Defense-related gene families in the model legume, *Medicago truncatula*: computational analysis, pan-genome characterization, and structural variation

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Peng Zhou

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisers: Dr. Nevin D. Young, Dr. Peter L. Tiffin

June 2015

Acknowledgements

I would like to thank my advisor Dr. Nevin D. Young for his support, guidance, advice and patience during my Ph.D. study. He has been an excellent mentor and a positive influence throughout my academic pursuits. I also thank Dr. Kevin A. T. Silverstein for continuously providing optimism and encouragement regarding my research and academic performance. My sincere thanks also go to other members of my thesis committee: Dr. James M. Bradeen, Dr. Peter L. Tiffin, Dr. H. Corby Kistler and Dr. Chad Myers, for their comments and suggestions regarding my research and thesis.

I thank the faculty and staff that have been helpful throughout my graduate studies. I thank my fellow graduate students, past and present, for providing encouragement and discussing solutions to problems posed by my research.

I thank my family for their enduring support throughout my studies. Their love, patience, support, encouragement, and understanding have been and continue to be an integral part of my life.

Finally, I would like to collectively thank, in no special order, Roxanne Denny, Joann Mudge, Jason Rafe Miller, Joseph Guhlin, Yong Bao, Liangliang Gao, Lian Lian, Diana Trujillo, Shaun Curtin, Sumitha Nallu, Robert M. Stupar, Andrew Farmer, Kelly Steele, Tim Paape, Roman Briskine, Arvind K. Bharti, Gregory D. May and many others for all the advice, support and help they provided during my graduate studies. I would like to acknowledge the computational resources provided by Minnesota Supercomputing Institute, and the funding provided by National Science Foundation.

Abstract

Medicago truncatula is a model for investigating legume genetics and the evolution of legume-rhizobia symbiosis. Over the past two decades, two large gene families in *M. truncatula*, the nucleotide-binding site leucine-rich repeat (NBS-LRR) family and the nodule-specific, cysteine-rich (NCR) gene family, have received considerable attention due to their involvement in disease resistance and nodulation, large family size, and high nucleotide and copy number diversity. While NBS-LRRs have been found in all plant species and therefore relatively well characterized at the sequence level, members of the cysteine-rich protein (CRP) families, including NCRs, have generally been overlooked by popular similarity search tools and gene prediction techniques due to their (a) small size, (b) high sequence divergence among family members and (c) limited availability of expression evidence. In this thesis, I first developed a homology-based gene prediction program (Small Peptide Alignment Detection Algorithm, i.e., SPADA) to accurately predict small peptides including CRPs at the genome level. Given a high-quality profile alignment, SPADA identifies and annotates nearly all family members in tested genomes with better performance than all general-purpose gene prediction programs surveyed. Numerous mis-annotations in the current Arabidopsis and *Medicago* genome databases were found by SPADA, most supported by RNA-Seq data. As a homology-based gene prediction tool, SPADA works well on other classes of small secreted peptides in plants (e.g., self-incompatibility protein homologues) as well as non-secreted peptides outside the plant kingdom.

I then comprehensively annotated the NBS-LRR and NCR gene families in the *Medicago* reference genome (version 4.0), and set out to characterize natural variation of these genes in diverse *M. truncatula* accessions. Previous studies using whole-genome sequence data to identify sequence polymorphisms (SNPs and short Insertion / Deletions) relied on mapping short reads to a single reference genome. However, limitations of read-mapping approaches have hindered variant detection, especially characterization of repeat-rich and highly divergent regions. As a result, studies of these large gene families

are also hindered due to high sequence similarity among family members along with high divergence among accessions. In this work I constructed high-quality *de novo* assemblies for 15 *M. truncatula* accessions. This allowed me to detect novel genetic variation that would not have been found by mapping reads to a single reference. This analysis led to a within-species diversity estimate 70% higher than previous mapping-based resequencing efforts, even using a smaller sample size. These results clearly demonstrate that *de novo* assembly-based comparison is both more accurate and precise than mapping-based variant calling in exploring variation in repetitive and highly divergent regions.

For the first time in plants, my results enable systematically identification and characterization of different types of structural variants (SVs) using a synteny-based approach. This analysis suggests that, depending on the divergence from the reference accession, 7% to 21% of the entire genome is involved in large structural changes, affecting 10% to 28% of all gene models. The results identify 64 Mbp of unique sequence segments absent in the reference, including 30 Mbp shared by at least 2 accessions and 34 Mbp of accessions-specific sequences, thus expanding the *Medicago* reference space (389-Mbp) by 16%.

Evidence-based annotation of the 15 *de novo* assemblies revealed that more than half of reference gene models were structurally diverse (lower than 60% sequence similarity) in at least one other accession. Not surprisingly, the NBS-LRR gene family harbors by far the highest level of nucleotide diversity, large effect single nucleotide changes, mean pairwise protein distance and copy number variation (levels comparable with transposable elements), consistent with the rapidly-evolving dynamics of disease resistance phenotypes. Characterization of deletion and tandem duplication events in the NBS-LRR and NCR gene families suggests accession-specific subfamily expansion / contraction patterns. This work illustrates the value of multiple *de novo* assemblies and the strength of comparative genomics in exploring and characterizing novel genetic variation within a population, and provides insights in understanding the impact of SVs on genome architecture and large gene families underlying important traits.

Table of Contents

List of Tables

List of Figures

Chapter 1. Detecting Small Plant Peptides Using SPADA

Small peptides encoded as one- or two-exon genes in plants have recently been shown to affect multiple aspects of plant development, reproduction and defense responses. However, popular similarity search tools and gene prediction techniques generally fail to identify most members belonging to this class of genes. This is largely due to the high sequence divergence among family members and the limited availability of experimentally verified small peptides to use as training sets for homology search and *ab initio* prediction. Consequently, there is an urgent need for both experimental and computational studies in order to further advance the accurate prediction of small peptides.

I present here a homology-based gene prediction program to accurately predict small peptides at the genome level. Given a high-quality profile alignment, SPADA identifies and annotates nearly all family members in tested genomes with better performance than all general-purpose gene prediction programs surveyed. Numerous mis-annotations were found in the current *Arabidopsis thaliana* and *Medicago truncatula* genome databases using SPADA, most of which have RNA-Seq expression support. I also show that SPADA works well on other classes of small secreted peptides in plants (e.g., self-incompatibility protein homologues) as well as non-secreted peptides outside the plant kingdom (e.g., the alpha-amanitin toxin gene family in the mushroom, *Amanita bisporigera*).

SPADA is a free software tool that accurately identifies and predicts the gene structure for short peptides with one or two exons. SPADA is able to incorporate information from profile alignments into the model prediction process and makes use of it to score different candidate models. SPADA achieves high sensitivity and specificity in predicting small plant peptides such as the cysteine-rich peptide families. A systematic application of SPADA to other classes of small peptides by research communities will greatly improve the genome annotation of different protein families in public genome databases.

Introduction

A major challenge in translating new genome sequences into useful community resources is the accurate annotation of genes and other functionally relevant features (Stein 2001). While there have been clear improvements in gene prediction algorithms (Yao et al. 2005), accurate prediction of small one and two-exon genes remains stubbornly problematic (Basrai, Hieter, and Boeke 1997). False-positive signals arising from the poor specificity of promoter motifs and other commonly-used signals employed by general purpose gene-finding algorithms are widespread (Lease and Walker 2006; Hanada et al. 2007; Yang et al. 2011; B. Pan et al. 2013). To address the flood of false-positive signals for small genes, many annotators filter out small-gene predictions lacking direct experimental expression evidence, resulting in a major problem of false negatives (Basrai, Hieter, and Boeke 1997; Lease and Walker 2006).

I propose here an alternative and complementary strategy for genome-wide annotation – a strategy that has as its strength predicting the small one- and two-exon genes that all-purpose gene-finding algorithms often fail to predict accurately. This approach focuses on finding all related paralogous genes within a target gene family and then using signals from the corresponding multiple sequence alignment to aid in refining the model predictions. I have implemented this approach in an open-source and freely available application called SPADA (Small Peptide Alignment Discovery Application). SPADA can be used directly with a user's own protein family alignments or with a comprehensive set of protein family alignments from public sources such as Pfam (Punta et al. 2011), InterPro (Hunter et al. 2012) or PROSITE (Sigrist et al. 2013), enabling the exhaustive discovery of essentially all members of the input families within a given genome sequence. Because these public resources continue to expand and include new and novel protein families, SPADA's ability to comprehensively identify arbitrarily large families of small peptides in genomes will steadily grow.

Here I describe the conceptual basis of SPADA and go onto test its performance with selected families of notoriously difficult genes to annotate properly – specifically, plant Cysteine-Rich Peptides (CRPs) in two model plant species (*Arabidopsis thaliana* and

Medicago truncatula), the S-Protein homologue (SPH) family in *A. thaliana*, and the alpha-amanitin toxin gene family in the mushroom *Amanita bisporigera*. In the case of CRPs, we examine the accuracy and recall compared to published composite test/training sets for these species based on previous semi-manual curation and subsequent experimental expression validation (Silverstein et al. 2007; Nallu et al. 2013; Tesfaye et al. 2013). I also compare SPADA's performance against a number of commonly used generic gene-prediction algorithms (Majoros, Pertea, and Salzberg 2004; Lomsadze et al. 2005; Blanco, Parra, and Guigó 2007; Keller et al. 2011), providing evidence of SPADA's advantage in identifying these challenging classes of small peptides.

Method

SPADA is a computational pipeline that, when provided with a multiple sequence alignment for a gene/protein family of interest, identifies all members of this family in a target genome. Technically, SPADA's pipeline is a general homology-based gene finding program with specifically enhanced power to detect and annotate small peptides with one or two exons. Unlike general-purpose gene prediction programs such as Fgenesh (Salamov and Solovyev 2000), SPADA works on an entire gene family at one time - with the goal of finding all family members in the genome. Unlike other homology-based gene predictors such as Genewise (Birney, Clamp, and Durbin 2004) and Exonerate (Slater and Birney 2005) that map a single protein sequence to the target genome, SPADA performs a similarity search using a profile alignment and identifies all homologs of the family. In addition, SPADA provides automated access to both similarity search tools (e.g., BLAST (Altschul et al. 1990) and HMMER (S R Eddy 1998) and *ab initio* gene predictors (e.g., Augustus), significantly improving the annotation efficiency of multi-member gene families. As shown in Figure 1.1, SPADA consists of four consecutive components: a) Pre-Processing, b) Motif Mining, c) Model Prediction and d) Model Evaluation & Selection.

In SPADA, HMMER (S R Eddy 1998) is first used to identify hits in the target genome sequence (translated in six reading frames) as well as in the target proteome (if

available) using a reasonably generous E-value (10). These hits are then tiled with regard to their genomic coordinates and merged into overlapping clusters. Finally, one best hit in each cluster is picked to generate a list of candidate hits.

The pipeline then allows the user to run one or more processes to predict gene structures for these potential genes. By default, SPADA runs Augustus (Keller et al. 2011) using hit locations as clues for "CDS regions" (coding sequence). In parallel, SPADA runs a custom pipeline optimized for predicting the exon boundaries of genes containing one or two exons using GeneWise (Birney, Clamp, and Durbin 2004), SplicePredictor (Brendel, Xing, and Zhu 2004) and custom Perl scripts.

In the next step, all gene structure predictions are combined to make a raw calling set, with each hit having one or more gene structure predictions. SPADA uses multiple statistics to assess the confidence of each candidate gene model, including an alignment score (mean pairwise score with known members in the original family-specific multiple sequence alignment), an HMM alignment score (sum of posterior probability scores in the Hmmsearch output file), the presence/absence of proper start/stop codons, as well as the SignalP D-score (Petersen et al. 2011) in the case of secreted peptides.

Finally, the best candidate gene model is picked for each hit and the resulting set is filtered using empirical cutoffs (hmmsearch E-value of 0.001) to remove false positives.

Pre-processing

Building family-specific multiple sequence alignments

The original motivation for developing this pipeline was accurately identifying and predicting Cysteine-Rich Peptides (CRPs) in plant genomes. For this purpose, SPADA comes with a complete set of manually-curated protein sequence alignments for plant Cysteine-Rich Peptide (CRP) families (Silverstein et al. 2007). In 2007, Silverstein *et al.* built multiple sequence alignments for most plant CRP families through iteratively scanning EST sequences from different plant species in TIGR's Gene Indices (now JCVI) (Silverstein et al. 2007). These alignments were re-aligned here using ClustalO (Sievers et al. 2011) and trimmed using trimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón

2009) to remove spurious sequences and poorly aligned positions. Finally, a profile Hidden Markov Model (HMM) was built for each CRP family using 'hmmbuild' in the HMMER package (Sean R. Eddy 2011).

As a general homology-based gene finding program, SPADA has been designed to work with any set of protein families. Users can start with a list of amino acid sequence alignments of their own interest, run the script "build_profile.pl" to generate custom HMM profiles, and initiate the pipeline using the new HMM(s). With this in mind, I have tested SPADA's performance on an additional protein family as a proof of concept (see Results section) and assessed its applicability to secreted protein families other than CRPs.

Processing genome sequence and annotation

In SPADA, genome FASTA sequences are translated in all six reading frames to amino acid sequences and then Open Reading Frames (ORFs) are extracted by breaking up these long amino acid sequences using stop codons. Here an ORF is defined as a segment of amino acid sequence with at least 15 residues and uninterrupted by stop codons. Extracting ORFs from the original translated genomic sequence reduces the target database size for the subsequent motif mining step and improves sensitivity. Using ORFs also ensures that no protein-coding exon spans stop codons in the middle of a sequence and that each exon will have a reasonable length. In theory, all protein-coding exons should locate within these ORFs, which will be discovered in the next motif-mining step. However, the exact exon boundaries are still unclear at this point of the search procedure.

If a gene annotation file in General Feature Format version 3.0 (GFF3) is available, SPADA can also read and process it, extracting the amino acid sequences of existing annotations and passing them onto the next motif mining step. In doing so, exon boundaries can be better refined, further improving the accuracy in the model prediction step.

Hard-masking of genome sequences is recommended (replacing repetitive sequences with 'N's) before running the pipeline. Some plant species have very large genomes with

highly repetitive content (e.g., Maize (Schnable et al. 2009)). By hard-masking the genome sequence, the target database size in the motif-mining step is effectively reduced, significantly improving the search sensitivity of the entire pipeline. However, if many family members locate in repeat-rich genomic regions (such as the fungi effector families (Rep and Kistler 2010)), the unmasked genome version should be used.

Motif mining

In SPADA, profile HMMs are used to search against translated genomic sequences (and known protein sequences, if available) using hmmsearch, a component of the HMMER package (v3.0) (Sean R. Eddy 2011). This program finds significant hits against a protein sequence database using one or more profiles as inputs. The output of the scan is a list of genomic intervals with significant sequence similarity to query profiles and amino acid sequences translated from these intervals. For single-exon genes, a contiguous stretch of amino acid sequence in the target databases will be discovered, roughly corresponding to the exon in the original genomic sequence. For genes containing two or more exons, partial amino acid sequence hits corresponding to different exons will be separated by introns (if they share a reading frame) or distributed in different target sequences (if in different reading frames). SPADA collects all these full and partial hits in translated protein sequences, recovers their original genomic coordinates, filters out low-significance hits (E-value higher than 0.1), selects the most significant hit for each genomic interval since multiple input profiles may hit the same region, and merges nearby partial hits. During this merging step, SPADA requires that each neighboring partial hit should hit a different segment (either upstream or downstream) in the input profile HMM. The merged genomic intervals (called “extended hits”) roughly correspond to the multiple exons in the underlying gene model - although the exact intron-exon boundaries and start/stop codon locations are yet to be refined at this stage of the procedure.

In parallel, SPADA searches against existing protein sequences (generated using the GFF3 annotation file), yielding a separate list of hits to the input profiles. These hits are

also treated as partial hits, i.e., mapped to their original genomic coordinates and then used to build "extended hits". This "hmmsearch against proteome" step is considered complementary to the abovementioned "hmmsearch against translated genome" step, since it improves prediction sensitivity by capturing otherwise non-significant partial hits in the translated genome search.

Model prediction

At this point, SPADA has generated a list of "extended hits" approximately corresponding to actual exon boundaries. For each extended hit the surrounding genomic sequence is extracted. By default, 2500 bp upstream from the hits are extracted, since the first exon (containing the signal peptide) is usually separated from the second exon (with the mature peptide) by an intron up to 1500 bp, as determined by manual curation and understanding of plant genomes. At the other end, 1500 bp downstream from the hit boundaries are extracted, since the correct stop codon can typically be found within 1000 bp downstream of the HMM hit. SPADA next runs one or more components (selected by the user) in parallel to determine gene structure in this region. A total of five prediction components are currently supported by the pipeline: Augustus (Keller et al. 2011), GeneWise (Birney, Clamp, and Durbin 2004), GlimmerHMM (Majoros, Pertea, and Salzberg 2004), GeneMark (Lomsadze et al. 2005) and GeneID (Blanco, Parra, and Guigó 2007). By default SPADA only runs two of these components (Augustus and GeneWise) since performance evaluation on a group of common plant peptides suggests that running all five of them does not offer a significant extra gain compare to running just two of them (see Results & Discussion).

The first component, which I denote "Augustus_evidence", runs Augustus (Keller et al. 2011) in its "evidence mode". The genomic sequence is used as input along with a "hint file" providing the program instructions for which part(s) of the input sequence are known to be part(s) of the coding sequence. In other words, location information of extended hits is incorporated in the prediction process. Augustus will then try to complete the gene model by looking for start/stop codons and canonical donor-acceptor splice sites

around the hits while preserving the open reading frame. The improvement in prediction accuracy and specificity by running Augustus in the "evidence mode" (as compared to the "Augustus de novo mode") is significant and will be discussed in Results & Discussion.

In parallel, SPADA runs a custom pipeline specifically designed to identify and predict genes with one or two exon(s) and with a leading signal peptide. This component, which we denote "Genewise+SplicePredictor", first runs GeneWise to align the extended hit sequence (translated to amino acid sequence) to genomic sequence and identifies compatible splice sites that preserve the hit ORFs. If GeneWise fails due to non-canonical splice sites, SPADA then runs SplicePredictor (Brendel, Xing, and Zhu 2004) to find all possible donor/acceptor splice sites and extracts compatible ones, extending the ORFs to the nearest start codon and stop codon. In practice, this "Genewise+SplicePredictor" approach works well as a complement to the "Augustus_evidence" approach (see Results and Discussion).

At this stage in the pipeline, SPADA reports all compatible gene models predicted by the two components. These candidate models are then passed on to the next step for evaluation in order to generate a best calling set.

Model evaluation & selection

For each extended hit, SPADA then evaluates the underlying candidate models using a number of measures and picks the most "confident" model for output. These evaluation statistics include the presence of start/stop codons at the beginning/end of the model, the presence of inframe stop codons, the SignalP score (Petersen et al. 2011) in the case of a secreted gene family, and in particular, the Multiple Sequence Alignment (MSA) score and the "Hmmsearch Probability" (HmmProb) score, as described below.

In theory, the correct gene model should encode an amino acid sequence that aligns to the original family-specific protein alignments better than any other candidate models. To calculate the MSA score, SPADA aligns the amino acid sequence of the candidate model to the profile alignment using ClustalO "profile-to-profile" mode (Sievers et al.

2011). SPADA then scores all pairwise alignments using BLOSUM80 scoring matrix (Henikoff and Henikoff 1992) and calculates a mean alignment score. The BLOSUM80 matrix is used instead of BLOSUM62 because the sequences that are being aligned tend to be fairly similar to each other, and a matrix with more conserved target frequencies such as BLOSUM80 should be more reasonable. Ideally, the candidate model with the highest MSA score should be the most probable model.

Nevertheless, the MSA score is not sufficient to pick the best model, since candidate models are sometimes too close to each other in sequence and the MSA scores may not vary appreciably among model alternatives. Therefore SPADA also calculates an "Hmmsearch Probability" Score for each candidate model. In theory, if hmmsearch is run using the original family HMM against all candidate models, the most significant hit in the output should then be the best model. In practice, the probability score in the hmmsearch output serves as a better predictor than the E-value itself, especially when a model contains more than one hit domain. The MSA score and the HmmProb score are used to evaluate each candidate model. SPADA then picks the best candidate model that meets the following criteria: (1) it has a SignalP D-score of no less than 0.4 (determined according to the software manual, this option could be turned off to allow prediction of non-secreted gene families); (2) it has proper start/stop codons and no premature stop codon; and (3) it has the highest (MSA score + HmmProb score).

SPADA uses a relatively relaxed E-value cutoff in the motif mining step (e.g., 10 for running hmmsearch) in order to increase specificity. This also results in numerous false positive hits. These hits will generally not have valid candidate gene models built for them in the model prediction step, and thus would not make it into the ultimate output. However, SPADA does employ a final filtering step based on hmmsearch E-value to refine gene models that are retained. Performance of the pipeline under different final E-value cutoffs (see Results and Discussion) are evaluated and the default cutoff is set to 0.001, which may be adjusted by the user to achieve customized search purposes. For all gene models passing the filter, SPADA outputs the sequences in FASTA format and gene coordinate information in GFF format. SPADA also generates for each gene family a

multiple sequence alignment including all predicted models and the family-specific consensus sequence. If a gene annotation file has been passed to the pipeline, SPADA will also report the comparison results of predicted models with existing annotation (e.g., the number of models with exactly the same exon boundaries, models with partial overlap, models in different reading frames, etc.).

Performance evaluation

Compilation of the test set

A test set of plant cysteine-rich peptides (CRPs) was compiled in two genomes: the model dicotyledon, *Arabidopsis thaliana*, and the model legume, *Medicago truncatula*. In previous work, Silverstein *et al*. (Silverstein et al. 2007) have exhaustively searched and curated all 516 CRP families (CRP0000 - CRP6250) in Arabidopsis. A large number of CRPs have also been identified and curated in an early release of the *Medicago* genome sequence (Silverstein et al. 2007). Recently, as a collaborative effort with J. Craig Venter Institute (JCVI), we expanded this list of CRPs in *M. truncatula* by manually inspecting and curating 136 CRP families (CRP0000 - CRP1530, focusing specifically on the Defensin-Like proteins or DEFLs) in *M. truncatula* (Nevin D. Young et al. 2011). This finally led to a complete list of CRP members for Arabidopsis and *Medicago* (742 for Arabidopsis and 725 for *Medicago*, Appendix File 1.1).

Evidence from multiple sources was collected to validate the expression of the models in the compiled test set. On the one hand, extensive RNA-Seq data were downloaded from NCBI Sequence Read Archive (Leinonen, Sugawara, and Shumway 2011) for both Arabidopsis and *Medicago*; on the other hand, I downloaded the AtMtDEFL microarray dateset (Nallu et al. 2013; Tesfaye et al. 2013) to seek additional support for expression of these gene families. The AtMtDEFL array include probe sets for 317 Arabidopsis DEFLs, 15 Arabidopsis DEFL-related Genes (MEGs), and 684 *Medicago* DEFLs, plus additional marker genes. In total, 583 (78.6%) out of the 742 CRPs in the Arabidopsis test set and 657 (90.6%) out of the 725 *Medicago* CRPs receive support from either RNA-Seq (FPKM >1) or microarray data (Appendix Table 1.3-1.4).

These carefully curated, high-quality CRP calls were then taken as our test set in evaluating the performance of different model prediction components in SPADA under different hmmsearch E-value cutoffs.

Evaluation procedure

A number of popular gene prediction programs were tested as SPADA model prediction components. In addition to the previously mentioned components, I also tested GeneID (v1.4.4) (Blanco, Parra, and Guigó 2007), GlimmerHMM (v3.0.1) (Majoros, Pertea, and Salzberg 2004), GeneMark (v3.9d) (Lomsadze et al. 2005), and Augustus (v2.6.1, *de novo* mode). The "Augustus_evidence" differs from the "Augustus_de_novo" component simply by the inclusion of a "hint file" (with hit location information) fed to the program. I evaluated the pipeline performance running these components (individually or in combination) based on our curated test dataset (see Results and Discussion), and decided to use the "Augustus_evidence" and "Genewise+SplicePredictor" approaches as default components in the SPADA model prediction step.

Prediction performance was measured at two different levels: coding nucleotide sequence and exonic structure. At each level, sensitivity and specificity for each component were calculated. I first define the true positives (TP, number of coding nucleotides that are correctly predicted as coding), true negatives (TN, number of noncoding nucleotides that are correctly predicted as noncoding), false negatives (FN, number of coding nucleotides predicted as noncoding) and false negatives (FP, number of noncoding sequences predicted coding). At the nucleotide level, Sensitivity (Sn) was then calcualted as the proportion of coding nucleotides that have been correctly predicted as coding (Sn=TP/(TP+FN)), while Specificity (Sp) was the proportion of predicted coding nucleotides that are actually coding (Sp=TP/(TP+FP)) (Burset1996). At the exon level, Sn was the proportion of actual exons in the input sequence that are correctly predicted, while Sp was the proportion of all predicted exons that are correctly predicted (Burset and Guigó 1996). Other measures such as Correlation Coefficient (CC) and Average Conditional Probability (ACP) were not evaluated since they require the calculation of

TN nucleotides/exons, which are noncoding regions that are predicted as noncoding. Unlike a general gene-finding program that tries to predict all coding genes in a given sequence, SPADA focuses only on coding genes that are significantly similar to a given profile, while ignoring all other genes. Consequently "TN" statistics is not straightforward to evaluate in this context.

Performance evaluation was done in both *A. thaliana* and *M. truncatula*. The extracted genomic sequences were used as input sequences. I evaluated the pipeline performance using each of the "GeneID", "Augustus_de_novo", "GlimmerHMM", "GeneMark", "GeneWise+SplicePredictor", "Augustus_evidence" component (individually), as well as "SPADA" (combination of "GeneWise+SplicePredictor" and "Augustus_evidence") and "All" (combination of all 6 individual components). All programs were installed and run locally on a GNU/Linux workstation. The appropriate parameter files, model files and training directories, if available, were used to run these programs in each species, otherwise the default parameter files (which are for Arabidopsis) were used. The outputs of these runs were parsed to derive a unique prediction for each test sequence.

RNA-Seq and microarray processing, data visualization

I mapped the RNA-Seq short reads (downloaded from NCBI SRA) to the reference using TopHat and summarized the results using Cufflinks (Trapnell et al. 2012). Cufflinks is able to estimate the expression value at the level of transcripts. I used a cutoff of FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) > 1 to determine if a model (either in the test set or in the SPADA prediction set) is expressed.

For the AtMtDEFL array, PMA (Present, Marginal and Absent) calls and normalized expression values of each probe set were obtained from the supplemental tables of two recent papers (Nallu et al. 2013; Tesfaye et al. 2013). I mapped the probe sequences to the transcript models in the test set as well as SPADA prediction. In many cases the annotated gene boundaries are not complete and lack portions of the 3'-UTR that is

prioritized in Affymetrix designs, and the probes designed in these regions would not be mapped. As a result, I require at least six probes in a probe set matching the target gene (with 23 or more identical nucleotides for each 25-mer oligo probe). Finally, the PMA calls of a probe set should be 'Present' in at least one tissue/treatment condition to indicate expression support for the transcript model it is mapped to.

In order to visualize some of the novel SPADA predictions as compared to the original genome annotation, as well as the underlying RNA-Seq read mapping support, I loaded the data (genome sequence file, annotation GFF file, SPADA prediction GFF file, RNA-Seq mapping BAM file) into IGV (Integrative Genomics Viewer) (Thorvaldsdóttir, Robinson, and Mesirov 2013), adjusted the width of each track, and made screenshots.

Results

Performance evaluation of SPADA on plant Cysteine Rich Peptide (CRP) families

SPADA performance under different search E-value thresholds

Using our manually curated high-quality CRP test set from Arabidopsis and *Medicago*, I first evaluated the performance of SPADA under different search E-value thresholds. Generally speaking, with a loose E-value threshold (e.g., 0.1), SPADA is able to predict almost all true models (i.e., achieving high sensitivity) while making many false predictions (i.e., specificity is low) (Appendix Figure 1.1). By setting the search threshold to a more stringent value, SPADA avoids making most of the false predictions, but also loses a small number of true models. In an effort to optimize search sensitivity (the ability to detect all true gene models) and specificity (preventing the detection of spurious false models, refer to the "Method - Performance evaluation - Evaluation procedure" section for a formal definition of sensitivity and specificity), I set the default search E-value threshold to 0.001. Users can also change the default E-value threshold to build custom searches (e.g., a very sensitive search using E-value cutoff of 1 to identify all potential hits).

Performance comparison of different gene prediction components

I then compared the performance of SPADA running different model prediction components: GeneID, Augustus ("*de novo*" mode as well as "evidence" mode), GlimerHMM, GeneMark, GeneWise+SplicePredictor as well as "SPADA" (combination of "Augustus_evidence" and "GeneWise+SplicePredictor") and "All" (combination of all 6 individual components) (Figure 1.2, Appendix Figure 1.1). The high specificities observed in all components are likely due to the model evaluation and selection step, where most false models are filtered. Prediction sensitivities, on the other hand, show substantial differences among components. In both genomes tested, "Augustus_evidence" and "GeneWise+SplicePredictor" gave the highest sensitivities among the six individual components. The default SPADA pipeline (denoted as "SPADA" in the figure) runs these two components and achieved even higher sensitivity. On the other hand, running all six individual components (denoted as "All" in the figure) gives the highest sensitivity, suggesting that search accuracy can still be improved by including more heterogeneous prediction programs in the pipeline. However, the gain in sensitivity offered by running all six components is marginal compared to running just two of them ("Augustus_evidence" and "GeneWise+SplicePredictor"), suggesting that a plateau in search accuracy could soon be reached and adding more prediction programs in the pipeline may not help much.

These results are expected as SPADA does not work as a general gene finding program but instead focuses on particular classes of genes with known profiles. Small genes are typically difficult to predict and often missed by genome annotation pipeline due to the intrinsic properties of many automatic gene finding algorithms (Haas et al. 2005). In my test with GeneID, Augustus and GlimmerHMM against the *Medicago* genome, the Arabidopsis training matrix was used since a *Medicago* specific one is not yet available. This explains to a large extent the extremely low sensitivity performance for these three programs in *Medicago*. Search specificities were generally quite high and did not vary much among different programs or genomes tested, indicating the relatively stringent search E-value (0.001) in effect allows few false positives.

Cysteine-rich peptides predicted by SPADA in Arabidopsis and Medicago

Using the default search E-value threshold and model prediction components, SPADA predicts 745 CRPs in Arabidopsis and 1170 (747 for CRP0000-CRP1530) in *Medicago* (Table 1.1, Appendix Table 1.1-1.2, Appendix Figures 1.2-1.3, Appendix File 1.2). These numbers are generally consistent with our manually curated CRP test sets (742 for Arabidopsis) and 725 for *Medicago* (Silverstein et al. 2007), with a sensitivity of 91%–93% and specificity of 85%–95% at the nucleotide level (Appendix Figure 1.1). Members within a sub-class typically show a conserved signal peptide and cysteine configuration (Appendix Figure 1.4 for example). I also checked the expression of these predictions using publicly available RNA-Seq data from NCBI: 570 (76.5%) out of the 745 Arabidopsis CRPs and 947 (80.9%) out of the 1170 *Medicago* CRPs receive either RNA-Seq or AtMtDEFL array expression support (Appendix Table 1.5-1.6). It should be noted that SPADA makes no attempt to predict pseudogenes as it filters out hits with in-frame stop codons. However, some pseudogenes with premature stop codons might still be predicted by SPADA as valid gene models if the in-frame part shows significant (though incomplete) similarity to the search HMM. This in part explains the higher number of SPADA predictions (747 for CRP0000-CRP1530) in *Medicago* than the test set (725) since pseudogenes were manually removed to obtain the test set.

The default E-value threshold of 0.001 is a compromise between sensitivity and specificity that generally works well for both organisms. For the purpose of identifying all potential small coding genes, a search with high sensitivity should be performed since it allows the user to see all potential hits and then determine for him/herself the boundary between false predictions and true predictions based on search scores. The users can then set the cutoff threshold empirically and select genes for experimental verification on their own. Thus, two CRP prediction sets by running SPADA using E-value threshold of 1 were also reported here (Appendix Table 1.7-1.8). In practice, users are encouraged to change the default E-value threshold to build custom searches.

According to the latest versions of genome annotation for Arabidopsis and *Medicago*, about 5% to 15% of SPADA predictions fall completely into intergenic

regions (i.e., are un-annotated, Table 1.2). Through manual inspection of these models, we found that some of the unannotated models turn out to be ORFs with premature stop codons (i.e., pseudogenes), while others had quite significant hmmsearch E-value and complete ORFs. In addition, some predicted models receive expression support from either existing EST sequence or RNA-Seq data. An example is shown in Figure 1.3 where the predicted CRP model is supported by RNA-Seq mapping, fits well in the family-specific alignment, but was missed by the genome annotation (*Medicago* genome annotation version 3.5) as well as by our test set. While such cases are infrequent (e.g., only 10 in Arabidopsis), I speculate the specificity of SPADA is likely to be underestimated.

I then performed manual inspection on these unannotated CRP models and tried to determine whether the calls are truly bad predictions (e.g., pseudogenes with pre-mature stop codons) or valid members of the family missed by current genome annotation (criteria being that the predicted model fits well in the family-specific alignment and has either RNA-Seq or Affymetrix expression support). The number of "novel" CRPs discovered in this fashion, is given in Table 1.2 (Appendix File 1.3). SPADA was able to identify 77 novel CRPs in *Medicago* that were missed by current genome annotation pipeline. The actual number of new CRPs in *Medicago* will be even higher since we only evaluated a subset of all CRP groups (CRP0000-CRP1530). This result is not unexpected given that the *Medicago* genome was released only recently and resources and efforts put into the genome annotation pipeline have been limited. On the other hand, only 3 novel CRPs were found in Arabidopsis, suggesting a relatively higher quality of gene calls in this extensively studied model organism.

Through examination of the novel CRPs, I also noticed that while some of the novel hits have a very significant hmmsearch E-value (e.g., 10^{-12}), most have moderate E-values (e.g., 10^{-4}-10^{-7}), suggesting that their sequence similarity to the input HMM profile is limited. While the input profile alignments were manually built and may not be exhaustive in capturing all groups of CRPs, I speculate that some of these novel CRPs might form new clades that define novel profile alignments, separate from the original

alignment. Consequently, a new round of genome scans using these novel profiles has the potential for capturing even more members that have been missed in the previous search.

Case study: the S-Protein Homologue (SPH) family in Arabidopsis

In addition to plant CRPs, SPADA is readily generalizable to other classes of putative secreted peptides by substituting an appropriate set of HMMs in place of CRP HMMs. Here, the SPH peptides (S-Protein Homologue) (Foote et al. 1994) were used as an example. A seed alignment including 45 plant self-incompatibility protein S1 sequences (PF05938) was obtained from the Pfam database. An HMM profile was built from this alignment and then used as input to scan the Arabidopsis genome (TAIR10) (Lamesch et al. 2012) by running SPADA. SPADA predicted 92 SPH peptides in total (Appendix Table 1.9, Appendix File 1.4). Forty-five (45) of these predictions are identical with TAIR10 annotation. Seventeen have minor discrepancies with TAIR10 gene models (coding regions all in the same reading frame but have a boundary conflict of less than 15 amino acids, probably resulting from different start codons or alternative splice sites). Nine are in major conflict with existing gene models (coding regions in different reading frames or having serious boundary conflict). I also discovered 21 new SPHs not present in TAIR10. Through manual inspection of gene models and sequence alignments with other family members, 3 out of the 15 major conflicts were found to reflect an error in TAIR10 (Figure 1.4A gives an example), while 19 out of the 21 models absent from TAIR10 are true members of the SPH family, missed by the current genome annotation (Figure 1.4B shows an example). As such, I demonstrate that SPADA accurately detects other classes of secreted peptides given a well-constructed profile alignment.

Case study: a fungal cyclic peptide family in Amanita bisporigera

In order to assess whether SPADA could be useful in searches for families of small non-secreted peptides outside the plant kingdom, I also examined the fungal cyclic peptides of Amanita mushrooms (Hallen et al. 2007). This family includes the amatoxins and phallotoxins, such as α-amanitin and phalloidin, respectively, which are synthesized

as proproteins of 34-35 amino acids. I began by creating a multiple sequence alignment via ClustalO of reported proproteins (Hallen et al. 2007) and executed SPADA as usual with the signal peptide filter turned off, using Arabidopsis as the training model for Augustus in searching the low-coverage genome contigs of *Amanita bisporigera*. As a negative control, I scanned the genome of *Amanita thiersii*, which is non-toxic and not known to produce this class of toxins.

SPADA identified five new peptides in the incomplete *A. bisporigera* genome with strong homology ($2.4\times10^{-17} < E < 5.8\times10^{-8}$) that fit well with the alignment of known proproteins (Figure 1.5). One additional hit ran off the end of the contig, producing the incomplete propeptide "MSDTNVMRLPFTTP". No additional predictions were made by SPADA with $E < 0.01$ beyond sequences that were already included in the original alignment. Further, SPADA did not identify any hits when scanning the genome of *A. thiersii,* as would be expected from that organism's non-toxic nature.

Discussion

Homology-based gene prediction

Unlike general-purpose gene predicting programs, SPADA works as a family-based gene finder. The major difference between SPADA and general gene predicting programs is that it incorporates prior information from the family profile in the prediction process. SPADA takes advantage of generic gene prediction programs, but goes a step further by suggesting where to look for family members. Through scanning the target genome using pre-built family-specific alignments, SPADA identifies and builds "extented hits" that serve as the backbone of the underlying exonic structure. This location information greatly improves prediction accuracy, as shown by the different performances of "Augustus_*de_novo*" and "Augustus_evidence" components in Figure 1.2. Among the six individual predicting components, the four that do not require additional information and make de novo predictions all yield low sensitivities. The other two approaches, "Augustus_evidence" and "GeneWise+SplicePredictor", make use of the location information and are able to predict most of the true positives.

Family-based gene prediction was first introduced in the AUGUSTUS package as the AUGUSTUS-PPX (Protein Profile eXtension) module (Keller et al. 2011). Although AUGUSTUS-PPX was shown to be more sensitive and accurate in predicting long, multi-exon gene family members than the standard AUGUSTUS algorithm, its approach is not suitable for small, divergent peptide families such as CRPs, SPHs or Amanita toxin-like peptides examined here. Rather than using the entire protein family alignment profile as input, AUGUSTUS-PPX makes use of conserved, ungapped blocks from the alignment to make a profile. This enables the algorithm to identify core match regions in the genome sequence which together act as a scaffold in the gene prediction. A modification of the standard AUGUSTUS gene-centric HMM is then used to fill in the pieces between scaffold elements with splice elements and other signals, ultimately emitting the most probable full gene structure. While this approach works well for families of typical genes with large numbers of conserved elements, it completely breaks down when applied to small, divergent peptide families like the CRPs, as these families tend to contain no conserved, ungapped regions of appreciable size to seed the initial scaffold. Indeed, when we applied AUGUSTUS-PPX to the CRPs we observed no improvement over "Augustus_*de_novo*".

Improving prediction accuracy by model evaluation

The default SPADA pipeline (running the "Augustus_evidence" and "GeneWise+SplicePredictor" components) achieves even higher sensitivity than the two individual components. This owes to the model evaluation & selection step. For each HMM hit, SPADA collects all candidate gene models built by its model predicting components, and in the model evaluation step, picks a best candidate model based on multiple evaluation statistics. True family members will probably get a high-scoring gene model, while most false positive hits will have no qualifying or only low-scoring gene models built. High-scoring gene models that passed the filter are more likely to be true models since they are the ones that best fit the family-specific alignment.

Pseudogenes and gene models without expression evidence may still have significant value

SPADA identifies paralogous gene family members throughout the genome. Many of these predictions currently lack expression evidence and some of the gene predictions have premature stop codons suggesting they may be psuedogenes. Nonetheless, it is important to identify all gene family members, regardless of their expression and pseudogene status, especially in evolutionarily dynamic gene families. The semi-automated approach that inspired SPADA's development identified hundreds of defensin-like genes in Arabidopsis which, at the time, had no expression evidence (Silverstein et al. 2005). Later, these genes turned out to be highly specifically expressed in reproductive tissues not previously examined with earlier genome-wide expression approaches (Jones-Rhoades, Borevitz, and Preuss 2007). Moreover, one must also be careful not to discard pseudogene predictions that are highly similar to other family members. A gene that appears as a pseudogene in the reference sequenced accession of a species may indeed be fully intact in other accessions, as observed among the defensin-like pollen-tube attractant, AtLUREs (. In their study, Takeuchi and Higashiyama observed half a dozen AtLUREs with disabling mutations in non-reference accessions, as well as putative functional and intact forms of AtLUREs 1.5 and 1.6, which are pseudogenes in the reference Col-0 genotype (Takeuchi and Higashiyama 2012).

Improving SSP annotation in current plant SSP databases

Previous work has sought to exhaustively identify small secreted peptides (SSP) in Arabidopsis (Lease and Walker 2006), rice (B. Pan et al. 2013) and *Populus deltoides* (Yang et al. 2011). These earlier studies only scanned short ORFs (25-250 amino acids) in translated genome sequence, though Pan *et al.* (B. Pan et al. 2013)did include multiple-exon gene predictions from *ab initio* gene predition programs such as Fgenesh and Augustus. However, with a primary focus on detecting all small secreted peptides, these studies did not utilize protein family information in the model building process since secreted peptides are so diverse. The Arabidopsis Unannotated Secreted Peptide Database

(AUSPD) only contains one-exon ORF predictions, and thus mis-annonates most (if not all) two-exon secreted peptides (Appendix Figure 1.5 for example). The OrysPSSP database (comparative Platform for Small Secreted Proteins from rice from rice and other plants) does contain multi-exon models predicted by Fgenesh (0.72%) and Augustus (1.16%) in addition to single-exon ORFs (B. Pan et al. 2013). However, since no prior information is incorporated into predictions by these *ab initio* gene predicting programs, multi-exon models in OrysPSSP are frequently in conflict with the true rice CRPs (Appendix Figure 1.6 for example). As a result, while most single-exon peptides in Arabidopsis and rice are captured in AUSPD and OrysPSSP respectively, a large portion of the two-exon and multi-exon genes (such as CRP0000-CRP1530) are clearly under-represented in these two databases. SPADA, on the other hand, used additional gene structure information obtained in the motif mining step and was able to correctly predict most of the CRP models (Appendix Table 1.10).

Complementarity of SPADA to generic gene prediction programs

SPADA is not designed to identify all genes in a genome. However, its applicability to new annotation projects steadily will increase due to the marked growth of protein sequence family signatures and alignments. InterPro release 43.0 contains 16,652 protein family signatures. In the last 3 years, the number of families characterized by InterPro has increased by 24%, compared with a 38% increase in the 3 years prior to that (Hunter et al. 2012). (Release 29.0 from October 2010 had 13,382 family entries; Release 16.1 from October 2007 had 9,729 entries.)

It should be noted that SPADA is unlikely to perform well with genes that have large numbers of exons due to the combinatoric explosion of potential splice donor and acceptor pair combinations to evaluate. For longer multi-exon gene families, AUGUSTUS-PPX should be used. Still, SPADA has been shown here to be extremely effective in predicting families of one- and two- exon genes often missed or excluded by standard gene prediction algorithms (Basrai, Hieter, and Boeke 1997; Lease and Walker 2006). Hence, it is anticipated that gene annotation pipelines would be improved by

routinely running SPADA to pick up small genes in addition to the standard generic gene prediction algorithms (e.g., Augustus) for larger genes.

Impact of better gene prediction algorithms on plant genomics

As sequencing costs have come down, there has been a commensurate expansion in the sequencing of multiple plant genomes within each species. Moreover, Genome Wide Association (GWA) studies are now routinely carried out in these populations. Gene annotation cannot be simply transferred across members of a species due to the myriad of SNPs and indels that alter gene structures. Gan *et al.* estimated that gene structural changes occurred in more than 30% of genes among the 18 Arabidopsis accessions they resequenced and assembled (Gan et al. 2011). Further, GWA studies have repeatedly implicated unannotated intergenic regions as having the most significant association with important agronomic traits (Brachi et al. 2010; Kump et al. 2011). While it is likely that many of these GWA peaks identify non-coding RNAs or regions in strong linkage disequlibrium with causative variants, we suspect that many of these sites may actually mark members of as yet unannotated families of small genes. Indeed, in a recent GWA study (Stanton-Geddes et al. 2013), many peaks turned out to coincide with NCR or other CRP family members that prior to our intensive family-based annotation studies had been un-annotated in *Medicago*.

Discovery of genes resembling Nodule-Cysteine-Rich (NCR) peptides in Arabidopsis

In striking contrast to Arabidopsis, the *Medicago* genome harbors a huge number (583 versus 3) of Nodule Cysteine-Rich peptides (NCRs, CRP1130-CRP1530) – Defensin-Like proteins with nodule-specific expression (Table 1.1). These NCRs are unique to *Medicago* (specifically, legumes in the Inverted Repeat-Lacking Clade) (Silverstein, Graham, and VandenBosch 2006) and have recently been shown to play vital roles in the communication between *Medicago* and symbiotic rhizobia (Wang et al. 2010; Van de Velde et al. 2010; Farkas et al. 2014).

Surprisingly, three CRPs were found in the Arabidopsis genome falling into the nodule-specific sub-families (CRP1130-CRP1530, or NCRs). Previously, NCRs were thought to be unique to *Medicago* and other IRLC legumes, playing a vital role in the legume-rhizobia symbiotic interaction (Mergaert et al. 2003). Looking closely at the sequence alignments (Appendix Figure 1.7), these "Arabidopsis NCRs" have all the conserved cysteine residues in the expected configuration, while also exhibiting substantial divergence from *Medicago* NCRs - and forming a separate Arabidopsis-specific clade. Furthermore, only one Arabidopsis NCR is predicted in each sub-class. It is possible, therefore, that these "Arabidopsis NCRs" are descendants from the most recent common ancestral genes that later evolved into *Medicago* NCRs. After the Arabidopsis-*Medicago* divergence, these ancient NCRs could have become increasingly divergent in the legume (*Medicago*) clade, eventually gaining new functions in nodule development and symbiosis, possibly through neo-functionalization, conferring a selective advantage and thus increasing rapidly in copy number through gene duplication.

Limitations of the SPADA pipeline

Because the model prediction step in the pipeline is not optimized for multi-exon gene models nor the extremely large introns present in animal genomes, I do not yet recommend SPADA to identify small peptides in animals (especially mammals). Also, SPADA is not expected to work well with bacterial genomes due to the absence of introns in their gene models. However, this pipeline will work well with organisms such as yeast, oomycete and fungi, since they have similar gene structures to plants (Galagan et al. 2005; Saxonov et al. 2000). In fact, it was recently found that oomycetes and fungi genomes encode large number of secreted effectors as a result of the evolutionary "arms-race" between pathogen and host (Haas et al. 2009; Spanu et al. 2010). Potentially, SPADA will be useful in effector discovery in these pathogen genomes given that a growing number of informative family alignments are becoming available.

The SPADA pipeline is useful beyond secreted peptides and outside plants

Although SPADA was initially designed to target secreted peptide families in plants, it can be used on non-secreted peptide families, especially in fungal systems. In the Results section, I tested SPADA using draft genome contigs of the mushroom *A. bisporigera* in search of additional members of a class of potent liver toxin pro-peptides characterized in earlier work (Hallen et al. 2007). Roughly 20 pro-peptides belonging to this family had been cloned and sequenced, with only about a dozen present among the draft genome sequence contigs. SPADA identified 5 new family members with convincing alignments and significant E-values ($2.4\times10^{-17} < E < 5.8\times10^{-8}$). Three of these were in contigs long enough that extensive homology in the 3'-UTR region characteristic of the family could be observed. The contigs of the remaining two hits ended shortly after the coding sequence preventing 3'-UTR homology characteristics from being confirmed. When the same input HMM constructed from the known 20 pro-peptides (Hallen et al. 2007) was used to scan against a related mushroom *Amanita thiersii* that is known not to produce toxic peptides, SPADA did not identify any candidate genes with $E < 0.01$.

Conclusions

SPADA is a homology-based gene prediction program to accurately identify and predict the gene structure for short peptides with one to a few exons. SPADA works well on small plant peptides such as the cysteine-rich peptide families. SPADA gives much more accurate and precise gene calls than traditional ab initio gene finding programs in tested genomes. Running SPADA on less well-annotated plant genomes (e.g., *Medicago*) reveals numerous mis-annotated and unannotated CRPs in the current genome annotation. Predictions made by SPADA constitute the most complete set of plant cysteine-rich peptides, and in this regard, will provide an invaluable resource for the research of small, secreted peptides in plants. The systematic application of SPADA to other classes of small peptides by communities will greatly improve the genome annotation of different protein families in public genome databases.

Table 1.1. Cysteine-Rich Peptides (CRPs) predicted in *A. thaliana* and *M. truncatula*

	Subgroup ID[#]	*A. thaliana*	*M. truncatula*
Defensin related	CRP0000-CRP0260,etc.	56	43
LCR/BET1 related	CRP0280-CRP0810,etc.	162	110
SCR related	CRP0830-CRP0880	32	6
Metallocarboxypeptidase inhibitor	CRP1004-CRP1030	0	1
CCP related	CRP1040-CRP1120	19	4
Nodule Cysteine-Rich peptide	CRP1130-CRP1530	3	583
Ripening related protein	CRP1600-CRP1605	0	21
Novel family	CRP1620,CRP2800,etc.	14	15
Miscellaneous	CRP1640-CRP1660,etc.	16	48
Rapid Alkalinization Factor	CRP1700-CRP2120	38	36
Thionin related	CRP2200-CRP2610	66	23
Root cap/late embryogenesis	CRP2820-CRP2850	5	7
Antimicrobial peptide MBP-1	CRP2900-CRP3000	1	2
Bowman Birk inhibitor	CRP3100-CRP3190	0	16
Pollen Ole e I	CRP3300-CRP3510	34	44
ECA1 gametogenesis related	CRP3600-CRP3740	124	17
Lipid transfer protein	CRP3800-CRP4962	127	127
2S Albumin	CRP4970-CRP5080	5	3
Glutenin/Giadin/Prolamin	CRP5090-CRP5270	0	0
Maternally-expressed gene/Ae1	CRP5300-CRP5520	20	2
Proteinase inhibitor II	CRP5545-CRP5600	6	2
Chitinase/Hevein	CRP5610-CRP5820	10	15
Kunitz type inhibitor	CRP6010-CRP6180	7	45
Total		**745**	**1170**

[#]CRP subgroup identifiers as assigned in Silverstein *et al.* (Silverstein et al. 2007).

Table 1.2. Manual inspection confirms novel CRP models predicted by SPADA.

	A. thaliana	*M. truncatula*
Total predictions	745	1170
Number of unannotated predictions[#]	5	125
Number of novel models[%]	3 (60%)	77 (62%)

[#]An unannotated prediction is a gene model predicted by SPADA but missed by current genome annotation.

[%]Novel models are unannotated predictions that are manually inspected to be true members of the family with evidence from family-specific alignment and/or RNA-Seq evidence.

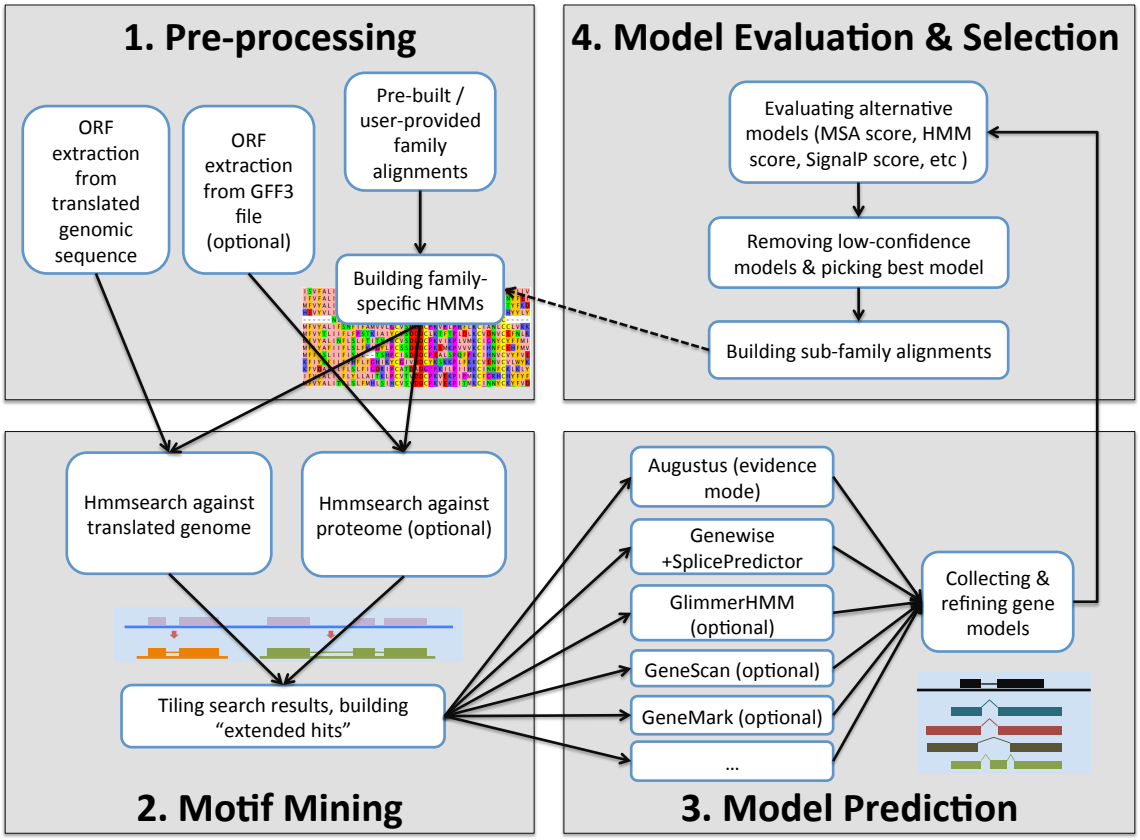


Figure 1.1. The SPADA workflow.

In the pipeline, SPADA first builds family-specific multiple sequence alignments and prepares genomic and protein sequences (Pre-processing). It then runs HMMer to identify hits in the target genome sequence and proteome sequence (Motif Mining). Next, SPADA runs several gene predicting programs (components) to generate candidate gene models (Model Prediction). Finally, SPADA picks the best candidate models for each hit based on multiple statistics (Model Evaluation & Selection).

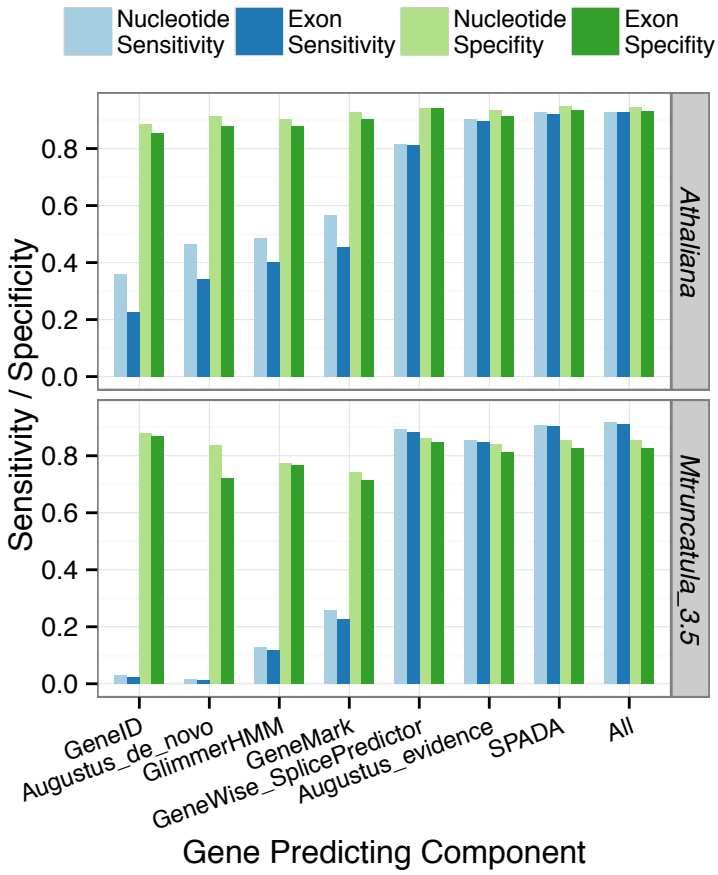


Figure 1.2. Performance comparison of different gene prediction components.

Search E-value threshold is set to 0.001 by default.

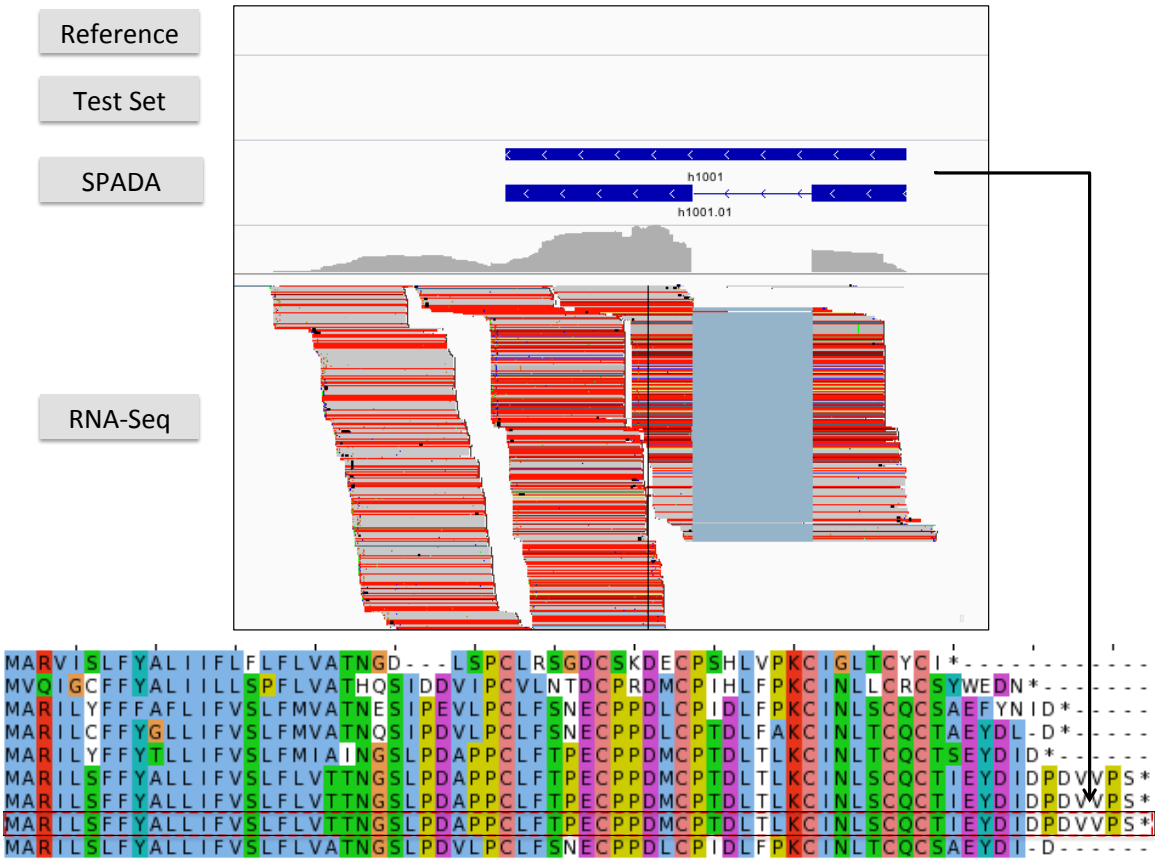


Figure 1.3. A novel gene model predicted by SPADA is missed by the current *Medicago* annotation.

A *Medicago* NCR (h1001.01, track "SPADA") and subgroup alignment of CRP1180 with h1001.01 shaded.

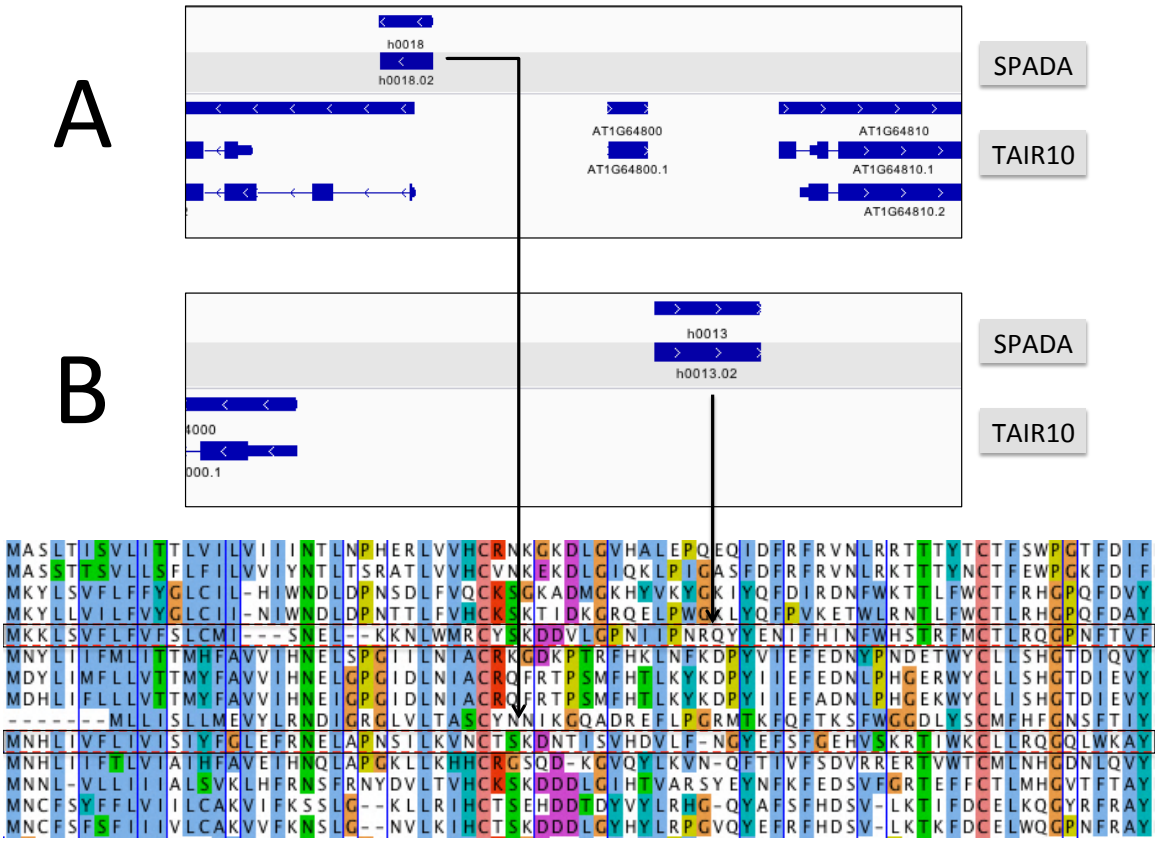


Figure 1.4. SPADA detects mis-annotated and novel SPH peptides in TAIR10.

(A) SPADA detects an SPH peptide (h0018.02) that is mis-annotated in TAIR10; (B) SPADA detects a novel SPH peptide (h0013.02) not present in TAIR10. Multiple sequence alignment of selected SPH peptides are shown below with h0018.02 and h0013.02 shaded.

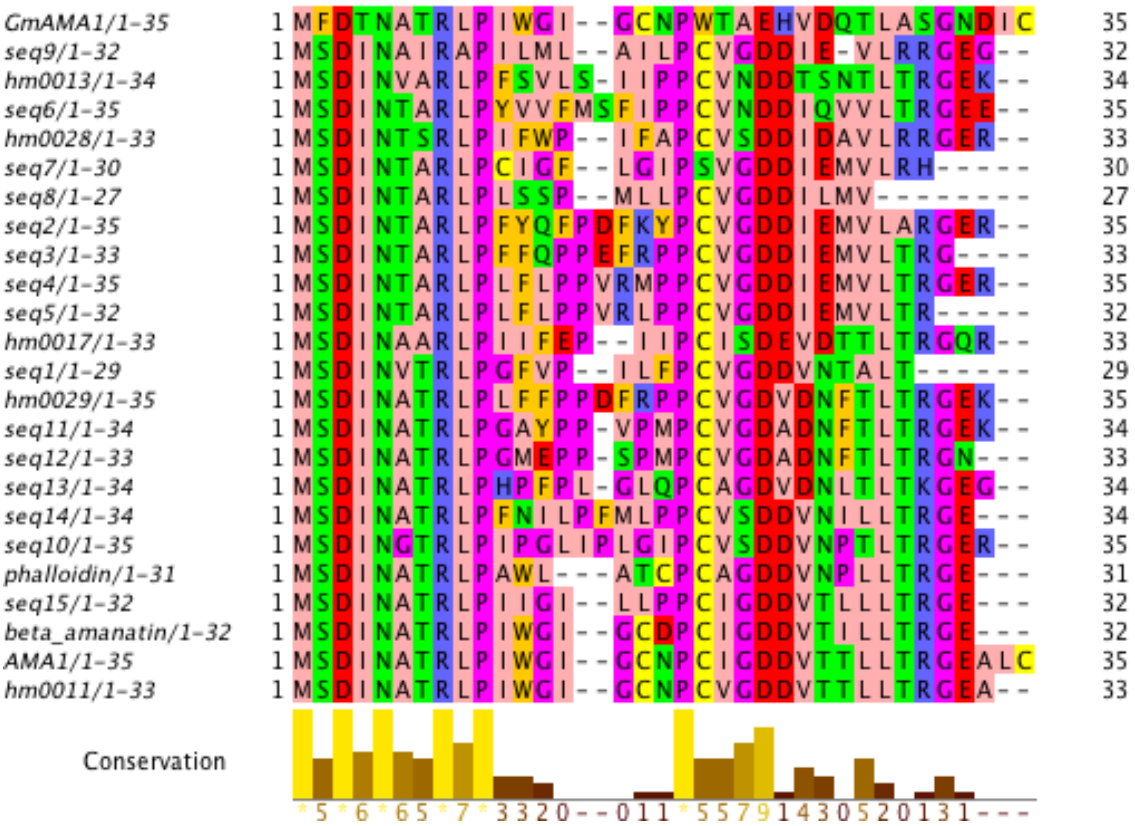


Figure 1.5. Multiple sequence alignment of Amanita toxin proproteins.

Sequences identified by SPADA are labeled as "hm****". All remaining sequences were obtained from Hallen *et al.* (Hallen et al. 2007), and were included in the initial alignment used as input for SPADA.

Chapter 2. The *Medicago* Pan-16 genome enables exploration of novel genetic variation

Medicago truncatula is a model for investigating legume genetics and the evolution of legume-rhizobia symbiosis. Previous studies using whole-genome sequence data to identify sequence polymorphisms (SNPs and short Insertion/Deletions) relied on mapping short reads to a single reference genome. However, limitations of read-mapping approaches have hindered variant detection in repeat-rich and highly divergent regions, as well as studies of large gene families potentially involved in disease resistance and plant-microbe symbiosis. *De novo* assembly and annotation of the genomes of 15 *M. truncatula* accessions allowed me to detect novel genetic variation that would not have been found by mapping reads to a single reference. This analysis leads to a within-species diversity estimate nearly 70% higher than previous mapping-based resequencing efforts, even based on a smaller sample size. The results clearly demonstrate that *de novo* assembly-based comparison is more accurate and precise than reference mapping-based variant calling in exploring variation in repetitive and highly divergent regions. For the first time in plants, this work systematically identified and characterized different types of SVs using a synteny-based approach. The results suggest that, depending on the divergence from the reference accession, as much as 7% to 21% of the entire genome is involved in large structural changes, altogether affecting 10% to 28% of all gene models. Based on genome-wide synteny alignments against the reference genome, these results identified 63 Mbp unique sequence segments absent in the reference, including 30 Mbp shared by at least two accessions and 34 Mbp of accession-specific sequences, thus expanding the published reference space for *Medicago* (389-Mbp) by 16%. Pan-genome analysis suggests a “restricted” pan-genome (450 Mbp) and core-genome size (270 Mbp) curve, both beginning to level off when ten or more accessions become sequenced. This work illustrates the value of multiple *de novo* assemblies and the strength of comparative genomics in exploring and characterizing novel genetic variation within a population,

providing insights into the impact of structural variation on genome architecture and large gene families underlying important traits.

Introduction

A better understanding about the nature and extent of genome variation within *M. truncatula* is critical for both practical and scientific reasons. The high quality, BAC-based sequence of the *Medicago* A17 genome has served as the "reference genome" for the *Medicago* research community. This reference genome assembly (version 4.0) covers ~80% of the overall genome (estimated at 465 Mbp) while capturing ~93% of all predicted gene models (Bennett and Leitch 2011; Tang et al. 2014). Recently, resequencing of additional *Medicago* accessions from diverse geographic locations has enriched the pool of sequence information available for *Medicago* (Branca et al. 2011; Stanton-Geddes et al. 2013). However, these studies relied on mapping short reads to a reference sequence in order to call polymorphic sites (i.e., read mapping-based approach). This introduces a potential bias due to significant structural differences between diverse *Medicago* accessions. Alignment to a single reference is most problematic when reads from a divergent *Medicago* accession, such as the functionally important R108, are incorrectly aligned to the A17 reference due to mis-alignment of paralogous regions. With a high within-species nucleotide diversity (genome-wide estimates of $\theta_w = 0.0063$ and $\theta_\pi = 0.0043$ bp^{-1} approximately three times more than found in soybean $\theta_{w\text{-cultivated}} = 0.0017$ bp^{-1} and $\theta_{w\text{-wild}} = 0.0023$ bp^{-1} populations), it may not be surprising that only a portion of the (reference) genomic regions can be confidently probed with read-mapping approaches. Diversity estimates may thus be underestimated due to the elimination of divergent (un-aligned) sequences. In addition, the length limitation of short read technologies such as Illumina leads to ambiguous mappings in repetitive and duplicated regions. Finally, the incompleteness of the reference genome limits our ability to detect variation in regions not present in the reference (e.g., assembly gaps) and leads to false alignments and SNP calls if reads from these regions align to their next best match. Taken together, over-reliance on read mapping-based approaches have hindered our understanding of the types and extent of genomic diversity in *Medicago*.

Structural variants (SV) include unbalanced copy number variation such as deletions, insertions and duplications, as well as balanced variants such as inversions and translocations (Weischenfeldt et al. 2013). In the case of humans, research suggested the number of nucleotide differences between individuals due to SVs is greater than that due to SNPs (Redon et al. 2006). Studies of model organisms such as Arabidopsis, Drosophila, humans, and maize have also revealed considerable levels of segregating structural polymorphism (Korbel et al. 2007; Emerson et al. 2008; Kidd et al. 2008; Swanson-Wagner et al. 2010; Cao et al. 2011; McVean et al. 2012). As a major contributor to genetic variation, genomic SV is thought to be an important factor in determining phenotypic variation for a wide range of traits, such as the digestion of starchy foods in humans (Perry et al. 2007), dwarfism and flowering time in wheat (*Triticum aestivum*) (S. Pearce et al. 2011; Díaz et al. 2012), insecticide and virus resistance in *Drosophila melanogaster* (Schmidt et al. 2010; Magwire et al. 2011) and resistance against soybean cyst nematode conferred by the *Rhg1* locus (Cook et al. 2012). Owing to inherent difficulties in their ascertainment, however, SVs have remained a relatively poorly understood form of genetic variation in comparison to SNPs.

The distribution of structural variation (SVs) in the *Medicago* and their impacts on genome architecture and important gene families remain largely unknown, yet essential in gaining insight into the genetic basis of legume-rhizobium symbiosis and identifying genes underlying phenotypic variation. The only practical way to understand the genomic diversity of *Medicago* fully and to capture the numerous members of divergent gene family members is to carry out whole genome sequencing and *de novo* assembly. Recent advances in NGS chemistry and computational approaches in sequence assembly have significantly improved the power and reliability of *de novo* assembly of NGS data. In this study, I analyzed *de novo* genome assemblies of 15 strategically chosen *Medicago* accessions. Working with colleagues, I constructed a *Medicago* Pan-16 genome and proteome, and fully characterized the genomic content and gene family repositories for these accessions. I performed whole genome alignment against the HM101 reference, constructed genome-wide synteny blocks and identified of extensive variation including

SNPs, Indels and complex SVs for each accession. The results showed that traditional variant detection approach based on mapping reads to a single reference genome is much less accurate than the synteny-based variant calling approach especially in repetitive and highly divergent genomic regions. We found that diversity estimates were significantly underestimated by previous mapping-based variant detection approach and were able to determine the extent and distribution of SVs that were not detectable by previous approaches. This work illustrates the value of multiple *de novo* assemblies in building and characterizing plant pan-genomes and provides insights in understanding the impact of different types of SVs on genome architecture and large gene families underlying important traits.

Methods

Plant material

Fifteen *M. truncatula* accessions from geographically distinct populations (Figure 2.1) broadly spanning the entire *Medicago* species range were chosen for deep sequencing and *de novo* assembly. These accessions were chosen for both biological interest and to facilitate evaluation of assemblies. In particular, 3 accessions were selected from the A17 (reference) clade, nine were selected from the France-Italy clade and 3 were picked from more distantly related clades. While most analyses were done on all 16 accessions including the reference HM101, some statistics sensitive to population structure were derived from a subset of 13 accessions (16 excluding the 3 distant accessions), which we refer to as "ingroup" accessions. Each accession was self-fertilized for 3 or more generations before growing seedlings for DNA extraction. Cloning and sequencing grade DNA was extracted from a pool of ~30 day old dark-grown seedlings by Amplicon Express (Pullman, WA) through Ultra Clean BAC Clone Prep followed by a CTAB liquid DNA prep.

Sequencing and genome assembly

Library preparation, sequencing and assembly described herein were performed at the National Center for Genome Resources (NCGR) in Santa Fe, NM. DNA sequencing was performed using Illumina HiSeq 2000 instruments. For each of the fifteen accessions, we made and sequenced one Short Insert Paired End library (SIPE) and either one or two Long Insert Paired End (LIPE) libraries following the requirements and recommendations of the ALLPATHS-LG whole genome assembler (Gnerre et al. 2011). The SIPE library is a 180 bp fragment library sequenced as 2 × 100 bp reads, while the two LIPE libraries are jumping libraries with insert sizes around 5 kbp and 9 kbp, and sequenced as 2 × 50 bp and 2 × 100 bp reads, respectively.

For the SIPE library the sample was mechanically fragmented by using the Covaris S2 System and then prepared based on the Kapa Hyper Prep Kit and ligated to standard Illumina paired-end adapters. Blue Pippin size selection was done to yield fragments of 300 nucleotides (180 nucleotides plus adapters).

For the LIPE libraries, either the Illumina or Nextera mate-pair library protocol was used. The DNA was size selected at 5 kb for the Illumina libraries or, using the Blue Pippin, at 9 Kb for the Nextera libraries. Nextera libraries were fragmented using the Covaris S2 System. Each of the libraries was sequenced to 40x to 100x sequence coverage, as recommended by the assembler algorithm (Appendix Table 2.1). ALLPATHS-LG assembly algorithm (version 49962) (Gnerre et al. 2011) was run on a linux server with default parameters to complete the assembly.

PacBio long read data was also generated to validate the identified structural variation in three accessions including HM034, HM056 and HM340. Sequencing was done using P4C2 chemistry and Smrtanalysis version 2.1 or P5C3 chemistry and Smrtanalysis version 2.3. Reads were filtered at minimum quality of 75, minimum sub-read length of 50 bp and minimum read length of 50 bp. Final coverage for each accession was estimated to be 18-20 fold.

Genome size estimates for Medicago accessions

Seeds of nine *Medicago* accessions (HM004, HM005, HM006, HM029, HM030, HM034, HM056, HM101 and HM324) were obtained from the *Medicago* HapMap project (http://www.medicagohapmap.org/). Seeds of known genome size standards were obtained from Dolezel *et al.* (Jaroslav Doležel and Bartoš 2005). Seedlings were grown in a growth chamber at 25°C under identical light and humidity conditions. Samples of leaf nuclei were prepared essentially following the procedure of Dolezel *et al.* (Jaroslav Doležel and Bartoš 2005) using Galbraith's buffer (Galbraith et al. 1983). Samples were analyzed on a BD FACSCalibur flow cytometer at the Biodesign Institute, Arizona State University. Mean DNA content was based on 15,000 nuclei, with peak means identified using CellQuest software (Becton Dickson). Calculation of genome size used the sample peak mean divided by the standard peak mean, multiplied by the genome size of the standard. Each plant accession was sampled three or more times on different days to minimize the effect of instrument drift and other variables. For most accessions three or more different plants were sampled, but occasionally only two plants were sampled. Values reported are averages of the samples from each accession. We used a value of 2.3 pg for the genome size of *Glycine max* Polanka rather than the 2.5 pg indicated in Dolezel *et al.* (Jaroslav Doležel and Bartoš 2005) based on our results obtained with other standards. Doležel *et al.* (J. Doležel, Doleželová, and Novák 1994) estimated 2C DNA amount of *Glycine max* 'Polanka' as 2**.**5 pg using human male leucocytes with 2C = 7.00 pg. However using *Raphanus sativus* Saxa at an estimated genome size of 1.11 pg, *M. truncatula* Jemalong (HM101) at an estimated genome size of 1.15 pg and *Lycopersicon esculentum* Stupicke at an estimated genome size 1.96 pg, we found that 2.3 pg was our average estimate of genome size for *Glycine max* Polanka using our methods and the same FACSCalibur flow cytometer used to sample all accessions.

Functional annotation

Repeat elements were masked using RepeatMasker (Smit, Hubley, and Green 1996) with the *M. truncatula* repeat library. AUGUSTUS (Stanke and Waack 2003) was used to

make *ab initio* gene predictions for each genome assembly with both RNA-Seq expression evidence and HM101 homology evidence. RNA-Seq data came from sequencing of four diverse accessions, HM034, HM056, HM101 and HM340, performed by colleagues in Dr. Robert Stupar's research group. Reads from HM034, HM056 and HM340 were directly mapped to their *de novo* assemblies using Tophat (Trapnell, Pachter, and Salzberg 2009) to generate intron hints for AUGUSTUS. For the remaining 12 accessions, we mapped the RNA-Seq reads from the closest available accession (one of HM034, HM056, HM340 or HM101) to the corresponding assembly and generated intron hints. We also transferred HM101 (Mt4.0 reference) annotation to each *de novo* assembly using synteny block (see next section: Comparative genomics analysis) information and generated exon hints for AUGUSTUS. Predicted protein sequences were scanned for PFAM domains (Pfam-A.hmm) (Finn et al. 2014) using HMMER (Sean R. Eddy 2011) and processed using custom scripts that I created. Domain categories were then assigned to each protein sequence according to the most significant Pfam hits. Among the resulting Pfam domains we curated 133 (Appendix Table 2.2) as being associated with transposable elements and grouped these into a large "TE" category. NBS-LRR genes were curated using sub-family HMMs (13 TNL subgroups and 22 CNL subgroups) built based on previous literature (Ameline-Torregrosa et al. 2008). I also ran SPADA (Small Peptide Alignment Discovery Algorithm) (Zhou et al. 2013), as described in detail in Chapter 1, on each assembly to refine annotation of 516 CRP gene families.

Comparative genomics analysis

Each *de novo* assembly was first aligned to the *Medicago* HM101 reference sequence (version 4.0) using BLAT (Kent 2002). Unaligned sequences (query sequences with no hit to the reference genome) were extracted and BLAT-aligned for a second time because it was found that BLAT tends to over-extend gap length whenever it encounters an assembly gap (stretches of 'N's). The resulting alignments were merged, fixed (removing non-syntenic or overlapping alignment blocks), cleaned (removing alignment blocks containing assembly gaps) using custom scripts. BLAT Chain/Net tools were then

used to obtain a single coverage best alignment net in the target genome (HM101), as well as a reciprocal-best alignment net between the two genomes. Finally, genome-wide synteny blocks were built for each *de novo* assembly (against HM101), enabling downstream analyses including variant calling, novel sequence identification and ortholog detection.

Synteny-based variant detection and structural variation identification

Based on the synteny blocks built for each *de novo* assembly and HM101, I identified SNPs (single base mismatches), short insertions and deletions (alignment gaps ≤ 50 bases), as well as large SVs. As illustrated in Figure 2.2 and in the Results, this enabled detection of large deletions (deleted sequence not present anywhere in the target genome), insertions (inserted sequence not present anywhere in the query genome), translocations (deleted sequence is inserted somewhere else in the genome and present only once) and copy number gains and losses (deleted or inserted sequence present elsewhere in the genome but belongs to another synteny block - i.e., present more than once). Variant calls from 15 accessions were converted to Variant Calling Format (VCF) and merged to a single VCF file using Bcftools (H. Li et al. 2009). Custom scripts were run to fill in missing genotypes (variants called in one accession but not another) where reference-genotype could be concluded according to synteny alignment information.

Novel sequence identification and Pan-16 genome construction

A raw set of novel sequences (i.e., sequences present in one or more *de novo* assemblies but absent in HM101) were obtained by subtracting all the aligned regions from the entire gap-removed assembly. Low-complexity sequences and tandem duplications were then scanned and removed using Dustmasker (Camacho et al. 2009) and Tandem Repeat Finder (Benson 1999). The remaining novel sequences were BLAST (Altschul et al. 1990) against the NCBI Nonredundant nucleotide (NT) database to determine their apparent species of origin, and classified into three categories: plant origin (having a best hit in *Medicago*, soybean, lotus or other plant species), foreign (best hit in non-plant species, e.g., endophytic bacteria, eukaryotic, etc.) and unknown (no

hits). Foreign sequences were filtered out in subsequent analyses. In order to understand the sharing status of novel sequences among different accessions, I ran Para-Mugsy (parallel version of MUGSY - the multiple whole genome aligner) (Angiuoli and Salzberg 2011) to build a multiple alignment of all novel sequence segments identified in 15 accessions. The resulting alignments were parsed and analyzed using custom scripts to determine how each segment is shared among accessions – e.g., private to one accession or shared by multiple accessions. We then constructed a *Medicago* Pan-16 genome which includes the Mt4.0 (HM101) reference as the backbone, as well as all non-redundant novel segments identified in the other 15 accessions. We further derived a pan-genome size curve and a core-genome size curve by adding one *de novo* assembly to the pool at a time and calculating the size of shared genomic regions (core-genome) and the size of total non-redundant sequences (pan-genome).

Comparison of variants (SNPs, short indels) identified by a reference-mapping approach and de novo *assembly-based approach*

Sequencing reads were also mapped to the Mt4.0 reference using GSNAP (Wu and Nacu 2010). Resulting alignments were converted to a BAM (H. Li et al. 2009) file which went through the standard GATK pipeline (DePristo et al. 2011) including duplication removal, indel realignment and base recalibration. SNPs and short indels were then called using GATK UnifiedGenotyper with default parameters. These mapping-based variants were evaluated against our synteny-based variant calls to generate a Venn Diagram for each accession, which includes an overlapping set (variants called by both approaches), a mapping-only call set (variants called by the GATK pipeline but not by the *de novo* assembly comparison) as well as an assembly-only call set. Since insertion/deletion positions are sometimes ambiguous, we left-aligned all indels from both approaches using the GATK LeftAlign module prior to the comparison. Nevertheless, multiple nucleotide polymorphisms (MNP) could still have different representations between the two approaches. As a result, the actual intersection of indels called from two approaches is probably underestimated. To see whether the assembly-

only SNP calls are enriched in more divergent genomic regions, we partitioned the genome into 1-kbp windows. For each window we then calculated sequence percent identity and the proportion of assembly-only SNPs calls (out of all SNPs called).

Results

Sequencing and de novo assembly

Each accession was sequenced with Illumina HiSeq2000 using a combination of short and long insert paired-end libraries with insert sizes of 180 bp and 9 kbp, for an average of 120 fold coverage (Appendix Table 2.1), and assembled using the ALLPATHS-LG whole genome assembler (see Methods). All fifteen genomes were well assembled: approximately 80%-94% of each of the genomes were assembled into scaffolds at least 100 kbp long, with scaffold N50 sizes ranging from 268 kbp to 1,653 kbp, and contig N50 sizes around 20 kbp (Table 2.1). Mate-pair libraries with large insert size significantly improved scaffolding, with the longest scaffold size reaching 16 Mbp - almost the length of a chromosome arm. Assembled genome sizes ranged from 388 Mbp to 428 Mbp, correlating well with the experimentally derived genome size estimates (Correlation coefficient = 0.83, P-value = 0.005, Figure 2.3). Gap-removed genome sizes are smaller than but comparable to the BAC-based Mt4.0 (HM101) reference genome (Table 2.1). About 20% of each assembly were annotated as repeat elements – slightly lower than 23% repetitive content in Mt4.0 reference (Table 2.1), an indication that repetitive sequences been missed / collapsed in these *de novo* assemblies to some extent. However, these assemblies essentially capture 87-96% of the unique contents in the reference genome space, including 90-96% of genic coding regions. As such, the current assemblies enabled accurate detection of variation and comparative analyses within genic regions.

Functional annotation, identification of CRPs and NBS-LRRs

Evidence-guided annotation integrating homology searches, RNA-Seq expression, and *ab initio* prediction yielded comparable numbers of coding genes (60,000 to 67,000)

for each of the 15 assemblies (Table 2.2). The number of transposable element (TE)-related genes identified in the 15 accessions were on average 20% lower than the HM101 reference – confirming that *de novo* assemblies have missed or collapsed some repetitive sequences. A closer look at the number of different TE categories suggests certain TE families were more likely to be missed / collapsed in *de novo* assemblies than others (Appendix Table 2.4), probably reflecting different sequence characteristics (GC content, mappability, etc.). Median protein length (TEs excluded) ranged from 218 to 228 amino acids – nearly equal to the estimate of 228 amino acids in HM101, an indication that long scaffold and contig N50s have helped keep the gene models intact. The quality of the annotation was supported by the observation that 77-87% of genes were either supported by an HM101 homolog or expressed (as determined by RNA-Seq) (Table 2.2). Large gene families such as NBS-LRRs and CRPs generally have consistent numbers of members among accessions (Table 2.2, Appendix Table 2.3); however, these gene families harbored complicated homology relationships with the accessions differing markedly in the size of specific sub-families (Appendix Table 2.5, 2.6). Further analysis (see Chapter 3) suggests family-specific expansion / contraction is a frequent phenomenon observed in large gene families.

Comparative analysis

I was able to align 92%-96% of each assembly with the HM101 reference (Table 2.3) using custom pipeline (see Methods). I also identified ~300 Mbp of sequences in syntenic blocks between each assembly and HM101, where SNPs, short indels and large SVs can be confidently determined. Global comparison revealed large syntenic blocks of conserved genomic regions as well as poorly aligned regions where structural changes frequently take place (Figure 2.4). The analysis suggests that synteny tends to break down near centromeric regions, where TE density is high, as well as regions enriched in highly variable gene families such as NBS-LRRs (Figure 2.4 and Figure 2.6, “covered bases” track). I found that chromosomes 3 and 6 have the highest level of non-centromeric structural changes (breaks in synteny) across all 15 accessions (Figure 2.4);

interestingly, these two chromosomes also harbor the highest number of NBS-LRRs, together accounting for 43% of the entire NBS-LRR family. Indeed, while ~80% of non-TE coding genes (genic regions) are located within conserved synteny regions, a loss of synteny between two compared accessions was often observed in regions containing either transposable elements (TEs, 53% in synteny blocks versus 47% outside) or rapidly evolving gene families such as NBS-LRRs (63% in synteny blocks versus 37% outside). It is therefore clear that rapidly evolving gene families, particularly TEs and NBS-LRRs, significantly increase local genome fragility and contribute to the overall genome architecture.

Global view of variation (SNPs, short indels, SVs)

In aligned genomic regions we found extensive variation including SNPs, short indels and large SVs. Altogether, I identified between 1.7 million (HM058) and 5.1 million (HM340) single nucleotide changes as compared to the HM101 reference (Table 2.4). As expected, SNP density correlates well with divergence from HM101 - with SNP bp^{-1} ranging from 0.63% of HM058 (closest to HM101) to 2.37% of HM340 (most distant from HM101). These estimates are much higher than previous reports based only on aligning next generation reads to the reference genome sequence. While sequencing errors, assembly errors and alignment ambiguities could all lead to elevated SNP rate estimates, many of the identified substitutions here are very likely to represent true diversity in regions previously overlooked by read-mapping approaches (see Discussion).

In addition to SNPs, I identified between 200,000 and 700,000 short insertions and deletions (size less than 50 bp), altogether affecting 1.5 - 5.3 Mbp. Large SVs including large insertions, deletions, translocations and copy number changes were also identified at base pair accuracy using synteny block information (Table 2.4; Figure 2.5). Although much rarer in occurrence, these large changes affect many more bases than do smaller changes (SNPs and short indels, Table 2.4) do. For example, there are 255k total small deletions in HM056 affecting 1.5 Mbp sequences, but 16.5k large deletions affecting 6.3 Mbp sequences. This is consistent with findings in other systems where large variants

typically have greater structural impacts on the genome (Redon et al. 2006; Conrad et al. 2010). My analysis identified nearly equivalent number of small insertions and deletions, which reflects the randomly arising nature of indels and is in strong contrast with traditional read mapping-based approach that typically calls more deletions than insertions (relative to reference sequence). When it comes to larger variants, however, the protocol used here still sees a limited power in detecting large insertions and copy number gains (CNG) – as evidenced by the number of large deletions and copy number loss (CNL) events being 30-50% higher (Table 2.4). Given that the high-quality BAC-based Mt4.0 reference is both more complete (8-10% larger) and continuous than the 15 *de novo* assemblies, sequences present in the reference but absent in other accessions (i.e., large deletions and CNLs) would have a higher chance to be picked up than large insertions and CNGs do.

A total of 7% (HM058) to 22% (HM022) of the entire genome content is affected by at least one type of structural changes (Table 2.4). Sequences in these regions are either highly divergent (and thus cannot be aligned), or deleted, inverted, duplicated or translocated to ectopic genomic locations (Figure 2.5). When using reference-mapping variant calling approaches, sequence affected by a simple deletion is simply removed in one accession, so this will manifest a sudden drop of read coverage in the deleted area. However, in the case of translocation and copy number changes, the actual sequence content is still present in the affected genome but at a different genomic position. Short read aligners will still try to place reads in these structurally different regions, resulting in erroneous SNP calls, which could be misleading for the downstream population genetics analysis and association studies. The synteny-based variant calling approach we employed, however, restricts SNP calling to syntenic blocks and thus provides a more accurate view of true genomic variation (see Discussion for detail).

Population genetics of identified variants

Based on the ~7 million SNPs I obtained genome-wide nucleotide diversity estimates (θ_π = 0.0073 bp^{-1} and θ_w = 0.0082 bp^{-1}, Table 2.5), values much higher than the

economically important legume *Glycine max* ($\theta_{w\ cultivated}$ = 0.0017 bp^{-1} and $\theta_{w\ wild}$ = 0.0023 bp^{-1}) (Lam2010). Approximately 70% of the SNPs were found in intergenic regions, which are also featured by the highest level of nucleotide diversity (θ_π = 0.0089 bp^{-1}). Diversity was much higher for synonymous than replacement polymorphisms in coding regions (Table 2.5). The minor allele frequency (MAF) spectrum also showed higher frequencies of rare variants (present in only one accession) for replacement and large effect SNPs (lost of start or stop codon, splice site variant, etc.) than other types of polymorphisms without an apparent functional impact (Appendix Figure 2.1). These findings are consistent with the expectation of stronger purifying selection acting at replacement sites, especially large-effect polymorphisms significantly changing the protein product (Nielsen 2005).

Our estimates of nucleotide diversity (θ_π = 0.0073 bp^{-1} based on 13 ingroup accessions) were 70% higher than previous results (θ_π = 0.0043 bp^{-1} based on 26 accessions) (Table 2.5) (Branca et al. 2011). This probably reflects a better estimate of the true genomic diversity within *M. truncatula* for the following reasons: (1) a better reference genome (Mt4.0 versus Mt3.0) allows more (potentially divergent) genomic regions to be assessed; (2) higher sequencing depth and newer sequencing chemistry (jumping libraries) help resolve repetitive (and highly variable) genomic regions and gene families; (3) directly comparing *de novo* assembly to a reference allows access to more repetitive and highly divergent genomic regions (down to 70% percent identity), which is not possible by read mapping-based approach and short read aligners (typically requiring >90% percent identity of a read to the reference) (see Discussion for detail).

Sliding window analysis suggests that SNP-based nucleotide diversity (θ_π) was generally higher at centromeric regions than at telomeric regions, with the exception that certain highly variable gene families such as NBS-LRRs also significantly elevate diversity level (Figure 2.6). While nucleotide diversity was negatively correlated with gene density (r = -0.212, P = 6.58e-47), it is positively correlated with TE density (r = 0.328, P = 0) and NBS-LRR density (r = 0.282, P = 0), and marginally with CRP density (r = 0.089, P = 0), confirming the high diversity of these large gene families (Appendix

Table 2.1). We also calculated a diversity statistic for short indels and large structural variation (Figure 2.6, Pi[InDel] and Pi[SV] tracks). Interestingly, these two statistics show trends very similar with SNP-based nucleotide diversity estimate, an indication of common local ancestries of different types of variation. Also, the diversity of short indels and SVs are negatively correlated with gene density and positively correlated with TE density and NBS-LRR density (Appendix Table 2.7), confirming the fragility and rapid evolution of these gene families.

Novel sequences identification and Pan-16 genome construction

During the comparison, we found extensive "novel" sequence in the 15 *de novo* assemblies that could not be aligned to the Mt4.0 reference even with a relaxed alignment stringency (70%-80% sequence percent identity). These sequences often exist in the form of novel insertions or complex substitutions, and sometimes as separate scaffolds. After filtering potential contaminant (foreign) sequences, we identified 9 - 22 Mbp novel segments (longer than 50 bp) in each of the 15 *de novo* assemblies (Table 2.3). In order to understand how these novel sequences are shared among 15 accessions, we made a multiple sequence alignment of all the novel segments and determined the presence and absence of each segment in each genome. Out of the total 63 Mbp non-redundant novel sequences identified, 47% are present in at least 2 accessions, with the remaining 53% being specific to a single accession (Figure 2.7A). The observation that sequencing 15 additional genomes adds at least 16% more unique content to the reference genome, and that nearly half of the novel sequences are found in more than one other accession, both indicate that having a single reference has seriously limited our understanding of the content of population gene pool and the level of genomic variation in natural *Medicago* populations.

I then obtained a size curve for both the pan-genome and the core-genome (Figure 2.7B), by randomly adding one genome to the population pool at a time. The core-genome size curve first drops quickly but then reaches a plateau after 10 accessions are added. Approximately 270 Mbp sequences were shared among all 16 accessions

including the Mt4.0 reference, representing most of the conserved coding and regulatory regions that presumably play vital house-keeping functions. On the other hand, another ~180 Mbp sequences were missing from at least one accessions ("dispensable"), reflecting the dynamic nature of genome content and prevalence of insertion / deletion, copy number variation and other structural changes in the *Medicago* populations. Similarly, the pan-genome size curve first sees steady increases each time when a new genome was added to the pool, and then reaches a stable value of approximately 450 Mbp (Figure 2.7B). Thus, our analysis of 16 *M. truncatula* accessions already suggests a "restricted" nature of the *Medicago* pan-genome: sequencing additional genomes will expand the current pan-genome by bringing in more (but minimal) novel sequences and variation.

Among the novel sequences identified, 1.3 – 2.5 Mbp per accession were predicted to be protein coding (Table 2.3). Enriched in these novel coding genes are TE-related genes and NBS-LRRs: while on average less than 2% of non-TE genes were identified as absent from the reference accession, as many as 2.8% of TEs and 6% of NBS-LRRs contribute to the "novel" gene pool (Appendix Figure 2.2). Interestingly, these gene families are also among the most poorly characterized genomic regions in our earlier discussion ("Results – Comparative analysis" section, Figure 2.4). In fact, based on the synteny alignments built for each *de novo* assembly we were able to cover 75% (31,048) of all non-TE reference gene models including 77% (554) of NCRs (≥ 80% coding regions in ≥ 10 accessions), while only 44% (5,418) TEs and 48% (414) NBS-LRRs were covered using the same criteria. Enrichment of these complex gene families in novel gene pool partially explains their poor coverage by synteny alignment. The rapid evolving nature and turnover of these genes greatly accelerate their divergence within the population, thus preventing both read mapping and syntenic anchoring. That said, the total number of NBS-LRRs predicted in each *de novo* assembly did not differ significantly from the reference annotation, indicating that instead of being missed by our assembly, these NBS-LRRs most likely were captured and assembled to relatively short and isolated contigs that could not be anchored to reference chromosomes.

Discussion

De novo assembly and comparative analysis enables exploration of both repetitive and highly divergent genomic regions that were previously overlooked by mapping-based strategies

My results show that SNP-based nucleotide diversity has been seriously underestimated by as much as 70%, in previous read mapping-based resequencing studies. This conclusion is explained by an improved reference genome (Mt4.0), higher sequencing depths, better sequencing chemistries, as well as a comparative variant discovery approach based on directly comparing *de novo* assemblies. It becomes essential then, to know better the gains and losses that our "synteny-based" variant calling provides versus the traditional "reference mapping-based" approach. To answer this question, we also ran the same set of original sequence reads through a read mapping and variant detection pipeline for selected accessions, and came up with a second set of variant calls (see Methods). While 75%-80% of the identified SNPs overlap between methods, a considerable fraction of variants (~20%) are only called by one approach (Figure 2.8A, Appendix Figure 2.4). We found that SNPs called only by mapping-based approach ("mapping-only" calls) are highly enriched in SV regions as determined by our assembly comparison: 75% mapping-only SNP calls were found in SV regions which account for ~15% of the entire genome. In other words, short read aligners have a tendency to place reads regardless of the genome context and therefore tend to make erroneous variant calls (see Figure 2.9 for illustration). Synteny-based comparisons, by contrast, are able to distinguish structural changes (translocation, etc.) from continuous alignments using synteny information, and know where SNPs should be called and where not. Indeed, further examination reveals that compared to the overlapping (validated) call set, these mapping-only SNP calls receive much lower read-depth support (typically just two reads, Figure 2.8C), very likely arising through read cross-mapping. I also investigated variants called by the synteny-based approach but missed by the mapping-based approach (i.e., "synteny-only" calls), and found that they are enriched in highly

polymorphic genomic regions (Figure 2.8D). As sequence divergence (SNP density) goes up, the mapping-based approach calls fewer and fewer SNPs while the proportion of synteny-only calls dramatically increases. In fact, 100% of SNP calls are made by the synteny-based approach alone when sequence similarity drops below 92%, consistent with the initial aligning parameters set for the short read aligner (a maximum of 8 mismatches per 100-bp read). Simply increasing the number of allowed mismatches does not solve the problem and only leads to more false positive SNP calls due to cross mapping, as evidenced by the enrichment of false positive SNP calls in SV regions. Moreover, the mapping-based approach generates considerable number of heterozygous calls (13-17% of all calls), most of which are not validated by assembly-comparison (Figure 2.8E) and are also enriched in highly polymorphic regions (Figure 2.8F) - a sign that allowing 8 mismatches is already creating serious cross-mapping issues. Such highly polymorphic regions and repetitive elements together lead to insurmountable difficulties for read mapping software and can only be effectively addressed by creating *de novo* assemblies for direct comparison. Reads are either discarded by aligners due to non-unique mapping or high number of mismatches and/or gaps, resulting in regions of low or zero (effective) coverage, or cross-mapped and mis-placed in structurally different regions if allowing too many mismatches (Figure 2.9). Synteny-based approach, on the other hand, accurately calls variants in such regions as long as syntenic alignments contiguously span the area. Future resequencing efforts with long read technology will be even more powerful in further resolving these repetitive and structurally different regions and should greatly improve our understanding of natural variation.

Synteny comparison offers accurate detection of both small and large variants

Beyond improved accuracy and precision in SNP discovery, the assembly comparison-based approach also enables better detection and characterization of insertion / deletion polymorphisms and complex structural changes generally. Resequencing studies typically tend to discover more deletions than insertions (relative to the reference) (Massouras et al. 2012). While there is no biological reason to expect more deletions than

insertions for a random sample, this ascertainment bias can largely be explained by the inherent limitation of short read aligners – a read with a deletion is easier to map than a read with an insertion. Directly comparing two assemblies, however, does not have this limitation and as a result, leads to a much more symmetrical size spectrum for small insertions and deletions (Appendix Figure 2.3). Insertion events longer than 10-bp are virtually invisible to read mapping-based approach, but can be effectively recovered by assembly comparison (Appendix Figure 2.3).

Nonetheless, ascertainment bias still exists for our assembly-based approach in the case of larger variants (≥ 50bp). The number of large deletions and copy number loss (CNL) events is on average 30-50% higher than that of large insertions and copy number gain (CNG) events (Table 2.4). This is not surprising given that the *de novo* assemblies are less complete (8-10% smaller) and more fragmented than the high-quality HM101 reference, and so many large insertion and CNG events are not captured due to insufficient synteny evidence (flanking sequence support) in a fragmented assembly. Improvement in sequencing and assembly technologies (especially longer reads) should lead to more complete and contiguous *de novo* assembles, enabling better detection and characterization of structural variants of all types.

While it is inherently difficult to detect large structural variants (from a few hundred bp to several thousand bp) with direct read mapping due to their relatively short read length (typically 100-bp), researchers have developed alternative approaches to detect structural variation by making use of special traces left in and around structurally changed regions during read mapping. Some of these algorithms compare empirical Depth-Of-Coverage (DOC) information with genome-wide average, identify regions with sudden drop or gain of read coverage and predict Copy Number Variation (CNV) events (Alkan et al. 2009; Yoon et al. 2009; Chiang et al. 2009). Other algorithms take advantage of the Paired-End Mapping (PEM) signatures left by inappropriately placed read pairs (soft clipped, orphan reads, abnormal insert size, etc.) around structurally altered regions, and then work to recover the actual break point of each SV (Korbel et al. 2009; Chen et al. 2009; Hormozdiari et al. 2009; Lee et al. 2009). While the PEM-based

approach is good at pinpointing SV breakpoints, the DOC-based approach is good at capturing large unbalanced sequence content changes such as CNVs and PAVs (Medvedev, Stanciu, and Brudno 2009; Alkan, Coe, and Eichler 2011). However, they both have their limitations. The DOC-based approach relies on a Poisson distribution of the read-depth signal, which is hard to satisfy in practice, being sensitive to local sequence characteristics such as GC content and repetitiveness (Bailey 2002). This makes DOC-based approaches less accurate in predicting CNVs of smaller size and blind to balanced sequence changes such as translocations and inversions. The PEM-based approach, on the other hand, is dependent on the insert size distribution, requiring a fixed cutoff of mapped insert size to be classified as "discordant", and also does not work in repetitive regions. Moreover, both DOC-based and PEM-based approaches absolutely rely on the correct mapping of reads in and around the SVs in the first place. They both fail where insufficient evidence of paired-end signatures can be found, such as highly polymorphic regions or repeat-rich regions.

Fortunately, these difficulties can all be bypassed by directly comparing a *de novo* assembly to the reference and calling variants using synteny alignment information. Both unbalanced variants (insertion, deletion, CNG, CNL) and balanced variants (translocation, inversion) can be detected and characterized with exact breakpoints identified at base-pair level. This analysis highlights the importance of comparative genomics using *de novo* assemblies and elevates our understanding of natural genomic variation to a completely new level.

Chromosomal-scale translocation revealed

Kamphuis *et al*. reported an aberrant rearrangement between *Medicago* linkage groups 4 and 8 (LG4 & LG8) in the reference accession A17 (Kamphuis et al. 2007). Based on synteny alignment against A17, we were able to confirm this large structural variation event comparing A17 with at least three accessions (HM004, HM034 and HM185, Figure 2.10). In addition, the exact breakpoints of the translocation were pinpointed to a single region on chromosome 4 and three regions on chromosome 8

(Figure 2.10 for illustration, Appendix Figures 2.5 and 2.6 for exact synteny relationship). Interestingly, each of the four breakpoints involves a gap (i.e., 'N's) in the reference (Mt4.0), with three 100 bp gaps and one 7.5 kbp gap - an indication that the regions in and around the rearrangement breakpoints are structurally unstable and difficult to assemble even using a BAC-by-BAC approach. We found numerous transposable element genes near the breakpoints, including a reverse transcriptase, a GAG-pre integrase and a cluster of 6 transferases near breakpoint 1 (Appendix Figure 2.7A), two helicases around breakpoint 2 (Appendix Figure 2.6B), two retrotransposons (UBN2) and two reverse transcriptases around breakpoint 3, and a MULE transposase right next to breakpoint 4 (Appendix Figure 2.6C). Interestingly, a cluster of at least 10 CC-NBS-LRRs was found both upstream and downstream breakpoint 2 (Appendix Figure 2.7B), and two CC-NBS-LRRs were also found right next to breakpoint 3, probably indicating an involvement of these variable gene families in shaping the overall genome structure. This translocation is mostly likely private to the reference A17 accession or clade, since we didn't find even one accession sharing the same haplotype structure with A17. In addition to the translocation, we noticed two large stretches of novel sequences (1.15 Mbp and 430 Kbp) downstream the translocation breakpoints on chromosome 4 and 8 that could not align anywhere in the reference space (Figure 2.10 red segments), which is potentially a deletion in A17 as the result of the chromosomal rearrangement. Sequencing additional close relatives of A17 may reveal the origin and history of this rearrangement.

Conclusion

In this study I built high quality *de novo* assemblies and systematically characterized natural variation in 15 *M. truncatula* accessions. This allowed me to detect novel genetic variation that would not have been found by mapping reads to a single reference. This analysis leads to a within-species diversity estimate much higher than previous mapping-based resequencing efforts, even using a smaller sample size. While this could be partially explained by a more complete reference genome and newer sequencing

chemistry in current study, my results clearly demonstrate that *de novo* assembly-based comparison is both more accurate and precise than mapping-based variant calling in exploring variation in repetitive and highly divergent regions. For the first time in plants, I was able to systematically identify and characterize different types of SVs using a synteny-based approach. My analysis revealed extensive structural variation in natural *Medicago* populations, exerting a larger impact on the overall genome architecture and population gene pool than smaller changes such as SNPs and short indels. Based on genome-wide synteny alignments against the reference (Mt4.0), I built a *Medicago* Pan-16 genome that considerably expands the reference space. Pan-genome analysis suggests a "restricted" pan-genome and core-genome size curve, both beginning to level off when ten or more accessions become sequenced. Improving existing assemblies and sequencing additional accessions may further reveal allelic variation and expand the gene pool within the population.

Table 2.1. Assembly statistics.

	Total Span	Total Bases	Scaffold Stats				Contig Stats				Repeat Elements	
			Number	N50	Median	Max	Number	N50	Median	Max	Bases	Percent
HM101	413,771,487	389,019,804									88,488,168	22.75
HM058	409,729,257	355,004,629	4,349	374,919	8,122	3,222,558	27,646	18,510	4,590	350,560	66,415,620	18.71
HM125	406,998,713	371,005,289	3,666	517,348	3,690	6,709,429	22,291	28,123	5,144	417,578	77,949,143	21.01
HM056	406,705,336	362,971,816	3,486	511,115	6,755	5,985,141	26,820	19,283	5,305	230,607	71,793,730	19.78
HM129	398,468,296	367,625,895	3,213	523,031	4,777	5,629,000	21,025	28,693	5,437	303,023	74,929,621	20.38
HM060	403,209,823	363,308,695	3,634	452,558	5,845	3,832,561	22,495	25,479	5,026	396,295	72,075,459	19.84
HM095	410,354,770	367,894,112	3,711	526,756	5,178	4,466,334	24,666	25,656	4,285	323,240	73,927,610	20.09
HM185	428,228,061	367,670,036	3,335	1,653,161	3,289	16,133,123	24,135	25,123	4,839	295,986	75,517,964	20.54
HM034	396,665,787	362,786,782	3,267	471,390	4,628	6,140,737	21,662	26,724	5,268	374,291	71,953,536	19.83
HM004	399,385,294	361,118,835	3,494	620,396	3,532	9,290,596	23,574	23,631	5,000	311,906	70,329,693	19.48
HM050	406,245,560	366,441,111	3,234	855,678	3,170	8,622,067	23,170	25,626	5,027	349,806	74,093,668	20.22
HM023	403,142,311	361,243,464	3,728	421,953	7,979	5,130,354	22,754	24,359	5,087	327,601	69,990,997	19.38
HM010	403,887,869	365,935,435	3,159	522,263	6,832	6,544,913	23,676	23,780	5,285	277,285	73,899,843	20.19
HM022	374,674,840	343,656,937	2,565	694,957	5,305	8,809,216	20,662	21,883	5,849	309,202	64,940,061	18.90
HM324	421,810,942	343,081,677	7,057	267,849	6,839	4,503,894	41,520	9,236	3,031	201,165	64,671,095	18.85
HM340	388,188,365	357,565,722	3,890	674,544	2,021	7,121,670	22,470	24,102	5,186	297,747	71,150,565	19.90

Table 2.2. Functional annotation statistics.

	# Total Genes	TE	non-TE	NBS-LRR	CRP[@]	Median Prot. Len.[*]	RNA-seq (%)[#]	Homology (%)[&]	RNA-seq + Homology (%)
HM101	67102	12312	54790	860	1428	228	39.6	-	39.6
HM058	64146	9540	54606	800	1280	222	-	85.2	85.2
HM056	65691	10379	55312	811	1304	220	36.4	84.9	86.9
HM125	67346	11175	56171	781	1280	218	-	84.4	84.4
HM129	65607	10455	55152	818	1278	222	-	83.6	83.6
HM034	64612	9958	54654	774	1266	222	36.0	83.1	85.2
HM095	65524	10197	55327	823	1283	222	-	82.8	82.8
HM060	64648	9967	54681	799	1272	223	-	83.7	83.7
HM185	65921	10666	55255	822	1274	222	-	83.4	83.4
HM004	64374	9680	54694	797	1278	224	-	82.8	82.8
HM050	65691	10282	55409	828	1289	224	-	82.6	82.6
HM023	64310	9661	54649	797	1281	222	-	83.2	83.2
HM010	65373	10339	55034	834	1304	223	-	83.2	83.2
HM022	59882	8193	51689	704	1295	227	-	78.6	78.6
HM340	64587	9387	55200	784	1287	228	36.1	75.1	77.6
HM324	60236	7751	52485	747	1213	221	-	76.6	76.6

[@]CRP: Cysteine rich protein family

[*]Median protein length (number of amino acids) was estimated using non-TE coding genes;

[#]RNA-Seq was done for four accessions using both un-inoculated root tissue and nodule; number indicates percentage of total predicted transcripts with FPKM > 0;

[&]Number indicates percentage of total predicted transcripts with at least one Mt4.0 ortholog (either syntenic ortholog or RBH-based homolog).

Table 2.3. Assembly comparison (with Mt4.0) statistics and novel sequences identified in 15 *M. truncatula* accessions

	Total Bases	Repetitive	Alignable to HM101	Bases in Synteny	Novel Sequences*		Novel Coding Seq	
HM058	355,004,629	66,415,620	343,325,053	323,600,625	9,024,826	2.50%	1,297,448	14.40%
HM125	371,005,289	77,949,143	357,808,394	327,976,297	9,863,489	2.70%	1,386,659	14.10%
HM056	362,971,816	71,793,730	350,554,159	326,377,315	9,377,865	2.60%	1,368,844	14.60%
HM129	367,625,895	74,929,621	351,879,700	320,623,943	11,755,160	3.20%	1,631,646	13.90%
HM060	363,308,695	72,075,459	347,211,386	317,763,091	12,044,190	3.30%	1,690,709	14.00%
HM095	367,894,112	73,927,610	351,219,964	317,035,668	12,460,933	3.40%	1,709,362	13.70%
HM185	367,670,036	75,517,964	351,105,536	317,771,859	12,390,758	3.40%	1,729,628	14.00%
HM034	362,786,782	71,953,536	346,005,399	317,021,576	12,433,286	3.40%	1,727,260	13.90%
HM004	361,118,835	70,329,693	344,257,187	315,304,988	12,721,129	3.50%	1,979,711	15.60%
HM050	366,441,111	74,093,668	349,114,197	317,140,191	13,022,686	3.60%	2,157,804	16.60%
HM023	361,243,464	69,990,997	344,455,479	315,834,571	12,376,960	3.40%	1,687,239	13.60%
HM010	365,935,435	73,899,843	348,722,480	316,833,225	12,639,855	3.50%	1,726,239	13.70%
HM022	343,656,937	64,940,061	315,921,590	275,649,889	19,732,777	5.70%	2,099,922	10.60%
HM324	343,081,677	64,671,095	312,883,637	266,427,488	21,757,962	6.30%	2,449,025	11.30%
HM340	357,565,722	71,150,565	326,264,086	279,190,368	20,778,471	5.80%	2,319,870	11.20%

*Novel sequences are segments not present in Mt4.0 (HM101) reference.

Table 2.4. Variants identified in 15 *M. truncatula* accessions by count (A) and affected base pairs (B).

(A)

	SNP[#]	SNP Density	Small Ins	Small Del	Large Ins	Large Del	CNG[*]	CNL[$]	Translocation
HM058	1,699,815	.0057	229,566	236,818	11,716	15,100	23,266	26,208	2,720
HM056	1,858,188	.0061	247,352	255,007	12,390	15,791	26,366	27,388	3,153
HM125	2,075,142	.0067	272,820	281,889	14,317	17,469	29,919	30,885	3,513
HM129	2,709,716	.0091	353,523	364,363	18,667	23,139	37,076	38,803	4,953
HM034	2,821,493	.0095	368,749	377,083	19,835	25,989	37,517	39,421	5,581
HM095	2,722,049	.0093	356,148	365,297	18,427	22,636	36,837	39,150	5,194
HM060	2,795,046	.0094	363,601	374,109	18,602	23,586	36,516	39,458	4,942
HM185	2,670,644	.0093	348,954	359,196	18,472	22,810	36,206	38,239	4,788
HM004	2,860,239	.0097	372,819	382,431	18,906	24,343	37,187	40,737	5,041
HM050	2,868,988	.0097	372,123	382,998	19,498	24,529	38,145	41,014	5,225
HM023	2,885,692	.0098	375,818	385,247	19,709	26,130	37,571	39,726	5,711
HM010	2,906,704	.0099	377,205	386,894	20,236	26,458	39,084	40,045	6,027
HM022	5,069,762	.0206	733,635	736,313	45,302	66,834	80,780	88,265	12,456
HM340	5,072,373	.0206	723,866	735,113	45,992	64,314	82,122	86,748	12,788
HM324	4,984,659	.0216	731,037	734,307	41,946	67,122	76,984	87,641	12,666

[#]Numbers listed here are all synteny-based variant calls;

[*]CNG: Copy number gain;

[$]CNL: Copy number loss.

(B)

	SNP	SNP Density	Small Ins	Small Del	Large Ins	Large Del	CNG	CNL	Translocation
HM058	1,699,815	.0057	1,472,729	1,433,967	2,907,955	4,793,101	6,444,318	20,368,937	3,562,521
HM056	1,858,188	.0061	1,604,545	1,557,103	2,981,372	4,857,991	8,113,896	21,380,128	4,226,924
HM125	2,075,142	.0067	1,748,580	1,717,576	3,496,909	5,635,199	10,027,202	24,014,458	4,944,633
HM129	2,709,716	.0091	2,208,085	2,241,253	4,614,069	7,342,439	13,060,326	28,844,685	6,815,118
HM034	2,821,493	.0095	2,363,055	2,356,532	4,882,635	8,945,373	12,472,627	27,903,777	7,324,491
HM095	2,722,049	.0093	2,234,135	2,242,821	4,473,529	7,056,774	12,803,351	28,792,835	7,798,090
HM060	2,795,046	.0094	2,285,418	2,312,236	4,494,901	7,367,260	11,875,765	28,864,192	6,531,081
HM185	2,670,644	.0093	2,210,653	2,216,582	4,688,974	7,036,603	11,944,244	28,452,857	6,883,387
HM004	2,860,239	.0097	2,350,976	2,375,956	4,564,540	7,826,327	11,743,812	29,647,325	6,444,519
HM050	2,868,988	.0097	2,350,028	2,373,850	4,705,884	7,586,259	12,515,726	29,768,759	6,715,431
HM023	2,885,692	.0098	2,425,521	2,405,538	4,771,137	8,801,487	11,836,264	28,012,528	7,383,912
HM010	2,906,704	.0099	2,429,096	2,423,439	4,925,567	8,920,031	12,921,565	28,068,159	7,407,099
HM022	5,069,762	.0206	5,132,433	5,133,963	10,211,293	21,744,266	25,079,221	50,472,526	13,098,504
HM340	5,072,373	.0206	5,079,289	5,107,633	10,466,002	20,071,674	27,438,010	49,833,972	14,307,541
HM324	4,984,659	.0216	5,077,977	5,325,368	9,070,313	20,128,282	22,585,126	46,987,452	13,245,864

Table 2.5. Coverage and diversity statistics by nucleotide class.

	Covered bases (bp)[#]	Total bases (%)	Polymorphic sites	π bp^{-1}	θ_w bp^{-1}
Total	279,689,505	-	7,043,505	0.0073	0.0082
Coding	49,106,309	0.18	897,243	0.0052	0.0060
Synonymous	7,219,248	0.03	190,416	0.0076	0.0086
Replacement	31,776,906	0.11	489,490	0.0044	0.0050
Introns	63,144,752	0.23	1,148,781	0.0053	0.0059
UTR 5'	3,505,093	0.01	43,689	0.0036	0.0040
UTR 3'	6,241,117	0.02	86,575	0.0040	0.0045
Intergenic	157,692,234	0.56	4,867,217	0.0089	0.0100

[#]Covered bases represent reference genomic regions covered by syntenic alignments in at least 10 (out of 13) in-group accessions.

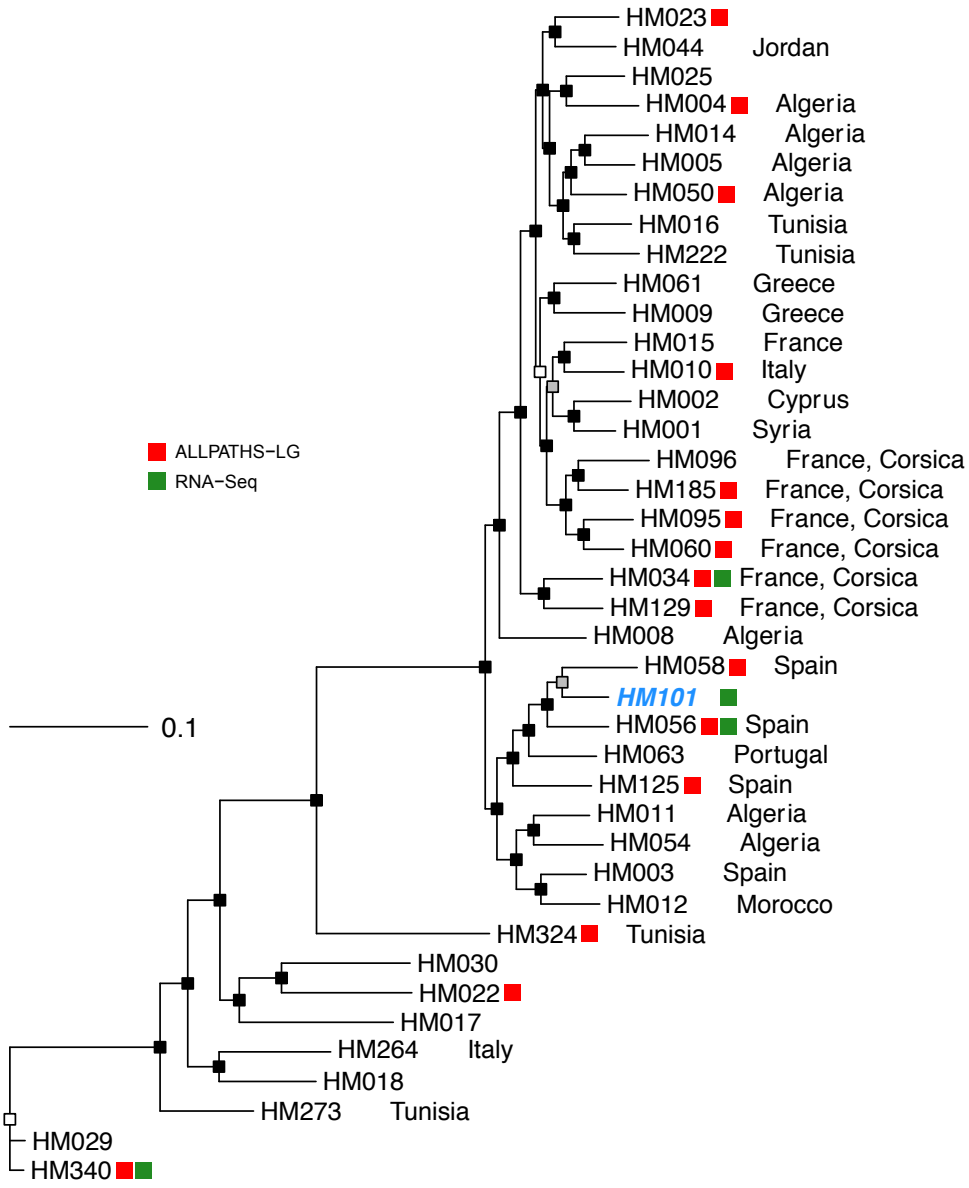


Figure 2.1. Phylogeny of selected *Medicago* accessions with their countries of origin.

Maximum likelihood tree built using 15,000 SNPs randomly sampled from all chromosome 5 SNPs called by the *Medicago* Hapmap project. Nodes with ML bootstrap support of more than 80% are indicated with filled rectangles. Red Rectangles: 15 accessions with *de novo* assemblies described in this thesis. Green Rectangles: 4 accessions with RNA-Seq data.

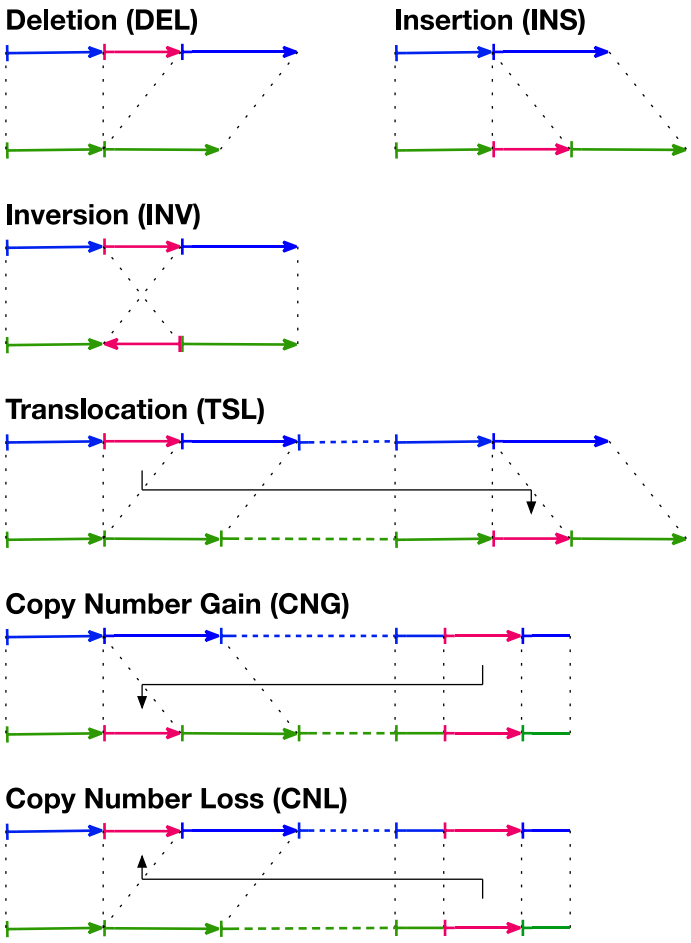


Figure 2.2. Illustration of synteny-based structural variant detection.

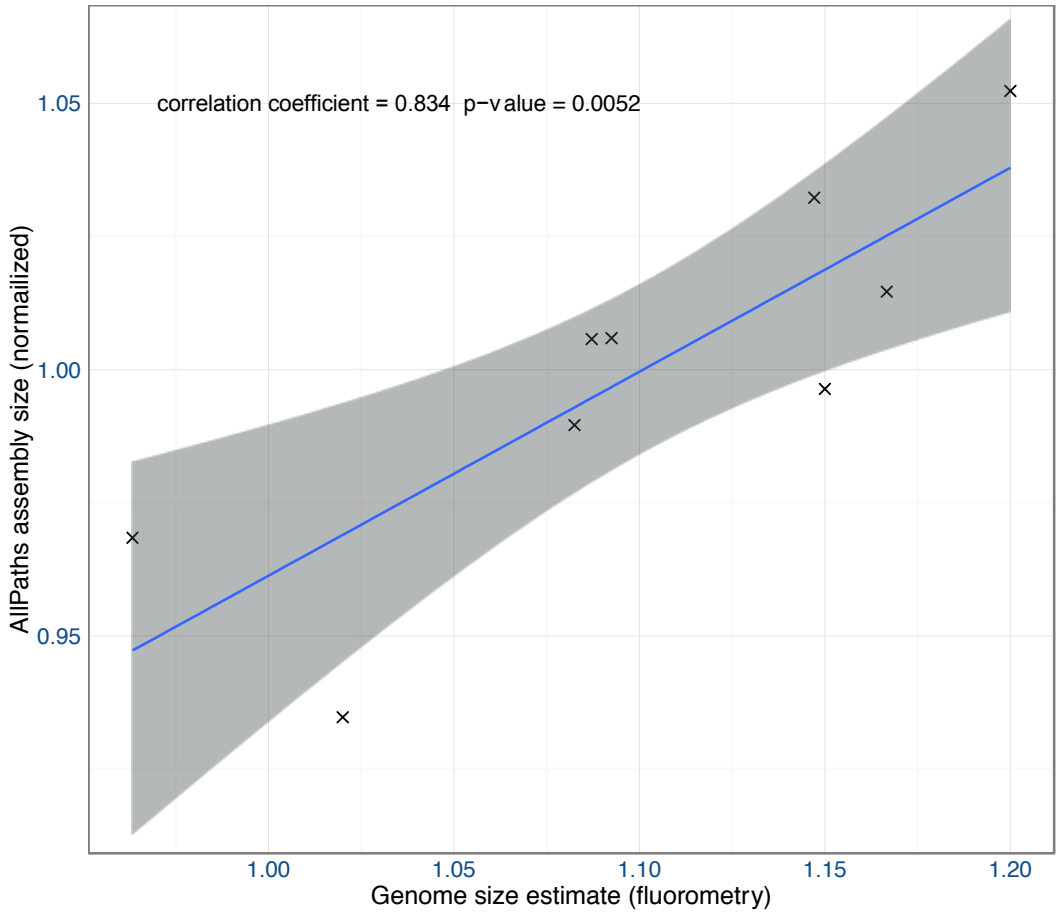


Figure 2.3. Correlation of assembled genome sizes (ALLPATHS) and fluorometry-based genome size estimates in nine *M. truncatula* accessions.

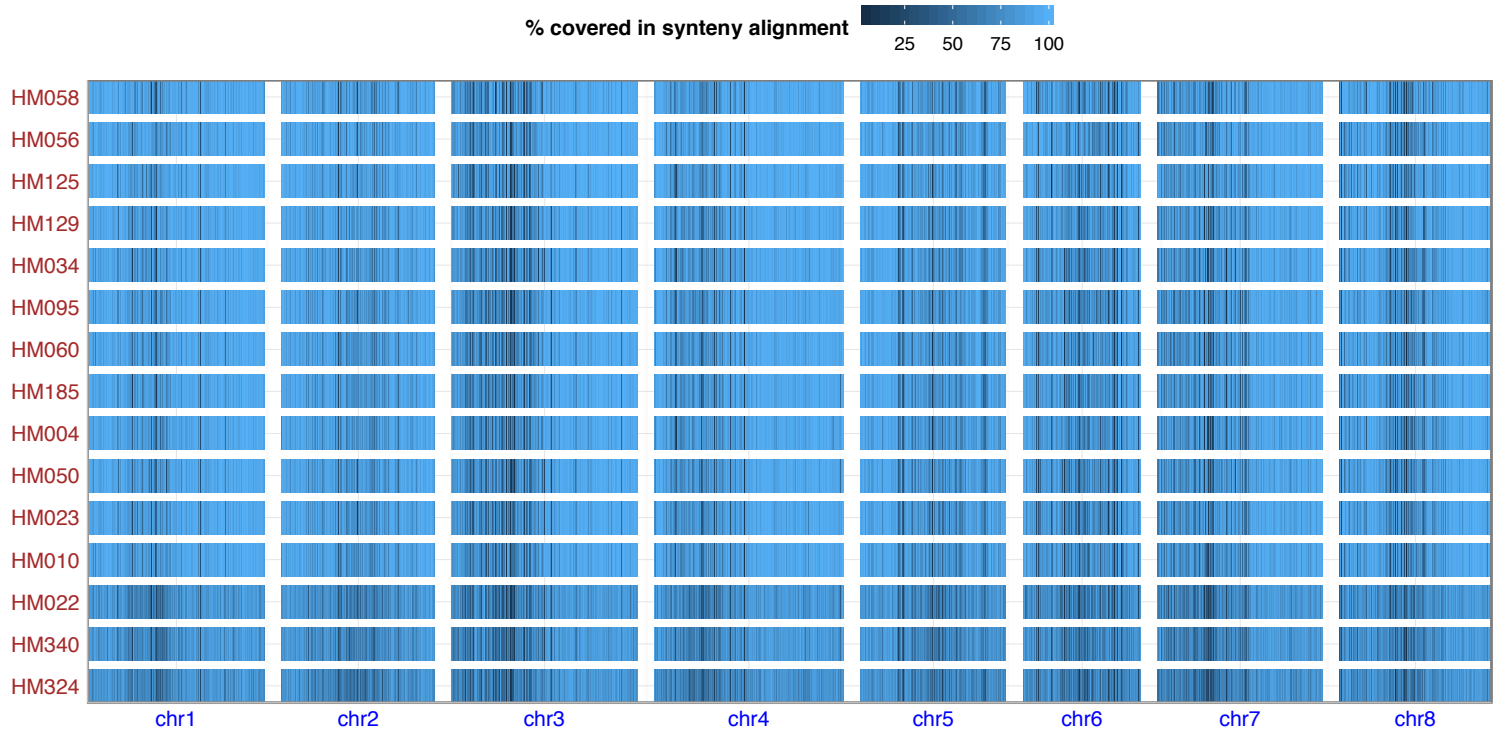


Figure 2.4. Heatmap showing percent covered by synteny alignment for each 100kb window in 15 *de novo M. truncatula* assemblies.

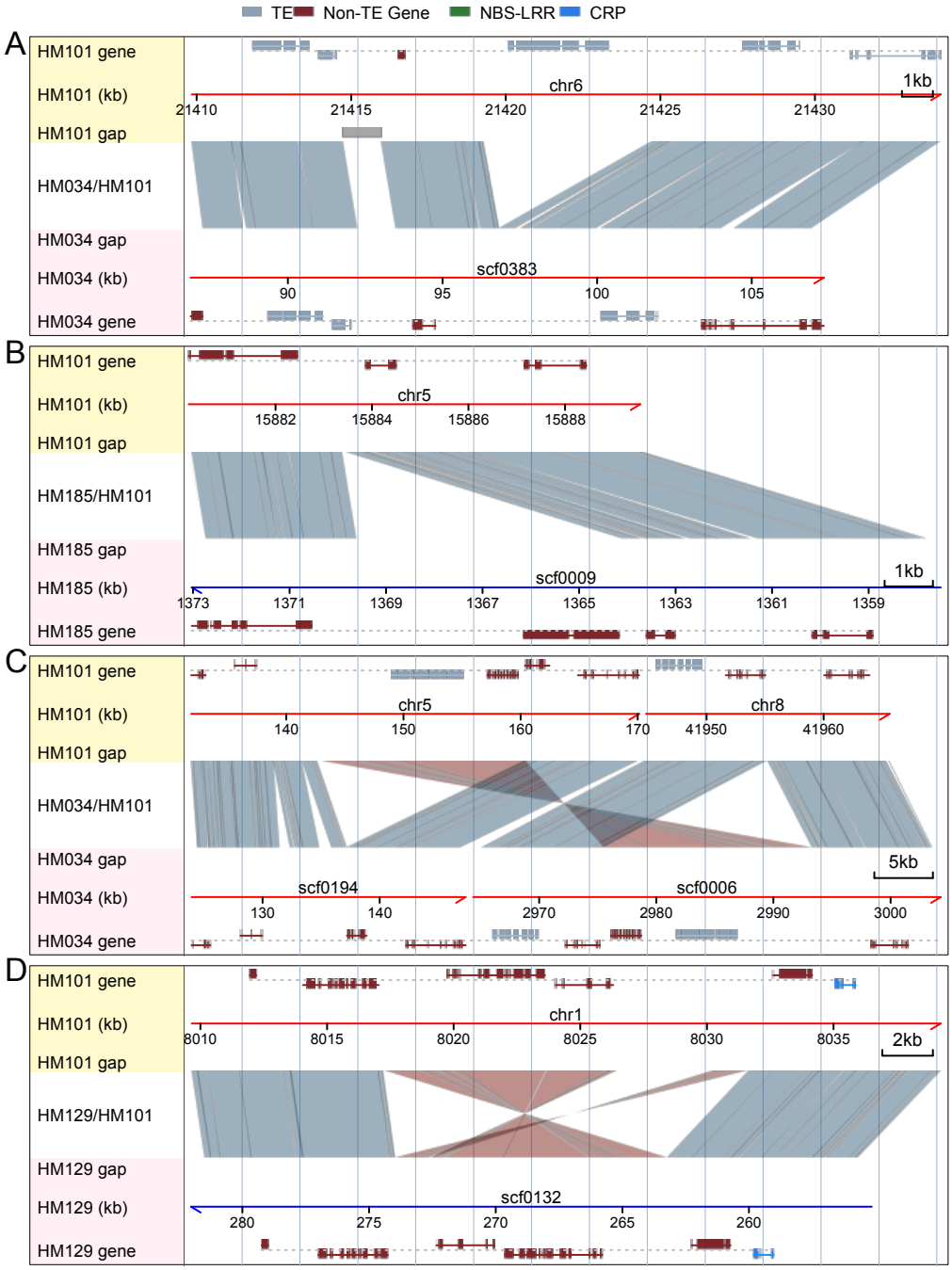


Figure 2.5. Illustration of different types of structural variants.

(A) 6-kbp deletion removing a transposable element (TE) in HM034 (relatively to HM101); (B) 6-kbp insertion adding a novel gene (cullin) in HM185; (C) 17-kbp translocation between chromosomes 5 and 8 involving a TE plus a nearby gene in HM034; (D) 20-kbp inversion followed by partial deletion involving several coding genes in HM129.

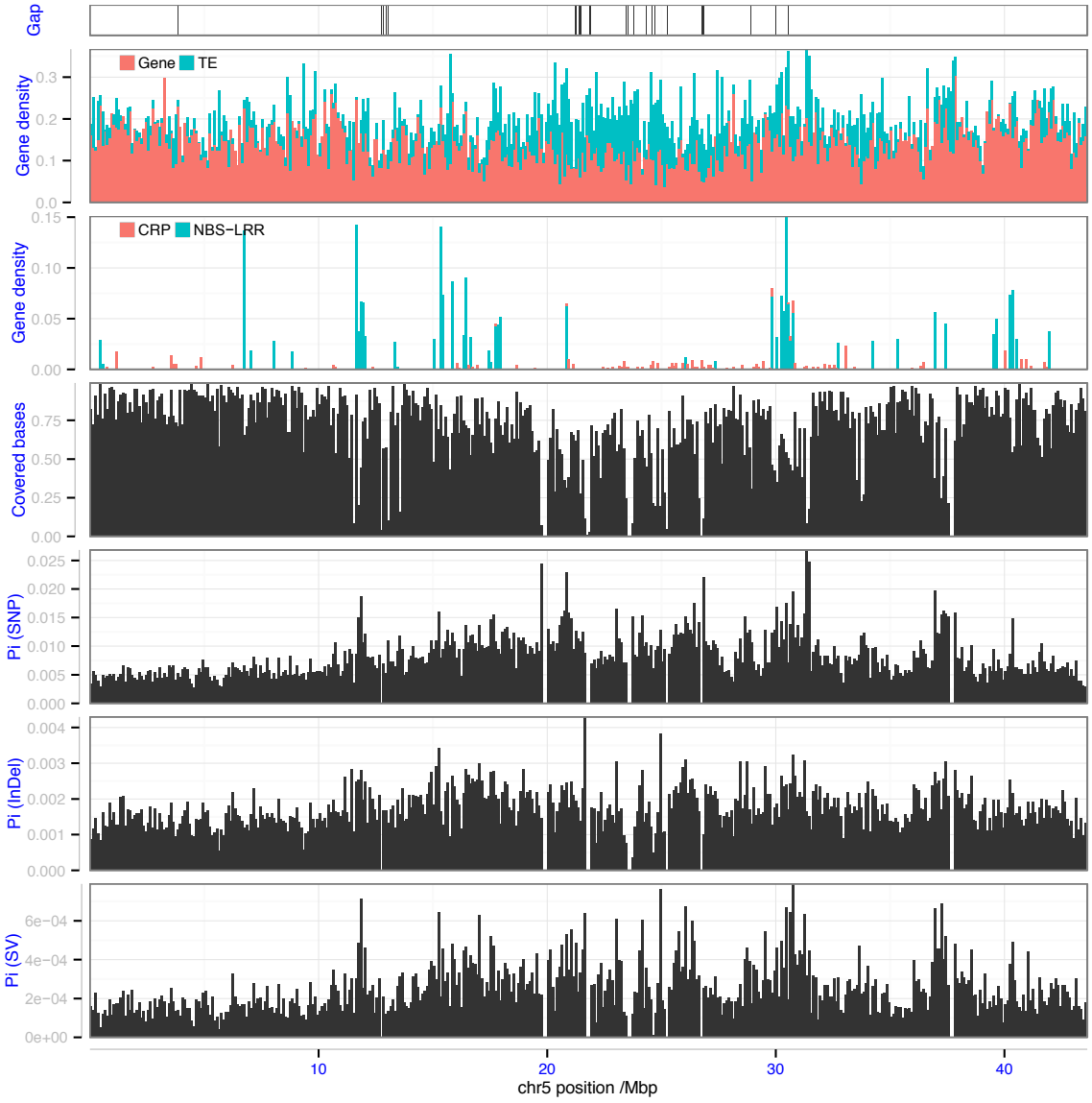


Figure 2.6. Sliding window analyses on chromosome 5 showing reference gap position, gene density of different categories (non-TE, TE, NBS-LRR, CRP), covered bases (bases covered by synteny blocks in at least 10 out 13 accessions), and nucleotide diversity (θ_π) for SNPs, short InDels (< 50bp) and large SVs (>= 50bp).

Nucleotide diversity (θ_π) estimates were calculated using only 13 "ingroup" *M. truncatula* accessions that are close to each other and form a tight clade in phylogeny.

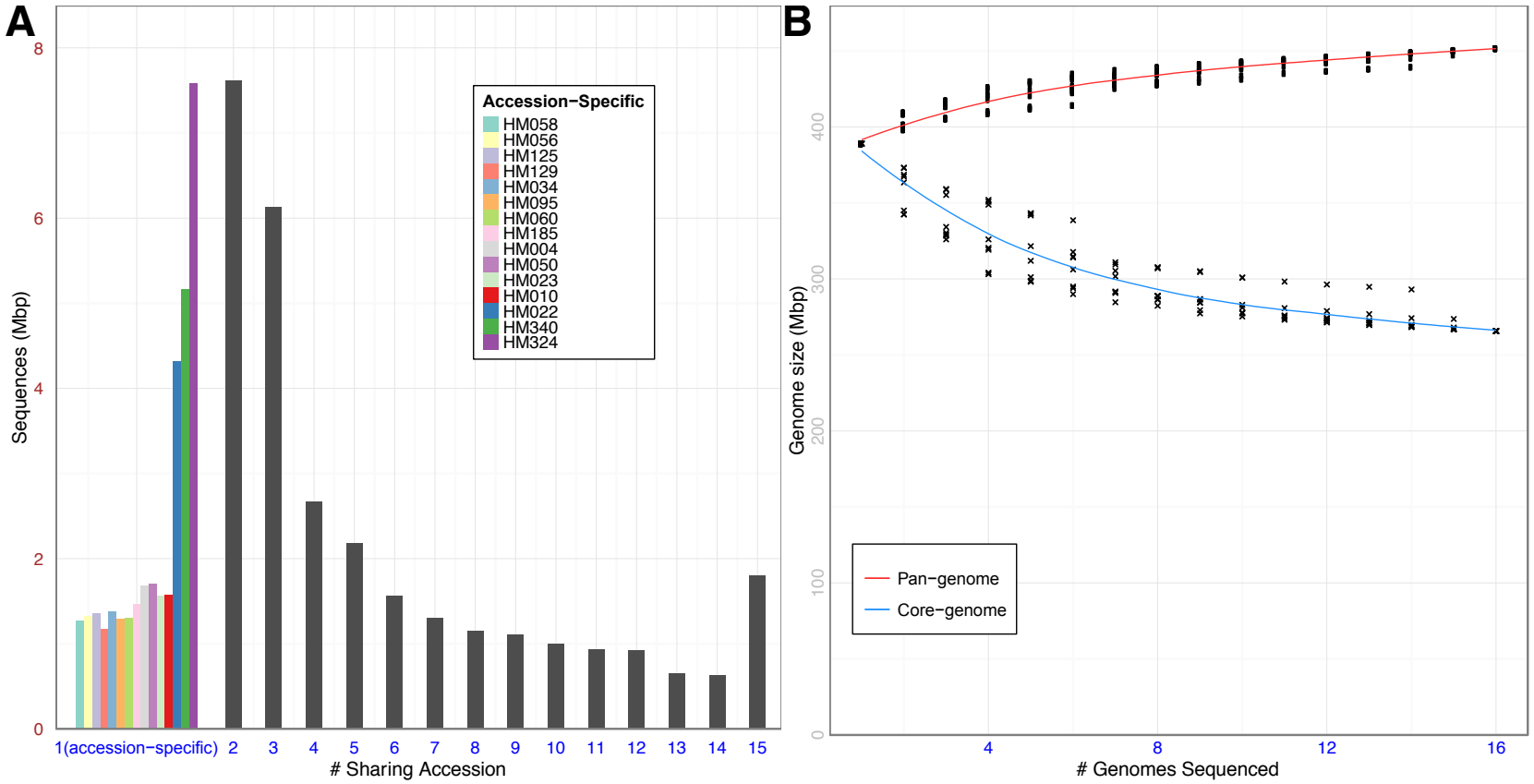


Figure 2.7. Novel sequences (absent in HM101) identified in 15 *M. truncatula* accessions (A) and the Pan-16 genome size curve (B).

The pan-genome size curve and core-genome size curve were derived by adding one *de novo* assembly to the pool at a time and calculating the size of shared genomic regions (core-genome, 'x' in the figure) and the size of total non-redundant sequences (pan-genome, 'o' in the figure). This process is repeated 8 times by shuffling the order of accessions.

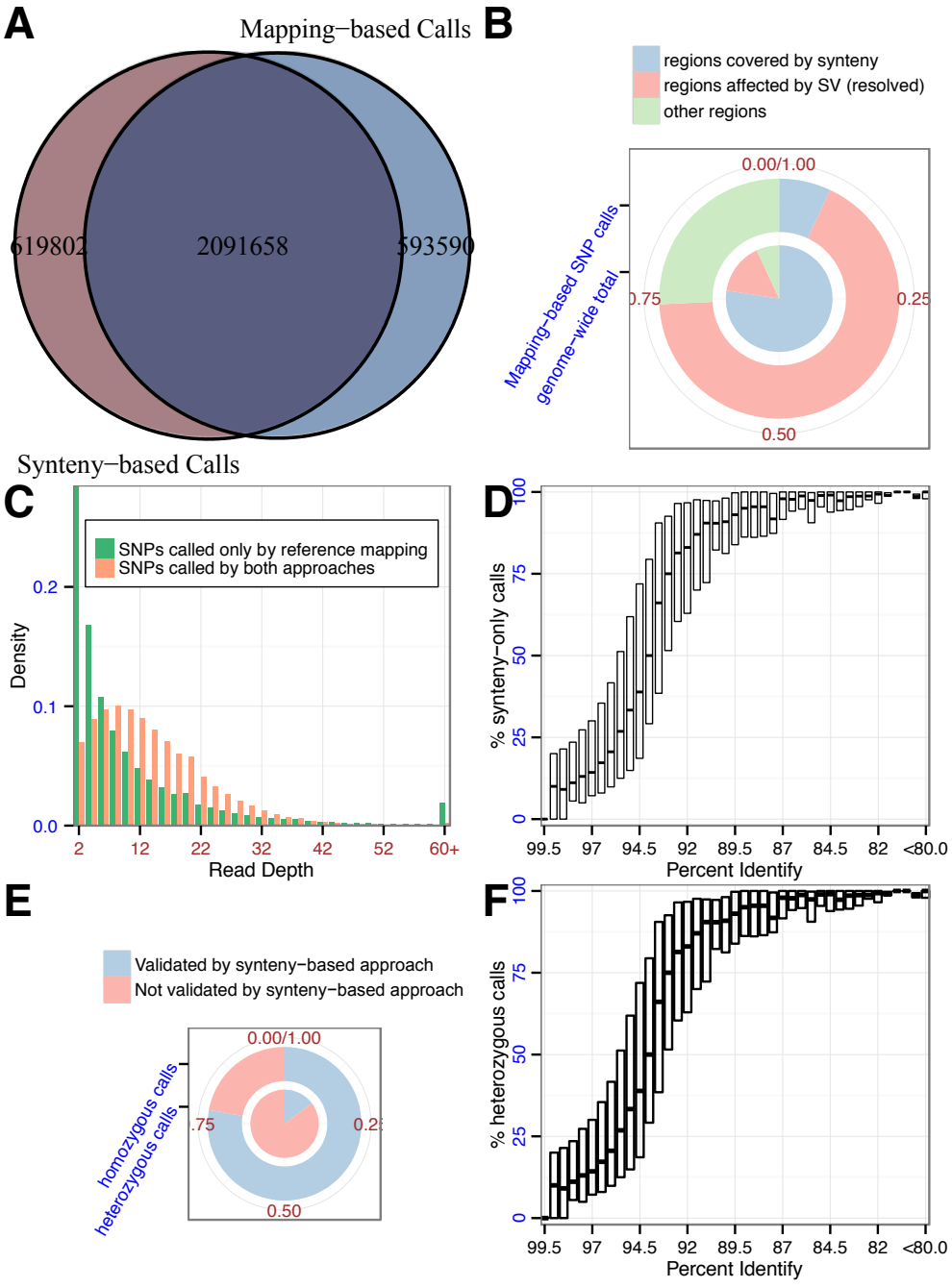


Figure 2.8. Comparison and characterization of SNP calling in HM010 from two different approaches.

(A) Venn-diagram showing overlap of mapping-based SNP call set and synteny-based call set; (B) Distribution of mapping-only SNPs in different genomic classes (outer ring) and genome-wide distribution of different genomic classes (inner piechart); (C) Distribution of read-depth support for "reference mapping-only" SNP calls and SNPs

called by both approaches (reference mapping + synteny comparison); (D) Proportion of synteny-only SNP calls binned by different sequence identity classes; (E) Proportion of heterozygous and homozygous SNP calls (mapping-based) validated by assembly-based approach; (F) Proportion of heterozygous SNP calls (out of all mapping-based calls) binned by different sequence identity classes. See Appendix Figures 2.4A and 2.4B for illustrations in two other accessions (HM004 and HM023).

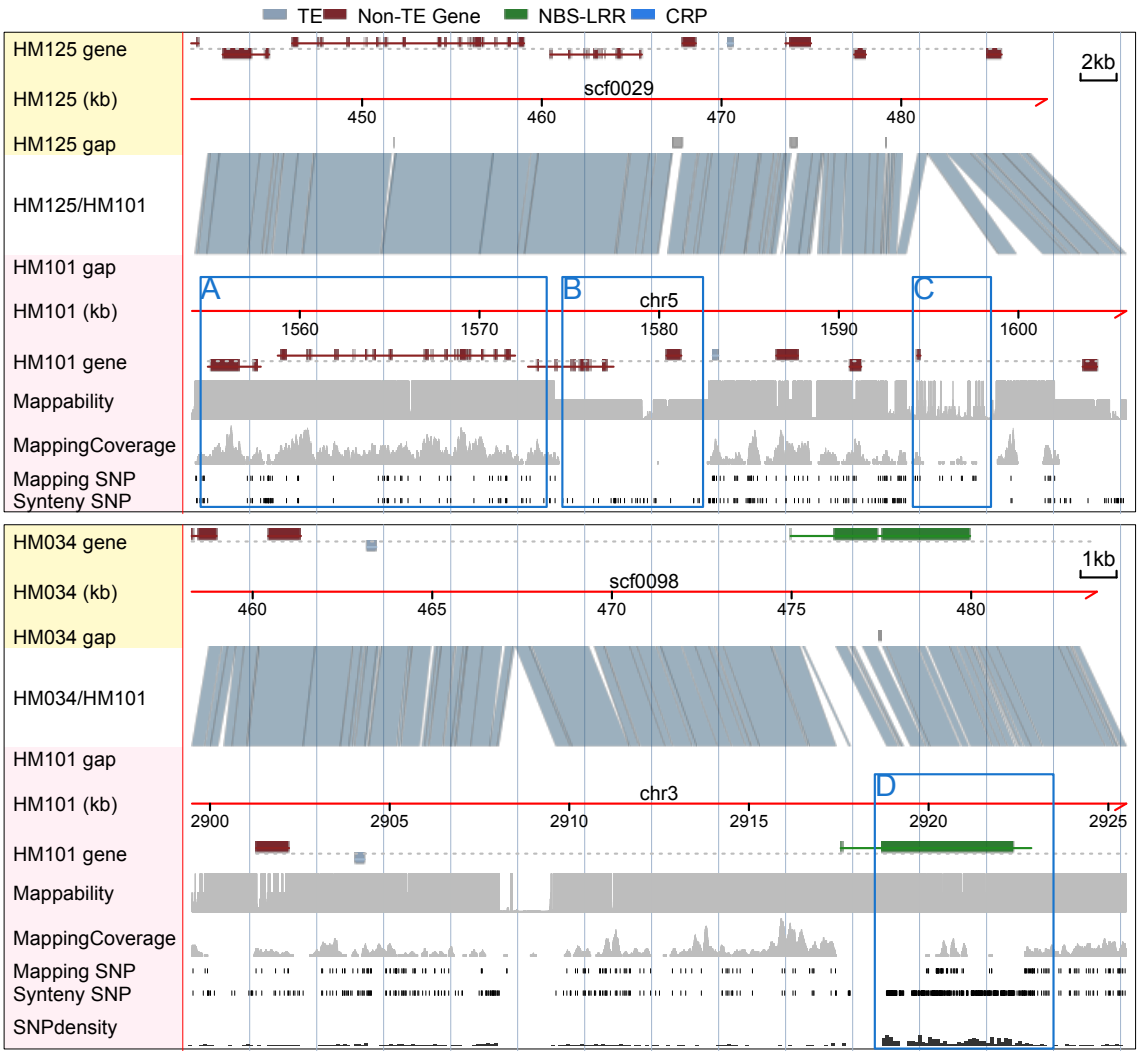


Figure 2.9. Illustration of SNPs called differently by two approaches.

(A) SNPs are called by both approaches (reference mapping-based and synteny-based) in most conserved genomic regions; (B) in repetitive regions, SNPs are only called by synteny-based approach and missed by mapping-based approach due to insufficient uniquely-mapped reads; (C) in structurally affected (deleted) regions SNPs are INCORRECTLY called by reference mapping-based approach due to cross-mapping of paralogous reads; (D) in highly divergent (i.e., high SNP density) regions SNPs are only called by synteny-based approach and missed by mapping-based approach due to too many mismatches in read alignment.

"Mappability" track (also known as "uniqueness") provides a measure of how often the sequence (60mer) found at the particular location will align within the whole genome with up to 2 mismatches (e.g., mappability of 1 means unique mapping and mappability of 0.5 means there are two copies of the sequence in the genome). "Coverage" track

shows coverage of HM125 reads mapped to HM101 reference. “Mapping SNP” track shows locations of SNPs called by read mapping and GATK-UnifiedGenotyper. “Synteny SNP” track shows locations of SNPs based on synteny comparison. “SNP density” track shows histogram of SNP density (number of synteny-based SNP calls per 1,000 bp window).

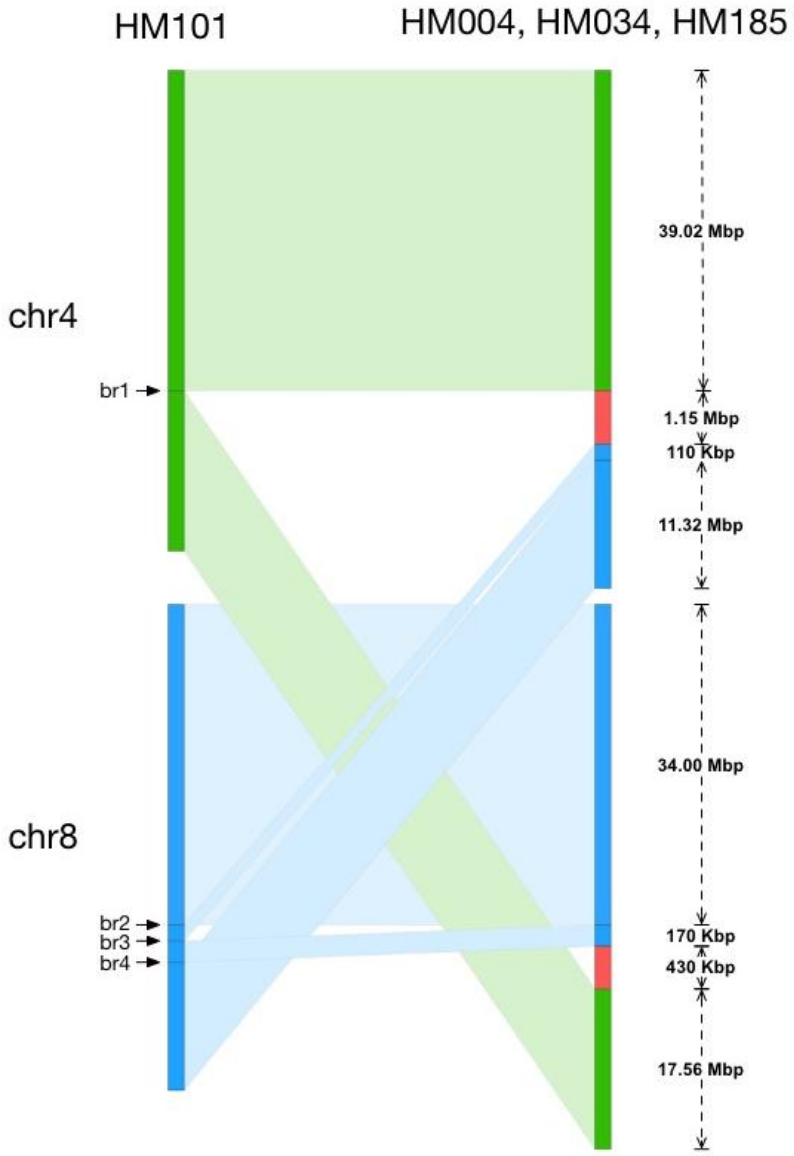


Figure 2.10. Schematic illustration of the rearrangement between chromosomes 4 and 8 in A17.

Green segments indicate chromosome 4 ancestry while blue segments indicate chromosome 8 ancestry (assuming A17 is the ancestor). Red segments indicate novel sequences (i.e., not present in the A17 reference). Breakpoint 1 (br1) is pinpointed to a 104 bp region (chr4:39,021,788-39,021,891) and includes a 100 bp gap. Breakpoint 2 (br2) is pinpointed to a 7,665 bp region (chr8:33,996,308-34,003,972) and includes a 7,663 bp gap. Breakpoint 3 (br3) is pinpointed to a 708 bp region (chr8: 34,107,285-34,107,992) and includes a 100 bp gap. Breakpoint 4 is pinpointed to a 277 bp region (chr8:34,275,249-34,275,525) and includes a 100 bp gap)

Chapter 3. Comparing multiple *Medicago* assemblies enable analysis of large gene families on a genome scale

Medicago truncatula is a model for investigating legume genetics and the evolution of legume-rhizobia symbiosis. Over the past two decades, two large gene families in *M. truncatula*, the nucleotide-binding site leucine-rich repeat (NBS-LRR) family and the nodule-specific, cysteine-rich (NCR) gene family, have received considerable attention due to their involvement in disease resistance and nodulation, large family size, and high nucleotide and copy number diversity. Previous studies using whole-genome sequence data to identify sequence polymorphisms (SNPs and short Insertion / Deletions; indels) relied on mapping short reads to a single reference genome. However, limitations of read-mapping approaches have hindered variant detection and characterization in both highly divergent and repeat-rich regions. As a result, studies of these large gene families are also hindered due to high sequence similarity among family members and high divergence among accessions. In the present study, I constructed high-quality *de novo* assemblies for 15 *M. truncatula* accessions. This allowed me to detect novel genetic variation that would not have been found by mapping reads to a single reference. Evidence-based annotation of the 15 *de novo* assemblies revealed that more than half of reference gene models were structurally different (lower than 60% sequence similarity) in at least one other accession. Not surprisingly, the NBS-LRR gene family harbors by far the highest level of nucleotide diversity, large effect single nucleotide changes, protein diversity and presence / absence variation (levels comparable with transposable elements), consistent with the rapidly evolving dynamics of disease resistance phenotypes. On the other hand, the one- or two-exon NCR family is less involved in gene structural changes but more frequently affected by copy number variation including both gains and losses in family members. Characterization of deletion and tandem duplication events in the NBS-LRR and NCR gene families suggests accession-specific subfamily expansion / contraction patterns, most of which were supported by PacBio long reads. This work illustrates the value of multiple *de novo* assemblies and the strength of comparative genomics in exploring and

characterizing novel genetic variation within a population. This work provides insights in understanding the impact of structural variants (SVs) on genome architecture and large gene families underlying important traits.

Introduction

Legumes comprise a diverse and ecologically significant plant family that serves as the second most important crop family in the world (P. H. Graham and Vance 2003). As a cool season legume, *Medicago truncatula* is closely related to many important crops such as alfalfa (*Medicago sativa*), red and white clover (*Trifolium pratense* and *T. repens*), pea (*Pisum sativum*), chickpea (*Cicer arietinum*), *Lotus japonicus* as well as soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*) (Lavin, Herendeen, and Wojciechowski 2005; Nevin Dale Young and Udvardi 2009). *M. truncatula* has been chosen as a model for studying legume biology due to its small genome size, simple diploid genetics, self-fertility, short generation time, amenability to genetic transformation and large collections of diverse ecotypes (Ronfort et al. 2006; Tadege et al. 2008; Nevin Dale Young and Udvardi 2009). Research interests on *M. truncatula* have focused on its symbiotic relationship with rhizobia and arbuscular mycorrhizae, root development, secondary metabolism and disease resistance (Oldroyd and Downie 2008; Nevin Dale Young and Udvardi 2009). Over the past two decades, two large gene families in *M. truncatula*, the nucleotide-binding site leucine-rich repeat (NBS-LRR) family and the nodule-specific, cysteine-rich (NCR) gene family, have received considerable attention due to their involvement in disease resistance and nodulation.

The majority of disease resistance genes in plants encode nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins. Plant NBS-LRR proteins (also called NB-LRR or NB-ARC-LRR proteins) are involved in the detection of diverse pathogens including bacteria, viruses, fungi, nematodes, insects and oomycetes (Ellis, Dodds, and Pryor 2000; Belkhadir, Subramaniam, and Dangl 2004; Jones and Takemoto 2004). This large, abundant gene family is encoded by hundreds of diverse genes per genome and can be subdivided into two subgroups based on sequence identity that precede the NBS domain: the TIR-NBS-LRR (TNL) proteins that contain a Toll-like domain, and CC-NBS-LRR (CNL) proteins characterized by a coiled-coil domain (Meyers et al. 1999; Q. Pan, Wendel, and Fluhr 2000; Meyers et al. 2003). There are approximately 150 NBS-LRR-encoding genes in *Arabidopsis thaliana*, over 300 in soybean and over 600 in rice

(Meyers et al. 2003; Kang et al. 2012; Shang et al. 2009). Huge differences exist among species in terms of the numbers and organization of different NBS-LRR subfamilies, with family-specific amplification occurring in legumes and Solanaceae (which includes tomato and potato) (Cannon et al. 2002; Plocik, Layden, and Kesseli 2004). Within the genome, NBS-LRR genes are organized either as isolated genes or, more frequently, as linked clusters of varying sizes that are thought to arise through both tandem and segmental duplications and could facilitate rapid R-gene evolution (Meyers et al. 2003; Monosi et al. 2004; Richly, Kurth, and Leister 2002; Leister 2004; Ameline-Torregrosa et al. 2008; Hulbert et al. 2001). While local tandem duplications are the main contributor to tightly linked tandem NBS-LRR clusters, a variety of events are thought to give rise to mixed NBS-LRR clusters that contain members from different subfamilies: ectopic duplication, transposition, as well as large-scale segmental duplication followed by subsequent local rearrangement (Meyers et al. 2003; Baumgarten et al. 2003; Kuang et al. 2004; McDowell and Simon 2006). The rate of evolution of NBS-LRR genes can be rapid or slow, with gene conversion events being frequent in some clades but rare in others (Kuang et al. 2004). This heterogeneous rate of evolution is consistent with a birth-and- death model of R gene evolution, in which gene duplication and unequal crossing-over can be followed by density-dependent purifying selection (McHale et al. 2006). As such, the uneven and clustered distributions of NBS-LRR genes and different selection pressures they experience have contributed to the generation of novel resistance specificities and to the expansion of this gene family through the mechanisms listed earlier (Marone et al. 2013).

Cysteine-rich peptides (CRP) are extremely abundant in plants, and are divided into many classes (M. a Graham et al. 2004; Silverstein et al. 2005; Silverstein et al. 2007). While different CRP groups differ in the configuration of conserved cysteine residues in the mature peptide, members within each group typically have striking similarities in their sequences, expression pattern and function (Broekaert et al. 1997; García-Olmedo et al. 1998). While classical CRP groups have active defense functions such as antimicrobial activity through the disruption of the pathogen's membrane (Shai 2002;

Thevissen et al. 2003), some CRP groups have passive defense functions that deter predation through allergenicity (Himly et al. 2003) and trypsin inhibition (Melo et al. 2002). In addition to defense roles, various CRP groups seem to be employed in flowers and seeds to play reproductive regulatory roles, such as the stigma-specific STIG1 family (Goldman, Goldberg, and Mariani 1994), the defensin-like S-locus cysteine-rich (SCR) proteins (Schopfer, Nasrallah, and Nasrallah 1999) and the pollen tube attraction polypeptides (LUREs) (Okuda et al. 2009). Other CRP groups have evolved functions to regulate plant growth and development, such as the rapid alkanization factor (RALF) proteins (G. Pearce et al. 2001) and lipid transfer protein (LTP)-like xylogens (Motose, Sugiyama, and Fukuda 2004).

During the last two decades, a large diverse family of CRPs was identified in the nodules of *M. truncatula* showing exclusive nodule-specific expression and a wide range of spatio-temporal patterns in the infected cells throughout nodule organogenesis (Fedorova et al. 2002; Mergaert et al. 2003; M. a Graham et al. 2004; Mergaert et al. 2006). To date, these nodule-specific cysteine-rich proteins (NCRs) have been found only in legumes belonging to the inverted repeat-lacking clade (IRLC) within the subfamily Papilionoideae, and absent from other non-IRLC legumes within the same subfamily, such as *L. japonicas* and *G. max* (M. a Graham et al. 2004; Alunni et al. 2007). Interestingly, in *M. truncatula* and other IRLC legumes, indeterminate nodules are formed where rhizobia undergo terminal differentiation into enlarged bacteroids and lose their ability to grow and divide, whereas *L. japonicas* and *G. max* form determinate nodules in which rhizobia retain the cell size, genome content and viability as free-living bacteria (Mergaert et al. 2006). Recent findings showing that NCR peptides act as symbiotic plant effectors to direct bacteroid differentiation have led to the speculation that NCRs are actually cysteine-rich antimicrobial peptides recruited by IRLC legumes in the context of symbiosis to dominate the endo-symbionts (Van de Velde et al. 2010; Farkas et al. 2014).

Previous work based on an earlier version of the *Medicago* genome (Mt1.0) identified 333 NBS-LRRs, including 177 CNLs and 156 TNLs (Ameline-Torregrosa et

al. 2008). Likewise, studies using EST sequences from public databases have led to the discovery of more than 300 NCR family members in *M. truncatula* (Fedorova et al. 2002; Mergaert et al. 2003; M. a Graham et al. 2004). The observation that clusters of these large gene families are frequently found in complex (e.g., assembly gaps that are difficult to fill) and repeat-rich regions indicated that they play structurally important role in the overall genome architecture (Ameline-Torregrosa et al. 2008). Recently, a high quality, BAC-based sequence of the *Medicago* A17 genome has become available as the "reference genome" for the *Medicago* research community (Tang et al. 2014). This reference genome assembly (version 4.0) covers ~80% of the overall genome (estimated at 465 Mbp) while capturing ~93% of all predicted gene models (Bennett and Leitch 2011; Tang et al. 2014). Based on the reference, I was able to compile a much more complete set of 860 NBS-LRRs and 717 NCRs using a profile-based approach (Chapter 1), and characterize their genomic distribution and phylogenetic pattern in this work.

The availability of a reference genome also enables population studies using next generation approaches, allowing for genome-scale analyses of nucleotide diversity and inferences on the underlying evolutionary forces shaping gene family diversity (Begun et al. 2007; Clark et al. 2007; McNally et al. 2009; Gore et al. 2009). In the case of NBS-LRR and NCR family, high levels of nucleotide and expression variation among different *Medicago* accessions have already been documented by EST, microarray and resequencing efforts (Tesfaye et al. 2013; Branca et al. 2011; Nallu et al. 2014; Stanton-Geddes et al. 2013). However, due to their large family size, high level of nucleotide diversity and frequent structural and copy number variation, characterizing these gene families at a genome and population scale has generally turned out to be difficult or even unsuccessful.

Traditional resequencing studies have relied on mapping short reads to a reference sequence in order to call polymorphic sites. This introduces a potential bias due to significant structural differences between diverse *Medicago* accessions. Alignment to a single reference is most problematic when reads from a divergent *Medicago* accession, such as the functionally important R108, are incorrectly aligned to the A17 reference due

to mis-alignment of paralogous regions. With a high within-species nucleotide diversity (genome-wide estimates of $\theta_w = 0.0063$ approximately three times more than found in soybean $\theta_{w\text{-cultivated}} = 0.0017$ bp^{-1} and $\theta_{w\text{-wild}} = 0.0023$ bp^{-1} populations), it may not be surprising that only a portion of the (reference) genomic regions can be confidently probed with read-mapping approaches. Diversity estimates may thus be underestimated due to the elimination of divergent (un-aligned) sequences. In addition, the length limitation of short read technologies such as Illumina leads to ambiguous mappings in repetitive and duplicated regions. Finally, the incompleteness of the reference genome limits our ability to detect variation in regions not present in the reference (e.g., assembly gaps) and leads to false alignments and SNP calls if reads from these regions are aligned to their next best match. Taken together, duplications that are present in some accessions but absent from others, high sequence similarity between recently duplicated family members, as well as high divergence between members among accessions (e.g., rapid structural changes in NBS-LRRs that often lead to very low sequence similarity between ortholog pairs) have seriously limited our understanding of the diversity and evolution of important large gene families such as NBS-LRRs and NCRs.

The distribution of structural variation (SVs) in the *Medicago* and their impacts on important gene families remain largely unknown. Yet this knowledge is essential in gaining insight into the genetic basis of legume-rhizobium symbiosis and identifying genes underlying phenotypic variation. The only practical way to understand the genomic diversity of *Medicago* fully and to capture the numerous members of such divergent gene families is to carry out whole genome sequencing and *de novo* assembly. Recent advances in NGS chemistry and computational approaches in sequence assembly have significantly improved the power and reliability of *de novo* assembly of NGS data. In this study, I worked with colleagues to develop *de novo* genome assemblies of 15 strategically chosen *Medicago* accessions. We constructed a *Medicago* Pan-16 genome and proteome, and fully characterized the genomic content and gene family repositories for these accessions. Based on these genome assemblies, I was able to systematically characterize all types of variation affecting different gene families: single-nucleotide

polymorphisms (SNPs), short insertion and deletions (indels) as well as large SVs that were not readily detectable by previous read-mapping approaches. I found that although SVs are commonly observed in rapidly-evolving large gene families, the way they impact family members differs: longer genes are affected by all types of SVs, while shorter genes seem to be predominantly affected by copy number changes through tandem or segmental duplication. This work illustrates the value of multiple *de novo* assemblies in building and characterizing plant pan-genomes and provides insights in understanding the evolution of large gene families underlying important traits.

Methods

Sequencing, assembly and functional annotation

Fifteen *M. truncatula* accessions from geographically distinct populations (Figure 2.1) spanning the entire *Medicago* species range were chosen for deep sequencing and *de novo* assembly. Sequencing was performed using Illumina HiSeq 2000 instruments. For each of the fifteen accessions, we made and sequenced one Short Insert Paired End library (SIPE) and either one or two Long Insert Paired End (LIPE) libraries following the recommendations of the ALLPATHS-LG whole genome assembler (Gnerre et al. 2011). Each accession was sequenced to an average of 120 fold coverage (Appendix Table 2.1) and assembled using the ALLPATHS-LG assembler algorithm (version 49962) (Gnerre et al. 2011). Assembled genome sizes ranged from 388 Mbp to 428 Mbp, values comparable to the reference HM101 genome. Approximately 80%-94% of each genome were assembled into scaffolds at least 100 kbp long, with scaffold N50 sizes ranging from 268 kbp to 1,653 kbp, and contig N50 sizes around 20 kbp (Table 2.1).

AUGUSTUS (Stanke2003) was used to make *ab initio* gene predictions for each genome assembly with both RNA-Seq expression evidence and HM101 homology evidence. Predicted protein sequences were scanned for PFAM domains (Pfam-A.hmm) (Finn et al. 2014) using HMMER (Sean R. Eddy 2011) and processed using custom scripts. Domain categories were then assigned to each protein sequence according to the most significant Pfam hits. Among the resulting Pfam domains we curated 133 as being

associated with transposable elements and grouped these into a large “TE" category. NBS-LRR genes were curated using sub-family HMMs (13 TNL subgroups and 22 CNL subgroups) built based on previous literature (Ameline-Torregrosa et al. 2008). We also ran SPADA (Small Peptide Alignment Discovery Algorithm) (Chapter 1) (Zhou et al. 2013) with default parameter on each assembly to refine annotation of 516 CRP gene subfamilies.

PacBio long reads was generated to validate the identified structural variation in three accessions including HM034, HM056 and HM340. Sequencing was done using P4C2 chemistry and Smrtanalysis version 2.1 or P5C3 chemistry and Smrtanalysis version 2.3. Reads were filtered at minimum quality of 75, minimum sub-read length of 50 bp and minimum read length of 50 bp. Final coverage for each accession was estimated to be 18-20 fold.

Comparative analysis and variant detection

Each *de novo* assembly was aligned to the *Medicago* HM101 reference sequence (version 4.0) using BLAT (Kent 2002). The resulting alignments were merged, fixed and cleaned using custom scripts. BLAT Chain/Net tools were used to obtain a single coverage best alignment net in the target genome (HM101), as well as a reciprocal-best alignment net between the two genomes. Genome-wide synteny blocks were then built for each *de novo* assembly (against HM101), enabling downstream analyses including variant calling, novel sequence identification and ortholog detection. Based on the synteny blocks built, I identified SNPs (single base mismatches), short insertions and deletions (alignment gaps $\leq$ 50 bases), as well as different types of structural variants (SVs). We were able to accurately detect and characterize large deletions, insertions, translocations and copy number gain and loss events at base pair resolution. Variant calls from 15 accessions were merged to a single VCF file using Bcftools (H. Li et al. 2009), with missing genotypes deducted where possible using synteny alignment information. We then partitioned the genome into 100-kbp sliding windows and calculated gene

density, TE density, NBS-LRR and CRP density, as well as nucleotide diversity (θ_π) estimates for SNPs, short InDels and SVs in each window.

Construction of a Medicago Pan-16 Proteome

For each *de novo* assembly, I first identified synteny orthologs (to HM101) using gene annotations from both genomes. For example, an ortholog pair can be determined if one gene in HM101 can be aligned to an HM004 gene through synteny at ≥80% sequence similarity and ≥70% sequence coverage. I also incorporated structural variant information (i.e., large insertions and deletions) in calling gene gains and losses as well as their locations. By repeating this for all 15 *de novo* sequenced accessions, I obtained a raw ortholog matrix with 16 columns representing 16 genomes and 70,000+ rows each representing an ortholog group. However, a gene gain event (insertion relatively to HM101) occurring in more than two accessions tended to independently introduce multiple row entries while actually representing a single event. I thus identified all unique insertion loci and did multiple sequence alignment of all inserted genes in each locus. Inserted genes were then clustered based on sequence similarity using MCL (Enright, Dongen, and Ouzounis 2002) and assigned to different ortholog groups. Since synteny blocks do not necessarily cover 100% of the reference gene space, there are numerous cells in the ortholog matrix with missing data, meaning that the ortholog status in those accessions cannot be determined through synteny. On the other hand, I noticed a considerable fraction of genes (10-20%) in one or more *de novo* assemblies residing on un-anchored (and short) scaffolds that are not in synteny with the reference genome. With the belief that some of the missing orthologs are actually among these orphan genes, I used a Reciprocal Best Hit (RBH) approach to assign orthologous relationships for genes without syntenic orthologs, and updated the ortholog matrix. Finally, for genes that could not be assigned to an ortholog group using either the synteny or RBH approach, I did an all-against-all Blast search and ran orthoMCL (L. Li, Stoeckert, and Roos 2003) and various custom scripts to generate ortholog groups. Large ortholog groups were partitioned so that each group has no more than one member from each accession. This

resulted in a Pan-16 proteome matrix where each ortholog group has between one (accession-specific) and 16 (shared by all accessions) members. The status in each cell can be “synteny-ortholog”, “RBH-ortholog”, “deleted”, “deleted-but-has-RBH-ortholog” or “missing data”. Each ortholog group was classified by the most frequent PFAM categories assigned to group members. Within each ortholog group, I built a multiple sequence alignment and calculated protein sequence distance matrix, from which the mean pairwise protein distance is obtained.

Gene family analysis

SNP-based nucleotide diversity were then estimated for the coding regions of each gene, and the distribution of (θ_π) for different gene families was obtained. Based on the Pan-16 proteome matrix, I generated an allele frequency spectrum (AFS) for each gene family using presence / absence information from each ortholog group. For each gene family, I also obtained the distribution of mean protein pairwise distance for each gene family by making a multiple sequence alignment and building a protein distance matrix. For selected gene families including CRPs and NBS-LRRs, I then built a score matrix with 16 accessions as rows and all family members as columns. Each cell in the matrix ranges from 0 (deleted) to 1 (present) representing the mean protein sequence similarity with orthologs from other accessions. Finally, I performed hierarchical clustering on the rows of the matrix and generated a dendrogram similar to the SNP-based phylogeny of 16 *Medicago* accessions.

Identification of gene family expansion / contraction events and validation using PacBio sequence

A list of gene gain (family expansion) and loss (family contraction) events were directly inferred from the Pan-16 proteome matrix. I characterized these events within different gene families with a special focus on the NBS-LRR and CRP gene families. For both gene gain and loss events PacBio long read sequences for three accessions (HM034, HM056 and HM340) were included to provide additional support. In the case of gene gains, my analysis specifically looked for tandem duplication events where there is an

identifiable ancestral gene (with >90% sequence similarity) in the vicinity (15 kbp) of the new gene, followed by scanning PacBio read alignments that span both the ancestral and newborn (novel) genes to provide support for the predicted duplication. For each gene gain and loss event, I also ran CLUSTALO (Sievers et al. 2011) to build a multiple sequence alignment, thereby obtaining a phylogeny to visualize any predicted changes in gene tree topology.

Results

Genome-wide identification of NBS-LRR and CRP gene families

I comprehensively scanned the *Medicago* reference genome (Mt4.0) for NBS-LRRs and CRPs using a combined approach (see Methods). A total of 860 NBS-LRRs and 1,428 CRPs (including 717 NCRs) were identified, annotated and assigned to a subfamily. Both gene families showed uneven and clustered distribution in the genome (Figure 3.1, 3.2). NBS-LRRs are mostly located on chromosomes 3 and 6, with members on the two chromosomes together accounting for 44% of the entire family. CRPs, on the other hand, are mostly clustered on chromosomes 2, 7 and 8, with several big clusters containing 15-20 tandem copies of a single subfamily on these chromosomes. Tandem duplication seems to be the primary mechanism driving the local expansion of many subfamilies (e.g., TNL0600 subfamily at the start of chromosome 6). Using a sliding window size of 100 kb, more than 80% of all NBS-LRRs reside in clusters of two or more, and more than 60% in clusters of five or more. Segmental duplication, on the other hand, also contributes to the large-scale expansion of several subfamilies, such as the CNL1600 clusters on chromosomes 5 and 8. I also noticed occasional ectopic duplication events, resulting in the transposition of foreign subfamily members to existing clusters (e.g., integration of a TNL0800 gene to a CNL0600 cluster on chromosome 5). These observations are consistent with a birth and death model in which tandem and segmental duplications are followed by density-dependent purifying selection (McHale et al. 2006).

Functional annotation of 15 Medicago accessions

In order to understand the natural variation affecting different gene families at a population level, my colleagues and I sequenced, assembled and annotated 15 additional *M. truncatula* accessions. We integrated homology evidence, RNA-Seq expression, and *ab initio* prediction results in the annotation process. The number of transposable element (TE)-related genes identified in the 15 *de novo* assemblies were on average 20% lower than the HM101 reference – an indication that the *de novo* assemblies have missed or collapsed some repetitive sequences. Total number of non-TE genes (52,000 to 56,000), however, is almost comparable to the reference (55,000, Table 3.1). Median protein length for non-TE genes ranged from 218 to 228 amino acids – nearly equal to the estimate of 228 amino acids in HM101, an indication that long scaffold and contig N50s have helped to maintain the intactness of gene models. The quality of the annotation was supported by the observation that 77-87% of all predicted genes (including TEs) were either supported by an HM101 homolog or by expression evidence (as determined by RNA-Seq) (Table 2.2). Large gene families such as NBS-LRRs and CRPs generally have consistent numbers of members among accessions (Table 2.2, Appendix Table 2.1. However, these gene families harbored complicated homology relationships with the accessions differing markedly in the size of specific sub-families (Appendix Table 2.5, 2.6). Further analysis suggests family-specific expansion / contraction is a frequent phenomenon observed in large gene families (see Discussion).

Population genetics analysis

I characterized the variability of different gene families using SNP-based nucleotide diversity (θ_π) in coding regions. Not surprisingly, the two broad categories of NBS-LRRs and TEs show the highest nucleotide diversity (Figure 3.3A). NCRs and other defensin-like genes (DEFLs, CRP0000-CRP1030) also harbor higher-than-average levels of diversity (Figure 3.3A). Other dynamic gene families such as FAR1, Exo_endo_phos, F-box and HSP70 are also among the most variable. I also characterized the effect of sequence variation with a focus on SNPs causing significant changes to the encoding

product (i.e., large-effect SNPs). While approximately 20% of all non-TE genes are affected by at least one large-effect SNP (in 12 "ingroup" accessions), as much as 70-80% of TEs and NBS-LRRs have reading frame changes (Figure 3.3B). The proportion of family members with reading frame changes basically follows the same trend as nucleotide diversity levels in different gene families, with the notable exception of CRP families (NCR, CRP0000-1030, CRP1600-6250), where only 10% are affected by large-effect SNPs, potentially as a result of the small size of these families (typically just 20-50 amino acids).

Novel coding sequences absent in the HM101 reference

When comparing each *de novo* assembly to the reference HM101 genome (Mt4.0), I found extensive "novel" sequence that could not be aligned to the reference even using a relaxed alignment parameter (70%-80% sequence percent identity). Among the novel sequences identified, 1.3 - 2.5 Mbp per accession were predicted to be protein coding (see Chapter 2 for detail). Enriched in these novel coding genes are TE-related genes and NBS-LRRs. While on average only 2% of coding non-TE genes were identified as absent from the reference accession, as many as 6% of NBS-LRRs contribute to the "novel" gene pool (Figure 3.4). NCRs, unlike other groups in the broader CRP family including classic DEFLs (CRP000-1030) and CRP1600-6250, contribute a significant amount (3.4%) to the novel gene pool. Considering the recently employed role of NCRs in directing endo-symbionts, it is possible that this large family is still undergoing a rapidly expanding and innovating process that leads to novel specificity in legume-rhizobia interaction.

Proteome diversity

To understand the effect of sequence variants on proteins, it is insufficient to study isolated DNA polymorphisms in the context of the reference annotation - especially at such high level of divergence (SNP density from 0.63% to 2.67%). Therefore, we fully annotated the 15 *de novo* assemblies and systematically identified ortholog groups among accessions, creating a pan-16 proteome. In addition to the 68k reference gene models, we

identified 78k ortholog groups with no HM101 members. Within this total of 146k groups, 31k ortholog groups were shared among all 16 accessions, and 94k were shared by at least two accessions (Figure 3.5A). A total of 52k singletons were also identified including as little as 2.3k HM004-specific to as much as 8.1k HM340-specific genes. The size curve of the pan-16 proteome resembles the pan-genome curve, with a stable core proteome (31k) plus a much larger dispensable proteome that still sees significant increase after inclusion of 15 *Medicago* accessions (Figure 3.5B).

We further investigated protein diversity in different gene families based on alignments of each ortholog group. On average, the distance of any two randomly selected protein orthologs was 1.8% (i.e., mean pairwise protein distance, Figure 3.6). However, the two groups of NBS-LRRs and TEs are extremely divergent, with approximately 10% difference between each ortholog pair. Other dynamic gene families such as FAR1, HSP70 and F-box, also show high levels of protein diversity. Interestingly, the NCR gene family did not show above-normal protein diversity level in contrast to its increased nucleotide diversity estimate (Figure 3.3, 3.6). This could be due to these reasons: 1) synonymous SNPs are prevalent in NCRs but do not contribute to diversity at the protein level; 2) gene structure predictions made on short-read-only *de novo* assemblies may sometimes be inaccurate due to sequencing and assembly errors, thus differentially inflating the protein distance estimates for larger genes based on protein sequence alignments; however, this is less an issue for smaller one- and two-exon genes such as NCRs. We speculate that point #2 may be more likely since there is no reason to see the protein diversity for non-TE coding genes going up beyond the level of F-box family.

Characterization of NBS-LRR gene family variation

The NBS-LRR gene family members are known for their high variability and rapid evolving nature. In addition to the highest point mutation rate and largest proportion of reading-frame changes (Figure 3.3), I also observed various types of structural changes targeting NBS-LRRs through mechanisms including gene truncation, domain swapping

and gene fusion (Figure 3.8). Indeed, among all gene families, NBS-LRRs contribute the most to the novel gene pool in the population (Figure 3.4), and harbors the highest level of protein diversity (Figure 3.5). Using the sequence identities of all NBS-LRRs in 16 *Medicago* accessions, I was able to reconstruct the original *Medicago* phylogeny through hierarchical clustering (Figure 3.9), an indication that in general, the gene tree of NBS-LRRs followed the evolutionary trajectory of different accessions within the population. However, subfamilies of NBS-LRRs are frequently affected by insertions, deletions and copy number changes (Figure 3.10). In particular, I noticed that NBS-LRR gene families contain large numbers of accession-specific genes (Figure 3.10), consistent with the highest proportion of novel genes among all gene families (Figure 3.4). It is possible that most of these accession-specific genes are actually members of existing ortholog groups that have accumulated too many point mutations – as evidenced by the highest nucleotide diversity – to be placed in existing groups. In other words, these gene families are evolving much faster than other conserved gene families as a response to the ever-changing pathogen environment.

Characterization of NCR gene family variation

As a rapidly expanding gene family in the *Medicago* species (Fedorova et al. 2002; Mergaert et al. 2003). the NCR family also shows evidence of high nucleotide diversity and high novel sequence content (Figure 3.3, 3.4). Similar to NBS-LRRs, hierarchical clustering of the protein sequence score matrix for CRPs also resulted in a tree resembling the *Medicago* phylogeny (Figure 3.11). In contrast to NBS-LRRs, the CRP genes are much shorter - typically in the range of 300 - 1000 bp (Silverstein et al. 2007). Presumably as result of being smaller targets, CRPs are less frequently affected by large SVs involving mechanisms such as gene truncation or domain swapping. Instead, removals, insertions or duplications of a complete CRP gene are more frequently observed, resulting in contraction or expansion of a certain sub-family clade (Figure 3.12 and 3.13). For example, out of the total of 274 CRP insertion / copy number gain events in HM340, I identified 74 cases of tandem duplications where the duplicated CRP has a

highly similar ancestor (>90%) in the vicinity (within 15 kbp). Using long PacBio sequence reads available for HM340, I was able to validate 57 (77%) of these tandem duplication events (See Figure 3.14 for illustration).

Discussion

Family-specific expansion / contraction is prevalent in large gene families

Typical gene families generally have consistent numbers of members across different accessions, with occasional insertion/deletion or translocation events (Figure 3.15). Orthologous relationships are simple and straightforward in these conserved gene families. However, large gene families such as NBS-LRRs and CRPs show very different scenario. While subfamilies generally have consistent numbers of members among accessions (Figure 3.10 and 3.13, Appendix Table 2.3), the orthologous relationships, especially for the NCR subfamilies, are much more complex with frequent indels, translocations and CNVs affecting family members (Figure 3.13). Different subfamilies also seem to have different evolutionary history (Figure 3.13). In addition, NBS-LRRs show considerable presence/absence variation among different accessions (Figure 3.9 and 3.10). As a result of high point mutation rates (i.e., highest nucleotide diversity), synteny-based orthologous relationships are rare for NBS-LRR subfamilies (Figure 3.10). Many family members exist as accession-specific singletons with no detectable orthologs (or too divergent to be detected) in other accessions – partly explaining their highest contribution to the novel gene pool of the population.

De novo assemblies capture "novel" gene pool in the population

Based on the novel sequences we identified in the 15 *de novo* assembly, we found that large gene families including NBS-LRRs, NCRs, TEs, HSP70, are the major contributors the "novel" gene pool in the *Medicago* population (Figure 3.4). Interestingly, these gene families were also found among the most poorly characterized genomic regions (see Figure 2.4 and Chapter 2 for detail). In fact, based on the synteny alignments built for each *de novo* assembly we were able to cover 75% (31,048) of all non-TE

reference gene models (≥ 80% coding regions in ≥ 10 accessions). By contrast, only 44% (5,418) TEs and 48% (414) NBS-LRRs were covered using the same criteria. Enrichment of these complex gene families in this novel gene pool partially explains their poor coverage by synteny alignment. The rapid evolving nature and turnover of these genes greatly accelerate their divergence within population, thus preventing both read mapping and syntenic anchoring. That said, the total number of NBS-LRRs predicted in each *de novo* assembly did not differ significantly from the reference annotation (Table 3.1), indicating that instead of being missed by our assemblies, these NBS-LRRs most likely were captured and assembled to relatively short and isolated contigs that could not be anchored to reference chromosomes (confirmed in Figure 3.10).

Different evolution patterns of NBS-LRRs and NCRs

The NBS-LRR gene family is known for its high variability and rapid evolving nature. In addition to the highest point mutation rate and largest proportion of reading-frame changes (Figure 3.3), NBS-LRRs are also the major contributor to novel gene pool in the population (Figure 3.4), and harbor the highest level of protein diversity (Figure 3.6 and 3.9). In fact, it is frequently found that NBS-LRR genes have multiple allelic forms in the population that are dramatically different (see Appendix Figure 3.1 for illustration). Protein ortholog analysis also reveals NBS-LRRs are frequently affected by PAVs (Figure 3.9 and 3.10). These PAVs exist as accession-specific singletons with no detectable orthologs (or too divergent to be detected) in other accessions, a result of the family's much higher point mutation rate. NCRs, on the other hand, shows a relatively different evolution pattern. While the NCR gene family is also notable for a high level of nucleotide diversity and significant contribution to the novel gene pool, it is less frequently affected by structural variation, probably due to the small size of the gene body (Figure 3.12). As a result, the actual protein diversity level is not particularly high for NCRs (Figure 3.6). Nevertheless, protein ortholog analysis revealed frequent gene gain and loss events for NCR subfamilies (Figure 3.13), suggesting the primary innovation for NCRs are subfamily expansion and contraction through tandem

duplication and deletion events followed by diversifying selection. Indeed, we detected and characterized the copy number changing events in the NCR subfamilies and validated the tandem duplication events with PacBio long reads (Figure 3.14). The exact mechanism how the NCR subfamilies expand and translocate still requires further investigation.

Conclusion

In this study we constructed high quality *de novo* assemblies and systematically characterized natural variation in 15 *M. truncatula* accessions. We built a *Medicago* Pan-16 genome and proteome, and fully characterized all types of variation affecting different gene families: single-nucleotide polymorphisms (SNPs), short insertion and deletions (indels) as well as large SVs that were not detectable by previous read-mapping approaches. We confirmed the long established view that the NBS-LRR gene family is by far the most variable and dynamic family, contributing significantly to population novel gene pool and playing a structural role in genome architecture and stability. We also found that unlike traditional DEFLs and CRPs, the NCR gene family shows higher levels of diversity and is frequently affected by gene gain and loss events. This work illustrates the value of multiple *de novo* assemblies in building and characterizing plant pan-genomes and provides insights in understanding the evolution of large gene families underlying important traits.

Table 3.1. Functional annotation statistics.

	# Total Genes	TE	non-TE	NBS-LRR	CRP	Median Prot. Len.*	RNA-seq (%)#	Homology (%)&	RNA-seq + Homology (%)
HM101	67102	12312	54790	860	1428	228	39.6	-	39.6
HM058	64146	9540	54606	800	1280	222	-	85.2	85.2
HM056	65691	10379	55312	811	1304	220	36.4	84.9	86.9
HM125	67346	11175	56171	781	1280	218	-	84.4	84.4
HM129	65607	10455	55152	818	1278	222	-	83.6	83.6
HM034	64612	9958	54654	774	1266	222	36.0	83.1	85.2
HM095	65524	10197	55327	823	1283	222	-	82.8	82.8
HM060	64648	9967	54681	799	1272	223	-	83.7	83.7
HM185	65921	10666	55255	822	1274	222	-	83.4	83.4
HM004	64374	9680	54694	797	1278	224	-	82.8	82.8
HM050	65691	10282	55409	828	1289	224	-	82.6	82.6
HM023	64310	9661	54649	797	1281	222	-	83.2	83.2
HM010	65373	10339	55034	834	1304	223	-	83.2	83.2
HM022	59882	8193	51689	704	1295	227	-	78.6	78.6
HM340	64587	9387	55200	784	1287	228	36.1	75.1	77.6
HM324	60236	7751	52485	747	1213	221	-	76.6	76.6

*Median protein length (number of amino acids) was estimated using non-TE coding genes;

#RNA-Seq was done for four accessions using both un-inoculated root tissue and nodule; number indicates percentage of total predicted transcripts with FPKM > 0;

&Number indicates percentage of total predicted transcripts with at least one Mt4.0 ortholog (either syntenic ortholog or RBH-based homolog).

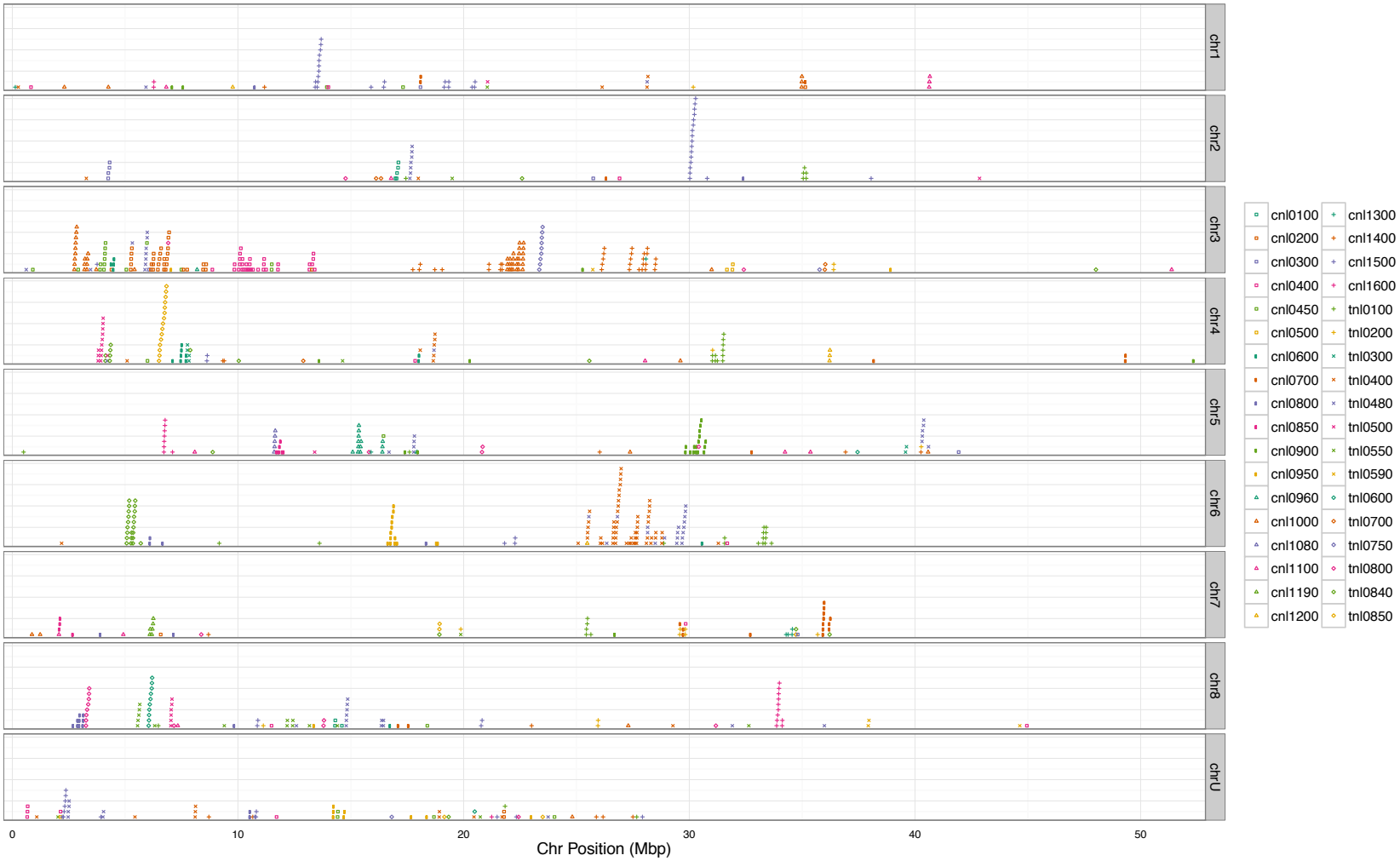


Figure 3.1. Genome distribution of NBS-LRR gene family in HM101 reference genome.

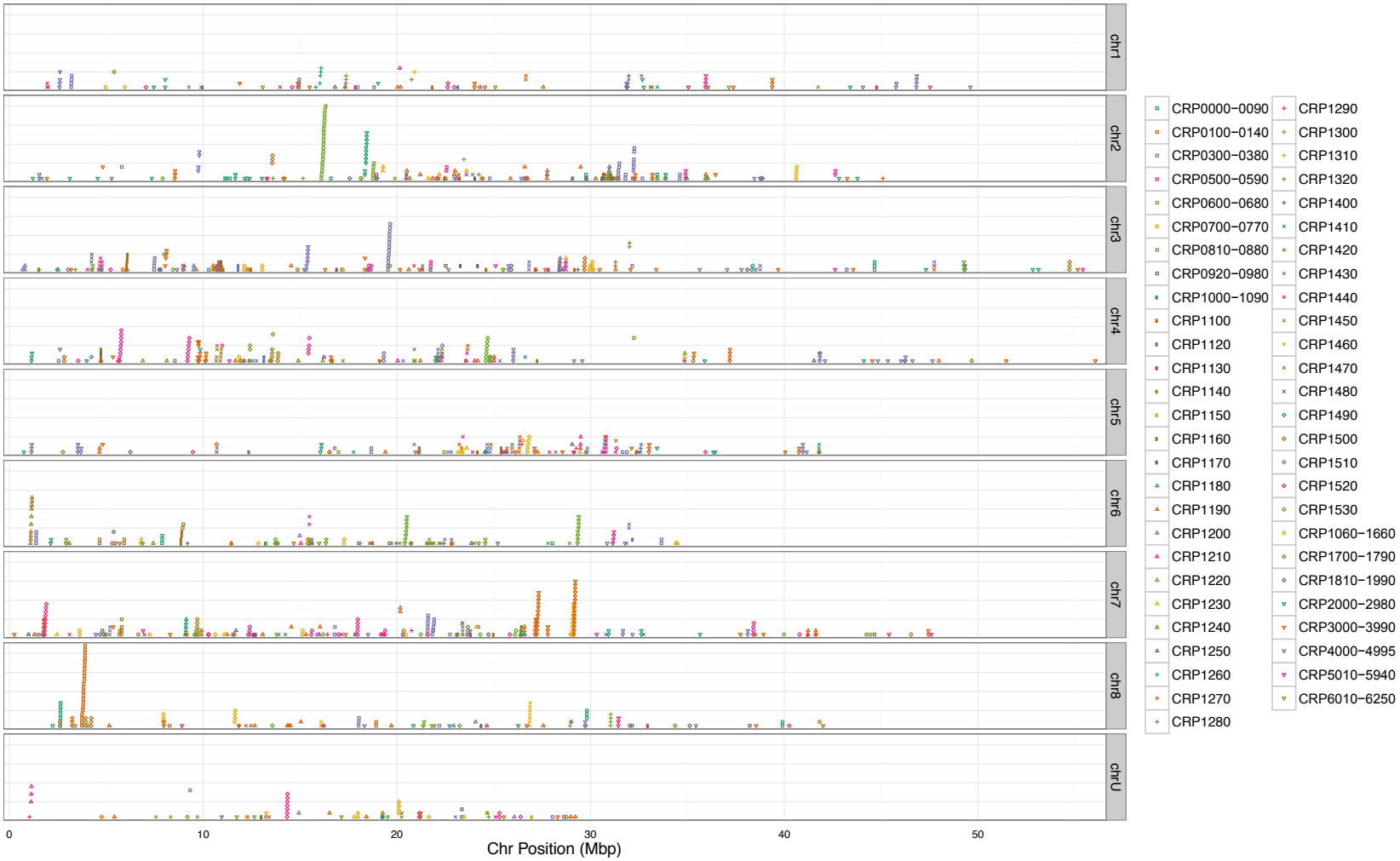


Figure 3.2. Genome distribution of CRP gene family in HM101 reference genome.

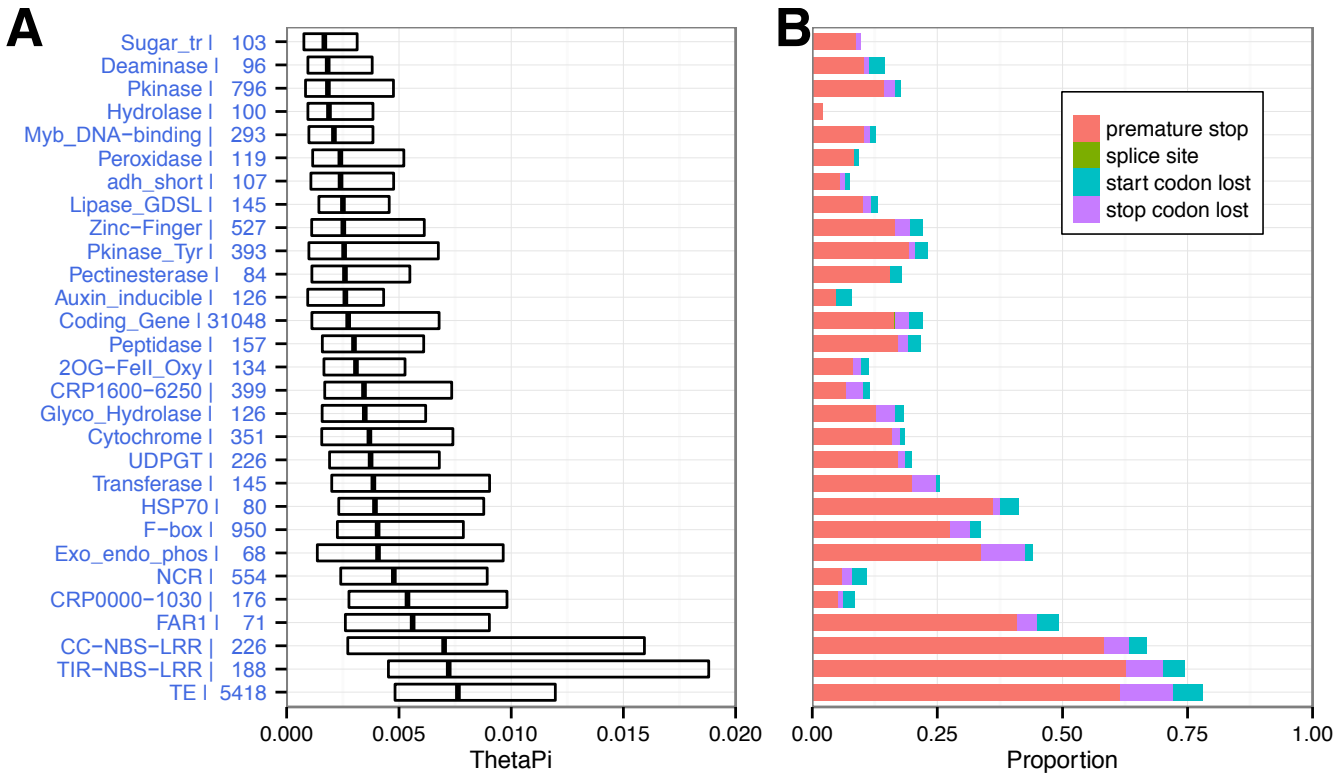


Figure 3.3. SNP-based nucleotide diversity estimates of different gene families (A) and proportion members affected by different types of large-effect SNPs (B).

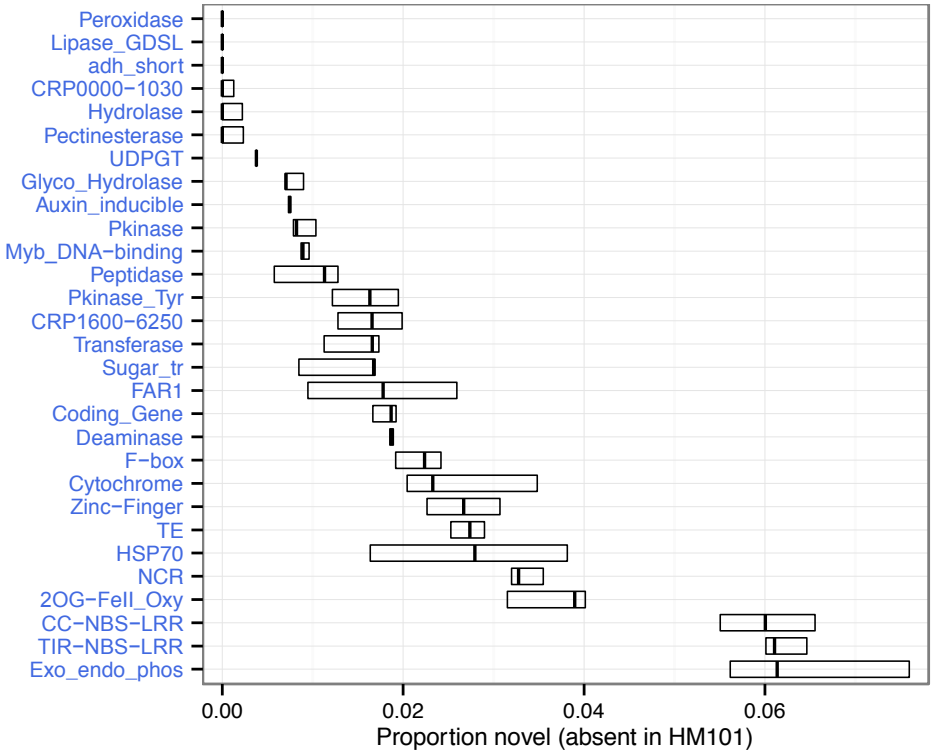


Figure 3.4. Proportion of novel genes identified in different gene families.

"Novel genes" are gene family members where 50% of the coding sequence regions are not present in the reference HM101 accession (i.e., cannot align anywhere in the reference genome). Error bars indicate 25%, 50% and 75% quantiles of the proportion distribution in 15 accessions.

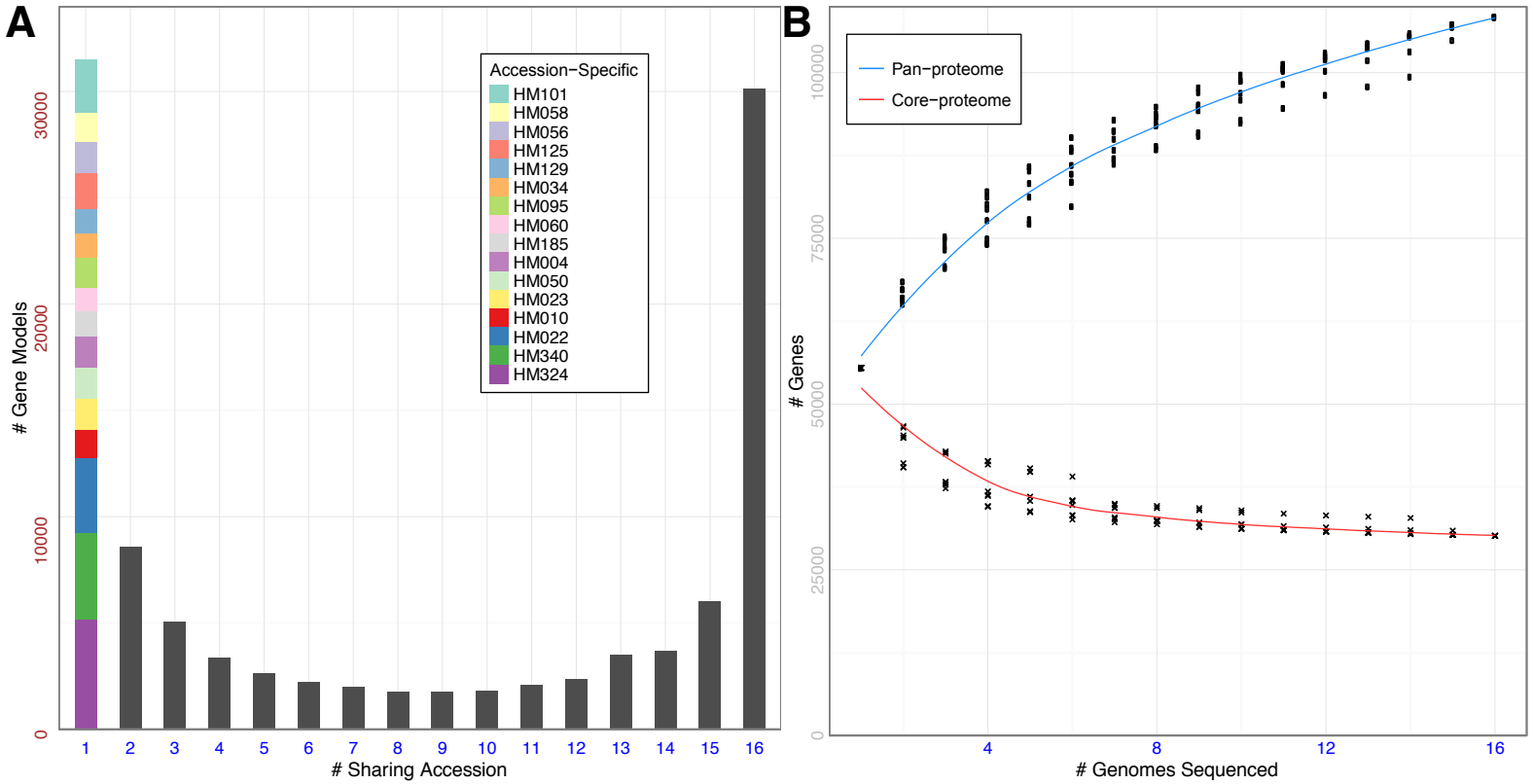


Figure 3.5. Allele frequency spectrum of ortholog groups identified in all 16 *M. truncatula* accessions (A) and the Pan-16 proteome size curve (B).

The pan-proteome size curve and core-proteome size curve were derived by adding one accession to the pool at a time and counting the number of shared ortholog groups (including members in all accessions, i.e., core-proteome, 'x' in the figure) and the total number of ortholog groups (pan-proteome, 'o' in the figure). This process was repeated 8 times by shuffling the order of accessions.

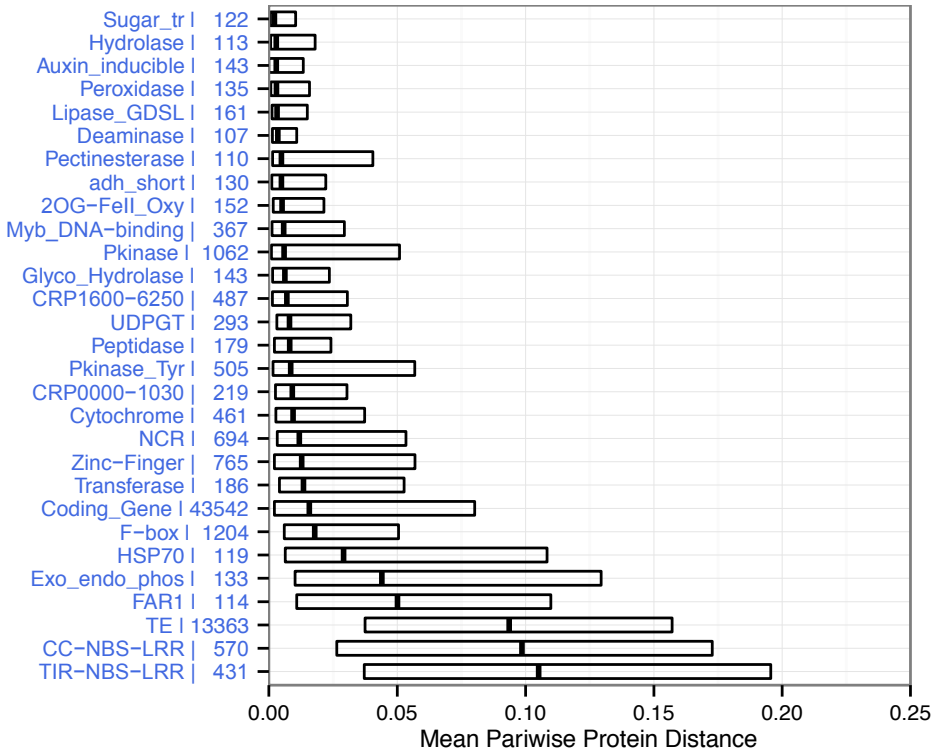


Figure 3.6. Distribution of mean pairwise protein distances in different gene families.

Barplot represents first quantile, median and third quantile of the distribution of mean pairwise protein distances.

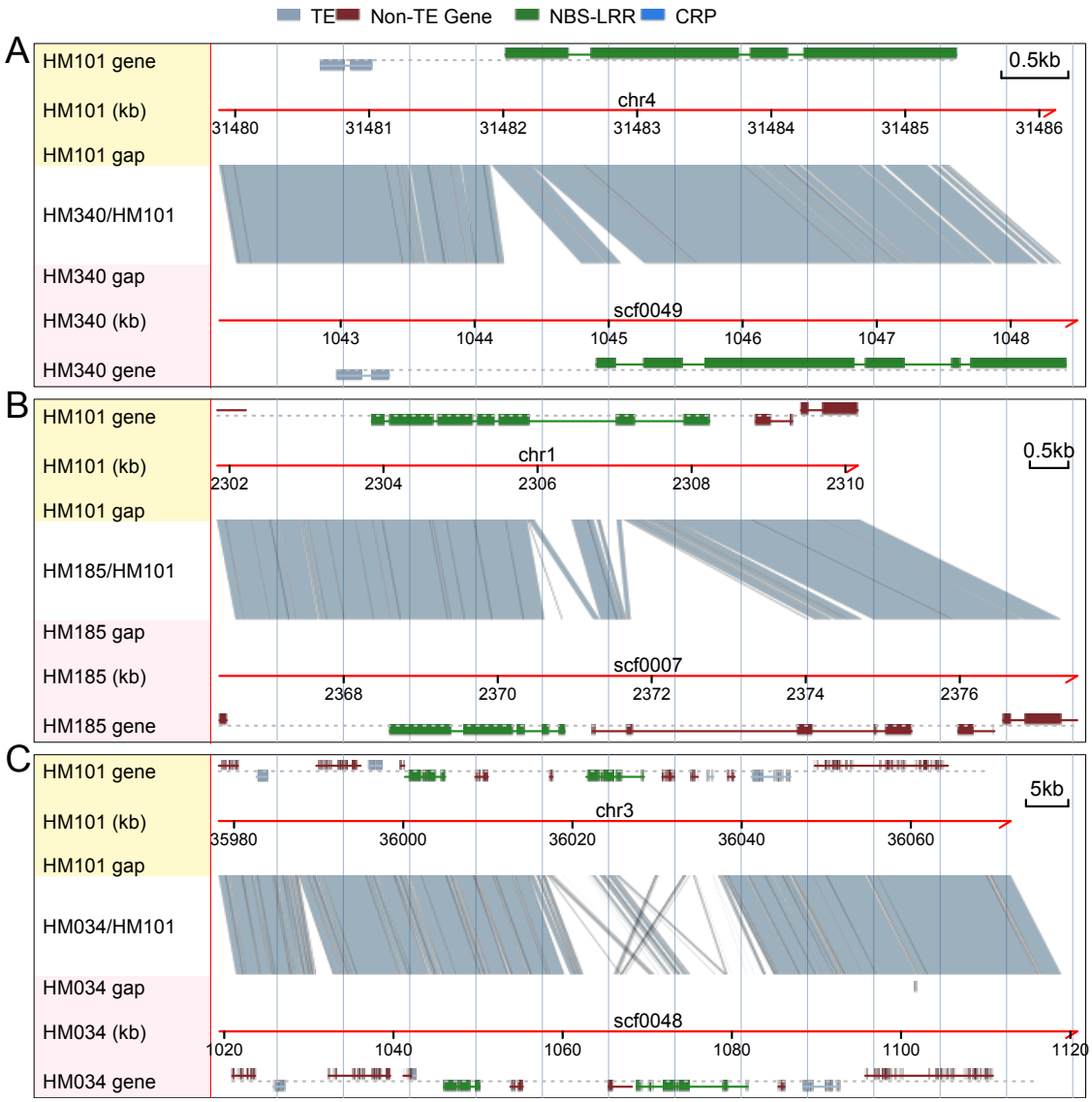


Figure 3.8. Illustration of NBS-LRR genes affected by different types of structural variants.

(A) Insertion in the first exon results in intron-exon structure change in the gene body in the HM340 ortholog (relatively to HM101); (B) a number of insertions and deletions result in truncation of the first half and fusion of the second half with a downstream gene in HM185; (C) synteny between HM034 and HM101 breaks at a NBS-LRR gene resulting in insertion / deletion / domain swapping / domain fusion of the ortholog pair.

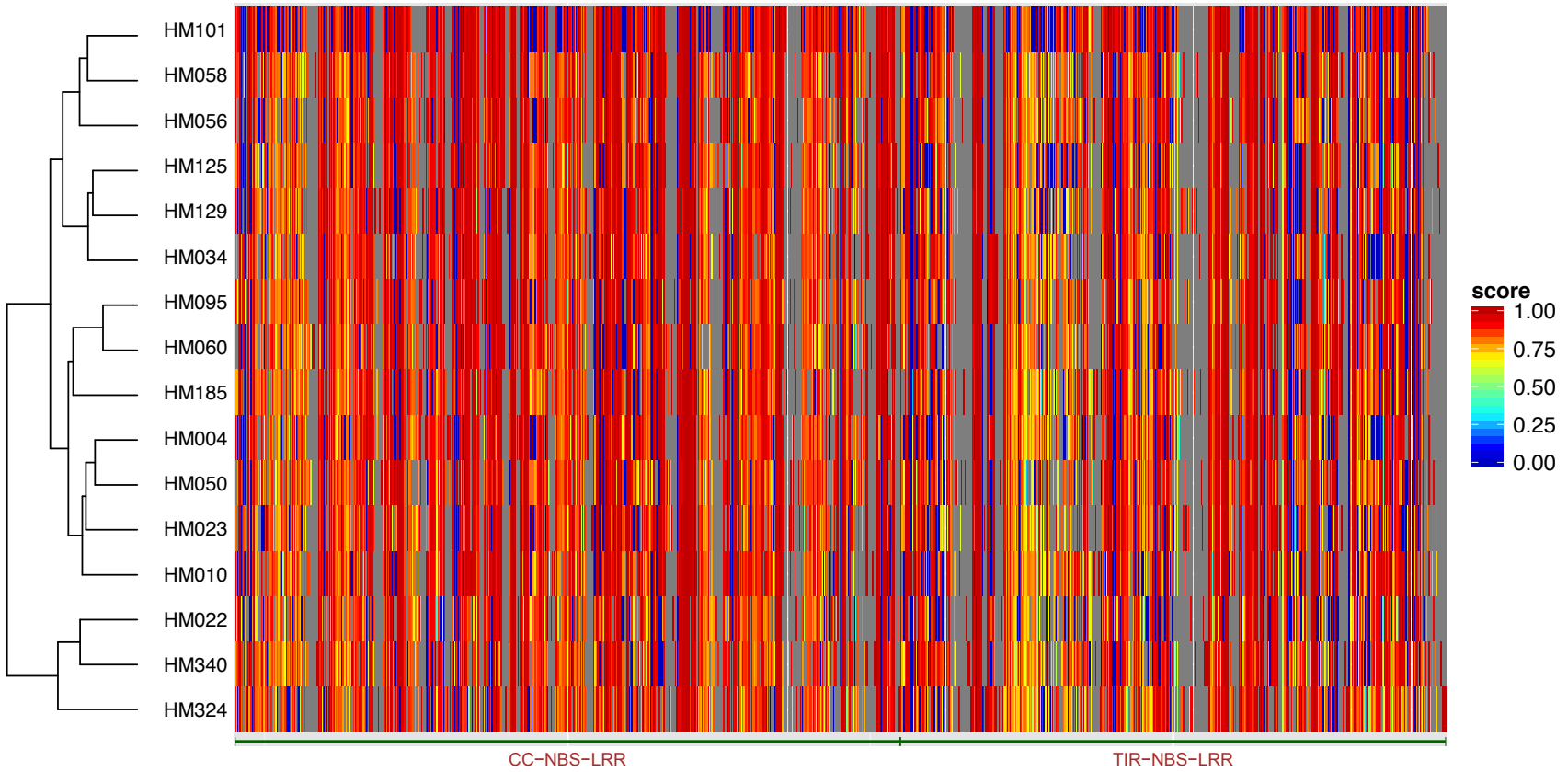


Figure 3.9. The NBS-LRR sequence identity matrix and hierarchical clustering.

Each cell in the score matrix ranges from 0 (gene deleted) to 1 (gene identical) representing the mean protein sequence similarity with orthologs from other accessions. Hierarchical clustering was performed on the rows of the matrix to generate a dendrogram similar to the *Medicago* phylogeny.

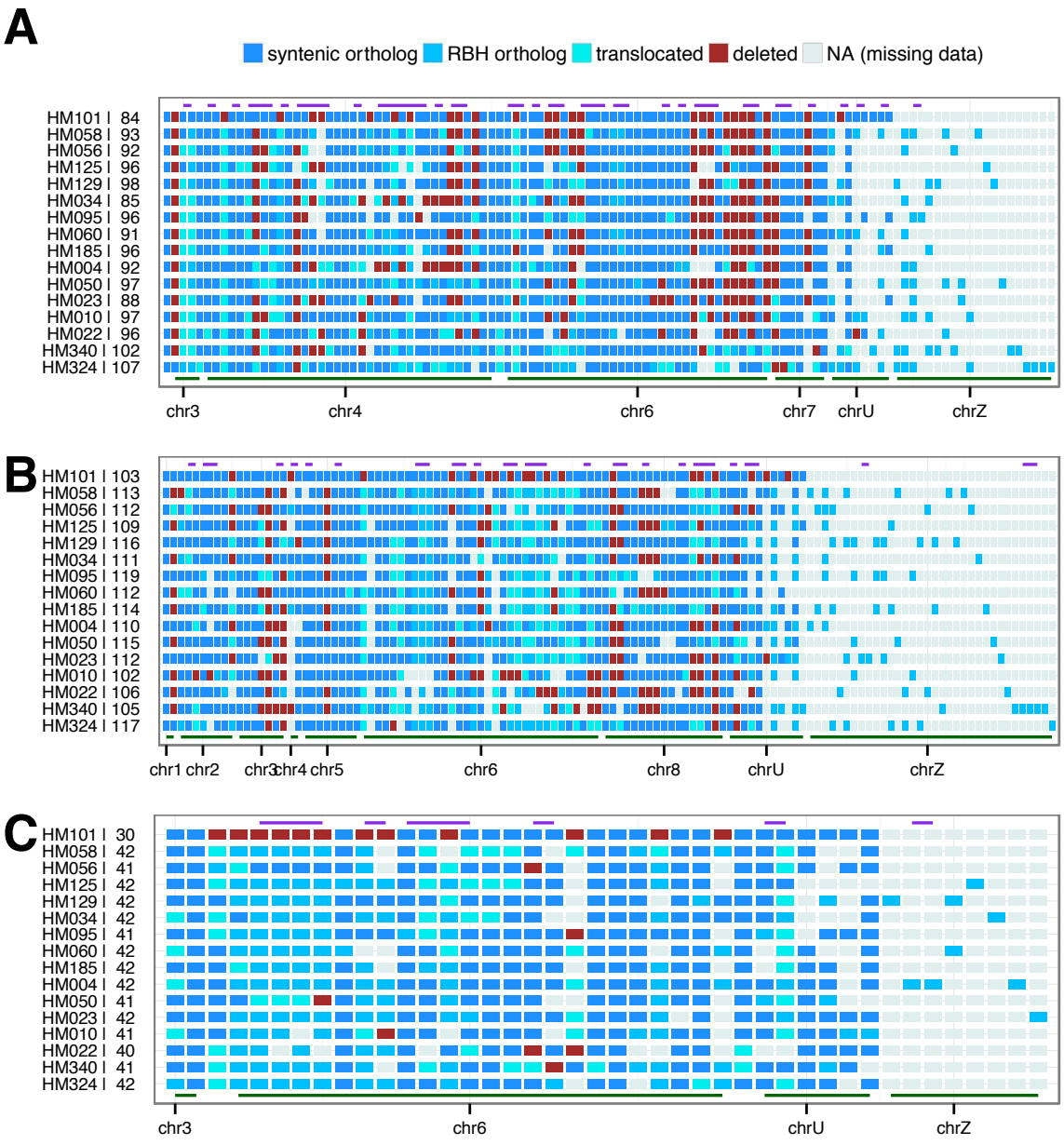


Figure 3.10. Ortholog status of selected NBS-LRR subfamilies: (A) TNL0850, (B) TNL0480 and (C) CNL0950.

"Syntenic ortholog": ortholog status can be determined by synteny alignment (with HM101); "RBH ortholog": no synteny ortholog can be found but reciprocal BLAST search identifies an ortholog (typically on short, separate scaffolds); "translocated": synteny ortholog is removed from original location and found in another location; "deleted": synteny ortholog is deleted from original location and not present elsewhere in

the de novo assembly; “NA”: insufficient information (synteny coverage, BLAST search) to infer the ortholog status at this locus.

Green lines at the bottom give their chromosomal locations with “chrU” standing for unanchored A17 (HM101) short contigs and “chrZ” standing for ortholog groups with no A17 members (i.e., not present in HM101). Purple lines on top indicate locations of gene clusters (within 15 kbp downstream or upstream to the next).

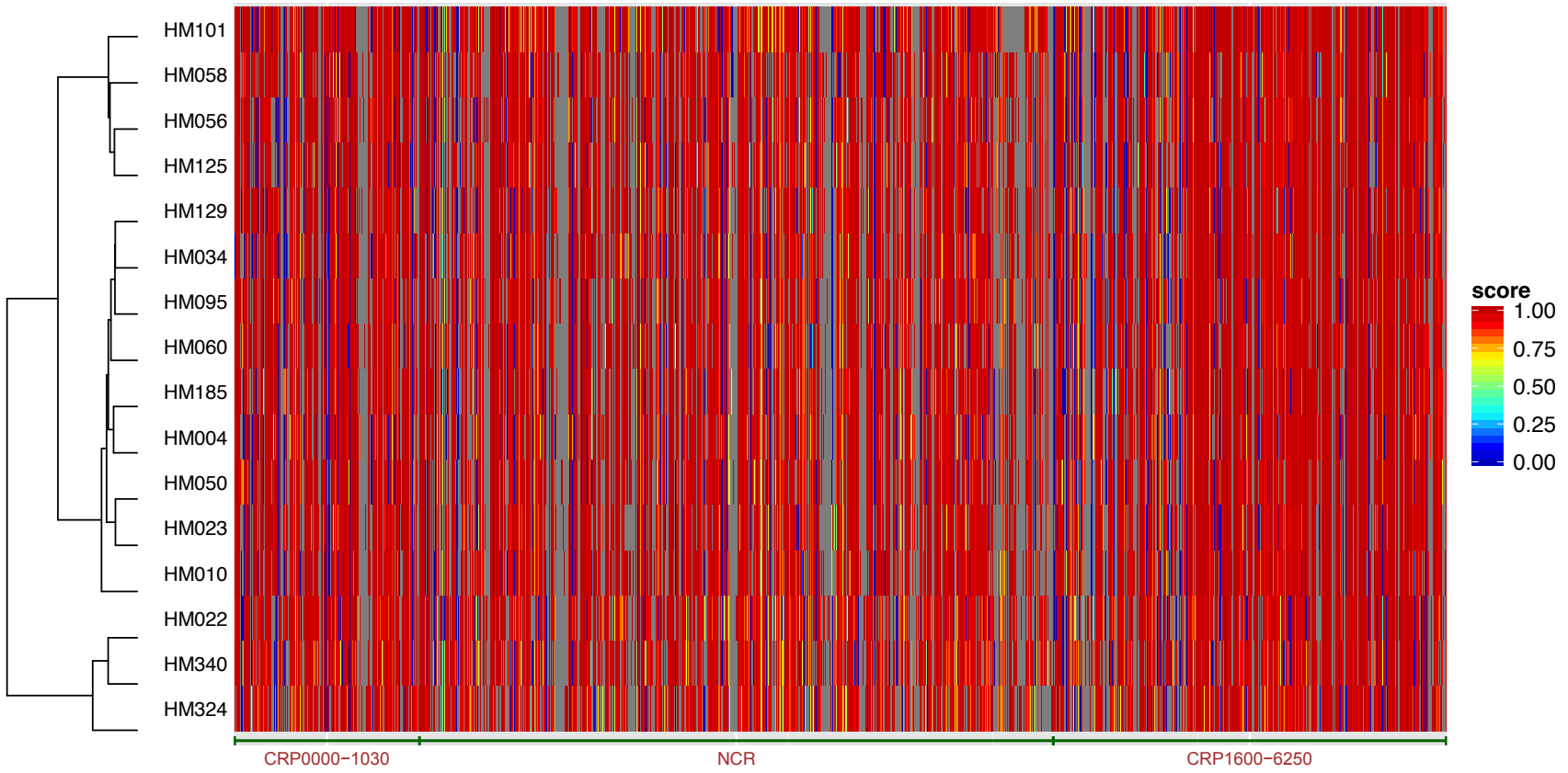


Figure 3.11. The CRP sequence identity matrix and hierarchical clustering.

Each cell in the score matrix ranges from 0 (gene deleted) to 1 (gene identical) representing the mean protein sequence similarity with orthologs from other accessions. Hierarchical clustering was performed on the rows of the matrix to generate a dendrogram similar to the *Medicago* phylogeny.

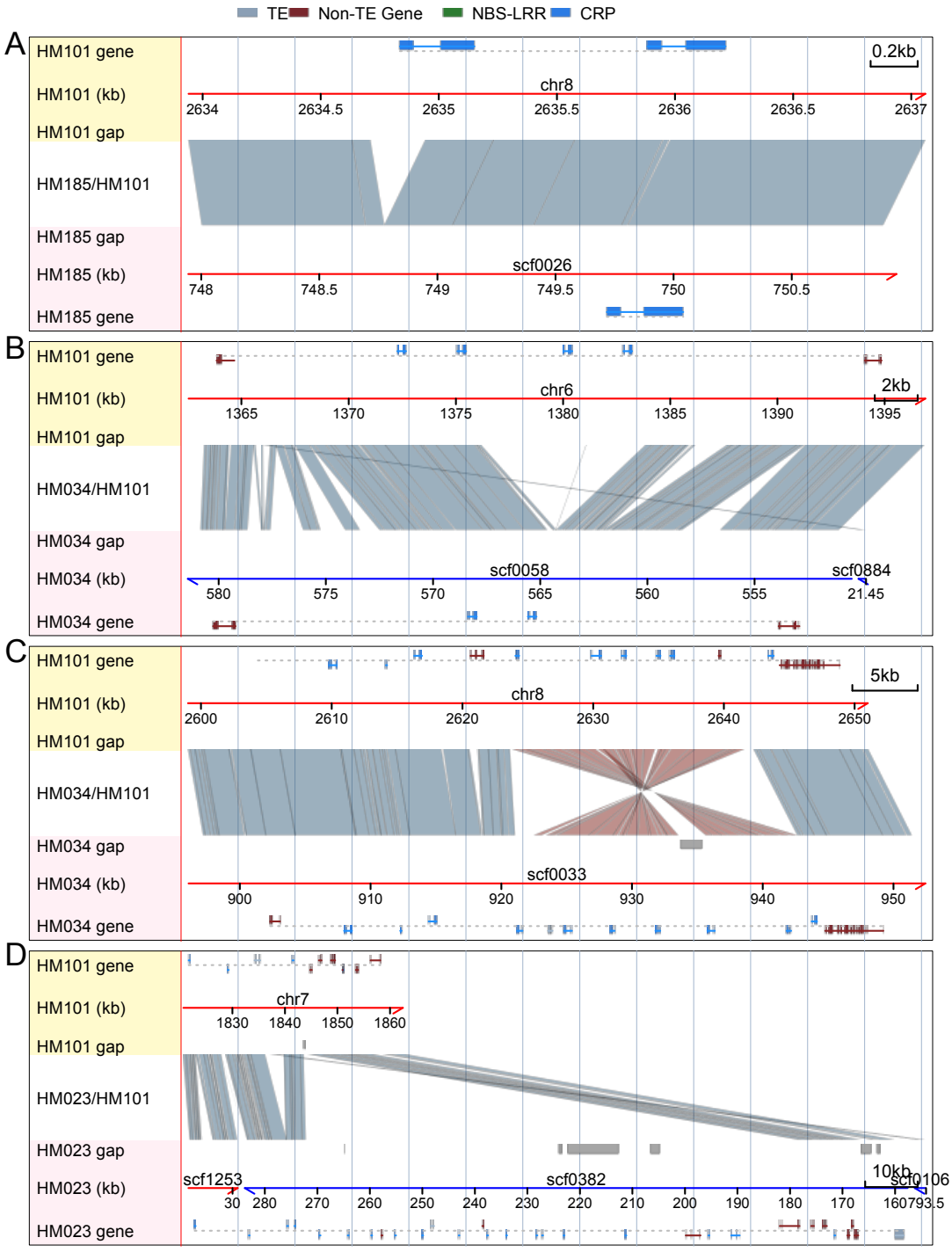


Figure 3.12. Illustration of NCR genes affected by different types of structural variants.

(A) Deletion of the signal peptide of an NCR gene leads to loss-of-function of the HM185 ortholog; (B) Deletion of two NCRs results in a cluster size change from 4 copies

in HM101 to 2 copies in HM034 (or more likely, a tandem duplication in HM101 relative to HM034); (C) Inversion of gene cluster containing 5 NCRs in HM034; (D) Tandem duplication of at least 17 copies of an NCR cluster in HM023.

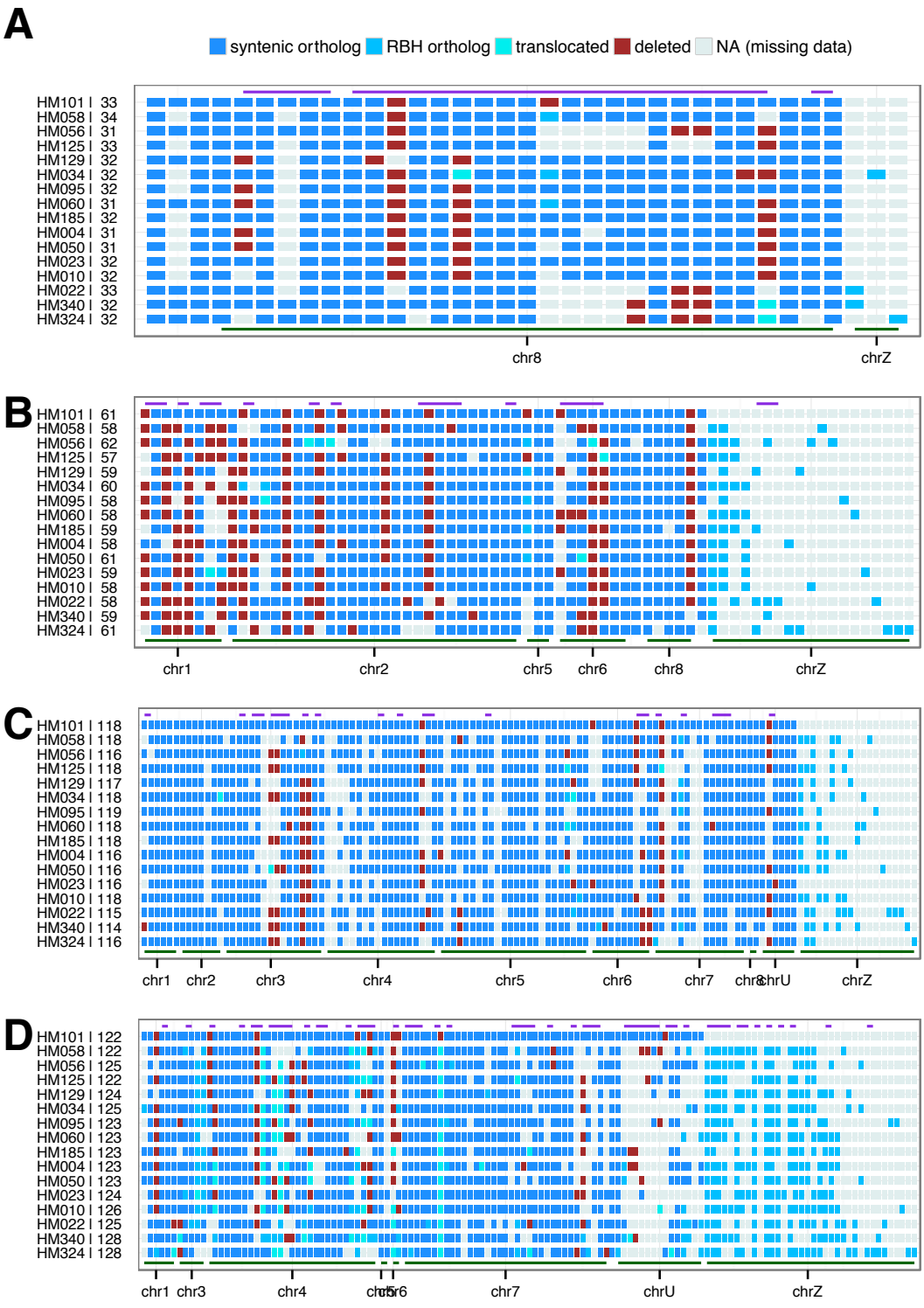

A
syntenic ortholog
RBH ortholog
translocated
deleted
NA (missing data)
HM101 | 33
HM058 | 34
HM056 | 31
HM125 | 33
HM129 | 32
HM034 | 32
HM095 | 32
HM060 | 31
HM185 | 32
HM004 | 31
HM050 | 31
HM023 | 32
HM010 | 32
HM022 | 33
HM340 | 32
HM324 | 32
chr8
chrZ
B
HM101 | 61
HM058 | 58
HM056 | 62
HM125 | 57
HM129 | 59
HM034 | 60
HM095 | 58
HM060 | 58
HM185 | 59
HM004 | 58
HM050 | 61
HM023 | 59
HM010 | 58
HM022 | 58
HM340 | 59
HM324 | 61
chr1
chr2
chr5
chr6
chr8
chrZ
C
HM101 | 118
HM058 | 118
HM056 | 116
HM125 | 118
HM129 | 117
HM034 | 118
HM095 | 119
HM060 | 118
HM185 | 118
HM004 | 116
HM050 | 116
HM023 | 116
HM010 | 118
HM022 | 115
HM340 | 114
HM324 | 116
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chrU
chrZ
D
HM101 | 122
HM058 | 122
HM056 | 125
HM125 | 122
HM129 | 124
HM034 | 125
HM095 | 123
HM060 | 123
HM185 | 123
HM004 | 123
HM050 | 123
HM023 | 124
HM010 | 126
HM022 | 125
HM340 | 128
HM324 | 128
chr1
chr3
chr4
chr5
chr6
chr7
chrU
chrZ

Figure 3.13. Ortholog status of selected CRP subfamilies: (A) CRP0110 (mycorhizal-specific defensin), (B) CRP0355 (reproductive-specific DEFL), (C) CRP1430 (6-cysteine NCR) and (D) CRP1520 (4-cysteine NCR).

"Syntenic ortholog": ortholog status can be determined by synteny alignment (with HM101); "RBH ortholog": no synteny ortholog can be found but reciprocal BLAST search identifies an ortholog (typically on short, separate scaffolds); "translocated": synteny ortholog is removed from original location and found in another location; "deleted": synteny ortholog is deleted from original location and not present elsewhere in the de novo assembly; "NA": insufficient information (synteny coverage, BLAST search) to infer the ortholog status at this locus.

Green lines at the bottom give their chromosomal locations with "chrU" standing for unanchored A17 (HM101) short contigs and "chrZ" standing for ortholog groups with no A17 members (i.e., not present in HM101). Purple lines on top indicate locations of gene clusters (within 15 kbp downstream or upstream to the next).

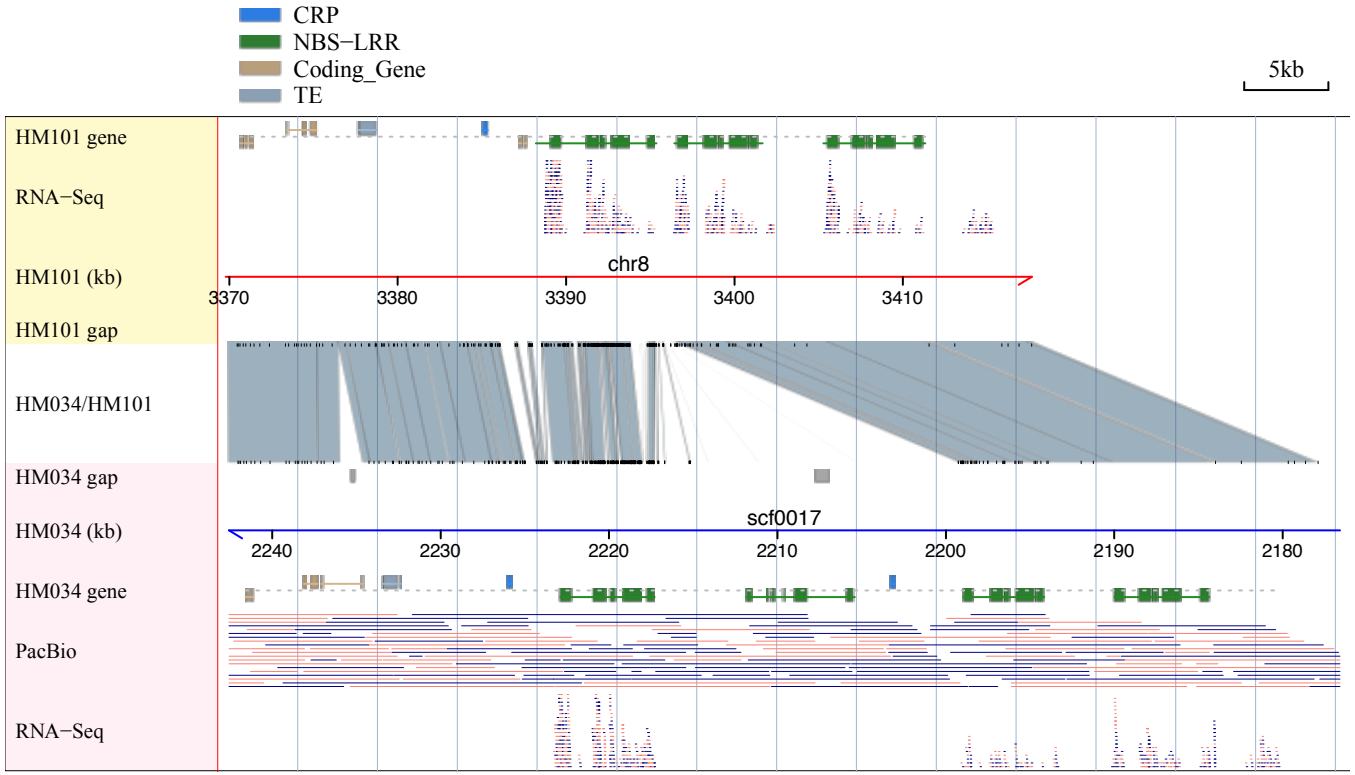


Figure 3.14. Tandem duplication of an NBS-LRR together with a CRP is supported long PacBio reads.

Blue and red segments represent reads in forward and reverse strands.

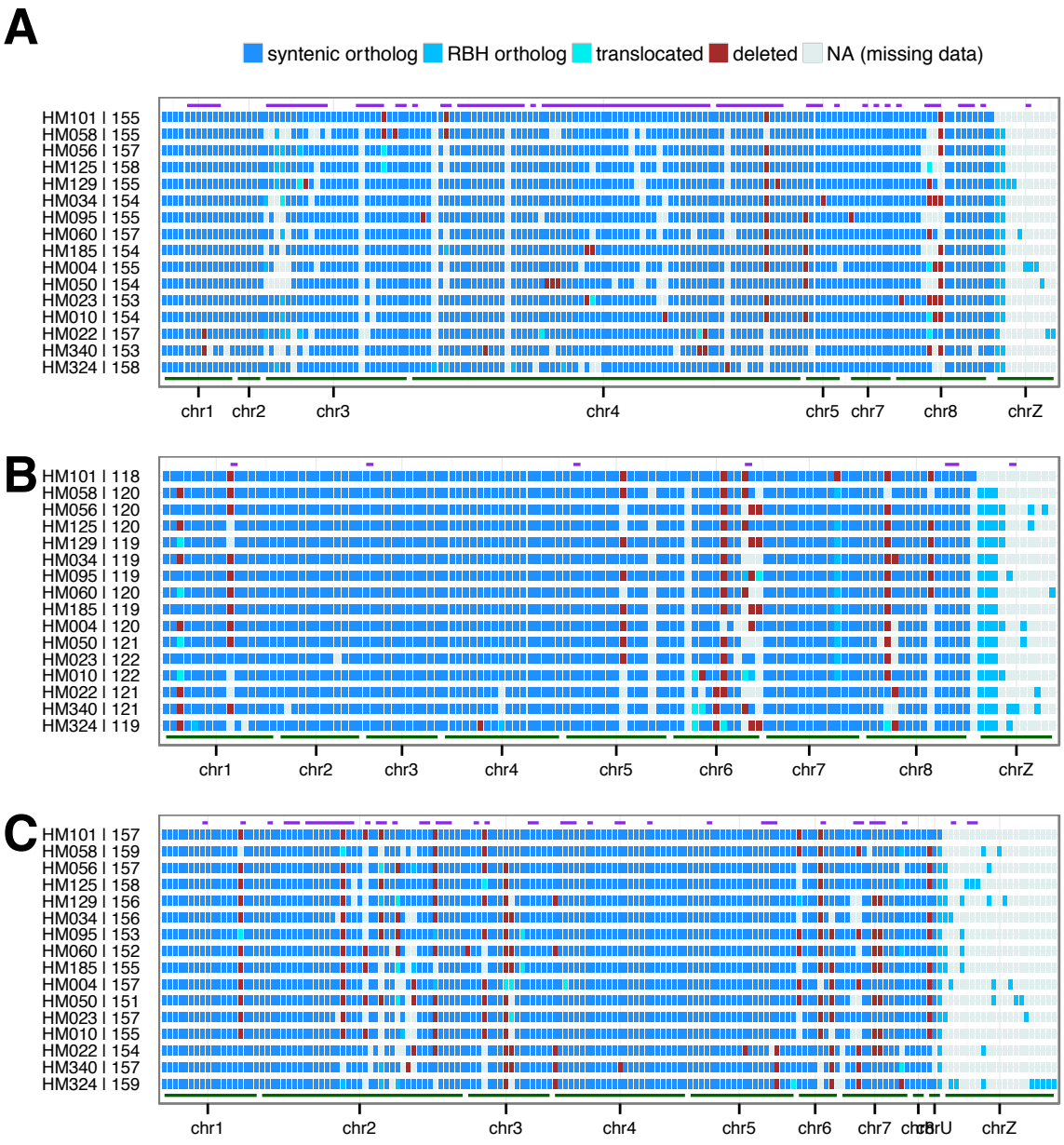


Figure 3.15. Ortholog status of selected (typical) gene families: (A) auxin_inducible, (B) deaminase and (C) peroxidase.

"Syntenic ortholog": ortholog status can be determined by synteny alignment (with HM101); "RBH ortholog": no synteny ortholog can be found but reciprocal BLAST search identifies an ortholog (typically on short, separate scaffolds); "translocated": synteny ortholog is removed from original location and found in another location; "deleted": synteny ortholog is deleted from original location and not present elsewhere in

the de novo assembly; “NA”: insufficient information (synteny coverage, BLAST search) to infer the ortholog status at this locus.

Green lines at the bottom give their chromosomal locations with “chrU” standing for unanchored A17 (HM101) short contigs and “chrZ” standing for ortholog groups with no A17 members (i.e., not present in HM101). Purple lines on top indicate locations of gene clusters (within 15 kbp downstream or upstream to the next).

Literature Cited

Alkan, Can, Bradley P Coe, and Evan E Eichler. 2011. "Genome Structural Variation Discovery and Genotyping." *Nature Reviews*. *Genetics* 12 (5). Nature Publishing Group: 363–76. doi:10.1038/nrg2958.

Alkan, Can, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, et al. 2009. "Personalized Copy Number and Segmental Duplication Maps Using next-Generation Sequencing." *Nature Genetics* 41 (10). Nature Publishing Group: 1061–67. doi:10.1038/ng.437.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Alunni, Benoit, Zoltan Kevei, Miguel Redondo-Nieto, Adam Kondorosi, Peter Mergaert, and Eva Kondorosi. 2007. "Genomic Organization and Evolutionary Insights on GRP and NCR Genes, Two Large Nodule-Specific Gene Families in Medicago Truncatula." *Molecular Plant-Microbe Interactions : MPMI* 20 (9): 1138–48. doi:10.1094/MPMI-20-9-1138.

Ameline-Torregrosa, Carine, Bing-Bing Wang, Majesta S O'Bleness, Shweta Deshpande, Hongyan Zhu, Bruce Roe, Nevin D Young, and Steven B Cannon. 2008. "Identification and Characterization of Nucleotide-Binding Site-Leucine-Rich Repeat Genes in the Model Plant Medicago Truncatula." *Plant Physiology* 146 (1): 5–21. doi:10.1104/pp.107.104588.

Angiuoli, Samuel V., and Steven L. Salzberg. 2011. "Mugsy: Fast Multiple Alignment of Closely Related Whole Genomes." *Bioinformatics* 27 (3): 334–42. doi:10.1093/bioinformatics/btq665.

Bailey, J A. 2002. "Recent Segmental Duplications in the Human Genome." *Science* 297 (5583): 1003–7. doi:10.1126/science.1072047.

Basrai, Munira a., Philip Hieter, and Jef D. Boeke. 1997. "Small Open Reading Frames: Beautiful Needles in the Haystack." *Genome Research*. doi:10.1101/gr.7.8.768.

Baumgarten, Andrew, Steven Cannon, Russ Spangler, and Georgiana May. 2003. "Genome-Level Evolution of Resistance Genes in Arabidopsis Thaliana." *Genetics* 165 (1). Genetics Soc America: 309–19.

Begun, David J., Alisha K. Holloway, Kristian Stevens, LaDeana W. Hillier, Yu Ping Poh, Matthew W. Hahn, Phillip M. Nista, et al. 2007. "Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila Simulans." *PLoS Biology* 5 (11): 2534–59. doi:10.1371/journal.pbio.0050310.

Belkhadir, Youssef, Rajagopal Subramaniam, and Jeffery L Dangl. 2004. "Plant Disease Resistance Protein Signaling: NBS–LRR Proteins and Their Partners." *Current Opinion in Plant Biology* 7 (4): 391–99. doi:10.1016/j.pbi.2004.05.009.

Bennett, M. D., and I. J. Leitch. 2011. "Nuclear DNA Amounts in Angiosperms: Targets, Trends and Tomorrow." *Annals of Botany* 107 (3): 467–590. doi:10.1093/aob/mcq258.

Benson, G. 1999. "Tandem Repeats: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2): 573–80.

Birney, Ewan, Michele Clamp, and Richard Durbin. 2004. "GeneWise and Genomewise." *Genome Research* 14 (5): 988–95. doi:10.1101/gr.1865504.

Blanco, Enrique, Genís Parra, and Roderic Guigó. 2007. "Using Geneid to Identify Genes." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 4 (1): Unit 4.3. doi:10.1002/0471250953.bi0403s18.

Brachi, Benjamin, Nathalie Faure, Matt Horton, Emilie Flahauw, Adeline Vazquez, Magnus Nordborg, Joy Bergelson, Joel Cuguen, and Fabrice Roux. 2010. "Linkage and Association Mapping of Arabidopsis Thaliana Flowering Time in Nature." *PLoS Genetics* 6 (5): 40. doi:10.1371/journal.pgen.1000940.

Branca, a., T. D. Paape, P. Zhou, R. Briskine, a. D. Farmer, J. Mudge, a. K. Bharti, et al. 2011. "Whole-Genome Nucleotide Diversity, Recombination, and Linkage Disequilibrium in the Model Legume Medicago Truncatula." *Proceedings of the National Academy of Sciences* 108 (42): E864–70. doi:10.1073/pnas.1104032108.

Brendel, Volker, Liqun Xing, and Wei Zhu. 2004. “Gene Structure Prediction from Consensus Spliced Alignment of Multiple ESTs Matching the Same Genomic Locus.” *Bioinformatics* 20 (7): 1157–69. doi:10.1093/bioinformatics/bth058.

Broekaert, W. F., B. P. a. Cammue, M. F. C. De Bolle, K. Thevissen, G. W. De Samblanx, and R. W. Osborn. 1997. “Antimicrobial Peptides from Plants.” *Critical Reviews in Plant Sciences* 16 (3): 297–323. doi:10.1080/713608148.

Burset, M, and R Guigó. 1996. “Evaluation of Gene Structure Prediction Programs.” *Genomics* 34 (3): 353–67. doi:10.1006/geno.1996.0298.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10: 421. doi:10.1186/1471-2105-10-421.

Cannon, Steven B., Hongyan Zhu, Andrew M. Baumgarten, Russell Spangler, Georgiana May, Douglas R. Cook, and Nevin D. Young. 2002. “Diversity, Distribution, and Ancient Taxonomic Relationships within the TIR and Non-TIR NBS-LRR Resistance Gene Subfamilies.” *Journal of Molecular Evolution* 54 (4): 548–62. doi:10.1007/s00239-001-0057-2.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, et al. 2011. “Whole-Genome Sequencing of Multiple Arabidopsis Thaliana Populations.” *Nature Genetics* 43 (10): 956–63. doi:10.1038/ng.911.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73. doi:10.1093/bioinformatics/btp348.

Chen, Ken, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, et al. 2009. “BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation.” *Nature Methods* 6 (9). Nature Publishing Group: 677–81. doi:10.1038/nmeth.1363.

Chiang, Derek Y, Gad Getz, David B Jaffe, Michael J T O’Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. 2009.

"High-Resolution Mapping of Copy-Number Alterations with Massively Parallel Sequencing." *Nature Methods* 6 (1). Nature Publishing Group: 99–103. doi:10.1038/nmeth.1276.

Clark, Richard M, Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, et al. 2007. "Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis Thaliana." *Science (New York, N.Y.)* 317 (5836): 338–42. doi:10.1126/science.1138632.

Conrad, Donald F, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, et al. 2010. "Origins and Functional Impact of Copy Number Variation in the Human Genome." *Nature* 464 (7289). Nature Publishing Group: 704–12. doi:10.1038/nature08516.

Cook, David E, Tong G Lee, Xiaoli Guo, Sara Melito, K. Wang, Adam M Bayless, Jianping Wang, et al. 2012. "Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean." *Science* 338 (6111): 1206–9. doi:10.1126/science.1228746.

DePristo, Mark a, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98. doi:10.1038/ng.806.

Díaz, Aurora, Meluleki Zikhali, Adrian S. Turner, Peter Isaac, and David a. Laurie. 2012. "Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum Aestivum)." *PLoS ONE* 7 (3). doi:10.1371/journal.pone.0033234.

Doležel, J., M. Doleželová, and F. J. Novák. 1994. "Flow Cytometric Estimation of Nuclear DNA Amount in Diploid Bananas (Musa Acuminata and M. Balbisiana)." *Biologia Plantarum* 36 (3): 351–57. doi:10.1007/BF02920930.

Doležel, Jaroslav, and Jan Bartoš. 2005. "Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size." In *Annals of Botany*, 95:99–110. doi:10.1093/aob/mci005.

Eddy, S R. 1998. “Profile Hidden Markov Models.” *Bioinformatics (Oxford, England)* 14 (9). Oxford Univ Press: 755–63. doi:10.1093/bioinformatics/14.9.755.

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195. doi:10.1371/journal.pcbi.1002195.

Ellis, Jeff, Peter Dodds, and Tony Pryor. 2000. “Structure, Function and Evolution of Plant Disease Resistance Genes.” *Current Opinion in Plant Biology*. doi:10.1016/S1369-5266(00)00080-7.

Emerson, J J, Margarida Cardoso-Moreira, Justin O Borevitz, and Manyuan Long. 2008. “Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in Drosophila Melanogaster.” *Science (New York, N.Y.)* 320 (5883): 1629–31. doi:10.1126/science.1158078.

Enright, a J, S V Dongen, and C a Ouzounis. 2002. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84.

Farkas, Attila, G. Maroti, H. Durg, Z. Gyorgypal, Rui M Lima, Katalin F Medzihradszky, Attila Kereszt, Peter Mergaert, and E. Kondorosi. 2014. “Medicago Truncatula Symbiotic Peptide NCR247 Contributes to Bacteroid Differentiation through Multiple Mechanisms.” *Proceedings of the National Academy of Sciences* 111 (14). highwire: 5183–88. doi:10.1073/pnas.1404169111.

Fedorova, Maria, Judith van de Mortel, Peter a Matsumoto, Jennifer Cho, Christopher D Town, Kathryn a VandenBosch, J Stephen Gantt, and Carroll P Vance. 2002. “Genome-Wide Identification of Nodule-Specific Transcripts in the Model Legume Medicago Truncatula.” *Plant Physiology* 130 (2). Am Soc Plant Biol: 519–37. doi:10.1104/pp.006833.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. “Pfam: The Protein Families Database.” *Nucleic Acids Research* 42 (D1): 30. doi:10.1093/nar/gkt1223.

Foote, H C, J P Ride, V E Franklin-Tong, E a Walker, M J Lawrence, and F C Franklin. 1994. “Cloning and Expression of a Distinctive Class of Self-Incompatibility (S)

Gene from Papaver Rhoeas L." *Proceedings of the National Academy of Sciences of the United States of America* 91 (6): 2265–69. doi:10.1073/pnas.91.6.2265.

Galagan, James E., Matthew R. Henn, Li Jun Ma, Christina a. Cuomo, and Bruce Birren. 2005. "Genomics of the Fungal Kingdom: Insights into Eukaryotic Biology." *Genome Research* 15 (12): 1620–31. doi:10.1101/gr.3767105.

Galbraith, D W, K R Harkins, J M Maddox, N M Ayres, D P Sharma, and E Firoozabady. 1983. "Rapid Flow Cytometric Analysis of the Cell Cycle in Intact Plant Tissues." *Science (New York, N.Y.)* 220 (4601): 1049–51. doi:10.1126/science.220.4601.1049.

Gan, Xiangchao, Oliver Stegle, Jonas Behr, Joshua G Steffen, Philipp Drewe, Katie L Hildebrand, Rune Lyngsoe, et al. 2011. "Multiple Reference Genomes and Transcriptomes for Arabidopsis Thaliana." *Nature* 477 (7365): 419–23. doi:10.1038/nature10414.

García-Olmedo, Francisco, Antonio Molina, Josefa M Alamillo, and Pablo Rodríguez-Palenzuéla. 1998. "Plant Defense Peptides." *Biopolymers* 47 (6): 479–91. doi:10.1002/(SICI)1097-0282(1998)47:6<479::AID-BIP6>3.0.CO;2-K.

Gnerre, Sante, Iain Maccallum, Dariusz Przybylski, Filipe J Ribeiro, Joshua N Burton, Bruce J Walker, Ted Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences of the United States of America* 108 (4): 1513–18. doi:10.1073/pnas.1017351108.

Goldman, M H, R B Goldberg, and C Mariani. 1994. "Female Sterile Tobacco Plants Are Produced by Stigma-Specific Cell Ablation." *The EMBO Journal* 13 (13): 2976–84.

Gore, Michael a, Jer-Ming Chia, Robert J Elshire, Qi Sun, Elhan S Ersoz, Bonnie L Hurwitz, Jason a Peiffer, et al. 2009. "A First-Generation Haplotype Map of Maize." *Science (New York, N.Y.)* 326 (5956): 1115–17. doi:10.1126/science.1177837.

Graham, Michelle a, Kevin a T Silverstein, Steven B Cannon, and Kathryn a VandenBosch. 2004. "Computational Identification and Characterization of Novel Genes from Legumes." *Plant Physiology* 135 (3). Am Soc Plant Biol: 1179–97. doi:10.1104/pp.104.037531.

Graham, Peter H, and Carroll P Vance. 2003. “Legumes: Importance and Constraints to Greater Use.” *Plant Physiology* 131 (3): 872–77. doi:10.1104/pp.017004.

Haas, Brian J, Sophien Kamoun, Michael C Zody, Rays H Y Jiang, Robert E Handsaker, Liliana M Cano, Manfred Grabherr, et al. 2009. “Genome Sequence and Analysis of the Irish Potato Famine Pathogen Phytophthora Infestans.” *Nature* 461 (7262): 393–98. doi:10.1038/nature08358.

Haas, Brian J, Jennifer R Wortman, Catherine M Ronning, Linda I Hannick, Roger K Smith, Rama Maiti, Agnes P Chan, et al. 2005. “Complete Reannotation of the Arabidopsis Genome: Methods, Tools, Protocols and the Final Release.” *BMC Biology* 3 (January): 7. doi:10.1186/1741-7007-3-7.

Hallen, Heather E, Hong Luo, John S Scott-Craig, and Jonathan D Walton. 2007. “Gene Family Encoding the Major Toxins of Lethal Amanita Mushrooms.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (48): 19097–101. doi:10.1073/pnas.0707340104.

Hanada, Kousuke, Xu Zhang, Justin O Borevitz, Wen-hsiung Li, and Shin-han Shiu. 2007. “A Large Number of Novel Coding Small Open Reading Frames in the Intergenic Regions of the Arabidopsis Thaliana Genome Are Transcribed and / or under Purifying Selection A Large Number of Novel Coding Small Open Reading Frames in the Intergenic Regions of .” *Genome Research*, no. 517: 632–40. doi:10.1101/gr.5836207.

Henikoff, S, and J G Henikoff. 1992. “Amino Acid Substitution Matrices from Protein Blocks.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (22): 10915–19. doi:10.1073/pnas.89.22.10915.

Himly, Martin, Beatrice Jahn-Schmid, Azra Dedic, Peter Kelemen, Nicole Wopfner, Friedrich Altmann, Ronald van Ree, et al. 2003. “Art v 1, the Major Allergen of Mugwort Pollen, Is a Modular Glycoprotein with a Defensin-like and a Hydroxyproline-Rich Domain.” *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* 17 (1): 106–8. doi:10.1096/fj.02-0472fje.

Hormozdiari, Fereydoun, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. 2009. "Combinatorial Algorithms for Structural Variation Detection in High-Throughput Sequenced Genomes." *Genome Research* 19 (7): 1270–78. doi:10.1101/gr.088633.108.

Hulbert, Scot H, Craig A Webb, Shavannor M Smith, and Qing Sun. 2001. "Resistance Gene Complexes: Evolution and Utilization." *Annual Review of Phytopathology* 39 (1): 285–312. doi:10.1146/annurev.phyto.39.1.285.

Hunter, Sarah, Philip Jones, Alex Mitchell, Rolf Apweiler, Teresa K. Attwood, Alex Bateman, Thomas Bernard, et al. 2012. "InterPro in 2011: New Developments in the Family and Domain Prediction Database." *Nucleic Acids Research* 40 (D1): D306–12. doi:10.1093/nar/gkr948.

Jones, David A, and Daigo Takemoto. 2004. "Plant Innate Immunity - Direct and Indirect Recognition of General and Specific Pathogen-Associated Molecules." *Current Opinion in Immunology*. doi:10.1016/j.coi.2003.11.016.

Jones-Rhoades, Matthew W., Justin O. Borevitz, and Daphne Preuss. 2007. "Genome-Wide Expression Profiling of the Arabidopsis Female Gametophyte Identifies Families of Small, Secreted Proteins." *PLoS Genetics* 3 (10): 1848–61. doi:10.1371/journal.pgen.0030171.

Kamphuis, Lars G, Angela H Williams, Nola K D'Souza, Theo Pfaff, Simon R Ellwood, Emma J Groves, Karam B Singh, Richard P Oliver, and Judith Lichtenzveig. 2007. "The Medicago Truncatula Reference Accession A17 Has an Aberrant Chromosomal Configuration." *New Phytologist* 174 (2). wiley: 299–303. doi:10.1111/j.1469-8137.2007.02039.x.

Kang, Yang, Kil Kim, Sangrea Shim, Min Yoon, Suli Sun, Moon Kim, Kyujung Van, and Suk-Ha Lee. 2012. "Genome-Wide Mapping of NBS-LRR Genes and Their Association with Disease Resistance in Soybean." *BMC Plant Biology* 12 (1): 139. doi:10.1186/1471-2229-12-139.

Keller, Oliver, Martin Kollmar, Mario Stanke, and Stephan Waack. 2011. “A Novel Hybrid Gene Prediction Method Employing Protein Multiple Sequence Alignments.” *Bioinformatics* 27 (6): 757–63. doi:10.1093/bioinformatics/btr010.

Kent, W. James. 2002. “BLAT - The BLAST-like Alignment Tool.” *Genome Research* 12 (4): 656–64. doi:10.1101/gr.229202. Article published online before March 2002.

Kidd, Jeffrey M, Gregory M Cooper, William F Donahue, Hillary S Hayden, Nick Sampas, Tina Graves, Nancy Hansen, et al. 2008. “Mapping and Sequencing of Structural Variation from Eight Human Genomes.” *Nature* 453 (7191): 56–64. doi:10.1038/nature06862.

Korbel, Jan O, Alexej Abyzov, Xinmeng Jasmine Mu, Nicholas Carriero, Philip Cayting, Zhengdong Zhang, Michael Snyder, and Mark B Gerstein. 2009. “PEMer: A Computational Framework with Simulation-Based Error Models for Inferring Genomic Structural Variants from Massive Paired-End Sequencing Data.” *Genome Biology* 10 (2): R23. doi:10.1186/gb-2009-10-2-r23.

Korbel, Jan O, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, et al. 2007. “Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome.” *Science (New York, N.Y.)* 318 (5849): 420–26. doi:10.1126/science.1149504.

Kuang, Hanhui, Sung-Sick Woo, Blake C Meyers, Eviatar Nevo, and Richard W Michelmore. 2004. “Multiple Genetic Processes Result in Heterogeneous Rates of Evolution within the Major Cluster Disease Resistance Genes in Lettuce.” *The Plant Cell* 16 (11). Am Soc Plant Biol: 2870–94. doi:10.1105/tpc.104.025502.

Kump, Kristen L, Peter J Bradbury, Randall J Wisser, Edward S Buckler, Araby R Belcher, Marco a Oropeza-Rosas, John C Zwonitzer, et al. 2011. “Genome-Wide Association Study of Quantitative Resistance to Southern Leaf Blight in the Maize Nested Association Mapping Population.” *Nature Genetics* 43 (2): 163–68. doi:10.1038/ng.747.

Lamesch, Philippe, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, et al. 2012. “The Arabidopsis

Information Resource (TAIR): Improved Gene Annotation and New Tools." *Nucleic Acids Research* 40 (D1): D1202–10. doi:10.1093/nar/gkr1090.

Lavin, Matt, Patrick S Herendeen, and Martin F Wojciechowski. 2005. "Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary." *Systematic Biology* 54 (4): 575–94. doi:10.1080/10635150590947131.

Lease, Kevin a, and John C Walker. 2006. "The Arabidopsis Unannotated Secreted Peptide Database, a Resource for Plant Peptidomics." *Plant Physiology* 142 (3): 831–38. doi:10.1104/pp.106.086041.

Lee, Seunghak, Fereydoun Hormozdiari, Can Alkan, and Michael Brudno. 2009. "MoDIL: Detecting Small Indels from Clone-End Sequencing with Mixtures of Distributions." *Nature Methods* 6 (7): 473–74. doi:10.1038/nmeth.f.256.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. 2011. "The Sequence Read Archive." *Nucleic Acids Research* 39 (SUPPL. 1): D19–21. doi:10.1093/nar/gkq1019.

Leister, Dario. 2004. "Tandem and Segmental Gene Duplication and Recombination in the Evolution of Plant Disease Resistance Gene." *Trends in Genetics : TIG* 20 (3): 116–22. doi:10.1016/j.tig.2004.01.007.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16). highwire: 2078–79. doi:10.1093/bioinformatics/btp352.

Li, L., Christian J. Stoeckert, and David S. Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13 (9): 2178–89. doi:10.1101/gr.1224503.

Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. 2005. "Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm." *Nucleic Acids Research* 33 (20): 6494–6506. doi:10.1093/nar/gki937.

Magwire, Michael M., Florian Bayer, Claire L. Webster, Chuan Cao, and Francis M. Jiggins. 2011. “Successive Increases in the Resistance of Drosophila to Viral Infection through a Transposon Insertion Followed by a Duplication.” *PLoS Genetics* 7 (10). doi:10.1371/journal.pgen.1002337.

Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. “TigrScan and GlimmerHMM: Two Open Source Ab Initio Eukaryotic Gene-Finders.” *Bioinformatics* 20 (16): 2878–79. doi:10.1093/bioinformatics/bth315.

Marone, Daniela, Maria Russo, Giovanni Laidò, Anna De Leonardis, and Anna Mastrangelo. 2013. “Plant Nucleotide Binding Site–Leucine-Rich Repeat (NBS-LRR) Genes: Active Guardians in Host Defense Responses.” *International Journal of Molecular Sciences* 14 (4): 7302–26. doi:10.3390/ijms14047302.

Massouras, Andreas, Sebastian M Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F Ayroles, Emmanouil T Dermitzakis, et al. 2012. “Genomic Variation and Its Impact on Gene Expression in Drosophila Melanogaster.” Edited by Brian Oliver. *PLoS Genetics* 8 (11). plos: e1003055. doi:10.1371/journal.pgen.1003055.

McDowell, John M., and Stacey A. Simon. 2006. “Recent Insights into R Gene Evolution.” *Molecular Plant Pathology*. doi:10.1111/j.1364-3703.2006.00342.x.

McHale, Leah, Xiaoping Tan, Patrice Koehl, and Richard W Michelmore. 2006. “Plant NBS-LRR Proteins: Adaptable Guards.” *Genome Biology* 7 (4): 212. doi:10.1186/gb-2006-7-4-212.

McNally, Kenneth L, Kevin L Childs, Regina Bohnert, Rebecca M Davidson, Keyan Zhao, Victor J Ulat, Georg Zeller, et al. 2009. “Genomewide SNP Variation Reveals Relationships among Landraces and Modern Varieties of Rice.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (30): 12273–78. doi:10.1073/pnas.0900992106.

McVean, Gil A., David M. Altshuler (Co-Chair), Richard M. Durbin (Co-Chair), Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, et

al. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422): 56–65. doi:10.1038/nature11632.

Medvedev, Paul, Monica Stanciu, and Michael Brudno. 2009. “Computational Methods for Discovering Structural Variation with next-Generation Sequencing.” *Nature Methods* 6 (11 Suppl). Nature Publishing Group: S13–20. doi:10.1038/nmeth.1374.

Melo, Francislete R, Daniel J Rigden, Oct�vio L Franco, Luciane V Mello, Maria B Ary, Maria F. Grossi de S�, and Carlos Bloch. 2002. “Inhibition of Trypsin by Cowpea Thionin: Characterization, Molecular Modeling, and Docking.” *Proteins: Structure, Function, and Genetics* 48 (2). wiley: 311–19. doi:10.1002/prot.10142.

Mergaert, Peter, Krisztina Nikovics, Zsolt Kelemen, Nicolas Maunoury, Danièle Vaubert, Adam Kondorosi, and Eva Kondorosi. 2003. “A Novel Family in Medicago Truncatula Consisting of More than 300 Nodule-Specific Genes Coding for Small, Secreted Polypeptides with Conserved Cysteine Motifs.” *Plant Physiology* 132 (1): 161–73. doi:10.1104/pp.102.018192.

Mergaert, Peter, Toshiki Uchiumi, Benoît Alunni, Gwénaëlle Evanno, Angélique Cheron, Olivier Catrice, A.-E. Mausset, et al. 2006. “Eukaryotic Control on Bacterial Cell Cycle and Differentiation in the Rhizobium-Legume Symbiosis.” *Proceedings of the National Academy of Sciences* 103 (13): 5230–35. doi:10.1073/pnas.0600912103.

Meyers, Blake C, Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W Michelmore. 2003. “Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis.” *The Plant Cell* 15 (4). Am Soc Plant Biol: 809–34. doi:10.1105/tpc.009308.

Meyers, Blake C., Allan W. Dickerman, Richard W. Michelmore, Subramoniam Sivaramakrishnan, Bruno W. Sobral, and Nevin D. Young. 1999. “Plant Disease Resistance Genes Encode Members of an Ancient and Diverse Protein Family within the Nucleotide-Binding Superfamily.” *Plant Journal* 20 (3): 317–32. doi:10.1046/j.1365-313X.1999.00606.x.

Monosi, B., R. J. Wisser, L. Pennill, and S. H. Hulbert. 2004. “Full-Genome Analysis of Resistance Gene Homologues in Rice.” *Theoretical and Applied Genetics* 109 (7): 1434–47. doi:10.1007/s00122-004-1758-x.

Motose, Hiroyasu, Munetaka Sugiyama, and Hiroo Fukuda. 2004. “A Proteoglycan Mediates Inductive Interaction during Plant Vascular Development.” *Nature* 429 (6994). nature: 873–78. doi:10.1038/nature02613.

Nallu, Sumitha, Kevin a T Silverstein, Deborah a. Samac, Bruna Bucciarelli, Carroll P. Vance, and Kathryn a. VandenBosch. 2013. “Regulatory Patterns of a Large Family of Defensin-Like Genes Expressed in Nodules of Medicago Truncatula.” Edited by Miguel A. Blazquez. *PLoS ONE* 8 (4): e60355. doi:10.1371/journal.pone.0060355.

Nallu, Sumitha, Kevin a T Silverstein, Peng Zhou, Nevin D. Young, and Kathryn a. Vandenbosch. 2014. “Patterns of Divergence of a Large Family of Nodule Cysteine-Rich Peptides in Accessions of Medicago Truncatula.” *Plant Journal* 78 (4): 697–705. doi:10.1111/tpj.12506.

Nielsen, Rasmus. 2005. “Molecular Signatures of Natural Selection.” *Annual Review of Genetics* 39 (January): 197–218. doi:10.1146/annurev.genet.39.073003.112420.

Okuda, Satohiro, Hiroki Tsutsui, Keiko Shiina, Stefanie Sprunck, Hidenori Takeuchi, Ryoko Yui, Ryushiro D Kasahara, et al. 2009. “Defensin-like Polypeptide LUREs Are Pollen Tube Attractants Secreted from Synergid Cells.” *Nature* 458 (7236). nature: 357–61. doi:10.1038/nature07882.

Oldroyd, Giles E D, and J Allan Downie. 2008. “Coordinating Nodule Morphogenesis with Rhizobial Infection in Legumes.” *Annual Review of Plant Biology* 59: 519–46. doi:10.1146/annurev.arplant.59.032607.092839.

Pan, Bohu, Jia Sheng, Weining Sun, Yinhong Zhao, Pei Hao, and Xuan Li. 2013. “OrysPSSP: A Comparative Platform for Small Secreted Proteins from Rice and Other Plants.” *Nucleic Acids Research* 41 (D1): 1–7. doi:10.1093/nar/gks1090.

Pan, Q, J Wendel, and R Fluhr. 2000. “Divergent Evolution of Plant NBS-LRR Resistance Gene Homologues in Dicot and Cereal Genomes.” *Journal of Molecular Evolution* 50 (3): 203–13. doi:10.1007/s002399910023.

Pearce, G, D S Moura, J Stratmann, and C A Ryan. 2001. “RALF, a 5-kDa Ubiquitous Polypeptide in Plants, Arrests Root Growth and Development.” *Proceedings of the National Academy of Sciences* 98 (22): 12843–47. doi:10.1073/pnas.201416998.

Pearce, S., R. Saville, S. P. Vaughan, P. M. Chandler, E. P. Wilhelm, C. a. Sparks, N. Al-Kaff, et al. 2011. “Molecular Characterization of Rht-1 Dwarfing Genes in Hexaploid Wheat.” *Plant Physiology* 157 (4): 1820–31. doi:10.1104/pp.111.183657.

Perry, George H, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, et al. 2007. “Diet and the Evolution of Human Amylase Gene Copy Number Variation.” *Nature Genetics* 39 (10): 1256–60. doi:10.1038/ng2123.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8 (10). Nature Publishing Group: 785–86. doi:10.1038/nmeth.1701.

Plocik, Alex, Jenn Layden, and Rick Kesseli. 2004. “Comparative Analysis of NBS Domain Sequences of NBS-LRR Disease Resistance Genes from Sunflower, Lettuce, and Chicory.” *Molecular Phylogenetics and Evolution* 31 (1): 153–63. doi:10.1016/S1055-7903(03)00274-4.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, et al. 2011. “The Pfam Protein Families Database.” *Nucleic Acids Research* 40 (D1): D290–301. doi:10.1093/nar/gkr1065.

Redon, Richard, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, et al. 2006. “Global Variation in Copy Number in the Human Genome.” *Nature* 444 (7118): 444–54. doi:10.1038/nature05329.

Rep, Martijn, and H. Corby Kistler. 2010. “The Genomic Organization of Plant Pathogenicity in Fusarium Species.” *Current Opinion in Plant Biology* 13 (4). Elsevier Ltd: 420–26. doi:10.1016/j.pbi.2010.04.004.

Richly, Erik, Joachim Kurth, and Dario Leister. 2002. “Mode of Amplification and Reorganization of Resistance Genes during Recent Arabidopsis Thaliana Evolution.” *Molecular Biology and Evolution* 19 (1): 76–84.

Ronfort, Joëlle, Thomas Bataillon, Sylvain Santoni, Magalie Delalande, Jacques L David, and Jean-Marie Prosperi. 2006. "Microsatellite Diversity and Broad Scale Geographic Structure in a Model Legume: Building a Set of Nested Core Collection for Studying Naturally Occurring Variation in Medicago Truncatula." *BMC Plant Biology* 6: 28. doi:10.1186/1471-2229-6-28.

Salamov, Asaf a., and Victor V. Solovyev. 2000. "Ab Initio Gene Finding in Drosophila Genomic DNA." *Genome Research* 10 (4): 516–22. doi:10.1101/gr.10.4.516.

Saxonov, S, I Daizadeh, a Fedorov, and W Gilbert. 2000. "EID: The Exon-Intron Database-an Exhaustive Database of Protein-Coding Intron-Containing Genes." *Nucleic Acids Research* 28 (1): 185–90. doi:10.1093/nar/28.1.185.

Schmidt, Joshua M., Robert T. Good, Belinda Appleton, Jayne Sherrard, Greta C. Raymant, Michael R. Bogwitz, Jon Martin, et al. 2010. "Copy Number Variation and Transposable Elements Feature in Recent, Ongoing Adaptation at the Cyp6g1 Locus." Edited by David J. Begun. *PLoS Genetics* 6 (6). Public Library of Science: 1–11. doi:10.1371/journal.pgen.1000998.

Schnable, Patrick S, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, et al. 2009. "The B73 Maize Genome: Complexity, Diversity, and Dynamics." *Science (New York, N.Y.)* 326 (5956): 1112–15. doi:10.1126/science.1178534.

Schopfer, C R, M E Nasrallah, and J B Nasrallah. 1999. "The Male Determinant of Self-Incompatibility in Brassica." *Science (New York, N.Y.)* 286 (5445): 1697–1700. doi:10.1126/science.286.5445.1697.

Shai, Yechiel. 2002. "Mode of Action of Membrane Active Antimicrobial Peptides." *Biopolymers* 66 (4): 236–48. doi:10.1002/bip.10260.

Shang, Junjun, Yong Tao, Xuewei Chen, Yan Zou, Cailin Lei, Jing Wang, Xiaobing Li, et al. 2009. "Identification of a New Rice Blast Resistance Gene, Pid3, by Genomewide Comparison of Paired Nucleotide-Binding Site-Leucine-Rich Repeat Genes and Their Pseudogene Alleles Between the Two Sequenced Rice Genomes." *Genetics* 182 (4): 1303–11. doi:10.1534/genetics.109.102871.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (539): 539. doi:10.1038/msb.2011.75.

Sigrist, Christian J a, Edouard De Castro, Lorenzo Cerutti, Béatrice a. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. 2013. "New and Continuing Developments at PROSITE." *Nucleic Acids Research* 41 (D1): D344–47. doi:10.1093/nar/gks1067.

Silverstein, Kevin a T, Michelle a Graham, Timothy D Paape, and Kathryn a VandenBosch. 2005. "Genome Organization of More than 300 Defensin-like Genes in Arabidopsis." *Plant Physiology* 138 (2). Am Soc Plant Biol: 600–610. doi:10.1104/pp.105.060079.

Silverstein, Kevin a T, Michelle a. Graham, and Kathryn a. VandenBosch. 2006. "Novel Paralogous Gene Families with Potential Function in Legume Nodules and Seeds." *Current Opinion in Plant Biology* 9 (2): 142–46. doi:10.1016/j.pbi.2006.01.002.

Silverstein, Kevin a T, William a. Moskal, Hank C. Wu, Beverly a. Underwood, Michelle a. Graham, Christopher D. Town, and Kathryn a. VandenBosch. 2007. "Small Cysteine-Rich Peptides Resembling Antimicrobial Peptides Have Been under-Predicted in Plants." *Plant Journal* 51 (2): 262–80. doi:10.1111/j.1365-313X.2007.03136.x.

Slater, Guy St C, and Ewan Birney. 2005. "Automated Generation of Heuristics for Biological Sequence Comparison." *BMC Bioinformatics* 6 (January): 31. doi:10.1186/1471-2105-6-31.

Smit, AFA, R Hubley, and P Green. 1996. "RepeatMasker Open-3.0." *RepeatMasker Open-3.0*.

Spanu, Pietro D, James C Abbott, Joelle Amselem, Timothy a Burgis, Darren M Soanes, K. Stuber, E. V. Loren van Themaat, et al. 2010. "Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism." *Science* 330 (6010): 1543–46. doi:10.1126/science.1194573.

Stanke, Mario, and Stephan Waack. 2003. “Gene Prediction with a Hidden Markov Model and a New Intron Submodel.” In *Bioinformatics*, 19:ii215–i225. doi:10.1093/bioinformatics/btg1080.

Stanton-Geddes, John, Timothy Paape, Brendan Epstein, Roman Briskine, Jeremy Yoder, Joann Mudge, Arvind K. Bharti, et al. 2013. “Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics in Medicago Truncatula.” Edited by Lewis Lukens. *PLoS ONE* 8 (5): e65688. doi:10.1371/journal.pone.0065688.

Stein, L. 2001. “Genome Annotation: From Sequence to Biology.” *Nature Reviews*. *Genetics* 2 (7): 493–503. doi:10.1038/35080529.

Swanson-Wagner, Ruth A., Steven R Eichten, Sunita Kumari, Peter Tiffin, Joshua C Stein, Doreen Ware, and Nathan M Springer. 2010. “Pervasive Gene Content Variation and Copy Number Variation in Maize and Its Undomesticated Progenitor.” *Genome Research* 20 (12): 1689–99. doi:10.1101/gr.109165.110.

Tadege, Million, Jiangqi Wen, Ji He, Haidi Tu, Younsig Kwak, Alexis Eschstruth, Anne Cayrel, et al. 2008. “Large-Scale Insertional Mutagenesis Using the Tnt1 Retrotransposon in the Model Legume Medicago Truncatula.” *Plant Journal* 54 (2): 335–47. doi:10.1111/j.1365-313X.2008.03418.x.

Takeuchi, Hidenori, and Tetsuya Higashiyama. 2012. “A Species-Specific Cluster of Defensin-Like Genes Encodes Diffusible Pollen Tube Attractants in Arabidopsis.” Edited by Ueli Grossniklaus. *PLoS Biology* 10 (12): e1001449. doi:10.1371/journal.pbio.1001449.

Tang, Haibao, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan, Shiguo Zhou, Laurent Gentzbittel, et al. 2014. “An Improved Genome Release (version Mt4.0) for the Model Legume Medicago Truncatula.” *BMC Genomics* 15 (1): 312. doi:10.1186/1471-2164-15-312.

Tesfaye, Mesfin, Kevin a T Silverstein, Sumitha Nallu, Lin Wang, Christopher J. Botanga, S. Karen Gomez, Liliana M. Costa, et al. 2013. “Spatio-Temporal

Expression Patterns of Arabidopsis Thaliana and Medicago Truncatula Defensin-Like Genes." *PLoS ONE* 8 (3): e58992. doi:10.1371/journal.pone.0058992.

Thevissen, Karin, Kathelijne K.A. Ferket, Isabelle E.J.A. François, and Bruno P.A. Cammue. 2003. "Interactions of Antifungal Plant Defensins with Fungal Membrane Components." *Peptides* 24 (11). sciencedirect: 1705–12. doi:10.1016/j.peptides.2003.09.014.

Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. 2013. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration." *Briefings in Bioinformatics* 14 (2): 178–92. doi:10.1093/bib/bbs017.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics* 25 (9): 1105–11. doi:10.1093/bioinformatics/btp120.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7 (3). Nature Publishing Group: 562–78. doi:10.1038/nprot.2012.016.

Van de Velde, Willem, Grigor Zehirov, Agnes Szatmari, Monika Debreczeny, Hironobu Ishihara, Zoltan Kevei, Attila Farkas, et al. 2010. "Plant Peptides Govern Terminal Differentiation of Bacteria in Symbiosis." *Science* 327 (5969): 1122–26. doi:10.1126/science.1184057.

Wang, Dong, Joel Griffitts, Colby Starker, Elena Fedorova, Erik Limpens, Sergey Ivanov, Ton Bisseling, and Sharon Long. 2010. "A Nodule-Specific Protein Secretory Pathway Required for Nitrogen-Fixing Symbiosis." *Science (New York, N.Y.)* 327 (5969): 1126–29. doi:10.1126/science.1184096.

Weischenfeldt, Joachim, Orsolya Symmons, François Spitz, and Jan O Korbel. 2013. "Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease." *Nature Reviews Genetics* 14 (2): 125–38. doi:10.1038/nrg3373.

Wu, Thomas D., and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics* 26 (7): 873–81. doi:10.1093/bioinformatics/btq057.

Yang, Xiaohan, Timothy J. Tschaplinski, Gregory B. Hurst, Sara Jawdy, Paul E. Abraham, Patricia K. Lankford, Rachel M. Adams, et al. 2011. “Discovery and Annotation of Small Proteins Using Genomics, Proteomics, and Computational Approaches.” *Genome Research* 21 (4): 634–41. doi:10.1101/gr.109280.110.

Yao, Hong, Ling Guo, Yan Fu, Lisa a. Borsuk, Tsui Jung Wen, David S. Skibbe, Xiangqin Cui, et al. 2005. “Evaluation of Five Ab Initio Gene Prediction Programs for the Discovery of Maize Genes.” *Plant Molecular Biology* 57 (3): 445–60. doi:10.1007/s11103-005-0271-1.

Yoon, Seungtai, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. 2009. “Sensitive and Accurate Detection of Copy Number Variants Using Read Depth of Coverage.” *Genome Research* 19 (9): 1586–92. doi:10.1101/gr.092981.109.

Young, Nevin D., Frédéric Debellé, Giles E. D. Oldroyd, Rene Geurts, Steven B. Cannon, Michael K. Udvardi, Vagner a. Benedito, et al. 2011. “The Medicago Genome Provides Insight into the Evolution of Rhizobial Symbioses.” *Nature*. doi:10.1038/nature10625.

Young, Nevin Dale, and Michael Udvardi. 2009. “Translating Medicago Truncatula Genomics to Crop Legumes.” *Current Opinion in Plant Biology* 12 (2): 193–201. doi:10.1016/j.pbi.2008.11.005.

Zhou, Peng, Kevin At Silverstein, Liangliang Gao, Jonathan D Walton, Sumitha Nallu, Joseph Guhlin, and Nevin D Young. 2013. “Detecting Small Plant Peptides Using SPADA (Small Peptide Alignment Discovery Application).” *BMC Bioinformatics* 14 (1): 335. doi:10.1186/1471-2105-14-335.

Appendices

Appendix Table 1.1. CRPs predicted by SPADA in *A. thaliana* using E-value threshold of 0.001.

Appendix Table 1.2. CRPs predicted by SPADA in *M. truncatula* using E-value threshold of 0.001.

Appendix Table 1.3. Expression support of the Arabidopsis CRP test set.

Appendix Table 1.4. Expression support of the *Medicago* CRP test set.

Appendix Table 1.5. Expression support of the CRPs predicted by SPADA in *A. thaliana*.

Appendix Table 1.6. Expression support of the CRPs predicted by SPADA in *M. truncatula*.

Appendix Table 1.7. CRPs predicted by SPADA in *A. thaliana* using E-value threshold of 1.

Appendix Table 1.8. CRPs predicted by SPADA in *M. truncatula* using E-value threshold of 1.

Appendix Table 1.9. SPH peptides predicted by SPADA in *A. thaliana*.

Appendix Table 1.10. Evaluation of SPADA, AUSPD and OrysPSSP using the manually-curated test set.

	A. thaliana		*M. truncatula*	
	AUSPD	SPADA	OrysPSSP	SPADA
Identical[*]	88	618	83	322
Minor conflict[#]	8	49	11	72
Major conflict[$]	43	38	125	71
Missed[^]	610	28	342	61

[*]Number of gene models identical with the test set;

[#]Number of gene models overlapping with the test set and in the same reading frame;

[$]Number of gene models overlapping with the test set but in a different reading frame;

[^]Number of gene models in the test set that were missed by AUSPD/OrysPSSP/SPADA.

Appendix Table 2.1. Sequencing statistics of 15 M. truncatula accessions.

	SIPE (Gb)	LIPE (Gb)	SIPE detail	LIPE detail	Coverage (fold)*
HM004	25.0	39.2	~50X KAPA	~45X Original LIPE, 34X Nextera	128.5
HM010	20.4	37.7	~41X KAPA	~31X Original LIPE, ~44X Nextera	116.3
HM022	24.5	27.9	~49X KAPA	~50X Nextera	104.9
HM023	24.2	41.4	~48X KAPA	~53X Nextera	131.3
HM034	27.1	32.6	~54X KAPA	~65X Nextera	119.5
HM050	20.4	34.2	~41X KAPA	~27X original LIPE, ~41X Nextera	109.2
HM056	78.6	56.8	~109X TruSeq v3, ~49X KAPA	~51X Nextera; ~33X original Illumina	270.8
HM058	32.2	35.3	~64X KAPA	~45X Nextera	135.0
HM060	24.3	24.1	~49X KAPA	~47X Nextera	96.9
HM095	26.6	29.7	~53X KAPA	~58X Nextera	112.5
HM125	21.8	24.0	~44X KAPA	~48X Nextera	91.6
HM129	24.8	22.2	~50X KAPA	~44X Nextera	93.9
HM185	21.2	28.6	~42X KAPA	~57X Nextera	99.5
HM324	25.0	41.3	~50X KAPA	~52X Nextera	132.6
HM340	33.2	45.3	~66X TruSeq v3	~33X original LIPE, ~58X Nextera LIPE	157.1

*Coverage was estimated using an assumed genome size of 500 million bases.

Appendix Table 2.2. Member counts of different gene families annotated in 15 *de novo* assemblies.

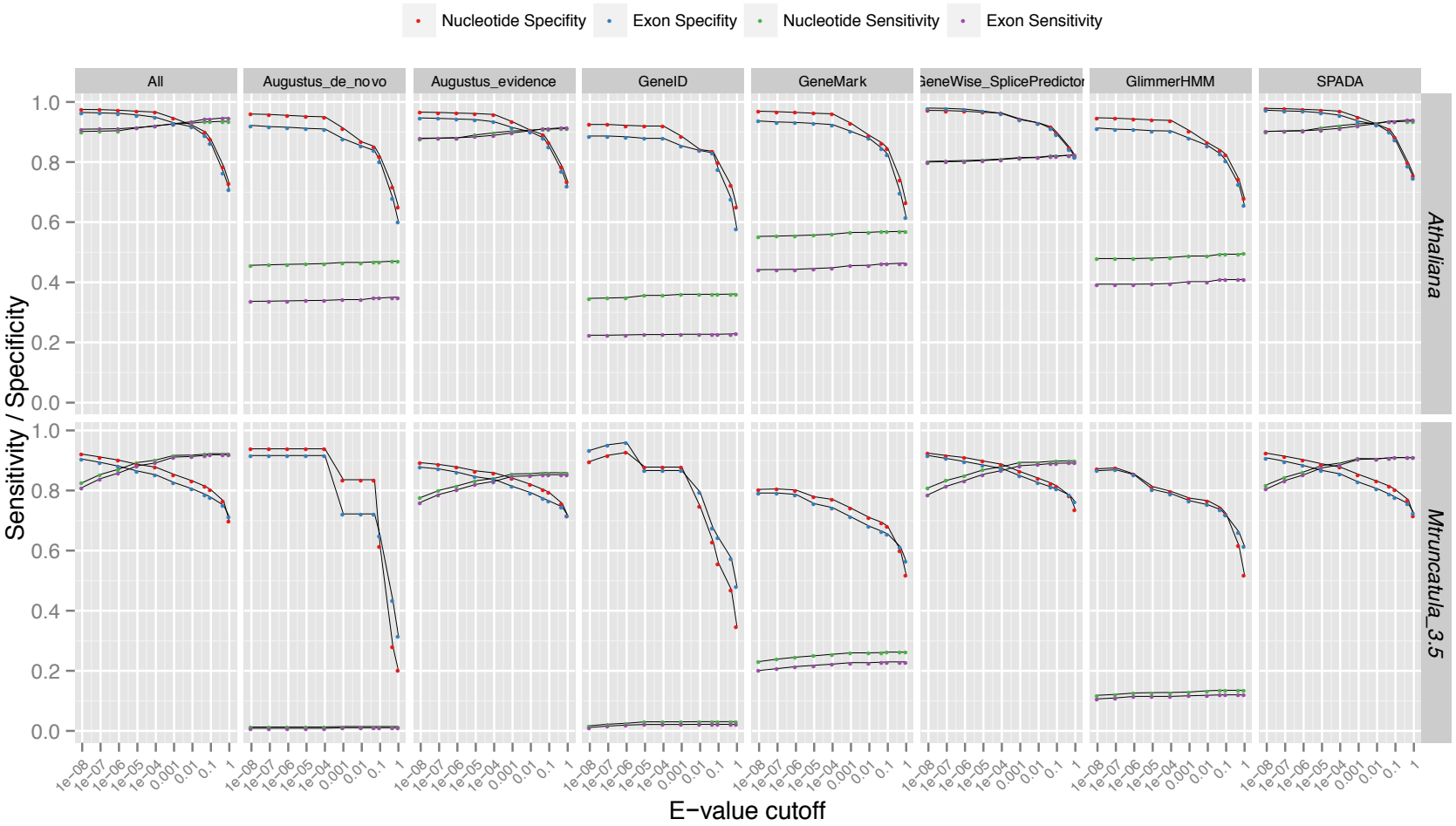
Appendix Table 2.3. Member counts of CRP subfamilies identified by SPADA in 15 *de novo* assemblies..

Appendix Table 2.4. Member counts of NBS-LRR subfamilies in 15 *de novo* assemblies.

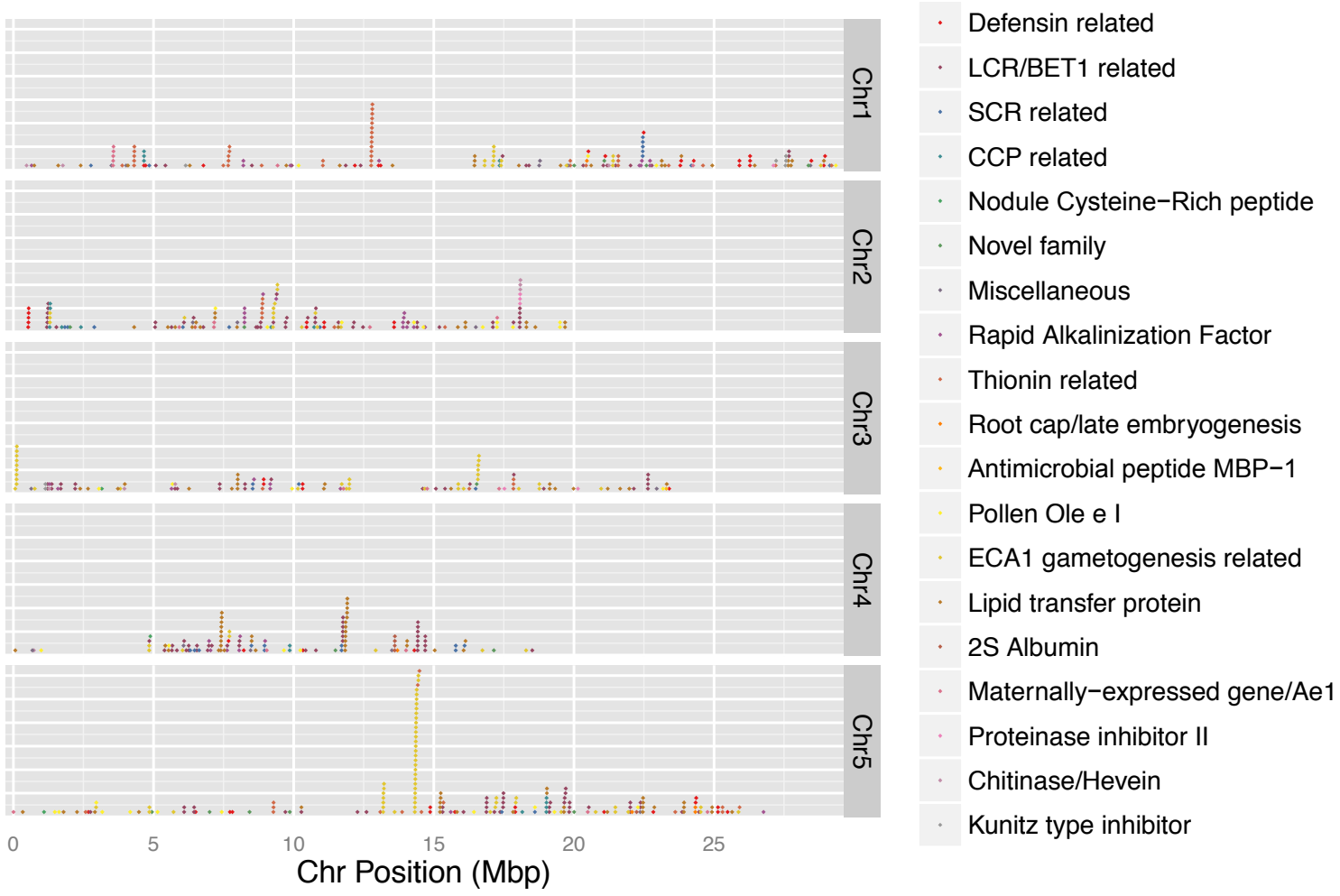
Appendix Table 2.5. Member counts of different TE subfamilies in 15 *de novo* assemblies.

Appendix Table 2.6. Correlation of nucleotide diversity estimates (SNPs, short Indels and large SVs) with non-TE genes, TEs, NBS-LRRs and CRPs.

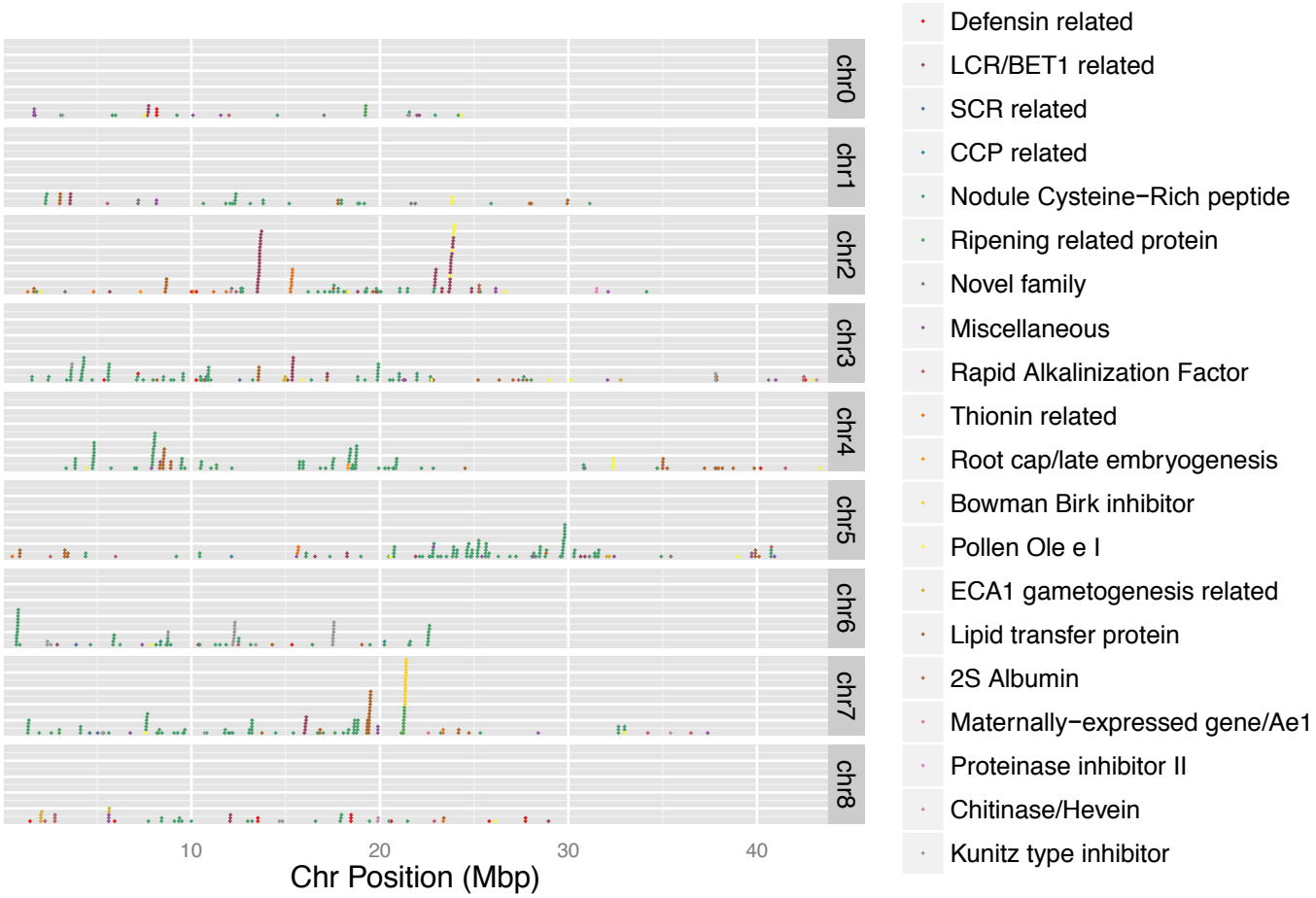
	pi_snp	**pi_indel**	**pi_sv**
Non-TE genes	r = -0.229, p = 6.58e-47	r = -0.212, p = 1.96e-40	r = -0.155, p = 3.27e-22
TEs	r = 0.328, p = 0	r = 0.330, p = 0	r = 0.319, p = 0
NBS-LRR	r = 0.282, p = 0	r = 0.193, p = 0	r = 0.254, p = 0
CRP	r = 0.089, p = 3.02e-08	r = 0.044, p = 0.00653	r = 0.037, p = 0.0223



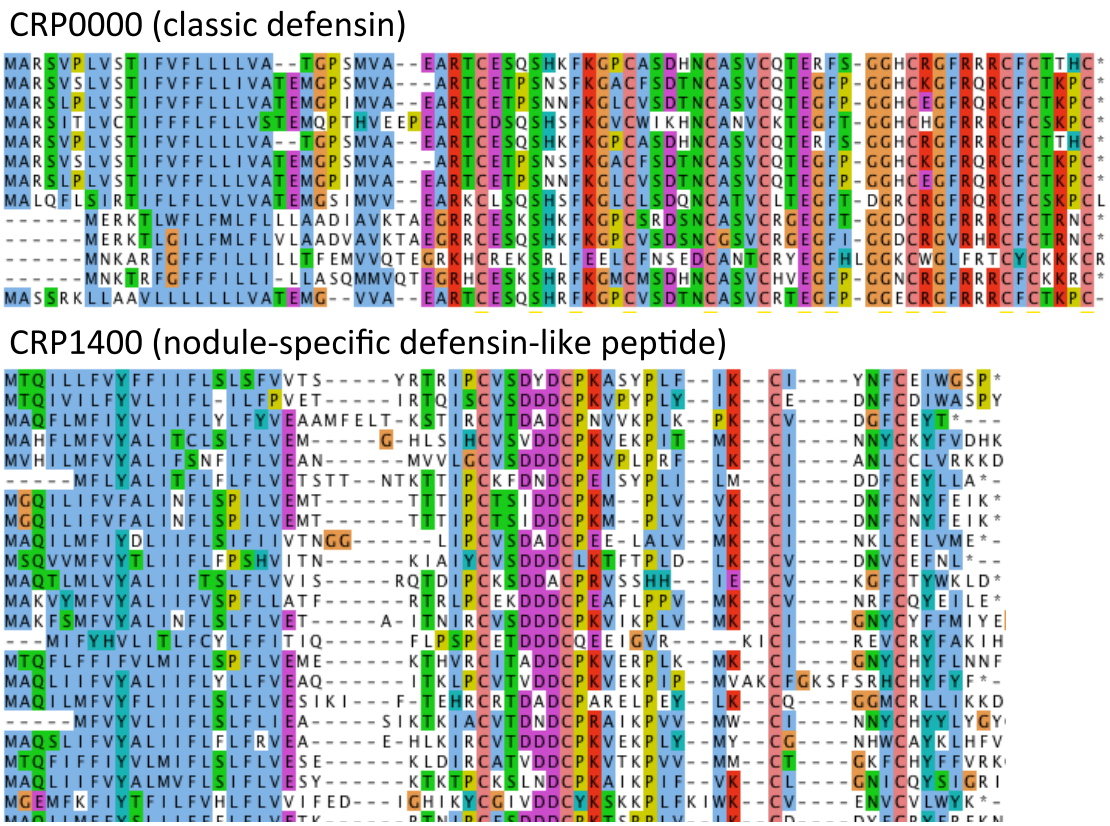
Appendix Figure 1.1. Performance comparison of five gene prediction components under different search E-value threshold



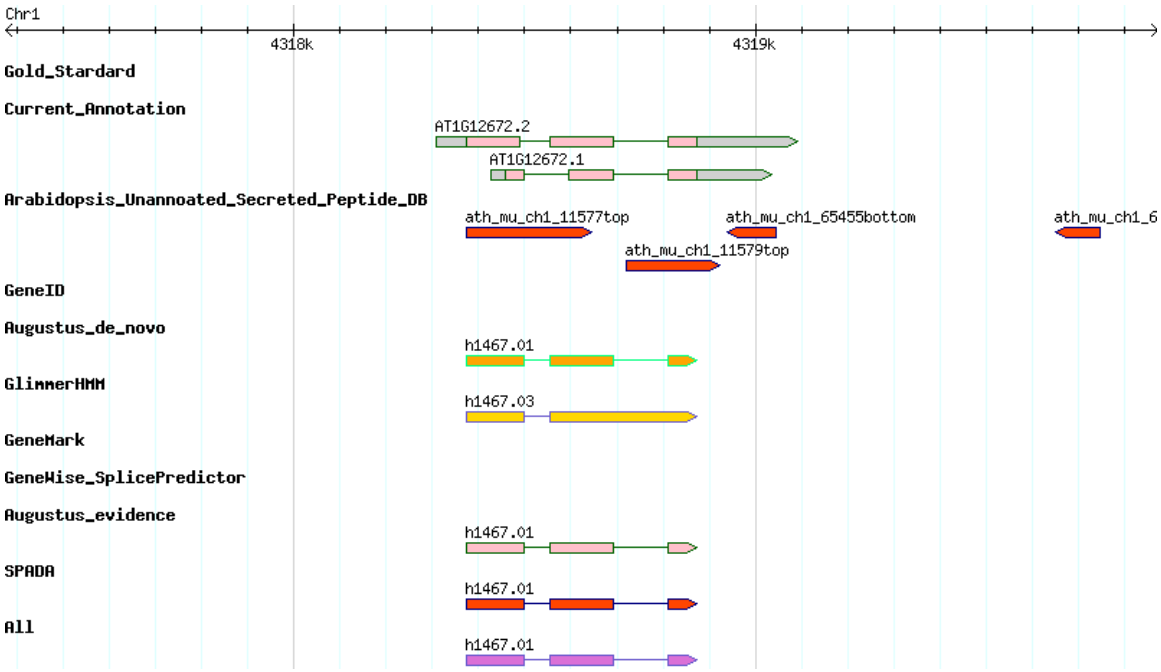
Appendix Figure 1.2. Genome distribution of CRPs predicted in *Arabidopsis thaliana*.



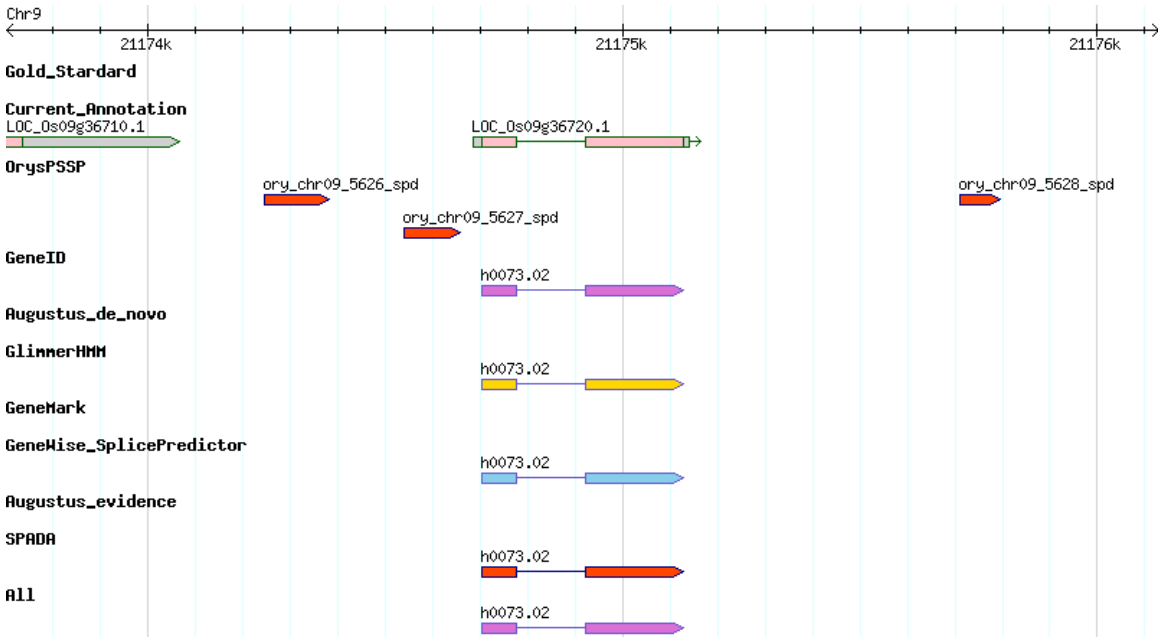
Appendix Figure 1.3. Genome distribution of CRPs predicted in *Medicago truncatula*.



Appendix Figure 1.4. Multiple sequence alignments of *Medicago* CRP sub-families CRP0000 and CRP1400.

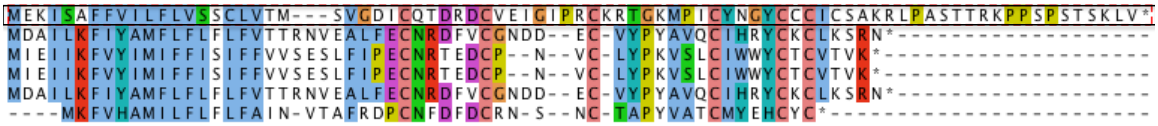


Appendix Figure 1.5. A typical Arabidopsis CRP mis-annonated in Arabidopsis Unannotated Secreted Peptide Database (AUSPD).



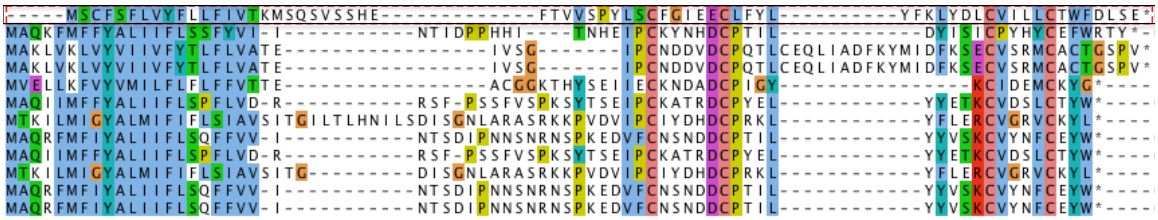
Appendix Figure 1.6. A typical rice CRP mis-annotated in OrysPSSP.

CRP1280



CRP1300

CRP1510



Appendix Figure 1.7. Sub-class alignments of three Arabidopsis NCRs with *Medicago* NCRs.

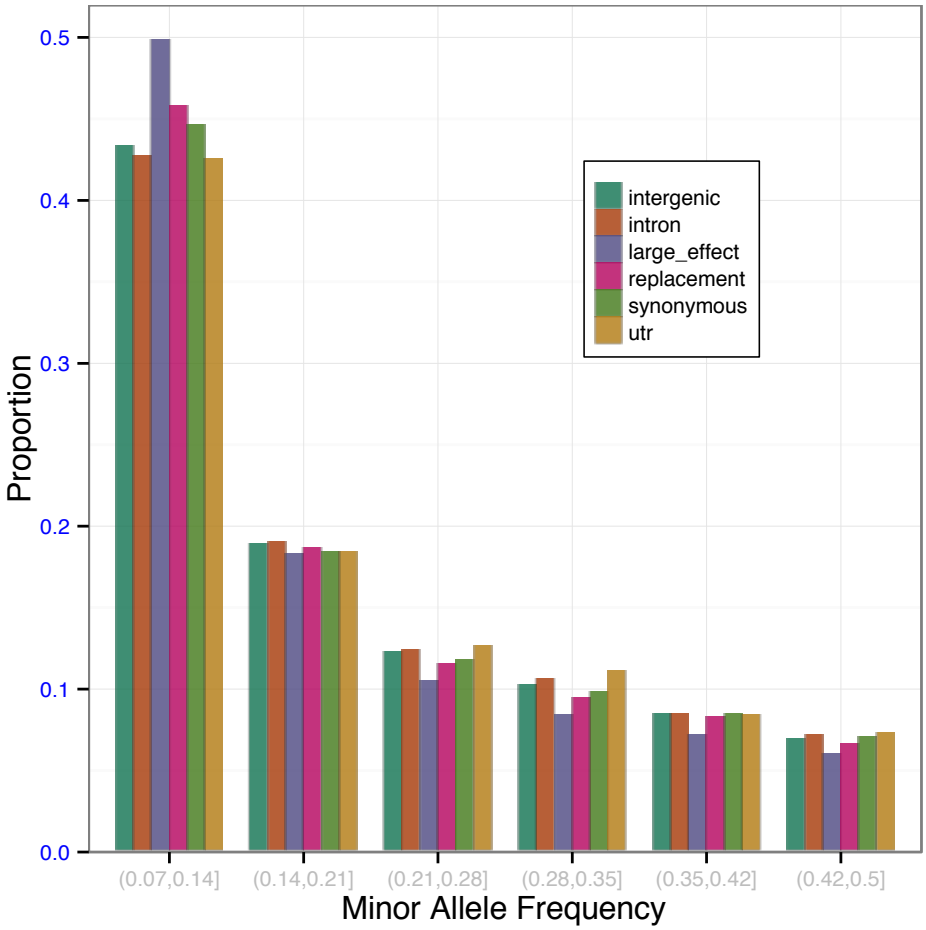
In each alignment the first sequence comes from Arabidopsis and the rest all come from *Medicago*.

Appendix File 1.1. Manually curated CRPs (test set) in *A. thaliana* and *M. truncatula* (in GFF3 format).

Appendix File 1.2. CRP predictions made by SPADA in *A. thaliana* and *M. truncatula* using search E-value threshold of 0.001 (in GFF3 format).

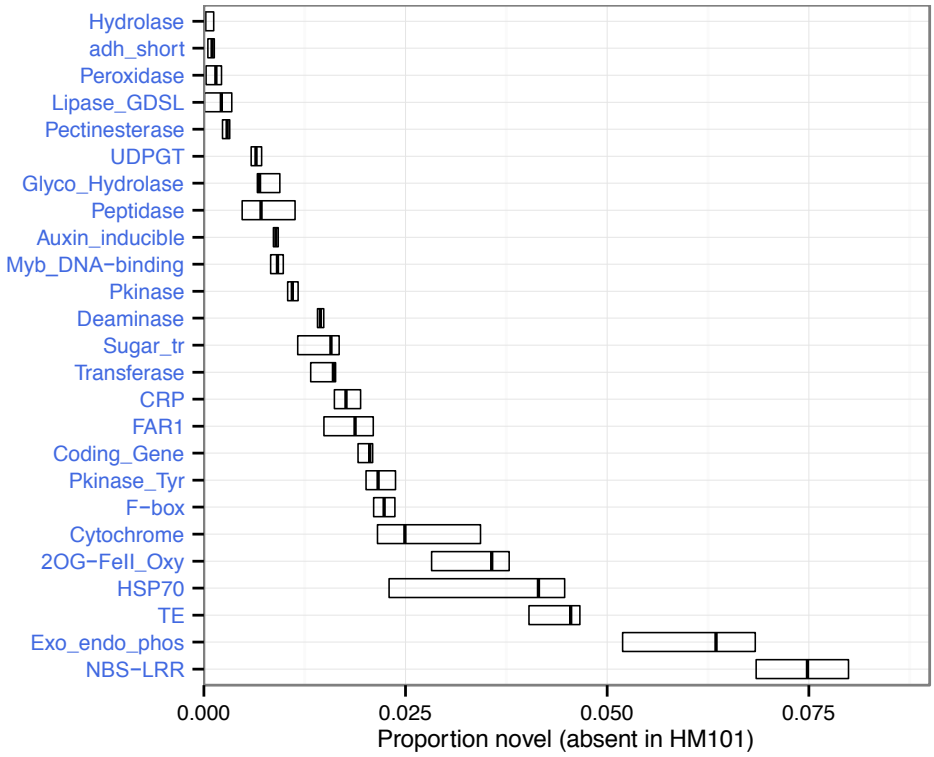
Appendix File 1.3. Novel CRP predictions made by SPADA in *A. thaliana* and *M. truncatula* as determined by manual inspection (in GFF3 format).

Appendix File 1.4. SPH predictions made by SPADA in Arabidopsis using search E-value threshold of 0.001 (in GFF3 format).



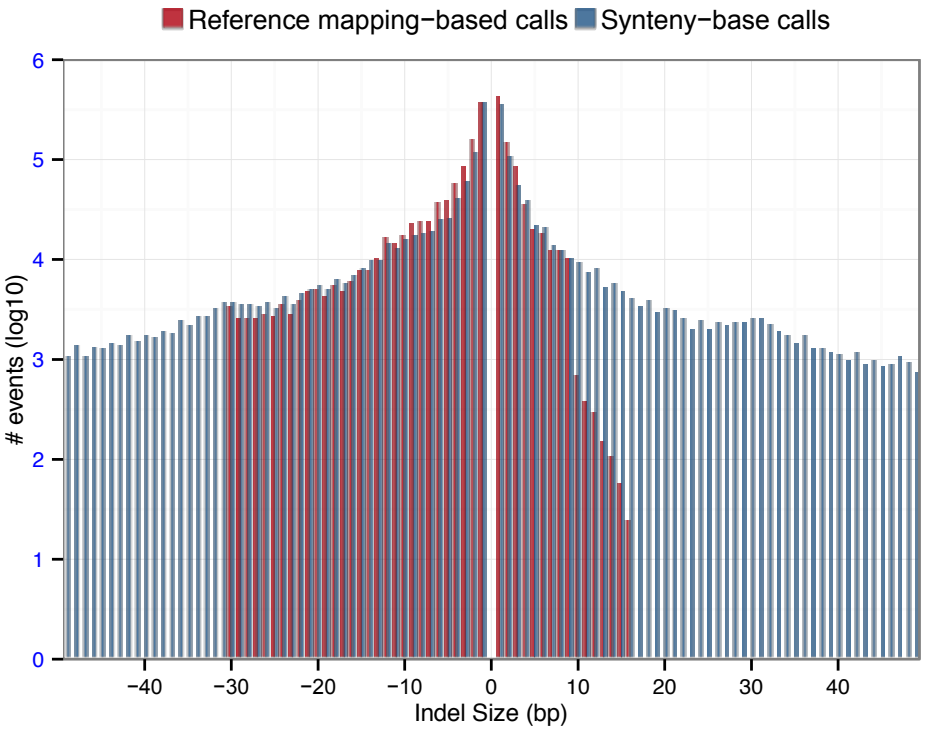
Appendix Figure 2.1. Minor allele frequency (MAF) spectrum of SNPs in different categories.

Large effect SNPs include lost of start or stop codon, gain of premature stop codon, as well as splice donor or acceptor variant.

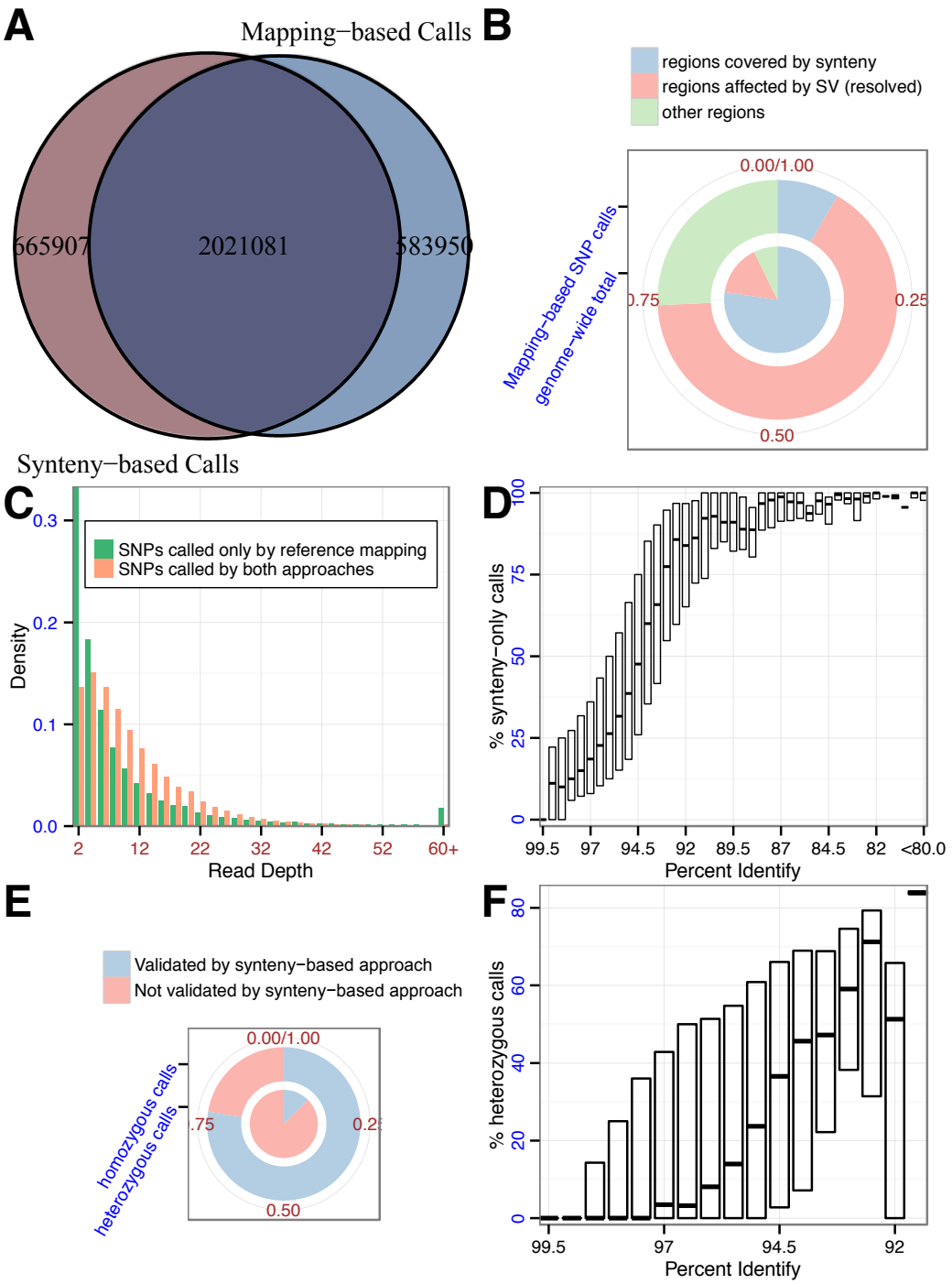


Appendix Figure 2.2. Proportion sequences identified as novel (absent in HM101) in different gene families.

Bars indicate 25%, 50% (median) and 75% quantiles of proportion novel sequence in 12 ingroup accessions.



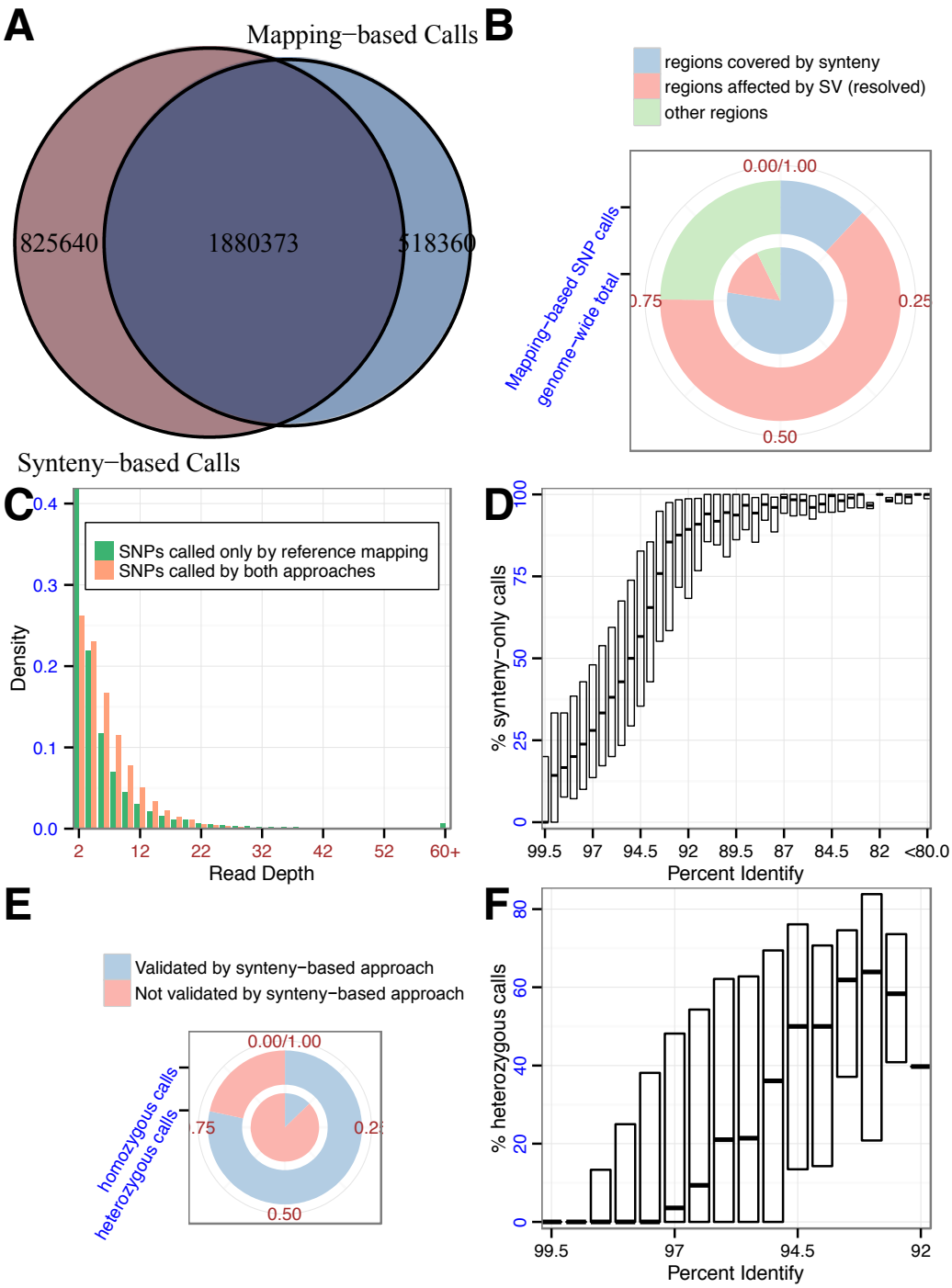
Appendix Figure 2.3. Size distribution of short InDels (less than 50-bp) called by the reference mapping-based approach and synteny-based approach.



Appendix Figure 2.4A. Comparison and characterization of SNP calling in HM004 from two different approaches.

(A) Venn-diagram showing overlap of mapping-based SNP call set and assembly-based call set; (B) Distribution of mapping-only SNPs in different genomic classes (outer ring) and genome-wide distribution of different genomic classes (inner piechart); (C)

Distribution of read-depth support for mapping-only SNP calls and overlapping SNP calls; (D) Proportion of assembly-only SNP calls binned by different sequence identity classes; (E) Proportion of heterozygous and homozygous SNP calls (mapping-based) validated by assembly-based approach; (F) Proportion of heterozygous SNP calls (out of all mapping-based calls) binned by different sequence identity classes.



Appendix Figure 2.4B. Comparison and characterization of SNP calling in HM023 from two different approaches.

(A) Venn-diagram showing overlap of mapping-based SNP call set and assembly-based call set; (B) Distribution of mapping-only SNPs in different genomic classes (outer ring) and genome-wide distribution of different genomic classes (inner piechart); (C)

Distribution of read-depth support for mapping-only SNP calls and overlapping SNP calls; (D) Proportion of assembly-only SNP calls binned by different sequence identity classes; (E) Proportion of heterozygous and homozygous SNP calls (mapping-based) validated by assembly-based approach; (F) Proportion of heterozygous SNP calls (out of all mapping-based calls) binned by different sequence identity classes.

Figure Link: https://www.dropbox.com/s/oa5glsq4fr3zwjb/figS2.5.pdf?dl=0

Appendix Figure 2.5. Synteny alignment confirms rearrangement of the long arms of chromosomes 4 and 8.

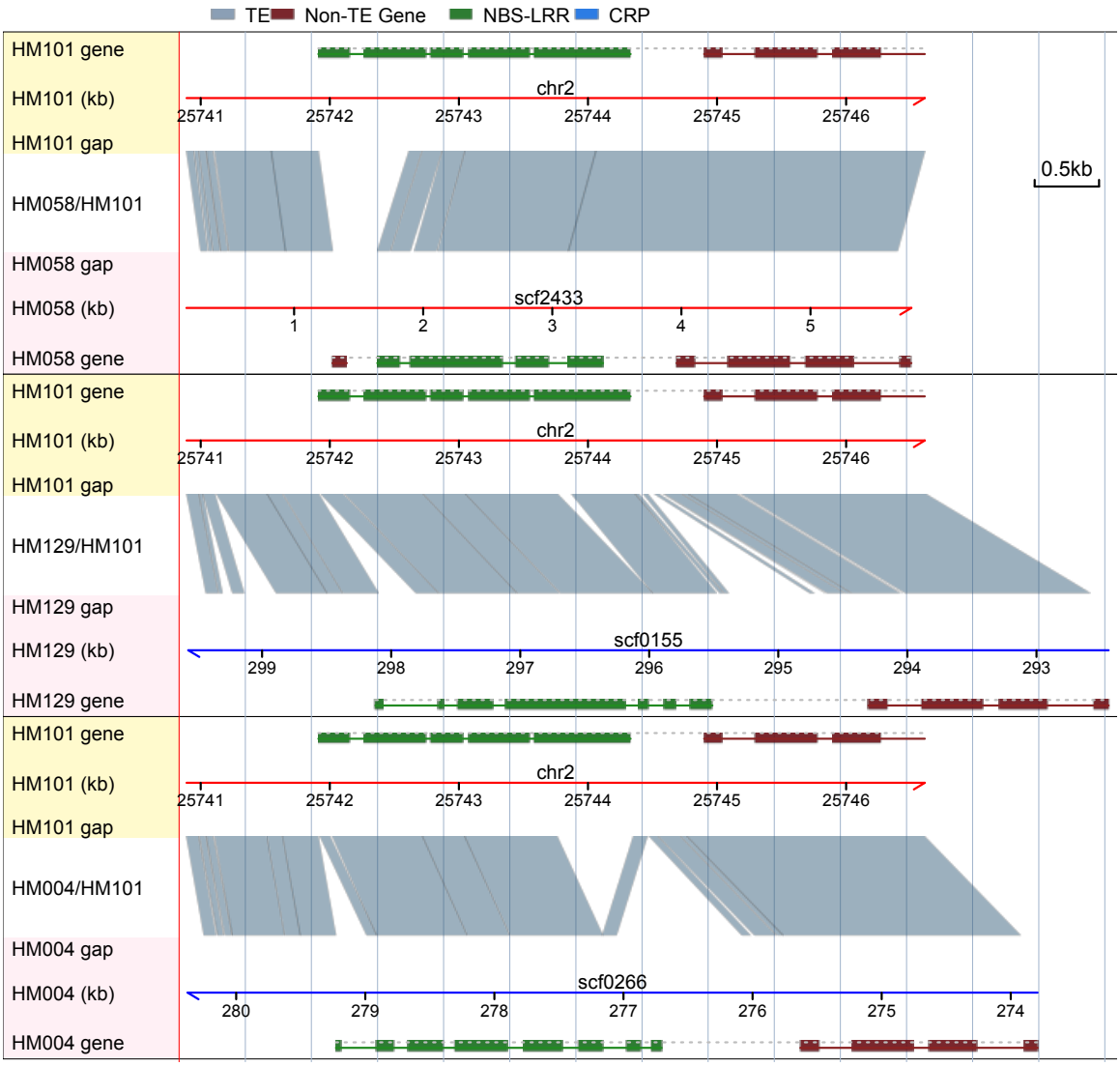
Four breakpoints (br1 – br4) are indicated by "x".

Figure Link: https://www.dropbox.com/s/cj3wzddi122dzfi/figS2.6.pdf?dl=0

Appendix Figure 2.6. Closer look at the chromosome 4/8 translocation breakpoints.

Figure Link: https://www.dropbox.com/s/xrrmij9zv8ujert/figS2.7.pdf?dl=0

Appendix Figure 2.7. Genome browser screenshots showing genes around the breakpoints of chromosome 4/8 rearrangement.



Appendix Figure 3.1. Illustration of an NBS-LRR gene with at least 4 allelic forms (i.e., haplotypes, including HM101) in the 15 accessions surveyed.