assembly due to the theoretical presence of large regions of repetitive DNA that cannot be assembled into scaffolds using current short-read, massively-parallel sequencing. GBS data was obtained using these rudimentary genome assemblies and as such also are missing potentially a large number of SNPs and alignments that cover the whole genome. New attempts at sequencing using third generation sequencing technology are planned with the hope of covering the remaining genome and ultimately unraveling SDR and identifying sexually antagonistic genes. The results of this study, including the limited MSR and sex-linked genes, are available at http://resource-hopbase.cgrb.oregonstate.edu/HopBase/v1.0/IHS/.

3.6 TABLES

Table 3.6.1, List of sex chromosome systems in *Humulus lupulus*

Name	**Sex Chromosomes**	**Description**
Winge	XX/XY	2:1 X-Y size ratio
New Winge	XX/XY	1.25:1 X-Y size ratio
Heteromorphic	XX/XY	Very small Y-chromosome
Sinoto	X-A-A-X/X-A-A-Y	14:12:10:7 XAAY size ratio
New Sinoto	X-A-A-X/X-A-A-Y	13:11:10:3 XAAY size ratio

Table 3.6.2. List of hop accessions utilized for genotyping-by-sequencing (GBS)

Type	Males	Females
Unnamed Cultivars	72	119
Named Cultivars	0	138
Dwarf pop	27	64
Downy Mildew pop	18	73
Total	117	394

3.7 FIGURES

Figure 3.7.1 Male and female flowers. Male flowers (left) typically shed pollen prior to female flowers (right) are receptive for pollination.

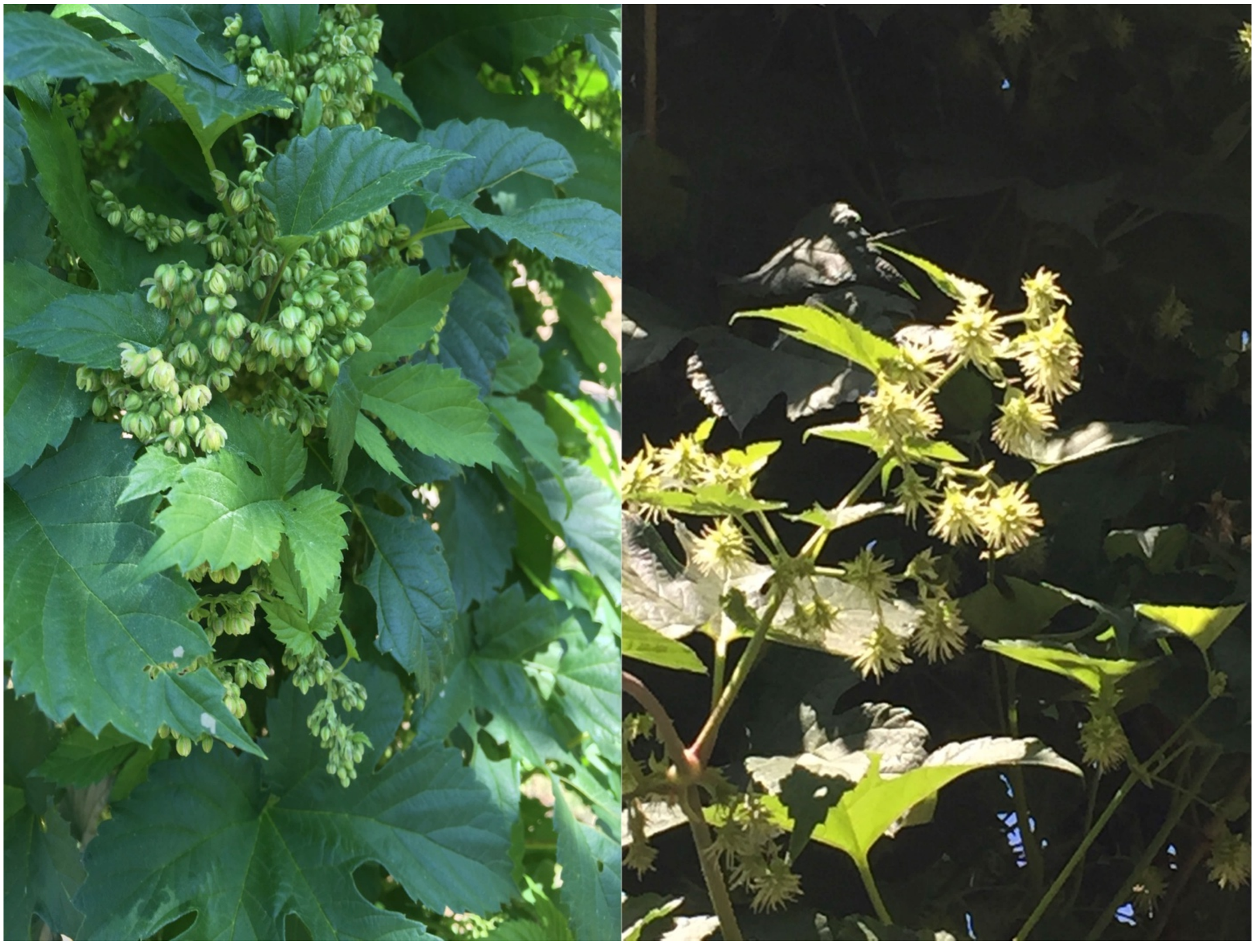


Figure 3.7.2 Manhattan plot of mixed linear model analysis from TASSEL 5.21 showing markers with significant association with sex (females coded as "0", males coded as "1"). Linkage group 4 was saturated with highly significant markers for sex.

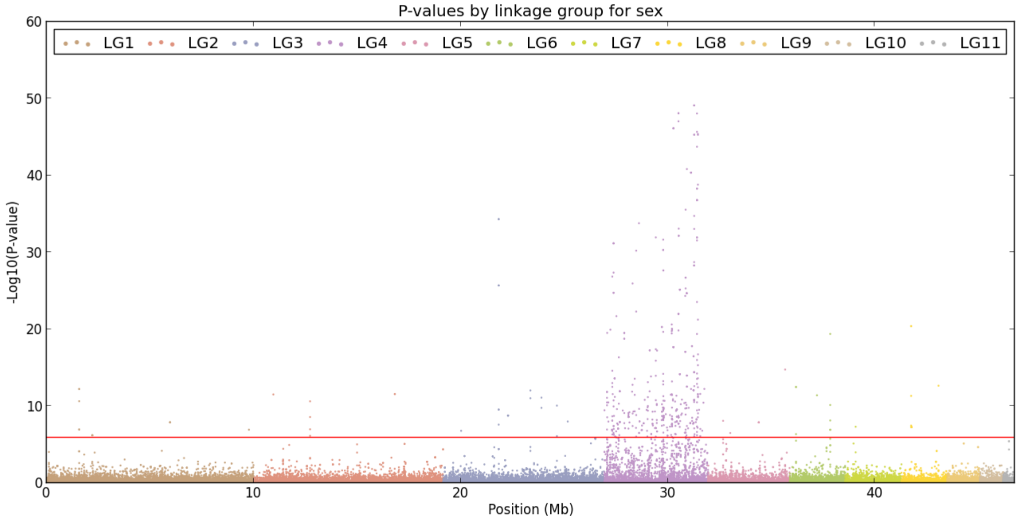


Figure 3.7.3 Sex linked markers found on LG4 from mapping population "Teamaker x 21422M" segregating for downy mildew resistance (Henning et al. 2015) as observed across the USDA-ARS world collection of hop germplasm

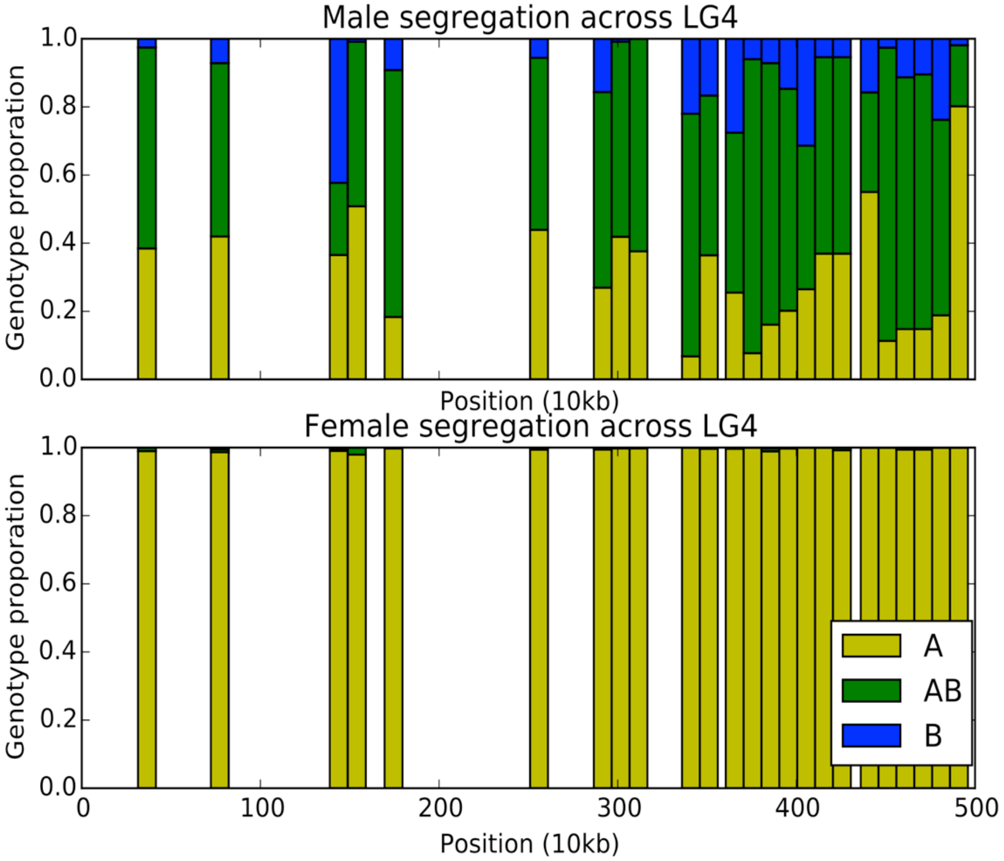
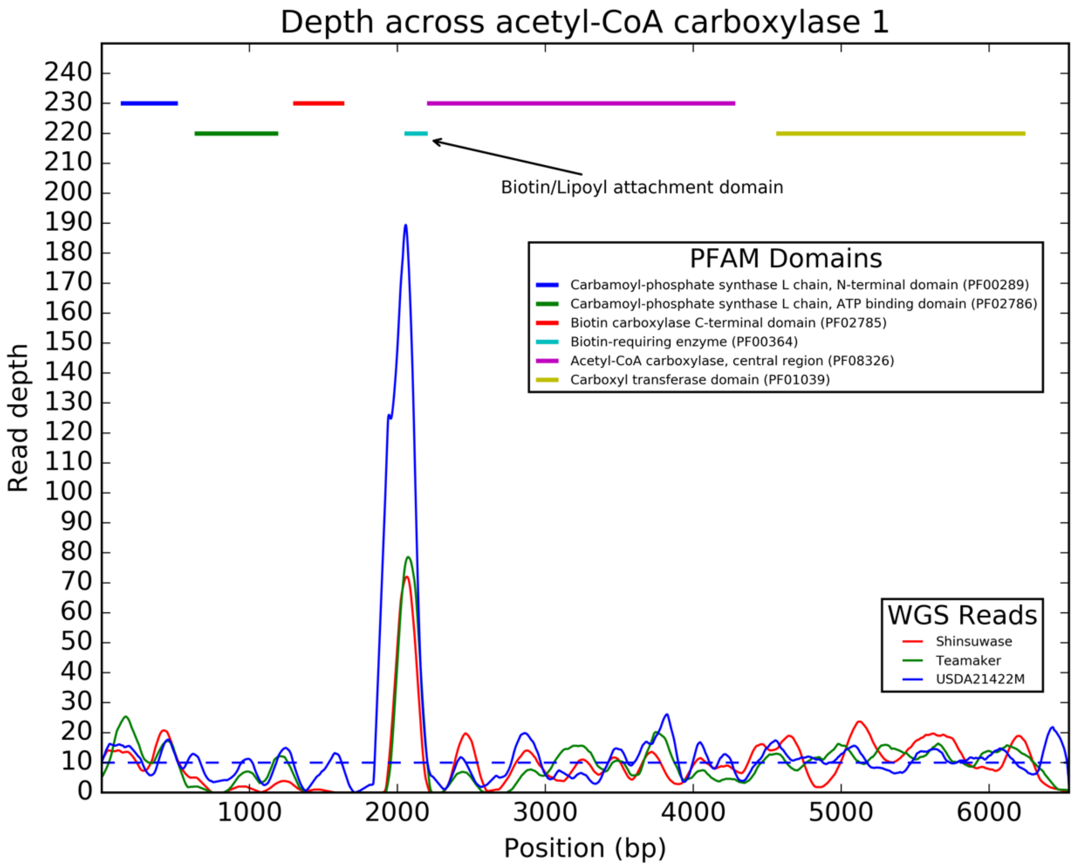


Figure 3.7.4 Depth of read coverage across the gene space of Acetyl-CoA carboxylase 1


Depth across acetyl-CoA carboxylase 1
Read depth
Position (bp)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
1000
2000
3000
4000
5000
6000
Biotin/Lipoyl attachment domain
PFAM Domains
Carbamoyl-phosphate synthase L chain, N-terminal domain (PF00289)
Carbamoyl-phosphate synthase L chain, ATP binding domain (PF02786)
Biotin carboxylase C-terminal domain (PF02785)
Biotin-requiring enzyme (PF00364)
Acetyl-CoA carboxylase, central region (PF08326)
Carboxyl transferase domain (PF01039)
WGS Reads
Shinsuwase
Teamaker
USDA21422M

Chapter 4: Featureless RNA coding prediction with Deep Learning

Steven T. Hill, Rachael Kuintzle, Erich Merrill III, Medisa Danaee, John Henning, David Hendrix

4.1 ABSTRACT

Differentiating protein-coding transcripts (mRNAs) from long intergenic noncoding RNAs (lincRNAs) is an area of growing interest in bioinformatics. LincRNAs have been shown to play an important role in gene regulation, however they are difficult to distinguish from classical mRNAs due to the presence of spurious sequences that bear similarity to open reading frames. Traditionally, lincRNA identification is performed by manually selecting features to consider during classification. This "feature engineering" leads to bias and poor performance on lincRNAs that are unusual. In the past year, natural language processing has made large strides by using deep neural networks and recurrent neural networks. Natural language and RNAs are both are represented by a sequence of characters; hence recurrent neural networks are a natural fit for analyzing RNA sequences. Here we present a coding prediction tool, DeepLinc, which uses recurrent neural networks with gated recurrent units (GRUs) for non-biased classification and modeling of lincRNAs and mRNAs. Neural nets bypass the bias introduced by manual feature selection. This is the first reported use of recurrent neural networks for DNA or RNA analysis. DeepLinc achieved state-of-the-art scores in accuracy in the classification of protein-coding RNAs. It also distinguished unusual lincRNAs from mRNAs, previously difficult due to bias. Furthermore, DeepLinc allows for identification of novel sequence features that may be biologically important for distinguishing lincRNAs from mRNAs in the cell, and could shed new light on the process of translation.

4.2 INTRODUCTION

The portfolio of features that distinguish coding transcripts (coding RNA sequences) from true noncoding transcripts is currently incomplete. By what physical mechanism do RNA sequences dictate whether or not it will be translated into a functional peptide or protein product? Some research groups have identified elements, such as the Kozak sequence and length of the open reading frame (ORF), to be important for translation in eukaryotes (Zur and Tuller 2013). However, these alone are insufficient for accurate RNA classification. Bioinformatics software cannot yet always unambiguously determine whether RNA encodes a protein based on sequence alone. Such ambiguity presents a problem for researchers trying to classify transcripts in organisms without extensive proteomic data (databases of real, observed protein and peptide sequences), and occasionally in well-annotated transcriptomes. For example, Wilhelm, et al. recently found 430 high quality peptides in massive proteomic data matching open reading frames (ORFs) from transcripts annotated as lincRNAs (Wilhelm et al. 2014). In addition, Calviello et al. discovered a number of actively translated ORFs in the annotated human ncRNA database. These examples suggest that prediction methods could be improved.

One such program, which attempts to classify transcripts as either lincRNAs or mRNAs is **C**oding **P**otential **A**ssessment **T**ool (CPAT). CPAT is among the newest software used in classification of protein coding genes (Wang et al. 2013). CPAT relies on a linear classification model using ORF coverage, ORF length, FICKETT score – which uses the combination of GC content and codon usage, and finally, Hexamer frequency. These features are thoroughly discussed in the CPAT publication. It is obvious that this set of features would struggle with extremely long lincRNAs, which may have a non-coding reading frame by chance, and short mRNAs that may contain a small ORF.

Natural language processing (NLP) and biological sequence analysis have much in common. Both fields rely on standard algorithms such as edit distance, grammar parsing, and language modeling. However, the two fields often innovate independently of each other. Transcript classification and modeling is similar to sentence modeling in classification; therefore, NLP methods may be valuable for application to biological sequence analysis. Traditionally, sentence classification has utilized bag-of-words (BOW) and n-gram models. BOW models can be simply a binary vector representing the presence or absence of words in a given vocabulary, or a vector of counts of the words. An n-gram model is equivalent to a k-mer model in bioinformatics, where a sentence of length n (or k) is the unit of analysis. That is, n-consecutive words are represented together. Often generative models such as Naïve Bayes (NB) or Hidden Markov Models (HMMs) are used with both representations.

Beginning in 2014 however, higher order models began to gain popularity. The landmark publication of the field was Alex Graves' dissertation, published in 2008 (Graves 2012). Recurrent neural networks with long short-term memory (LSTM) schemes couples with gradient learning began to gain popularity over more simple models such as Hidden Markov Models (HMMs) (Hochreiter and Schmidhuber 1997). These models are used for tasks such as machine translation, parts-of-speech tagging, sentiment analysis, and language modeling. Perhaps the most important invention was that of word embeddings. Word embeddings take a one-hot word vector, or a sparse vector in which each word has a unique vector of all zeros except in index *i*, where *i* represents the word, and converts them into a lower dimensional dense vector. The training of these embeddings allowed word vectors to have an idea of context. An important example involves using vector math with embedding vectors, "king" – "man" + "woman" ≈ "queen". The combination of LSTMs and word embedding lead directly to state of the art accuracy in sentiment analysis and other sequence labeling tasks.

In bioinformatics, we often also want to label and model sequences. Grammar parsing is used to predict RNA folding and protein structure prediction. Approaches for part of speech tagging are similar to gene finding. Finally, sentiment analysis is similar to transcript classification. Sentiment analysis performs nearly as well with character level models as with word embedding models. We can expect transcript classification to also succeed using character (nucleotide) level models.

Although neural networks have been gaining popularity in bioinformatics, they are typically shallow networks using a convolutional layer instead of a multiplication layer. Recently, DeepBind has been developed as a general-purpose model used for detecting RNA – DNA binding motifs in a sequence (Liu 2012; Bahdanau et al. 2014; Pennington et al. 2014; Dyer et al. 2015). A model was trained for each motif with a single layer convolutional neural network to predict if a transcript would bind or not to a given DNA sequence. The authors then perturbed the DNA sequence through mutation, under the model that mutations that disrupt the score indicate information relevant to the binding motif. DeepSEA (Zhou and Troyanskaya 2015) used a similar idea to predict the effect of noncoding variants, that is, single nucleotide variation in the DNA sequence outside of translated gene regions. The authors train a convolutional neural network to predict the chromatin state of each noncoding sequence, and then perturb the model to identify the effect of a single SNP. This idea of training a model, and then perturbing the input data provides powerful means of understanding the high level features represented in a neural network model. Although HMMs have had much success with applications in bioinformatics, recurrent neural networks have not been frequently used in bioinformatics, and they have never been used at the nucleotide level.

4.3 RESULTS

4.3.1 Network model.

The DeepLinc model can be simply described a Recurrent Neural Network (RNN). However, the model has 3-layers, and the recurrent portion is layered through time. The network can be described as follows:

$$P(s = mRNA \mid s) = sigmoid(NN(GRU(Embedding(s)))$$

$$Score(s = mRNA \mid s) = NN(GRU(Embedding(s))$$

The embedding layer, or embedding, is a one-to-one mapping of a one-hot encoding of a transcript to a higher dimensional vector. This layer enables the model to learn more weights associated with each nucleotide. The embedding layer is followed by the recurrent layer, which is referred to as the GRU.

A recurrent neural network (RNN) can be thought of as a conventional feed-forward neural network with some modifications that allow it to accept a variable-length sequence input and place a dependence on the concept of time. This is made possible by the presence of a recurrent hidden state whose output, sometimes called ”activation”, at each time depends on the previous time’s recurrent hidden state.

The Gated Recurrent Unit (GRU) is a modification to recurrent neural networks that allows for the gradient to travel farther through time (Chung et al. 2014). The GRU computes the next hidden state, h_{t+1}, by interpolating between the candidate hidden state, $\widetilde{h_t}$, and the previous state, h_{t-1} (Equation 1). Two “gates” control this interpolation, by scaling the contribution from the hidden and candidate hidden states. The first is an update gate, *z,* which directly controls the interpolation of the previous state and candidate state. The second “gate”, *r*, is a reset gate, which is used to compute the candidate hidden state by interpolating between the previous state and a set of network weights. Each portion of the gated recurrent unit has their own set of network weights that are shared through the time dimension. The final layer of the network is the classification layer. It is simply a fully connected neural network, where the weights of the final hidden state in the RNN are fully connected to a single output. This output is put through a sigmoid activation function in order to represent the final output as a probability, ranging from 0 to 1. The "score", noted from here out, is simply the output before the activation function, it is bounded by 32-bit floating point precision between -19 and 19 and will more accurately represent changes in the networks prediction.

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$

$$\tilde{h}_t = \tanh\big(W_h x_t + U_h (r_t \times h_{t-1})\big)$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times h_t$$

$$s_t = W_s h_t + b_s$$

$$p_t = \sigma(s_t)$$

Equation 1: Recurrent neural network with Gated Recurrent Units. Here the symbol "×" denotes element-wise multiplication.

4.3.2 Training DeepLinc.

For training, we developed a implementation strategy such that DeepLinc used equal sized samples of mRNAs and lincRNAs. Transcripts in the training set were required to be *at most* 1000 nucleotides in length and the length distribution between mRNAs and lincRNAs must be as similar as possible. These restrictions force the model to learn something other than transcript length and enables to model to be trained efficiently (Figure 4.6.1). To train the models, a voting ensemble of five classifiers were trained using 10,000 randomly selected mRNAs and lincRNAs from the Ensembl human annotations v38 release 82. The testing and validation select were sampled before training was sampled. The test set contained 200 transcripts of each class and the validation set contained 800 transcripts of each class. The model was trained with stochastic gradient descent using the ADAM update algorithm (Bottou 2010; Kingma and Ba 2014).

4.3.3 State of the art accuracy – performance metrics.

DeepLinc achieved a 6% improvement on true positive rate among human samples. While this may not seem like a large number, when accuracy is high, it becomes much more challenging to improve. Performance on mRNAs with short ORFs and lincRNAs with long ORFs improved drastically. This improvement is clearly seen in Table 4.5.1, as accuracy on mRNAs with short ORFs went from 11% (CPAT) to 80% (DeepLinc). DeepLinc also performed consistently, almost every individual model in the ensemble outperformed CPAT.

DeepLinc has diminished performance on other organisms when trained on humans. This can be expected to some degree since different species have differences in codon usage, and other sequence features that may affect accuracy. One of the fundamental assumptions of machine learning models is that the training and test datasets are independently and identically distributed. Deep learning models are sensitive to very minute differences between datasets, and the rules for translation may change quite a bit among organisms. However, these types of models were designed for use on larger datasets, and a linear or rule-based model would be much better for any non-model organism.

4.3.4 Model learns length.

Early on during development of the DeepLinc implementation, it was apparent that the model had learned that longer transcripts were mRNAs and shorter transcripts were lincRNAs (Figure 4.6.1). While this may seem remarkable considering that DeepLinc was only trained on sequences shorter than 1000 nucleotides and there is a true difference between mRNAs and lincRNAs, there is a significant variation of lengths between the two classes of transcripts even when restricted to 1000 nucleotides or less. The model could represent this difference by simply scaling all of weights in the model down. This sort of transformation would favor longer sequences to be classified as mRNAs. Since learning length as a feature could bias classification toward longer sequences, the length normalization during training must be enforced. Length normalization improved the classifiers accuracy on the test set for sequences of varying lengths.

4.3.5 Interpreting DeepLinc.

In order to interpret what the DeepLinc model is “thinking” while reading a sequence, we computed the score at each position during prediction. We selected transcript ENST000379359 for its compact length, presence of a single open reading frame, and the fact that it was in our testing set. By mutating individual features of the transcript, we can ask the model what opinion it has on different mutations. While no individual mutation lowered the score enough for the model to think the transcript was a lincRNA, it clearly has a brief dip in score for each of the major mutations added. To investigate this further, we systematically mutated every base in the transcript to each other possible base.

4.3.6 Model learns the reading frame.

The mutations that decrease the score the most both introduce an in-frame stop codon near the second methionine codon. In addition, mutations occurring after the true stop codon and before the initial start codon appear to have little effect on the score of the transcript. After applying one of the mutations, the stability of the transcript decreases drastically, after applying the second mutation; the transcript is no longer predicted to be an mRNA.

4.3.7 Shuffling Analysis.

To measure the effect of different mRNA regions on coding score, we shuffled different parts of 10,000 mRNAs in the 5’ UTR, CDS, 3’ UTR, 5’ UTR + CDS, 3’ UTR + CDS, and 5’ UTR + 3’ UTR. We compared this to a background model of a randomly shuffled mRNAs. The model is clearly more sensitive to the shuffling of the whole sequence than any of the individual parts of the transcript. However, shuffling a single UTR along with the CDS has a higher impact than shuffling the entire sequence (Table 4.5.2).

4.3.8 Computational performance.

In order for DeepLinc to be useful, the model must scale well to longer sequences and larger datasets. A prediction from an RNN can be done in linear time O(n), where n is the length of the transcript. In addition, the use of a graphics processing unit (GPU) speeds up the matrix multiplication by more than 100 fold. This allows the model to perform each multiplication in parallel, placing the upper bound of performance purely on the length of the transcript. In addition, prediction can be done in batches, further parallelizing the predictions. Batches can be scaled until the VRAM is entirely used on the hardware being used. For the purposes of our experiments, all predictions and training were done on two NVIDIA GeForce 980s

4.4 METHODS

4.4.1 Network description.

The DeepLinc model was implemented using Passage (https://github.com/IndicoDataSolutions/Passage), a recurrent neural network library built on top of the expression language, Theano (Team et al. 2016). For input, the DeepLinc accepts a sequence of indices, corresponding to sparse one-hot vectors where A = 1, T = 2, C = 3, G = 4, N=0. In this case, the zero-vector is used to represent a no-op. It can be used to train in batches with variable length transcripts. These vectors are transformed by the embedding layer into a 256 dimensional sparse vector. Finally, these vectors are passed into the recurrent layer. The recurrent layer is a gated recurrent unit with a hard-sigmoid inner activation (gate transformation) and a *tanh* outer activation. To get the final score and coding probability, the hidden state within the GRU is passed into a fully connected neural network. The score is given by the output of this linear equation, and the coding probability is given by the sigmoid activation applied to the score (Equation 1). The output of each layer passed through a dropout layer, which randomly chooses to ignore (replace with a 0) certain network weights. Dropout has been shown to reduce overfitting (Srivastava et al. 2014).

4.4.2 Network parameter selection.

For hyper parameter tuning, each model was trained for 30 epochs on the same dataset. We optimized the following parameters:

- Embedding layer size, s $\in$ [5, 12, 128, 256, 512]

- Size of hidden layer, h $\in$ [64, 128, 256, 512, 1024]

- Dropout probability, pd $\in$ [0.1, 0.2, 0.3, 0.4, 0.5]

After many days of training each model, we found the best results were:

- Embedding layer size, 256

- Size of hidden layer, 512

- Dropout probability, 0.4

In particular, we observed that the sizes of the embedding and hidden layers would do no better or would decrease in performance from the values we have selected. Dropout values lower than 0.2 would often not converge, however in some cases they performed well. A dropout probability value of 0.4 had the best accuracy on the test set while maintaining time to convergence and train accuracy.

4.4.3 Training – Dropout, mini-batches, updater.

During full training, each model was trained for 100 epochs. The model was then evaluated against a small subset of the test set, called the validation set, in order to evaluate the model after various epochs. The model with the best accuracy was kept. Mini-batches of size 32 were used during training. Although larger mini-batches can speed-up the computational performance, they also increase the bias of the model. The ADAM update rule was used during stochastic gradient descent (Bottou 2010; Kingma and Ba 2014).

4.4.4 Performance for different sized datasets.

To understand how much data was needed for good performance, we evaluated the model on datasets of size 1000, 2500, 5000, and 10000. Large datasets performed better in general compared to smaller datasets. This aligns with common machine learning theory. To learn a high dimensional representation, the model needs large amounts of data (Table 4.5.3).

4.4.5 Pre-training evaluation.

If DeepLinc is going to be useful for genome annotation, either there must be a way to incorporate outside datasets or useful features must be extracted from the model. To circumvent the curse of dimensionality on the Zebrafish dataset, we pre-trained the model using the weights from the best Human model. After loading these weights, training continued for 20 epochs using Zebrafish data. The result did not improve accuracy, however it did find a better balance between true positive and true negative rate (81% and 82%). Although the performance on zebrafish using pre-training alone is poor, accuracy greatly improves when the score from DeepLinc is used as a feature in the logistic regression classifier. In order to effectively evaluate this, we trained the logistical regression classifier on the test set for Zebrafish, holding out an additional 800 transcripts to be used when testing the logistic regression classifier. It is important to consider that the classifier was only trained seeing 800 transcripts in both instances, as this could account for the poor accuracy using logistic regression. Accuracy without using DeepLinc's score as a feature (DLScore) is 83.5%, accuracy when using DLscore is 87.7%

4.4.6 K-mers analysis.

K-mer models correspond to using a k-sized window of nucleotides instead of single nucleotides. These types of models are useful when you need to capture contextual information within your model, and they are a staple in traditional lincRNAs classification. However, recurrent models handle this with the way the model is constructed itself. The hidden state captures the relevant contextual information. Without surprise, the model performed worse on 3-mers and 6-mers than it did on single nucleotides.

4.4.7 Ensembling.

The stochastic nature of training neural networks lends favorably to ensembling. An easy way to improve a model is to build a small voting-ensemble of slightly different models. To do this, you simply train several models and when making predictions choose the class label which had the majority of votes. Ensembling the DeepLinc model increased the accuracy from roughly 90% to

93%. We also explored ensembling up to 20 classifiers. The results peak at about 94% accuracy with an ensemble size of five.

5.1 DISCUSSION

Transcript classification is perhaps the simplest use of recurrent neural networks in bioinformatics. Recurrent neural networks are a natural fit for sequence analysis. Similar to how we read a sentence or spell words, the RNN-based models look at the sequence throughout the time dimension. The success of DeepLinc can be used as a model for other applications of recurrent neural networks, both in discriminative and generative contexts. While this paper explores using RNNs as a discriminative classifier, much work has been done using them as a generative model. In particular, adversarial generative networks have been gaining much popularity (Goodfellow et al. 2014; Radford et al. 2015).

Discriminative versions of RNNs can also be used for sequence-to-sequence learning. This is a natural fit for *ab initio* gene finding. It is easy to imagine a model which simply takes in a genic region, and outputs a sequence of equal length labeling each nucleotide with exon, intron, UTR, or intergenic. A natural follow-up to the transcript classification done in this paper would be to try and classify mRNAs based on their functional annotation. The underlying model would represent features that discriminate the different functional classes. Taking this even further, sequences of these transcript vectors could be used to model time course gene expression.

Although DeepLinc achieves state of the art accuracy on our human test dataset, most organisms, including model organisms, lack the quality and quantity of data to make deep learning viable as a stand-alone classifier. We have shown that combining deep learning with a logistic regression classifier can help improve this performance – even in the presence of limited data. Although we trained the two models separately, it is also possible to train this sort of network end-to-end, using the standard CPAT features combined with the score from DeepLinc to produce a final classification module as the last layer of the network.

In addition, CPAT is not designed to explore new features for transcript classification. To circumvent this, we have put together a python package based loosely on CPAT. The software contains both a command-line front end with pre-trained models, and a programmatic backend for adding new features to the classifier. This model could be combined with downstream work to further interpret the models produced by DeepLinc.

While this work used recurrent neural network models that are commonly used by data scientists, their use in practice is rarely applied to something as long as an RNA transcript. The long-term dependency problem, while solved in some settings, could still cause problems for something as long as transcripts (Hochreiter and Schmidhuber 1997). The computational performance is bounded by the length of the transcript, and parallelism is mostly lost when working on a single long transcript. Future work should be done to explore more efficient ways to train and use RNNs for very long sequences.

4.5 TABLES

Table 4.5.1: Performance of DeepLinc and CPAT on different datasets when trained on Human.

Test Set	DeepLinc Accuracy	CPAT Accuracy
Human mRNAs	**0.931**	0.873
Human lincRNAs	**0.949**	0.943
Human mRNAs with short ORFs	**0.805**	0.111
Human lincRNAs with long ORFs	**0.837**	0.91
Mouse mRNAs trained on Human	0.917	**0.951**
Mouse lincRNAs trained on Human	0.814	**0.915**
Zebrafish mRNAs trained on Human	**0.93**	0.845
Zebrafish LincRNAs trained on Human	0.62	**0.915**

Table 4.5.2: Results of shuffling different parts of mRNAs.

	Mean	Median	Std	Z-score
AllBases	-9.806693702	-10.55759062	5.300471231	-1.850155066
CDS	-11.38982757	-11.85603901	6.335669261	-1.797730768
CDS+3'UTR	-11.95715383	-12.42043036	6.038115398	-1.980279117
5'UTR+CDS	-11.2854429	-11.74899201	6.184087012	-1.824916577
3'UTR	-0.904051686	-0.6787737	2.580438923	-0.350348027
5'UTR+3'UTR	-0.66165051	-0.379995	3.251621147	-0.203483272

5'UTR	-0.096465968	0	2.735610334	-0.035263051

Table 4.5.3 Accuracy based on dataset size.

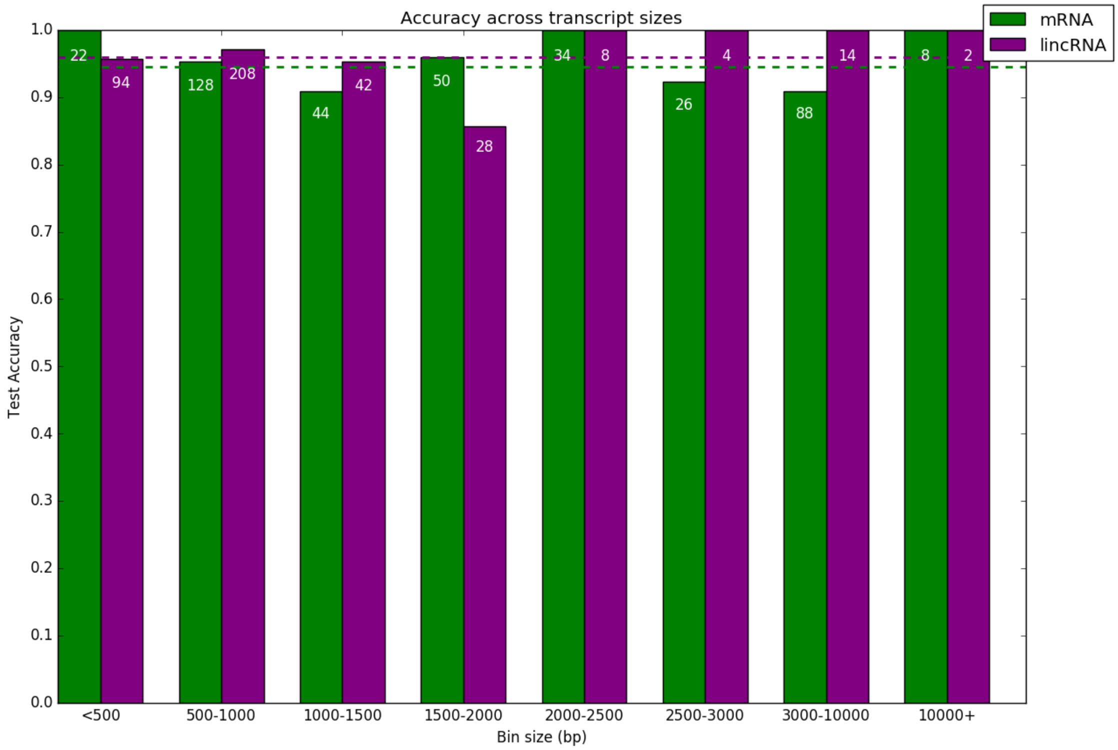
Training Size (per class)	**Accuracy**
1000	0.565
2500	0.8075
5000	0.84
10000	0.92

Table 4.5.4 Model accuracy when using K-mers.

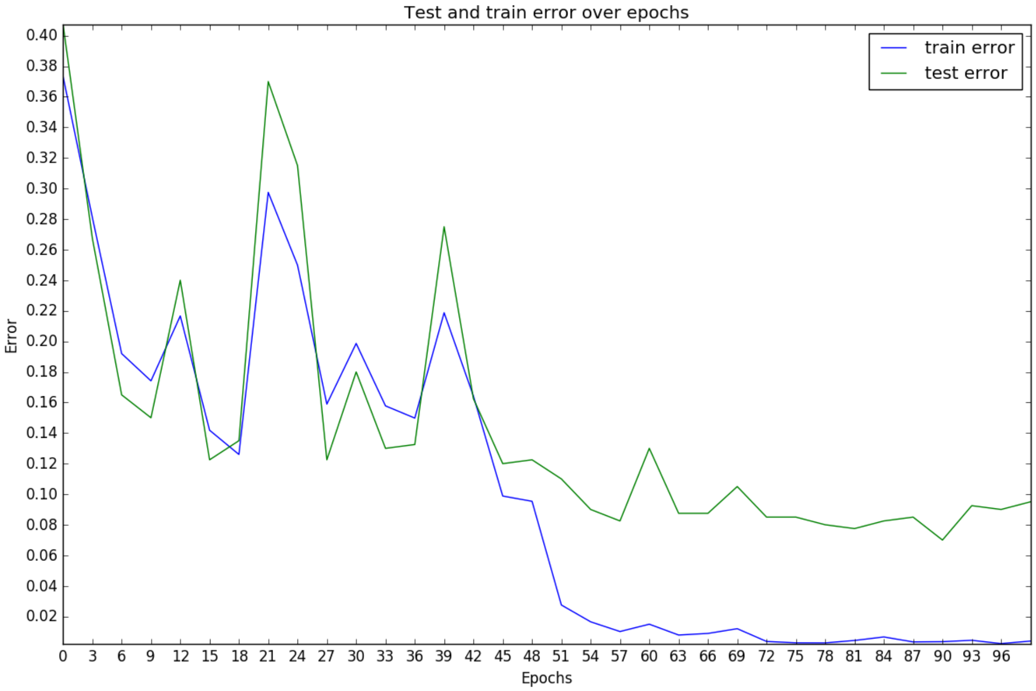
Size of k-mers	**mRNA Accuracy**	**lincRNA Accuracy**
k = 1	0.88	0.92
k = 3	0.85	0.83
k = 6	0.75	0.76

4.6 FIGURES

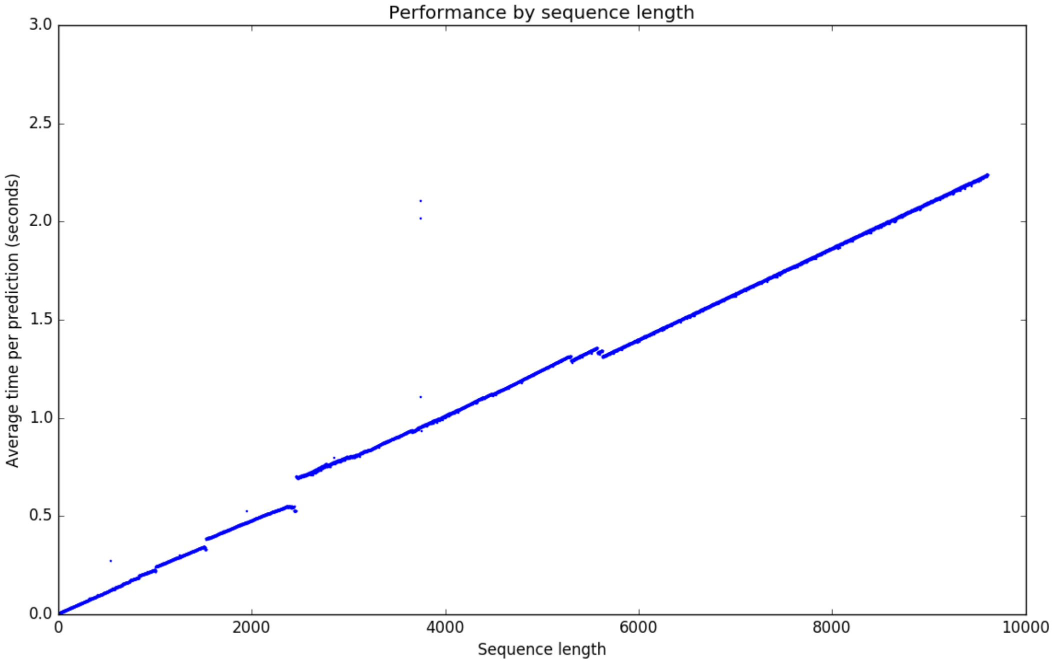
4.6.1 DeepLinc’s accuracy as a function of transcript length.



4.6.2 DeepLinc converges on both training and testing set.



4.6.3 DeepLinc predicts in linear time with respect to transcript length.



Chapter 5: General Conclusions

In this thesis there are three main contributions to the scientific community. A draft hop genome was produced, along with an annotation for two different cultivars. A web resource is now available for scientists interested in working on hop. Using these resources, we showed the ability to data mine and interpret previously unknown biology. For the first time, the hop sex chromosomes were characterized at the genomic level.

In addition, we provide a new form of transcript annotation. Recurrent Neural Networks have not yet been used in any annotation context. By using them on a relatively simple task, we show the power and natural ability of these models to operate on genomic and transcriptomic sequences. It is possible not only to use the models as we have, but also to interpret these models and gain biological insight. Although we provided an annotation for putative hop protein coding genes, not a single lincRNA is characterized. By using the random forest classifier we presented in chapter 4, it is possible to bootstrap a method of lincRNA annotation when no such data exists. We propose the following process for hop:

1) Identify the "obvious" lincRNAs. These are transcripts with very short ORFs (less than 50 amino acids), no homology to known protein coding domains, and no homology to known plant proteins.

2) Train a classifier to predict mRNAs and lincRNAs using the now annotated mRNAs with the features, Hexamer frequency, and Fickett score.

3) Make predictions on the remaining transcripts with know known protein coding domain or homology to plant proteins. Use a very conservative threshold for predicting lincRNAs

4) Bootstrap steps 1-3 until a reasonably sized dataset is achieved (See chapter 4 for details), and train a DeepLinc model. Finally, use DeepLinc to predict the remaining unannotated putative noncoding RNAs.

With the conclusion of this work, a framework is in place for more sophisticated sequencing technologies. HopBase will function with any genome provided, with any future update to the assembly, including a new cultivar or a more complete assembly, the website's functionality will stay the same.

BIBLIOGRAPHY

Abrusán G, Grundmann N, DeMester L, Makalowski W (2009a) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329–1330.

Abrusán G, Grundmann N, DeMester L, Makalowski W (2009b) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329–1330.

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning.

Anantharam V, Varaiya P, Walrand J (1987) Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. Autom Control IEEE Trans On 32:968–976.

Aron PM, Shellhammer TH (2010) A discussion of polyphenols in beer physical and flavour stability. J Inst Brew 116:369–380.

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, pp 177–186

Boutain J (2014) Nanopore DNA Sequencing a Native North American Hop (Humulus lupulus var. lupuloides) with Implications for Research on Cannabaceae Collections. University of Hawaii at Manoa

Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Bradnam KR, Fass JN, Alexandrov A, et al (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:1–31.

Brady JL, Scott NS, Thomas MR (1996) DNA typing of hops (Humulus lupulus) through application of RAPD and microsatellite marker sequences converted to sequence tagged sites (STS). Euphytica 91:277–284. doi: 10.1007/BF00033088

Brooks S, Horner C, Likens S, Zimmermann C (1972) Registration of Cascade Hop1 (Reg. No. 1). Crop Sci 12:394–394.

Cerenak A, Satovic Z, Javornik B (2006) Genetic mapping of hop (Humulus lupulus L.) applied to the detection of QTLs for alpha-acid content. Genome 49:485–494.

Chopra R, Nayar S, Chopra I (1986a) Glossary of Indian medicinal plants (including the supplement). Council of Scientific and Industrial Research. New Delhi 2–79.

Chopra R, Nayar S, Chopra I (1986b) Glossary of Indian medicinal plants (including the supplement). Council of Scientific and Industrial Research. New Delhi 2–79.

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling.

Clark SM, Vaitheeswaran V, Ambrose SJ, et al (2013a) Transcriptome analysis of bitter acid biosynthesis and precursor pathways in hop (Humulus lupulus). BMC Plant Biol 13:1.

Clark SM, Vaitheeswaran V, Ambrose SJ, et al (2013b) Transcriptome analysis of bitter acid biosynthesis and precursor pathways in hop (Humulus lupulus). BMC Plant Biol 13:1.

Crescenzi P, Kann V (1997) Approximation on the web: A compendium of NP optimization problems. Springer, pp 111–118

Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Danilova T, Danilov S, Karlov G (2003) [Molecular-genetic polymorphisms of cultivars of common hops (Humulus lupulus L.) using ISSR-PCR analysis]. Genetika 39:1484–1489.

Delyser D, Kasper W (1994) Hopped beer: the case for cultivation. Econ Bot 48:166–170.

Gnerre S, MacCallum I, Przybylski D, et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci 108:1513–1518.

Goodfellow I, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial nets. pp 2672–2680

Graves A (2012) Supervised sequence labelling. In: Supervised Sequence Labelling with Recurrent Neural Networks. Springer, pp 5–13

Hamel PB, Chiltoskey MU (1975) Cherokee plants: their use. A 400 year history.

Haunold A (1972) Polyploid breeding with hop Humulus lupulus L. Tech Q Masters Brew Assoc Am 9:36–40.

Haunold A, Horner C, Likens S, et al (1976a) Registration of Willamette Hop1 (Reg. No. 6). Crop Sci 16:739–739.

Haunold A, Likens S, Horner C, et al (1976b) Registration of Columbia Hop1 (Reg. No. 5). Crop Sci 16:738–739.

Hecker M, Lambeck S, Toepfer S, et al (2009) Gene regulatory network inference: data integration in dynamic models—a review. Biosystems 96:86–103.

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 423–447.

Henning JA, Coggins J, Peterson M (2015a) Simple SNP-based minimal marker genotyping for Humulus lupulus L. identification and variety validation. BMC Res Notes 8:542.

Henning JA, Haunold A, Townsend MS, et al (2008) Registration of “Teamaker”hop.

Henning JA, Townsend MS, Gent DH, et al (2011) QTL mapping of powdery mildew susceptibility in hop (Humulus lupulus L.). Euphytica 180:411–420.

Henning J, Gent D, Twomey M, et al (2015b) Precision QTL mapping of downy mildew resistance in hop (Humulus lupulus L.). Euphytica 202:487–498.

Hill S, Coggins J, Liston A, et al Genomics of the hop pseudo-autosomal regions. Euphytica 1–9.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780.

Jakse J, Cerenak A, Radisek S, et al (2013) Identification of quantitative trait loci for resistance to Verticillium wilt and yield parameters in hop (Humulus lupulus L.). Theor Appl Genet 126:1431–1443.

Jiang H, Lei R, Ding S-W, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:1.

Johnson M, Zaretskaya I, Raytselis Y, et al (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36:W5–W9.

Karlov G, Danilova T, Horlemann C, Weber G (2003a) Molecular cytogenetics in hop (Humulus lupulus L.) and identification of sex chromosomes by DAPI-banding. Euphytica 132:185–190.

Karlov G, Danilova T, Horlemann C, Weber G (2003b) Molecular cytogenetics in hop (Humulus lupulus L.) and identification of sex chromosomes by DAPI-banding. Euphytica 132:185–190.

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360.

Kingma D, Ba J (2014) Adam: A method for stochastic optimization.

Koie K, Inaba A, Okada Y, et al (2004) Construction of the genetic linkage map and QTL analysis on hop (Humulus lupulus L.). pp 59–66

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:1.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. pp 1097–1105

Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517.

Landis PH (1939) The Hop Industry, A Social and Economic Problem. Econ Geogr 15:85–94. doi: 10.2307/141007

Li H, Handsaker B, Wysoker A, et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Listgarten J, Lippert C, Kadie CM, et al (2012) Improved linear mixed models for genome-wide association studies. Nat Methods 9:525–526.

Li X, Waterman MS (2003) Estimating the repeat structure and length of DNA sequences using ℓ-tuples. Genome Res 13:1916–1922.

Luo R, Liu B, Xie Y, et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:1–6.

Manichaikul A, Mychaleckyj JC, Rich SS, et al (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873.

McAdam EL, Freeman JS, Whittock SP, et al (2013) Quantitative trait loci in hop (Humulus lupulus L.) reveal complex genetic architecture underlying variation in sex, yield and cone chemistry. BMC Genomics 14:1.

Miranda C, Stevens J, Helmrich A, et al (1999) Antiproliferative and cytotoxic effects of prenylated flavonoids from hops (Humulus lupulus) in human cancer cell lines. Food Chem Toxicol 37:271–285.

Natsume S, Takagi H, Shiraishi A, et al (2014a) The draft genome of hop (Humulus lupulus), an essence for brewing. Plant Cell Physiol pcu169.

Natsume S, Takagi H, Shiraishi A, et al (2014b) The draft genome of hop (Humulus lupulus), an essence for brewing. Plant Cell Physiol pcu169.

Neve RA (2012) Hops. Springer Science & Business Media

Nussbaumer T, Martis MM, Roessner SK, et al (2013) MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res 41:D1144–D1151.

Ono T (1962a) The wild hop native to Japan.

Ono T (1962b) The wild hop native to Japan.

Patzak J, Henychova A, Krofta K, Nesvadba V (2012) Study of molecular markers for xanthohumol and DMX contents in hop (Humulus lupulus L.) by QTLs mapping analysis. Brew Sci 65:96–102.

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830.

Pertea M, Pertea GM, Antonescu CM, et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295.

Pillay M, Kenny ST (1996) Random amplified polymorphic DNA (RAPD) markers in hop, Humulus lupulus: level of genetic variability and segregation in F1 progeny. Theor Appl Genet 92:334–339. doi: 10.1007/BF00223676

Priyam A, Woodcroft BJ, Rai V, et al (2015) Sequenceserver: a modern graphical user interface for custom BLAST databases. bioRxiv 033142.

Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504.

Quang D, Xie X (2015) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. bioRxiv 032821.

Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks.

Rhee SY, Beavis W, Berardini TZ, et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 31:224–228.

Salmon E (1917) The Value of Hop-Breeding Experiments. J Inst Brew 23:60–97.

Seaton G, Haley CS, Knott SA, et al (2002) QTL Express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics 18:339–340.

Seefelder S, Ehrmaier H, Schweizer G, Seigner E (2000) Male and female genetic linkage map of hops, Humulus lupulus. Plant Breed 119:249–255.

Skinner ME, Uzilov AV, Stein LD, et al (2009) JBrowse: a next-generation genome browser. Genome Res 19:1630–1638.

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Smit AF, Hubley R, Green P (1996) RepeatMasker Open-3.0.

Smith J, Kinman ML (1965) The use of parent-offspring regression as an estimator of heritability. Crop Sci 5:595–596.

Spindel J, Begum H, Akdemir D, et al (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11:e1004982.

Srivastava N, Hinton GE, Krizhevsky A, et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958.

Stanke M, Keller O, Gunduz I, et al (2006a) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439.

Stanke M, Keller O, Gunduz I, et al (2006b) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439.

Stevens JF, Ivancic M, Hsu VL, Deinzer ML (1997a) Prenylflavonoids from Humulus lupulus. Phytochemistry 44:1575–1585.

Stevens JF, Ivancic M, Hsu VL, Deinzer ML (1997b) Prenylflavonoids from Humulus lupulus. Phytochemistry 44:1575–1585.

Sung B, Chung JW, Bae HR, et al (2015) Humulus japonicus extract exhibits antioxidative and anti‑aging effects via modulation of the AMPK‑SIRT1 pathway. Exp Ther Med 9:1819–1826.

Team TTD, Al-Rfou R, Alain G, et al (2016) Theano: A Python framework for fast computation of mathematical expressions.

Timp W, Comer J, Aksimentiev A (2012) DNA base-calling from a nanopore using a Viterbi algorithm. Biophys J 102:L37–L39.

Tomlan MA (2013) Tinged with gold: hop culture in the United States. University of Georgia Press

UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195.

USDA (2016) Hop Acreage Strung for Harvest, 2016.

Wang L, Park HJ, Dasari S, et al (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41:e74–e74.

Wilhelm M, Schlegl J, Hahne H, et al (2014) Mass-spectrometry-based draft of the human proteome. Nature 509:582–587.

Wright A (1976) The significance for breeding of linear regression analysis of genotype-environment interactions. Heredity 37:89–93.

Xie Y, Wu G, Tang J, et al (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829.

Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12:931–934.

Zur H, Tuller T (2013) New universal rules of eukaryotic translation initiation fidelity. PLoS Comput Biol 9:e1003136.