GENOME STRUCTURE COMPARISONS ACROSS PLANT TAXA AND APPLICATIONS FOR STUDYING DOMESTICATION IN CULTIVATED SORGHUM

by

HAIBAO TANG

(Under the Direction of Andrew H. Paterson)

ABSTRACT

The wealth of sequence data from many flowering plant species offer unique opportunities for us to understand the dynamic genomic changes that have occurred in many plant lineages. Such changes are evident through comparisons from both within the same genome and across multiple related genomes. I have developed improved methodology to identify and interpret synteny patterns that are more suitable for comparing plant genomes. The novel computational tool helps to infer ancient whole genome duplications in both the eudicot and monocot lineages and also provides clearer correspondences between representative taxa across the two divergent lineages. In the second part of my dissertation, I fine mapped the sorghum grain shattering (seed dispersal) locus *Sh1*, which was previously mapped to a ~1Mb genomic region by linkage study. In order to associate the shattering trait with specific DNA polymorphisms, I carried out extensive resequencing in the region using a diversity panel consisting of shattering and non-shattering sorghum individuals. The second study suggests a few candidate DNA changes for further functional confirmations, among which might underlie a key genetic transition from wild to domesticated sorghum.

INDEX WORDS: synteny, paleopolyploidy, domestication, association mapping, linkage disequilibrium, grain shattering, sorghum, *Sh1*

GENOME STRUCTURE COMPARISONS ACROSS PLANT TAXA AND APPLICATIONS FOR STUDYING DOMESTICATION IN CULTIVATED SORGHUM

by

HAIBAO TANG

B.S., Fudan University, China, 2004

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2009

GENOME STRUCTURE COMPARISONS ACROSS PLANT TAXA AND APPLICATIONS FOR STUDYING DOMESTICATION IN CULTIVATED SORGHUM

by

HAIBAO TANG

Major Professor:	Andrew H. Paterson
Committee:	Shu-Mei Chang Jessica C. Kissinger Russell Malmberg Rodney Mauricio

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2009

DEDICATION

I would like to dedicate this work to my family and friends that are supportive through the five years of my graduate study in Athens. Foremost to my dear parents, although far away in China, are always encouraging my work, and have helped me making the right decisions at the many crossroads in my life. I am also deeply indebted to my love, Ann Hao, who during these years always stood behind me, took good care of me and tolerated my sometimes erratic personality. Finally, I would like to thank my friends that have brought joy and happiness to my otherwise mundane and reclusive life.

ACKNOWLEDGEMENTS

I would like to thank Andy for providing the best mentorship a graduate student can ask for. Andy is always supporting and helping my research and writings, while providing the right amount of guidance to my premature ideas. Next, I would like to thank John Bowers for extensive training of my research skills and frequent exchange of scientific ideas, and Xiyin Wang for the collaborative work behind many computational projects in the Paterson lab. I would like to thank the UGA Graduate School for the dissertation completion award in the last year of my study. Finally, my special thanks to the graduate committee members for precious time and input to my research and training.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER 1 INTRODUCTION

1.1 Overview

This thesis consists of two major research projects that are central to my Ph.D. study. The first two chapters (Chapter 2 and 3) focus on the first project, which describe computational results for comparing the gene orders both within genomes (to identify ancient genome duplications) and across genomes (to identify orthologous regions).

Chapter 2 introduces the algorithmic foundations for inference of conserved gene orders. The development of the main algorithms, implemented in the computer program MCscan, borrows heavily from theories behind biological sequence alignment, ranging from the alignment criteria, to extension from pairwise to multiple alignments. The increased sensitivity of the alignments permits a high resolution gene-based synteny map across multiple sequenced plant genomes, and provides a powerful tool to infer positional conserved homologs both within and across genomes. An NSF-funded Plant Genome Duplication Database (PGDD), created as an extension of this work is also briefly described.

Chapter 3 describes ancient whole genome duplication events (paleopolyploidy) from the comparisons of some sequenced flowering plant taxa. The major analyses follow two threads, building on previous work in *Arabidopsis* (Bowers et al. 2003a) and rice (Paterson et al. 2004). I describe in depth about these ancient events in both the eudicot and monocot lineage, respectively. The better understanding of the genome redundancy in both lineages naturally leads to a better analysis of the eudicot-monocot comparisons. The work is a significant improvement over previous efforts and reveals deep synteny comparisons across divergent plant lineages.

The next two chapters are relevant to studying functional genomics and sorghum domestication, which focus on my second project – association mapping of the sorghum grain shattering gene *Sh1*.

Chapter 4 is a literature review and sets some theoretical ground for chapter 5. The chapter reviews our current knowledge on the genetic processes underlying crop domestication, as well as the various genomic tools and mapping methods to study the domestication. In particular, the comparisons between association mapping and linkage mapping methods are described in more details. Grain shattering (as controlled by *Sh1* in sorghum) is an important agronomic trait that was selected and eventually fixed during domestication of cereal crops.

Chapter 5 describes the fine mapping effort of the sorghum shattering locus *Sh1*. Previous linkage mapping study has pointed the locus to a ~1Mb region on sorghum chromosome 1. We sequenced the corresponding region in a wild sorghum species (*S. propinquum*) and compared the sequence to the public genome sequence of domesticated sorghum (*S. bicolor*). The comparisons between the two sorghum species set the ground for the association mapping of the grain shattering locus. I compiled a sorghum diversity panel consisting of shattering and non-shattering individuals. Using the resequencing data on the sorghum diversity panel, I characterized the pattern of linkage disequilibrium and performed statistical tests on the segregating sites within the target region. I found a few sites that show significant marker-trait association and are promising candidates for further functional characterizations.

Chapter 6 is a chapter that reviews major conclusions in previous chapters and suggests future research directions.

Most chapters are organized in similar structure, first introducing the basic concept and brief review of recent progress in the field, followed by original research work, with the exception of Chapter 4, which is largely a literature review by itself.

1.2 Related publications

The following lists published papers related to my research, on which I have co-authored. Some papers are other's work, but in which I have contributed. Nonetheless, in this dissertation I try to be careful so that I only present my own original research results, with research results from co-authors excluded.

The publications are organized in relevance to the chapters of this thesis. For the published materials that are partly re-used in this thesis, copyright permissions to use were granted by the journals (with licenses requested).

CHAPTER2: GENE ORDER COMPARISONS IN RELATED PLANT SPECIES

- Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and Collinearity in Plant Genomes. *Science*, **320**, 486-488.
- Lyons, E, Pedersen, B, Kane, J, Alam, M, Ming, R, Tang, H., Wang, X, Bowers, J, Paterson, A.H., Lisch D, Freeling, M. (2008) Finding and Comparing Syntenic Regions among Arabidopsis and the Outgroups Papaya, Poplar, and Grape: CoGe with Rosids. *Plant Physiology*, **148**, 1772-81.
- Bowers, J.E., …, and *29 others*, Tang, H., Wing, R.A. and Paterson, A.H. (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses, *PNAS*, **102**, 13206-13211.

CHAPTER 3: INFERENCE OF PALEO-POLYPLOIDY IN MAJOR PLANT LINEAGES

- Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M. and Paterson, A.H. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps, *Genome Research,* **18**, 1944-1954.
- Wang, X., Tang, H., Bowers, J.E., Feltus, F.A. and Paterson, A.H. (2007) Extensive concerted evolution of rice paralogs and the road to regaining independence, *Genetics*, **177**, 1753-1763.

- Wang, X., Tang, H., Bowers, J.E., and Paterson, A.H. (2009). Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res*.
- Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L., Salzberg, S.L., Feng, L., Jones, M.R., Skelton, R.L., Murray, J.E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., ..., and *62 others* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, **452**, 991-996.
- Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., ..., and *32 others* (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551-556.
- Kim, C., Tang, H. and Paterson, A.H. (2009). Duplication and Divergence of Grass Genomes: Integrating the Chloridoids. *Tropical Plant Biology*. **2**, 51-62.

CHAPTER 4: STUDY OF DOMESTICATION IN THE POST-GENOMICS ERA

- Charles, M., Tang, H., Belcram, H., Paterson, A.H., Gornicki, P. and Chalhoub, B. (2009). Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after their divergence from Panicoideae. *Mol Biol Evol*.
- Jang, C.S., Kamps, T.L., Tang, H., Bowers, J.E., Lemke, C., and Paterson, A.H. (2009). Evolutionary fate of rhizome-specific genes in a non-rhizomatous Sorghum genotype. *Heredity* **102:** 266-273.

CHAPTER 2 GENE ORDER COMPARISONS IN RELATED PLANT SPECIES

2.1 Introduction

Eukaryotic genomes differ in the degree to which genes remain on corresponding chromosomes (synteny) and in corresponding orders (collinearity) over time (Coghlan et al. 2005). For example, most eutherian (mammalian) orders have incurred only moderate reshuffling of chromosomal segments since descent from common ancestors ~130 million years ago (Ferguson-Smith et al. 2007). Indeed, karyotype evolution along major vertebrate lineages appears to have been slow since an inferred whole-genome duplication occurred ~500 million years ago (Nakatani et al. 2007). Accordingly, accurate identification of orthologs across eutherian taxa is relatively routine, and deduction of synteny and collinearity is often straightforward with "best-in-genome" criteria (Miller et al. 2007), identifying one-to-one best matching chromosomal regions in pairwise genome comparisons.

Angiosperm genomes fluctuate remarkably in size and arrangement even within close relatives, with recurring whole genome duplications occurring over the past ~200 million years accompanied by wholesale gene loss that has fractionated ancestral gene linkages across multiple chromosomes (Bowers et al. 2003b). Angiosperm genome sizes span more than 1000-fold (Bennett et al. 1991), with much of the difference between some well-studied genomes in heterochromatin (Bowers et al. 2005). Additionally, the reshuffling of short DNA segments by mobile elements nearly eliminates large-scale collinearity in heterochromatic regions (Bowers et al. 2005).

Despite recurring whole-genome duplications, angiosperm chromosome numbers are more static than genome size, mostly within a range of less than 50-fold (Bennett et al. 1991). Condensation of two chromosomes into one is known in many lineages; a particularly striking

case involved the demonstration that *n*=10 (chromosome number) members of the *Sorghum* genus are ancestral to *n*=5 members of the genus (Spangler et al. 1999). Indeed, *Sorghum bicolor* (sorghum) and *Zea mays* (maize) have the same chromosome number (*n*=10), although maize has been through a whole-genome duplication since their divergence (Swigonova et al. 2004), whereas the most recent duplication in sorghum is shared with all other cereals (Paterson et al. 2004). The occurrence of several condensations may explain why single arms of several maize chromosomes (10 and 5) correspond to entire sorghum chromosomes (6 and 4) (Bowers et al. 2003a). The comparison of the botanical model *A. thaliana* to other angiosperms is complicated by additional 9 to 10 chromosomal rearrangements in the past few million years since its divergence from *A. lyrata* and *Capsella rubella*, including condensation of six chromosomes into three, bringing the chromosome number from *n*=8 to *n*=5 (Yogeeswaran et al. 2005).

Synteny can be identified through the clustering of neighboring matching gene pairs; however, differences in gene density and tandem gene arrays among species may cause statistical artifacts. Collinearity, a more specific form of synteny, requires common gene order. Collinearity and synteny have traditionally been identified by looking for one-to-one (pairwise) conservation between species. To take better advantage of new genomic resources as they become available, multiway collinearity analyses are needed, with progressive alignments accompanied by statistical evaluation and iterative refinement (Miller et al. 2007). In angiosperms, such multiple alignments offer the further advantage of unraveling ancient genome duplications (Chapter 3).

There are several limitations of existing algorithms for comparing plant genomes, due to the unique architecture of plant genomes. Algorithms (e.g. BLASTZ/CHAINNET pipeline, and LAGAN/SUPERMAP pipeline) commonly used in vertebrate genome alignments focus on identifying orthologous regions while largely ignoring paralogous regions (Kent et al. 2003). A general theme for detection of distant synteny relationships is to use "all versus all" BLASTP

searches as inputs, and model the matches in a homology matrix representation (or genomic dot plot) where synteny is uncovered by clustering neighboring matches inside the matrix. Such an approach is central to ADHORE (Vandepoele et al. 2002) and DiagHunter (Cannon et al. 2003), and influences other algorithms (Calabrese et al. 2003). Two recent methods DAGchainer (Haas et al. 2004) and ColinearScan (Wang et al. 2006) formulate the problem by dynamic programming and use empirical or statistical strategies which effectively improve sensitivity and specificity of inferring chromosomal homology. However, each method still only predicts pairwise collinearity patterns. A key need is to combine pairwise collinear segments into one inferred order which utilizes multiple collinearity.

In this chapter, I will first describe the algorithms that identify the synteny patterns, both the pairwise and multiple alignment case. These algorithms are at the core of the computer program – MCscan that I wrote to find the synteny patterns across several plant genomes. I stored my identified synteny patterns in 9 sequenced plant genomes in a public database (PGDD) with interactive web interface, to facilitate more common use by other plant researchers. In the future, we will continue to support the updates and improvements of PGDD through NSF funding.

2.2 Algorithm for aligning pairwise gene orders

There are many levels of resolution when we infer conserved synteny, ranging from the genetic marker level, gene level and base pair level. In this study, I define the synteny at the "gene" level, treating individual genes as the smallest unit. Most of the analyses are based on "gene pairs" -- a pair of gene models that show high similarity in BLAST search. In order to identify synteny patterns, we often look for the gene pairs that are closer to one another on the chromosomes (or "chaining" of gene pairs). This is an important basis for the algorithms that follow.

Computational inference of synteny correspondences is closely analogous to the nucleotide or protein sequence alignment, albeit with a much larger alphabet size (nucleotide 4,

protein 20, and in the case of gene order comparisons – the number of gene families in the genome). However, it is somewhat simpler than aligning residues in that we do not consider mismatches between the "letters" in the alphabet, because mutation from one gene family member to another is unlikely over the evolutionary scale that I focus on. I can draw rich algorithmic results from the sequence alignment literatures, and often with efficient and re-usable implementations. I will discuss the algorithmic foundations for aligning two gene orders, and then extend to the multiple cases when dealing with more than one genome or subgenomes (in the case of polyploidy).

2.1.1 Problem formulation

When we represent each gene along the chromosome with a unique gene family identifier (each as an integer), there is the following formulation.

Input: two sequences of integers $\{a_1, a_1, ..., a_m\}$ *and* $\{b_1, b_1, ..., b_n\}$*;*

Output: good alignment between the two sequences of integers.

The problem lies in what constitutes a 'good' alignment, which also applies to biological sequence alignment. Different criteria exist, ranging from the models that are aesthetic to the researcher, to the "parsimony" model favoring the alignments with the highest score (or least cost), and the "probabilistic" model that to approximates an evolutionary model of how sequence *a* changes into sequence *b*. I will deal with both the parsimony and probabilistic model in the following text.

2.2.1 Parsimony method - gapped alignment

The gapped alignment methods for DNA and protein sequence alignment differ in two versions – Needleman-Wunsch (global alignment) and Smith-Waterman (local alignment). The central

idea is to build up the optimal alignment by re-using the solutions for smaller and smaller sequences, or called "dynamic programming". The alignment proceeds by recursively filling up a matrix with the cell (i, j) containing the best score up to a_i and b_j. The same idea later was extended in (Haas et al. 2004; Wang et al. 2006) to align the gene orders and infer chromosomal homology.

It is important to note that in the case of gene order alignment, it is computationally wasteful to use the two-dimensional matrix to store the scores, since the "matching" states are sparse. It suffices to just use a one-dimensional array with the size equal to the number of matching gene pairs to store the scores. This is also asymptotically faster than the two-dimensional case, because for each gene pair, we only need to search for the succeeding pair within a specified distance *d*. This gives a linear complexity to the number of homologous gene pairs. There is the following recurrence condition, assuming two gene pairs *u* and *v* are on the "chaining" path where *u* precedes *v*,

$$ChainScore(v) = MatchScore(v) + \max_u\{ChainScore(u) + GapPenalty(u, v), 0\}$$

The elements of the one-dimensional array *ChainScore* are the gene pairs, sorted with the relative order on both chromosomes. My typical scoring (but can be easily modified in the MCscan implementation) to evaluate the synteny pattern is the following scheme, +50 bonus for a matching gene pair, and a linear gap penalty -3 for each gap in between the pairs and report all pairwise segments with scores above 300. The gene order alignments are then retrieved through the backtracking of the dynamic programming arrays.

The statistical significance of the pairwise alignments can be evaluated using a formula derived in (Wang et al. 2006).

$$E = 2P_N^m \prod_{i=1}^{m-1}(\frac{l_{1i}}{L_1} \cdot \frac{l_{2i}}{L_2}),$$

where N is the number of matching gene pairs between two chromosomal regions defined by the inferred synteny block; m is the number of collinear gene pairs; L_1 and L_2 are respective lengths of the two chromosomal regions; and l_{1i} and l_{2i} are distances between two adjacent collinear gene pairs in the syntenic block. The expectation multiplies by 2 since there are two possible orientation configurations between two collinear segments. This is only an approximation to more rigorous yet computationally expensive permutation test (Van de Peer 2004) and Monte Carlo methods (Hampson et al. 2005), however computational experiments and analytical results (Wang et al. 2006) suggests that this gives a reasonable estimate for the significance of the syntenic blocks.

2.2.2 Probabilistic modeling of gene order evolution

There are several apparent drawbacks of the parsimony procedure. First of all is the lack of evolutionary model (e.g. how realistic is the parsimony model), and also limited interpretability of the scoring scheme and lack of rationale for choosing specific parameters (e.g. why choose +50 for a match and -3 for a gap). Second, we would like to measure the reliabilities for different parts of the alignment. Third, we wish to consider not only a few optimal alignments based on parsimony principle, but to weigh all alternative alignments (sub-optimal alignments) probabilistically.

Ideally there are many different types of gene order mutations that need to be modeled (**Figure 2.1**). Note for gene order alignments, there are no *substitutions* and we can focus on only *insertion* and *deletions*. I do not yet attempt to model rearrangements operations such as *inversions* and *transpositions*, since these operations are computationally more difficult to solve with a much larger solution space to consider.

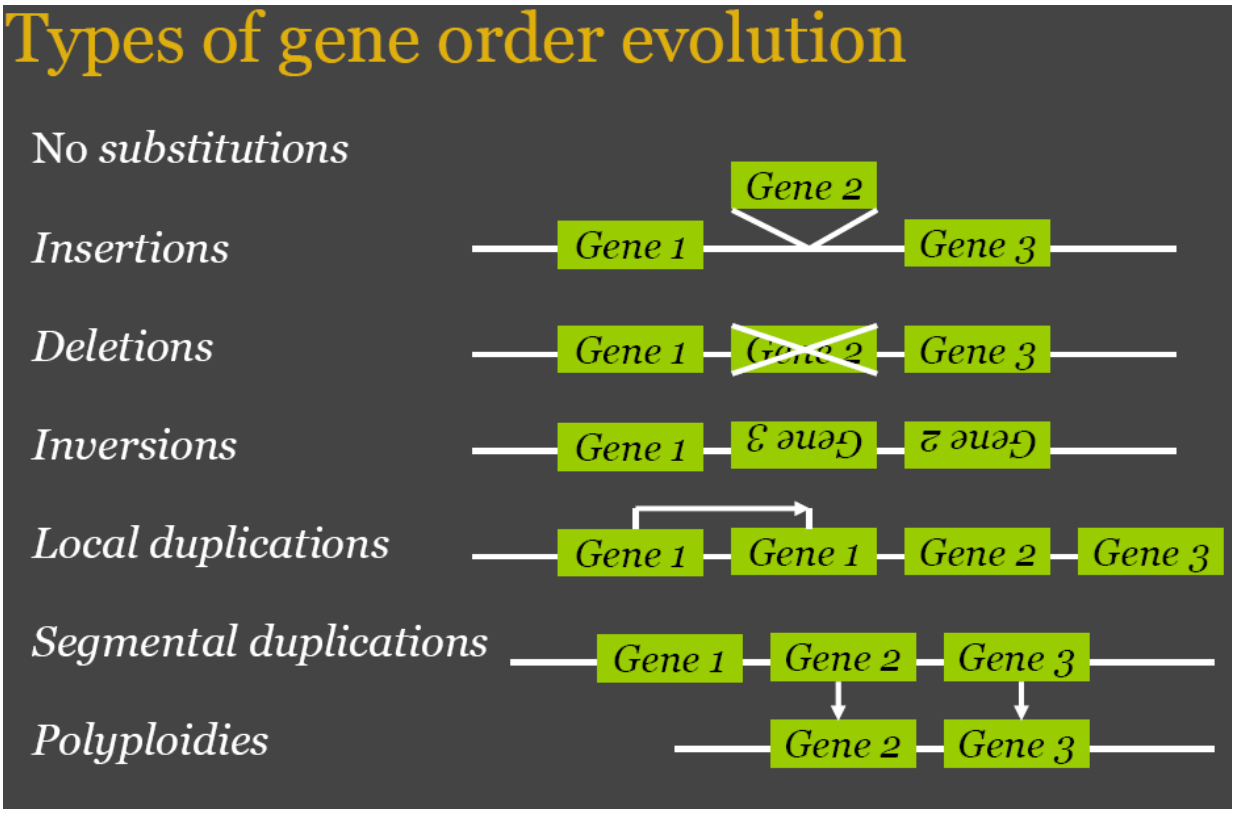

Figure 2.1: Different types of gene order evolution.

To model the evolution of two gene orders sharing a common ancestor, it is convenient to first consider a *time-reversible* model. A time-reversible model is useful since it is obviates the need to sum over all possible ancestral states (Thorne et al. 1991). A relevant insertion-deletion model (TKF91) was originally proposed in (Thorne et al. 1991), which models the insertions and deletions with a simple birth-and-death process with imaginary *links* between the residues. The links and the residues are created at a rate of λ and deleted at a rate of μ. The TKF91 model was later simplified in (Hein et al. 2000) and again in (Holmes et al. 2001). In particular, Holmes and Bruno converted the model into a paired hidden markov model (PHMM), with three emitting states H (homologous), I (insertions), D (deletions) and a few silent states to factor out the transition probabilities. The transition probabilities are shown in **Figure 2.2**. The labeled probabilities are functions of birth rate (λ), death rate (μ) and time (t). The following

parameters in the PHMM were derived by solving the differential equations in (Thorne et al. 1991),

$$\alpha = e^{-\mu t},\ \beta = \frac{\lambda(1-e^{(\lambda-\mu)t})}{\mu-\lambda e^{(\lambda-\mu)t}},\ \gamma = 1-\frac{\mu(1-e^{(\lambda-\mu)t})}{(1-e^{-\mu t})(\mu-\lambda e^{(\lambda-\mu)t})},\ g = \frac{\lambda}{\mu}$$

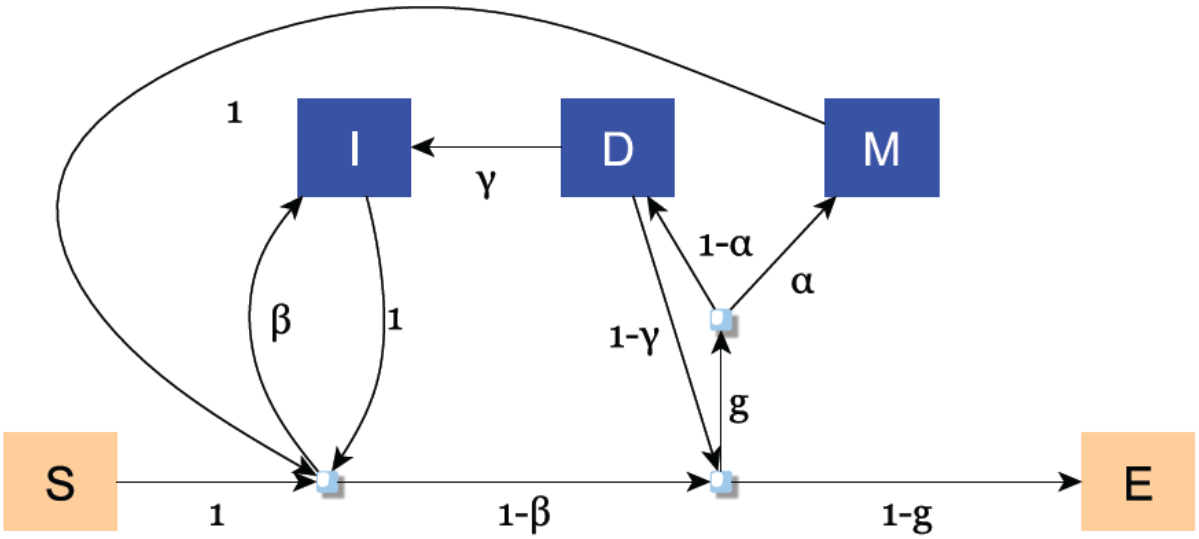

Figure 2.2: TKF91 paired HMM formulated in (Holmes et al. 2001). There are five emission states in this HMM model, blue emission states (I, D, M) and three silent states for factoring the transition probabilities.

The TKF91 model and its PHMM formulation allow us to track the evolution of the gene orders and evaluate the likelihood of the alignment. The full probability of *a* and *b* summing over all paths, $P(a,b)$, can be evaluated by the standard forward algorithms for PHMMs. Values of the parameters (λ, μ, *t*) can then be found which numerically maximize the likelihood function. I performed the numerical optimization using a downhill simplex algorithm implemented in the python package *scipy.optimize*.

With the PHMM form, we can also calculate the expected accuracy of the alignment, which was not possible in the previous score-based (parsimony) method. Following Durbin et al.

(Durbin 1998), I use the notation $a_i \diamond b_j$, which means that a_i is aligned to b_j. We have, through conditional probability,

$$P(a,b,a_i \Diamond b_j) = P(a_{1..i}, b_{1..j}, a_i \Diamond b_j) P(a_{i+1..m}, b_{j+1..n} \mid a_{1..i}, b_{1..j}, a_i \Diamond b_j)$$
$$= P(a_{1..i}, b_{1..j}, a_i \Diamond b_j) P(a_{i+1..m}, b_{j+1..n} \mid a_i \Diamond b_j)$$

Note that on the right side of the equation, the first term can be calculated using the forward algorithm and the second term can be calculated using the backward algorithm (Durbin 1998). With $P(a,b,a_i \Diamond b_j)$ calculated we can then used the Bayes' rule and obtain,

$$P(a_i \Diamond b_j \mid a,b) = \frac{P(a,b,a_i \Diamond b_j)}{P(a,b)},$$

which becomes the posterior probability that gene a_i is aligned to b_j.

The whole procedure is illustrated with an example (**Figure 2.3**). In this testing example, we consider a pairwise alignment between two chromosomes, each with more than 1000 genes (the numbers on the axis denote the ranks of genes along the chromosome). Before the algorithm, the dot plot looks quite "noisy", yet with strong homology close to the short arm terminal regions of both chromosomes. The TKF91 PHMM was then performed and the maximum likelihood estimates were obtained for the three parameters of the model (with a maximized log-likelihood of -2268). Finally each matching gene pair was evaluated for the posterior probability, and color-coded to reflect the values. Only the pairs with large probability are visible after the posterior decoding. We note that close to the center of the alignments (roughly 400th to 800th gene on both chromosomes), the reliability of the alignment is quite poor (circled in **Figure 2.3**). This is reflected in the low values of posterior probability of the aligned gene pairs. In this region, the paths are splitting in a few places, therefore "sharing" the probability on several different sub-optimal paths.

For brevity, I omit the development of recursive functions for both forward and backward algorithms, but they can be readily found in the Durbin book (Durbin 1998).

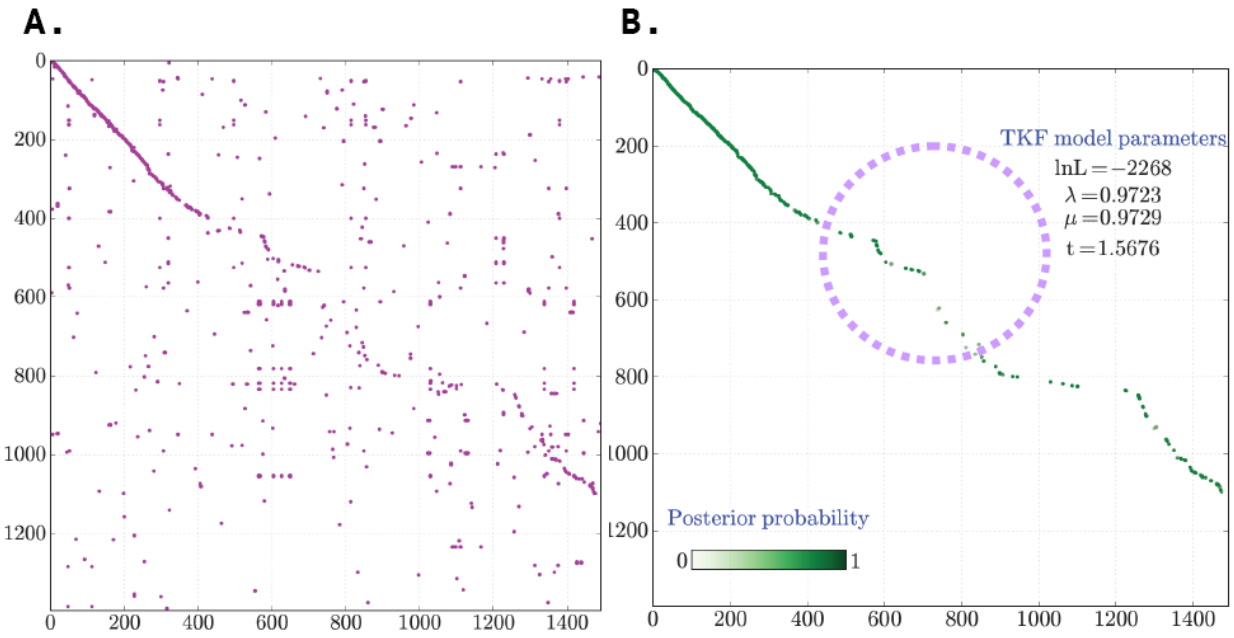

Figure 2.3: Dot plot before and after the TKF PHMM run. (A) Original (unfiltered) dot plot between two gene orders. (B) The maximum likelihood alignment found by the TKF91 model.

2.3 Algorithm for aligning multiple gene orders

The multiple gene order alignment problem is computationally intractable since it is a variant of the "shortest common super-sequence" problem. This is a similar situation for the multiple sequence alignment. To circumvent the intractability, we employ a heuristic method which at each iteration of the algorithm, new matching gene orders are added to a consensus order. Multiple chromosomal regions are aligned progressively by adding one closest-related region at a time by dynamic programming. **Figure 2.4** is a flow chart diagram of the MCscan algorithm. The key step is step 3, where the multi-way view is constructed through the stacking of many related pairwise alignments. In early version of MCscan, the consensus method was used, maintaining a merged linear order at each iteration (**Figure 2.5**).

The most recent version of MCscan instead uses partial order alignment (**Figure 2.5**). The partial order graph alignment (Lee et al. 2002) is sometimes an improvement over the consensus method (Rodelsperger et al. 2008). In this method, the gene orders are represented

as directed acyclic graph (DAG) structures. The distances between matching nodes are determined through depth first search (DFS) traversal over the graphs. Successive rounds of alignments can benefit from the incorporation of more gene orders, allowing a higher sensitivity than previous methods. A similar concept was also reviewed in (Van de Peer 2004) and implemented in the software program i-ADHORE (Simillion et al. 2008).

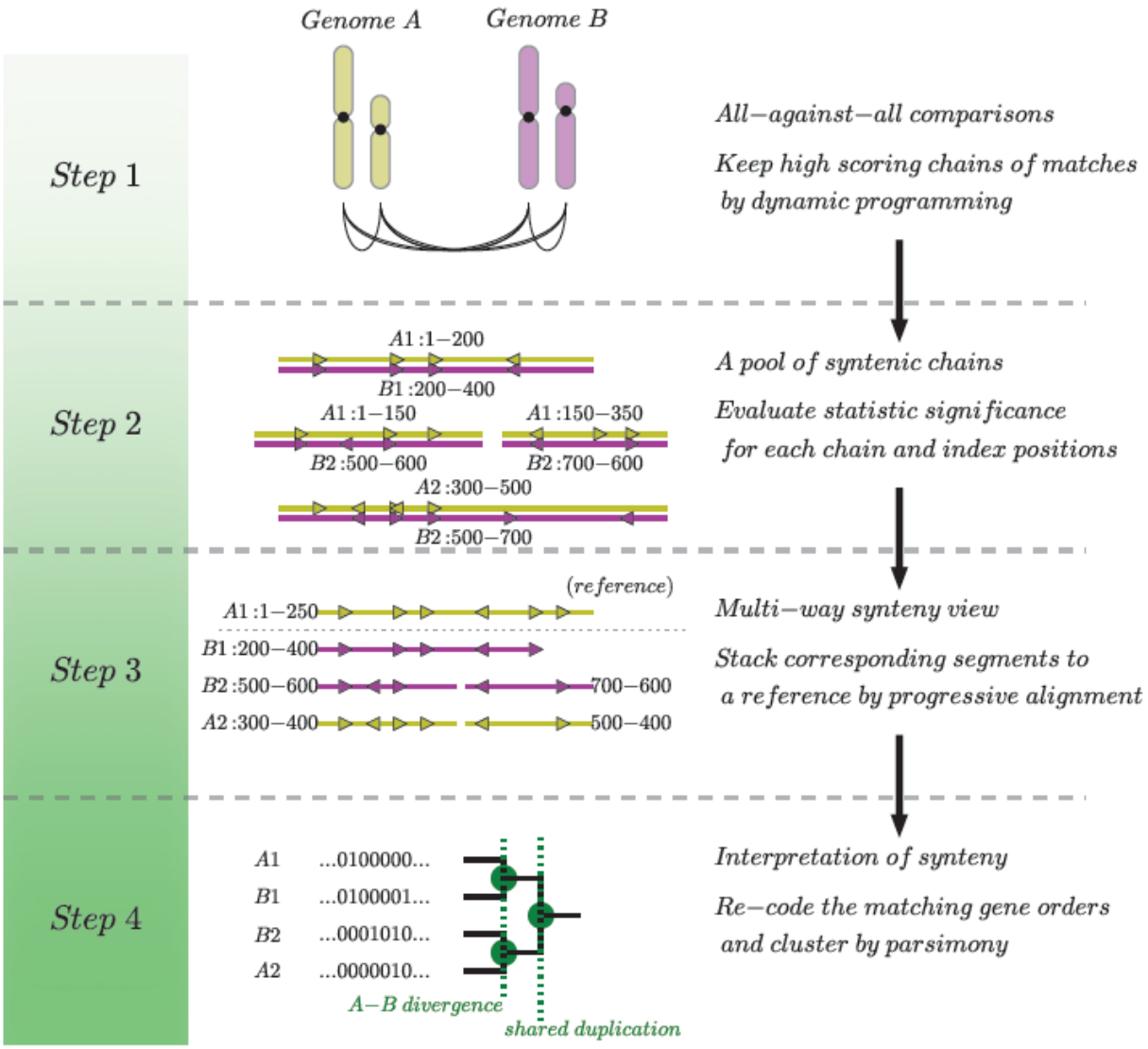

Figure 2.4: Flow-chart of MCscan core algorithm.

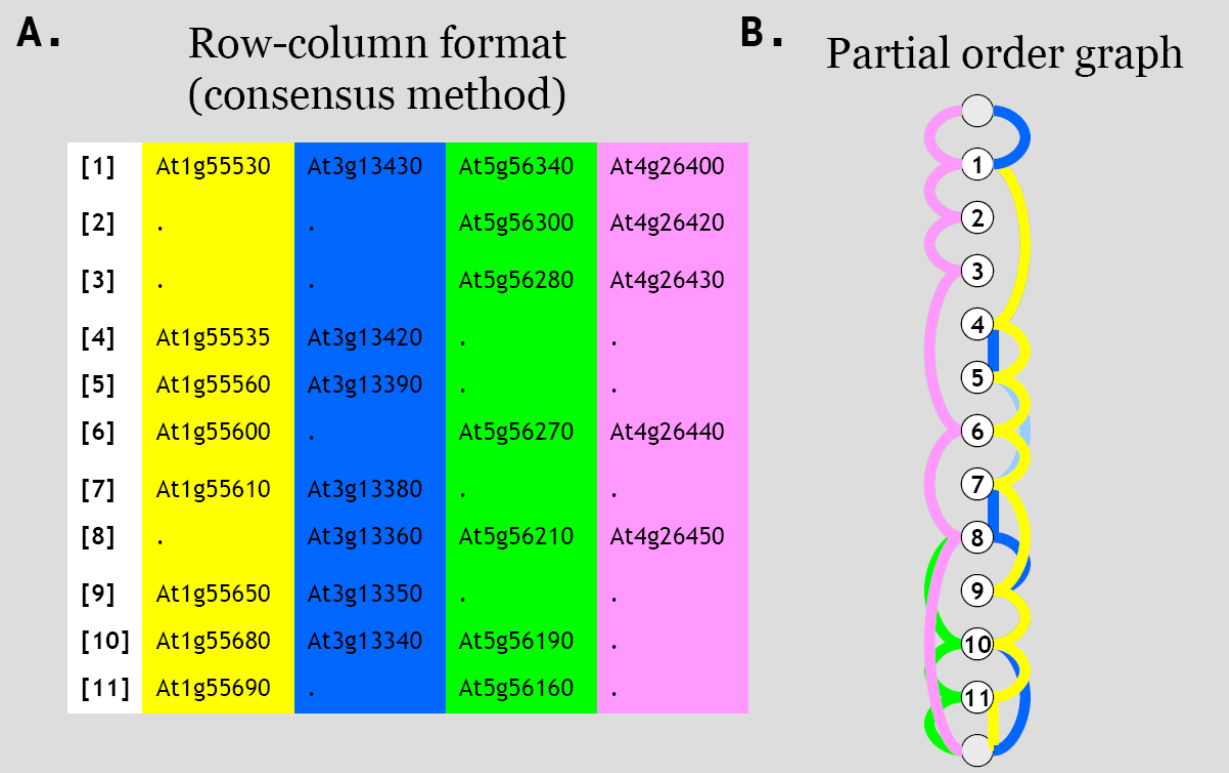

Figure 2.5: The internal representation of multi-alignment data structure. (A) The consensus representation. (B) The partial order graph representation. The two representations are often interchangeable.

The MCscan program is implemented in ANSI C++ with source code publicly available (http://chibba.agtec.uga.edu/duplication/mcscan/). A copy of the MCL program (http://www.micans.org/mcl/, version 08-157) was used in the pipeline and dispatched with the MCscan package. The program takes two input files – a file containing BLASTP results and a file describing gene coordinates – and outputs both pairwise syntenic blocks and the multi-aligned gene orders threaded by a reference genome. There are several parameters to configure according to the user's need (see documentation on the software site). For example, the significance cutoff would reduce sensitivity but increase specificity. General advice on the parameter choices are given and the default values are optimized for large plant genomes but some parameters may still need to be optimized (e.g. those that depend on the average gene density).

2.4 Plant Genome Duplication Database (PGDD)

Although many plant genome sequences are sequenced or soon to be sequenced, public database and websites characterizing their synteny patterns are not readily available. Herein, we stored the synteny blocks inferred by MCscan in the Plant Genome Duplication Database (PGDD), which provides a central data repository for plant researchers to search the orthologs and paralogs that are supported by positional conservations.

Table 2.1: Plant genomes included in PGDD, and more in the pipeline.

Species name	Release version	Reference
Arabidopsis thaliana	TAIR 8.0 (Aug. 2008)	(AGI 2000)
Carica papaya	EVM (Jul. 2007)	(Ming et al. 2008)
Populus trichocarpa	JGI 1.1 (Dec. 2004)	(Tuskan et al. 2006)
Medicago trunculata	Release 2.0 (Feb. 2008)	--
Glycine max	Release 1 (Dec. 2008)	--
Vitis vinifera	Genoscope (Aug. 2007)	(Jaillon et al. 2007)
Brachypodium distachyon	Release (May 2009)	--
Oryza sativa	RAP 2.0 (Nov. 2007)	(IRGSP 2005)
Sorghum bicolor	Sbi 1.4 (Dec. 2007)	(Paterson et al. 2009)

The web URL for PGDD is (http://chibba.agtec.uga.edu/duplication) and it is regularly maintained. To date, PGDD contains the collinearity patterns from 9 sequenced angiosperm genomes – *Arabidopsis*, papaya, poplar, *Medicago*, soybean, grape, *Brachypodium*, rice and sorghum (**Table 2.1**). The data included in the database are updated upon the availability of new versions of gene annotations as well as newly sequenced plant genomes (once "Ft Lauderdale" restrictions on whole-genome analyses are lifted).

2.4.1 Main functionalities of PGDD

Display of macro-scale synteny blocks. Traditional dot-plots of synteny relationships are provided (**Figure 2.6**). Users can directly click on any region of the dot-plot to zoom in on the fine structure of syntenic segments of interest. Synteny blocks can be filtered with regard to chromosomal positions (to focus on specific rearrangements) and *Ks* distances between gene pairs (reducing noise from extraneous duplications, to focus on the "signal" of a specific duplication event as illustrated).

Display of the fine structures of syntenic regions of interest. This is a query service where a user enters a locus ID for a gene model and the server interactively displays the syntenic genes as well as the chromosomal segments on which they belong (**Figure 2.6**).

The *BLAST-View* program allows the users input their own sequences and BLAST against the predicted gene set of the included plant genomes. The output of the BLAST is then visualized to place all the BLAST hits on the chromosomes. This is useful when a researcher has some sequences and wishes to see where the homologous sequences are located in related genomes, or study the distribution of a particular gene family.

2.4.2 Data preparation

MCscan was used to pre-calculate the synteny patterns between every pairwise genome comparisons and self-comparisons. For homologs inferred from the synteny alignments, we aligned the protein sequences using CLUSTALW and used the protein alignments to guide coding sequence alignments. We used Nei-Gojobori method implemented in *yn00* program in the PAML package (Yang, 1997) to calculate *Ks*. Log-Gaussian mixture models are fitted to the *Ks* distributions using GMM with Bayes Factors (http://astro.u-strasbg.fr/~fmurtagh/mda-sw/), to reveal the underlying components within the *Ks* distribution (Cui et al. 2006). The *Ks* distributions between any two genomes are shown on the website and can be used to guide the user to select specific age range between homologs. In-house python scripts are used to pipeline

all the calculations (available with documentations at

http://chibba.agtec.uga.edu/duplication/data/syn_calc.zip).

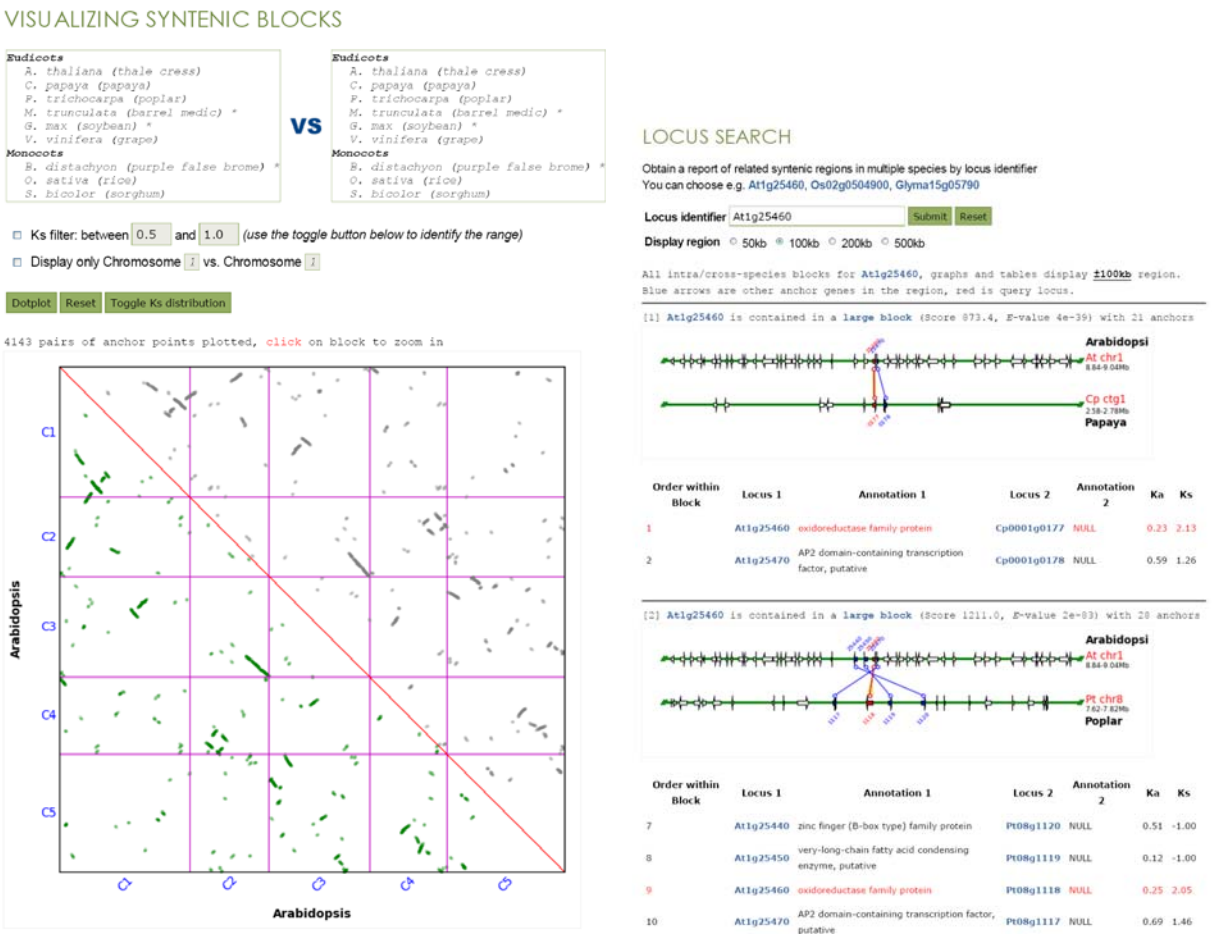

Figure 2.6: Screenshots of PGDD web interface. (left) Dot-plot visualization of syntenic relationships between genomes (in this example showing *Arabidopsis* self-comparison). Users can directly click on any region on the dot-plot to zoom in and view fine structure of syntenic segments, and download specific homologous gene pairs. (right) Locus search of a specific gene within- and cross-genomes shows structural changes in the homologous chromosomal segments.

2.4.3 Database structure and web interface

The database mainly consists of two pieces of information stored in a MySQL relational database (**Figure 2.7**). Information on gene coordinates, functional annotations and gene sequences are stored in the *loci* table. Inferred homologous pairs (from MCscan) are stored in a

separate *block* table. The web interface is written in the Python programming language, and hosted by Apache server through *mod_python*. The server-side Python scripts (for graphics and processing of queries) are called by jQuery (http://jquery.com) AJAX requests. All the homologous blocks and related data (genetic distances like *Ks* and *Ka*) are available for download.

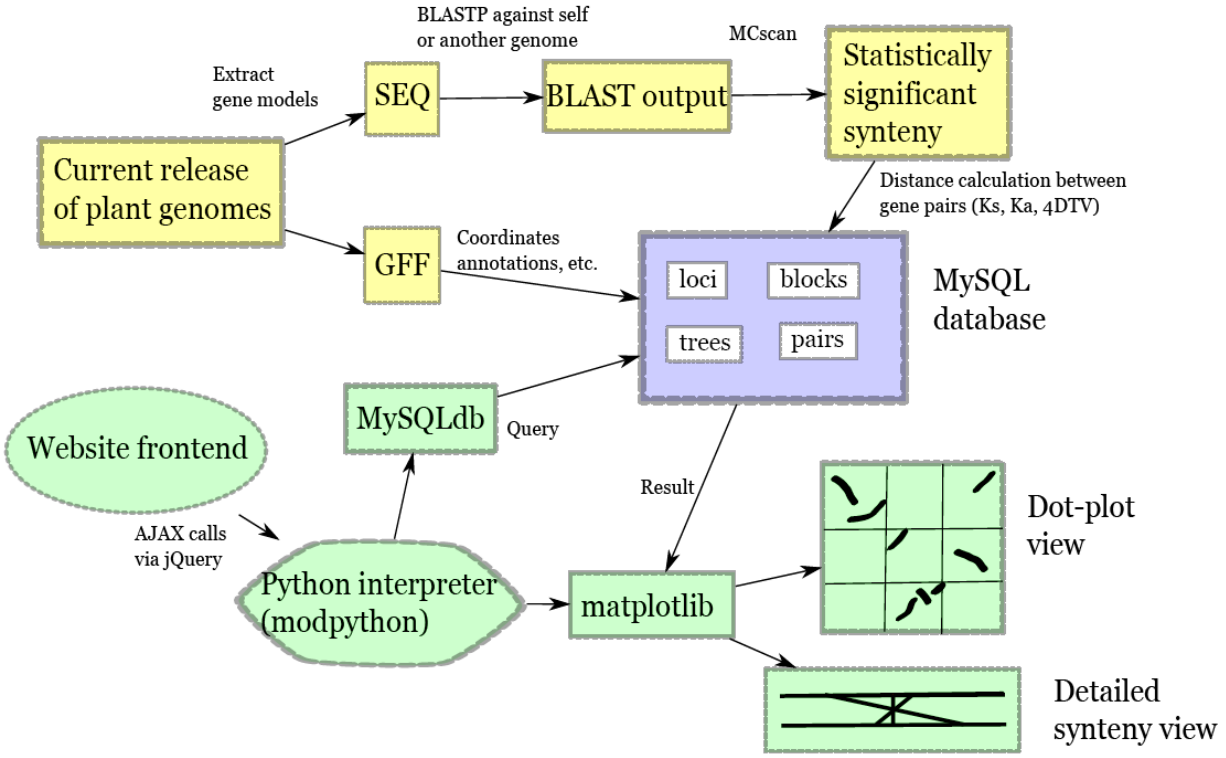

Figure 2.7: Organization of the PGDD database.

2.4.4 Community service and future plans

We have plans to improve the now NSF-funded PGDD in both data content and user experience. We plan to include more plant genome sequences as they are published, and also integrating various physical and genetic marker information, to permit earlier but often important comparison within and between plants. The genome-level visualizations can be improved with better zooming and panning.

Integration of multiple data sources into the existing PGDD framework will enhance its value as a platform upon which to study many evolutionary questions. We have plans to integrate with the VISTA (Frazer et al. 2004), GEvo (Lyons et al. 2008) and other major databases. Recently, links to the PGDD query service are provided in the *Arabidopsis* information resource (TAIR) (Swarbreck et al. 2008) and *Populus* Genome Integrative Explorer (PopGenIE) (Sjodin et al. 2009), which allows easier access for many plant researchers working on different organisms.

2.5 Conclusion

The synteny and collinearity patterns between closely related plant lineages are clear, yet computational identification and enumeration of synteny blocks between relatively divergent plant lineages remain difficult. This is complicated by many rounds of shared and non-shared whole genome duplication events, subsequent DNA loss and other genomic rearrangements in specific plant lineages. I formulated the synteny identification problem and implemented score-based and likelihood-based method to infer the synteny blocks. The identified synteny patterns shed light on genome evolution and expansion of plant gene families. All data are publicly available in a NSF-funded database (PGDD) and will benefit the plant research community by providing a valuable platform to perform comparative and evolutionary genomics exploration.

CHAPTER 3 INFERENCE OF PALEO-POLYPLOIDY IN MAJOR PLANT LINEAGES

3.1 Introduction

Ancient whole genome duplications (WGD) are evident in lineages of fungi (Kellis et al. 2004), animals (Aury et al. 2006; Jaillon et al. 2004) and plants (Bowers et al. 2003b; Cui et al. 2006), offering opportunities for the evolution of new (Spillane et al. 2007) or modified (Hittinger et al. 2007) gene functions, altering gene dosages, and creating new gene arrangements. Reciprocal gene loss following WGD can contribute to reproductive isolation through divergent resolution of duplicate copies (Bikard et al. 2009), and foreshadow the diversification of species (Lynch et al. 2000; Scannell et al. 2006; Soltis et al. 2009). Although controversial, some studies also suggested a possible link between polyploidy and the likelihood of a plant lineage to survive mass extinction events (Fawcett et al. 2009; Van de Peer et al. 2009b), taking the evidence that many ancient polyploidy events in plants appeared to occur around a time close to Cretaceous-Tertiary (K-T) extinction events.

Nonetheless, the frequencies of polyploidy in plant lineages are indeed higher than in other lineages. For example, *Arabidopsis thaliana* has undergone three paleo-polyploidies, including two doublings (Bowers et al. 2003b) and one tripling (Jaillon et al. 2007), resulting in ~12 copies of its ancestral chromosome set in a ~160Mb genome. The two most recent paleo-polyploidies affecting *Arabidopsis* α and β, following the usage in (Bowers et al. 2003b), now appear to have occurred within the crucifer lineage (Jaillon et al. 2007; Ming et al. 2008). *Populus trichocarpa* (poplar) underwent a duplication specific to its own salicoid lineage (Tuskan et al. 2006) and shares only one of the three paleo-polyploidies (γ) affecting *Arabidopsis*. *Vitis vinifera* (grape) (Jaillon et al. 2007) and *Carica papaya* (papaya) (Ming et al. 2008), the latter within the same taxonomic order (Brassicales) as *Arabidopsis*, each have only

γ and no subsequent polyploidies (**Figure 3.1**). There is also a shared whole genome duplication event (ρ) predating the diversification of major cereal species including rice and sorghum (Paterson et al. 2004) (**Figure 3.1**). In addition to the genome sequences, recent analyses of ESTs in many basal angiosperm lineages suggest that virtually all angiosperms are paleopolyploids (Blanc et al. 2004; Cui et al. 2006), with the possible exception of the basal angiosperm *Amborella* (Cui et al. 2006).

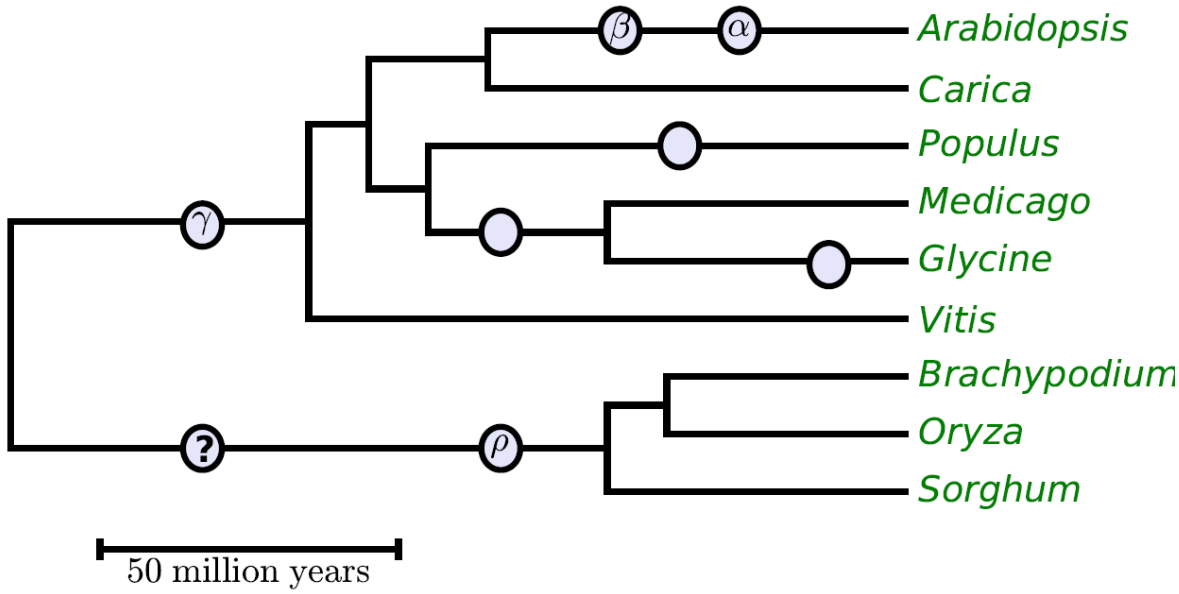

Figure 3.1: Currently known polyploidies in representative angiosperm lineages. The question mark is previously unknown monocot paleo-polyploidy events that are new findings in this work.

Traces from past whole genome duplication events can often be detected from pairwise synteny segments, including two sets of retained paralogs that have maintained relative genomic locations on syntenic chromosomes. In angiosperms, genome duplications are recurring in many lineages, generating large numbers of paralogous loci. Gene loss at duplicated loci fractionates ancestral linkage patterns and reduces the density of continuous stretches of "paleologous" gene pairs which are the remaining signatures of paleo-polyploidy (Thomas et al.

2006). Depending on the level of gene loss, the remaining signatures of duplication are sometimes so eroded that the homologous segments can no longer be identified based only on similarity to one another. The problem is multiplied when the species in question has undergone several genome duplications, with recent duplications tending to obscure synteny from more ancient events as is found in most angiosperm genomes. Such highly degenerate duplicated segments have been referred to as "ghost duplications", and can often be resolved by comparison to an appropriate outgroup genome that did not experience polyploidy or undergo massive gene loss (Van de Peer 2004). For example, bridging of ghost duplications using outgroups has clarified the history of polyploidy in both *Saccharomyces* and *Tetraodon* (Jaillon et al. 2004; Kellis et al. 2004; Scannell et al. 2007).

One partial solution for inferring ancestral gene orders in angiosperms has been a "bottom-up" approach, in which the most recently duplicated segments are interleaved to generate hypothetical intermediates that are further recursively merged (Aury et al. 2006; Bowers et al. 2003b). However, this approach requires an additional cycle of deductions for each duplication event and compounds any errors.

An alternative "top-down" approach requires only one cycle of deduction by simultaneously searching for and aligning all structurally similar segments across multiple genomes and subgenomes. The top-down approach should be more sensitive because it can incorporate transitive homology (Van de Peer 2004), in which segments *A* and *B* have undergone reciprocal gene loss and no longer show correspondence to each other but both correspond with a third segment *C*. Relationships among such degenerated duplicated regions, easily missed by a bottom-up approach, can often be resolved by comparison to another genome that does not have the duplication or that underwent independent gene loss. Such comparisons have clarified synteny among yeast species (Kellis et al. 2004). The top-down approach is conceptually more attractive in that it only requires one cycle of deduction – first searching for pairwise synteny and then combining the resulting pairs to form a multi-way correspondence

among structurally similar chromosomal segments. The efficacy of the top-down approach, however, depends on the searching strategy because of the degenerate synteny resulting from post-duplication gene loss.

In this chapter, I continue the discussion of the synteny and collinearity identification problem, but now with a focus on the problem of identifying ancient polyploidy in plants. In particular, the above two methods are repeatedly used in my research – "bottom-up" (intra-genomic) and "top-down" (inter-genomic) method. The dual methods are complementary and both useful to reveal the more ancient duplications in eudicot and the monocot lineages. In eudicots, the most ancient duplication is now known to be a hexaploidy event (called γ) (Jaillon et al. 2007; Ming et al. 2008) while in the monocots I call the more ancient duplications σ (Tang et al. in review), and likely two genome doublings. Understanding these more ancient events is essential to uncover the correspondence between eudicot and monocot genomes, which I discuss in the last section of the chapter.

3.2 Characterization of a paleo-hexaploidy event in the eudicot lineage

3.2.1 Patterns of synteny conservation across several eudicot genomes

Using the gene order alignment software MCscan, we can show a high degree of collinearity between *Arabidopsis*, *Carica* (papaya), and *Populus* (poplar) (Ming et al. 2008). Application of the MCscan algorithm to the *Vitis* genome validated the reconstructed order and inferred triplicated structure of a common *Arabidopsis-Carica-Populus* ancestor. *Vitis* is a eudicot outside of the two eurosid clades that contain *Arabidopsis-Carica* (eurosids II) and *Populus* (eurosids I) (Soltis et al. 2005), therefore providing an independent lineage suitable to test the gene order alignments. When the *Arabidopsis-Carica-Populus* consensus is aligned to *Vitis*, the two independently inferred triplication patterns correspond closely (**Figure 3.2**). Thus, top-down gene order alignment revealed genome triplication that eluded prior detection in *Arabidopsis* (Bowers et al. 2003b) and *Populus* (Tuskan et al. 2006) and also supported the

conclusion that the triplication occurred in a common ancestor of *Vitis*, *Arabidopsis*, *Carica*, and *Populus* (Ming et al. 2008).

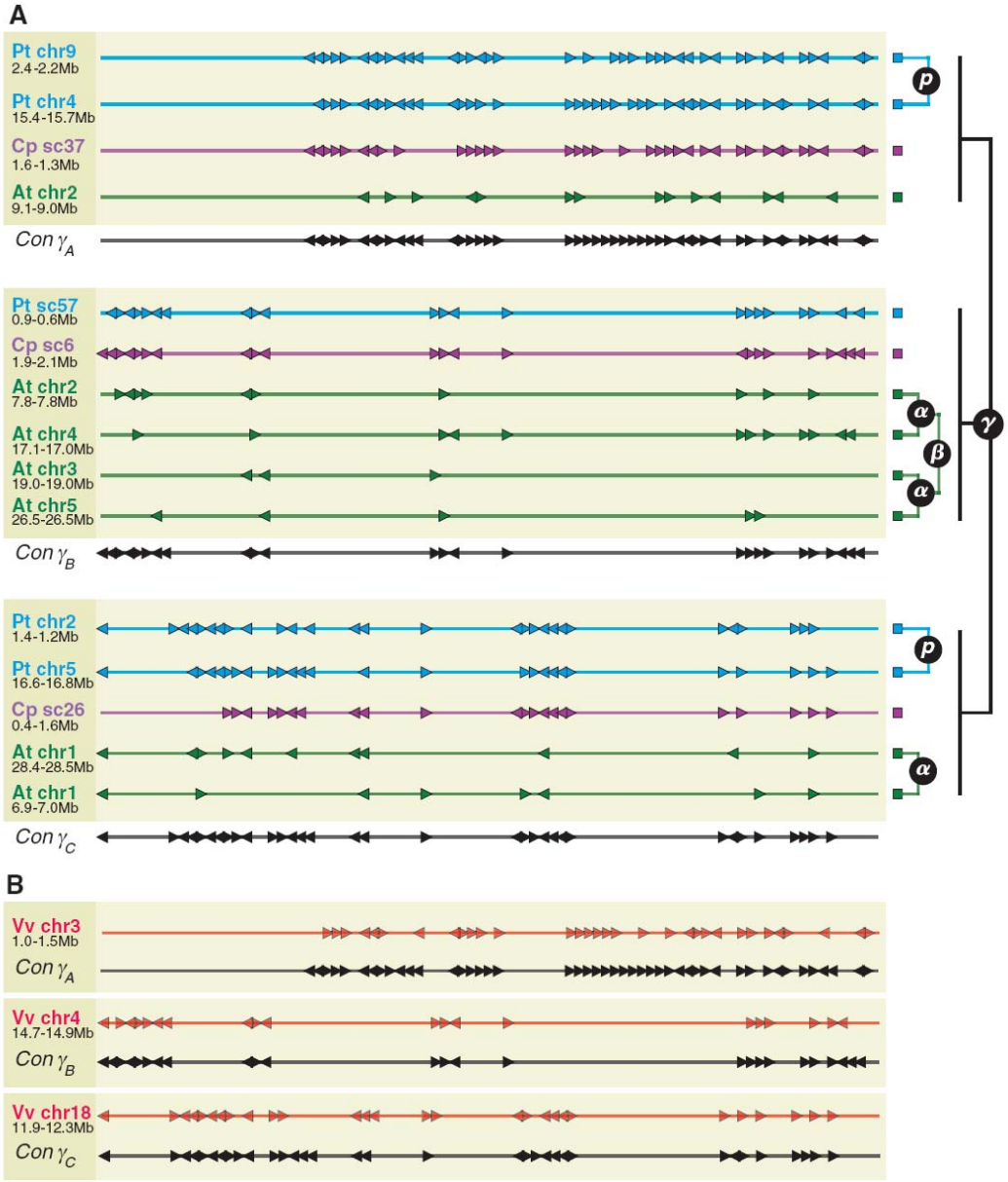

Figure 3.2: Typical view of multiple collinear regions among several eudicot genomes, affected by many rounds of polyploidy. Triangles represent individual genes and reflect their transcriptional orientations. Genes with no syntenic matches to the selected regions are not plotted. (A) Alignment among *Arabidopsis* (*At*), *Carica* (*Cp*), and *Populus* (*Pt*) chromosomal regions. The regions are grouped into three consensus γ-subgenomes (Con γ_A, γ_B, γ_C) on the basis of parsimony. Aligned genes within each γ subgenome are merged into an inferred order by consensus. (B) The inferred γ partitions are validated with the *Vitis* genome (*Vv*) because each γ-subgenome clustered in (A) has only one closely matching *Vitis* chromosomal region.

In addition to the triplication of chromosomal regions, the triplication of gene loci is also evident from the analyses in **Table 3.1**. For example, we found that 88 aligned loci in *Carica* have multiplicity levels of three (triplication γ), with only one aligned locus exceeding a multiplicity of three; 54 aligned loci in *Populus* have the expected multiplicity level of six (triplication γ × duplication *p*), but only 3 loci exceed six. The loci that exceed the expected multiplicity level are likely produced by additional small-scale (single gene or segmental) duplications in each lineage.

Table 3.1: Number of clustered groups of genes at different multiplicity levels in five angiosperm species. The statistics are based only on groups that contain genes from at least two different species, as constructed from syntenic alignments. (*) denotes expected level of multiplicities for *Carica*, *Populus*, *Vitis*; The multiplicities for *Arabidopsis* is 12 (yet no gene groups retained all 12 copies), and equivocal for *Oryza*.

Species	Multiplicity level							# of genes (%)
	1	2	3	4	5	6	≥7	
Arabidopsis	6742	2642	868	282	80	32	13	16451 (61%)
Carica	9118	942	88*	1	0	0	0	11270 (44%)
Populus	5147	6362	763	618	96	54*	3	23457 (51%)
Vitis	9926	1671	239*	15	2	0	0	14055 (46%)
Oryza	2197	685	140	35	9	2	0	4184 (14%)

3.2.2 Further circumscribing the γ duplication event

The γ duplication event was dated to have occurred after the monocot-dicot separation but before the expansion of the rosids (Jaillon et al. 2007). We investigated the lower boundary of this claim by sampling genomic regions from other eudicots outside the rosids for which long,

contiguous sequences (BACs) were available in GenBank, including tomato (*Solanum lycopersicum*) and banana (*Musa acuminata*).

We first mapped tomato unigenes onto 194 sequenced tomato (*Solanum lycopersicum*) BACs as preliminary gene annotations and inspected synteny to *Vitis*. Among the 78 *Solanum* BACs that have more than 10 distinctively mapped unigenes, 72 have more than 50% of genes showing primary synteny to a single *Vitis* chromosome. Each individual tomato BAC corresponds closely to only one of the triplicate regions rather than showing equal matches to each of the three γ paleo-homeologous chromosomes in *Vitis*. **Figure 3.3** shows one example of a *Solanum* BAC that aligns to the *Vitis* gene order. Although the *Solanum* BACs that we inspected only represent about 2.5% of the genome (*Solanum* genome was estimated to be ~1000Mb), the evidence so far strongly supports the hypothesis that the γ triplication occurred in a common ancestor of asterids and rosids. Under this scenario, each *Solanum* segment would be expected to have up to four primary syntenic segments in *Arabidopsis*, as has been suggested (Ku et al. 2000).

We also examined synteny to *Vitis* for chromosomal regions from a monocot species that is basal to the cereals – banana (*Musa acuminata*). On average, the levels of synteny between *Musa* BACs and *Vitis* chromosomes are 50% lower than synteny between *Solanum* and *Vitis*, reflecting the longer evolutionary distance of *Musa*-*Vitis*. Furthermore, in contrast to the one-to-one primary synteny pattern of *Solanum* and *Vitis*, *Musa* BACs show roughly equal matches to any of the three γ-homeologs in *Vitis* (**Figure 3.3**), a pattern similar to *Oryza*-*Vitis* (Jaillon et al. 2007). However, failure to detect one-to-one (as opposed to one-to-three) correspondence between monocot regions and *Vitis* cannot be viewed as strong evidence that γ occurred after the eudicot-monocot split. An alternative but equally plausible scenario is that the monocots and eudicots share γ but diverged soon after γ occurred. Under this scenario, the gene arrangements between two orthologous chromosomes would share very little synteny because of stochastic,

independent gene losses in both lineages – leading to similarly-low levels of correspondence of chromosomes in one taxon to each of its three γ paralogs in another taxon.

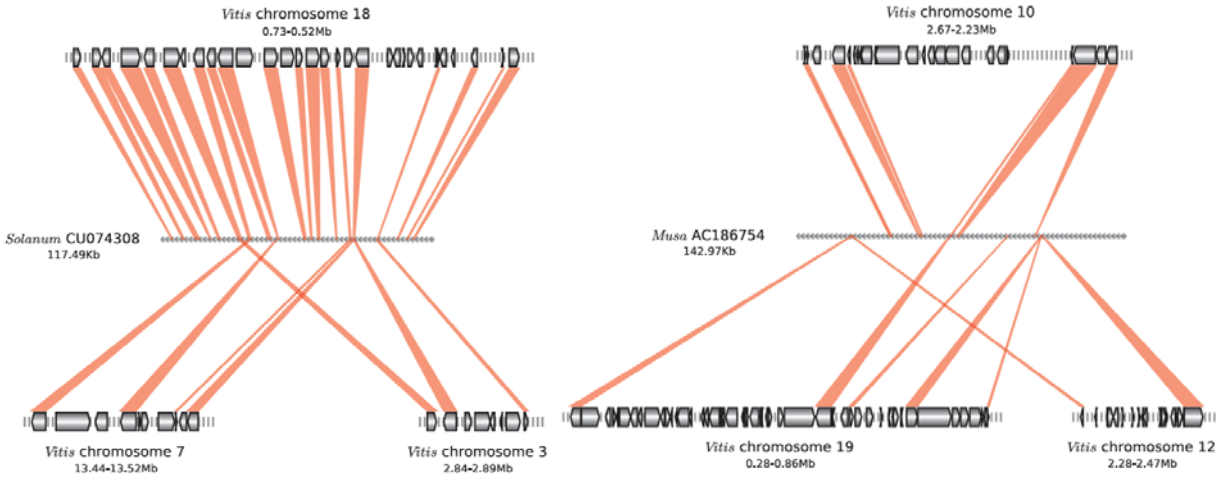

Figure 3.3: Collinearity between triplicate *Vitis* γ-homeologous regions with BAC sequences from *Solanum* (left) and *Musa* (right). Black glyphs represent genes with the tip showing the transcriptional direction, red shades represent synteny matches between a *Vitis* gene and *Solanum* or *Musa* sequences.

While highly-specific one-to-one synteny is indicative that two lineages share the γ triplication, frequent one-to-three synteny is *not* necessarily indicative that one lineage lacks the triplication. So far we can only confidently place the γ triplication before the asterid-rosid split and consider the status of the paleo-hexaploidy in the monocot lineage to be unclear. It is still difficult to test the hypothesis that the γ triplication predated the divergence of monocots and eudicots. For example, additional data from an outgroup genome such as *Amborella* would help, but does not necessarily solve the placement of the triplication if γ is found absent in that outgroup. Much of the uncertainty is rooted in the fact that the γ triplication is an ancient event that at least predated the asterids-rosids, and comparisons across this evolutionary distance are often less effective. Therefore we need broader and more judicious sampling of plant taxa. Indeed, fortuitous discoveries of genomes like grapevine that have close-to-ancestral karyotypes facilitate comparisons across major angiosperm clades. Similarly, additional karyotypically-

conserved monocot or basal angiosperm genomes that are free of recent polyploidies might better elucidate the scenario.

3.2.3 Rate variations between paleologs within four eudicot species

Deduction of a consensus gene order for multiple taxa permits us to directly compare estimates of the ages of gene duplications based on rates of nucleotide substitution per synonymous site (*Ks*) between paleolog pairs (syntenic paralogs), filtering out the inevitable influence of "background" (i.e. single gene) duplications which superimpose an L-shaped curve on the relics of whole-genome duplications (Blanc et al. 2004; Cui et al. 2006). By excluding the single gene duplications, we were able to analyze the *Ks* distributions with less ambiguity.

The actual distributions of *Ks* between paleologs can be modeled as mixtures of log-transformed exponentials and normals, representing single gene duplications and whole genome duplications, respectively (Cui et al. 2006). Since I excluded all the single-gene duplication, the *Ks* distributions that I derived can be modeled as mixtures of log-normal components representing multiple rounds of genome duplications, using the EMMIX software (http://www.maths.uq.edu.au/~gjm/emmix/emmix.html). *Ks* values that are less than 0.005 were discarded to avoid fitting a component to infinity (Cui et al. 2006), and the mixed populations were modeled with one to five components. We selected one best mixture model for each paleolog distribution based on Bayesian Information Criterion (BIC) (Cui et al. 2006).

Although γ apparently occurred in a common ancestor of *Carica*, *Populus* and *Vitis*, the median *Ks* between *Vitis* γ-paleologs (1.22) is much lower than that of *Carica* (1.76) and *Populus* (1.54) (**Figure 3.4A**; **Table 3.2**). The median values of *Ks* among γ duplicates in these three genomes show highly significant difference (Kruskal-Wallis one-way ANOVA, $P=2.25\times10^{-142}$).

The *Ks* distributions analyzed with mixture models show the expected number of components for each species, except for *Arabidopsis*, where we can find only two instead of three distinct components. This two-peak distribution (**Figure 3.4B**) is similar to the results of

a previous study (Maere et al. 2005) even though MCscan provides better deductions about the identities of paleologs. We postulate that more rapid substitutions occur at synonymous sites in *Arabidopsis* than in the other three eudicot species, with *Arabidopsis* γ paleologs being saturated with synonymous substitutions. Therefore within *Arabidopsis*, *Ks*-based distances between paralogs cannot differentiate γ duplicates from either the tail of the distribution of β duplicates, or from noise, or both. The median *Ks* values between *Arabidopsis* β and γ duplicates are close to saturation (2.00), much larger than those of the γ-duplicates in the other three species. Repeating the analysis using a more conservative genetic distance – transversion rate at four-fold degenerate sites (4DTV) (**Figure 3.4C**) shows almost the same pattern as using *Ks*, suggesting that the saturation effect of DNA substitutions may have also similarly affected 4DTV distance.

Table 3.2: Mixture model estimates for distributions of *Ks* between paleologs in each species.

Species	Sample size	# of mixture components	Median	Variance	Proportion
Arabidopsis	7435	2	0.86 2.00	0.08 0.20	0.51 0.49
Carica	907	1	1.76	0.32	1
Populus	13113	2	0.27 1.54	0.01 0.24	0.62 0.38
Vitis	2288	1	1.22	0.16	1

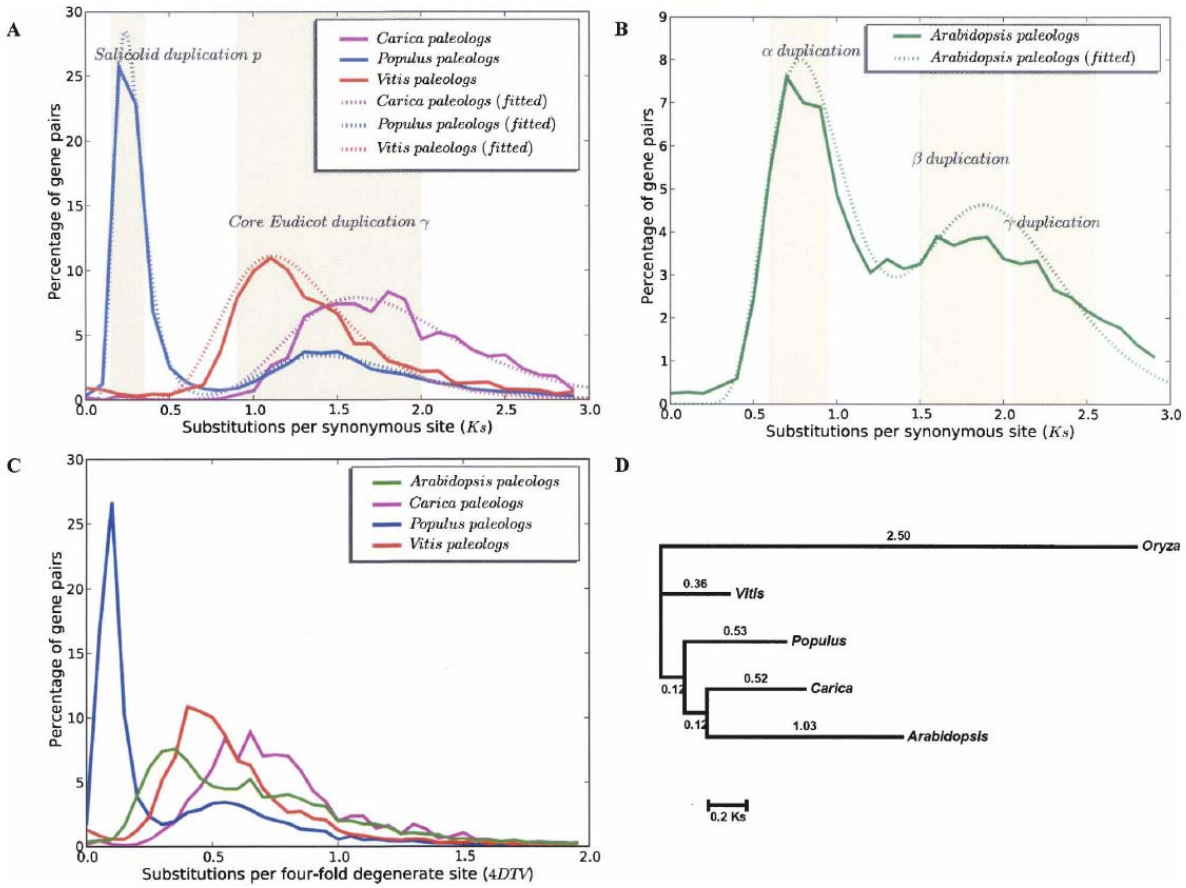

Figure 3.4: *Ks* analyses of homologous genes. (A, B) Distribution of *Ks* distances among *Carica*, *Populus*, *Vitis* and *Arabidopsis* paleologs. *Ks* values are grouped into bins of 0.1 intervals. Certain *Ks* intervals are highlighted as they correspond to several presumed whole genome duplication events. Dotted lines are fitted mixtures of log-normal distributions for the paleolog *Ks* distributions. (C) Distribution of 4DTV distance among paleologs in the same four eudicot lineages. (D) Phylogeny of single-copy ortholog set used in relative rate estimates. A total of 47 orthologous genes that are single copy in all five species were used in the analysis. Protein alignments for each ortholog group were constructed and then used to guide DNA alignments. The alignments are then concatenated, with 53856 aligned nucleotide positions. Per-site *Ks* values on each branch were estimated by *codeml* in PAML package (Yang 2007) using a constrained topology that reflects organismal relationships.

Differences in the median values of distances between the paralogs that are derived from the common γ event can be explained by different substitution rates among the four rosid lineages. We constructed a phylogenetic tree with per-branch *Ks* estimates, based on orthologous gene groups that are strictly single copy in all five species (**Figure 3.4D**). The same trend was found, with increasing evolutionary rates in branches leading to *Vitis*, *Populus*, *Carica*

and *Arabidopsis* respectively, suggesting that the variations of substitution rates are not confined to populations of duplicate genes but are rather lineage-specific. A similar range of nuclear rate variation in flowering plants has been documented in previous studies, and is often associated with life history (Gaut et al. 1996; Koch et al. 2000; Smith et al. 2008). The short generation time in the annual *Arabidopsis* might have contributed to the fast substitution rates compared with *Populus* or *Vitis* which are perennials. In general, this correlation between rate of molecular evolution and life history is commonly observed in major clades of flowering plants (Smith et al. 2008).

Because substitution rates vary among lineages, timing of duplication or speciation events is hard to determine using genetic distance measures alone. For the same reason, dating of ancient events based on phylogenetic trees (Bowers et al. 2003b; Tuskan et al. 2006) can produce incongruous results since the drastic differences in rates may lead to incorrect trees that are artifacts due to long-branch attractions (Felsenstein 2004).

One phylogenetic model placed *Vitis* within the eurosid I clade (Jaillon et al. 2007), in contrast with the prevailing view of the Vitaceae as sister to both eurosid I and eurosid II (Davies et al. 2004; Soltis et al. 2005). Indeed, *Populus* and *Vitis* do show small *Ka* or *Ks* values for substitutions between inferred orthologs (**Table 3.3**). However, the seemingly smaller distance between *Populus* and *Vitis* genes should be interpreted with caution since both species appear to have relatively slow evolutionary rates. The striking differences in evolutionary rates among these taxa at the DNA sequence level, may in part explain the controversial placement of *Vitis* inside the eurosids by some workers (Jaillon et al. 2007). Indeed, we found that if we use *Arabidopsis* as reference point, the increasing *Ks* distances from *Carica*, *Populus* and *Vitis* appear to support the view that *Vitis* is an outgroup to the rosids (**Table 3.3**).

Table 3.3: *Ks* and *Ka* values for syntenic orthologs of five sequenced plant genomes. For each syntenic group, the smallest *Ks* or *Ka* value among all orthologous pairs was retrieved to represent the value. The lower triangle shows median *Ks* values and the upper triangle shows median *Ka* values. Numbers in brackets correspond to the number of syntenic groups used in each comparison. *Ks* values between *Oryza* and four eudicots show saturated substitutions and high variances, therefore should not be considered reliable estimates and are excluded.

	Arabidopsis	*Carica*	*Populus*	*Vitis*	*Oryza*
Arabidopsis	--	0.24	0.23	0.25	0.37
Carica	1.57 (6913)	--	0.17	0.19	0.35
Populus	1.64 (8366)	1.08 (8504)	--	0.16	0.31
Vitis	1.72 (7381)	1.12 (7920)	0.98 (10143)	--	0.32

3.3 Characterization of multiple polyploidy events in grass lineage

3.3.1 Current knowledge of ancient WGD in the grass lineage

It is well established that one WGD (hereafter denoted as **ρ**) occurred in the cereal lineage an estimated 70 million years ago, and is thought to have preceded the radiation of the major cereal clades by 20 million years or more (Paterson et al. 2004; Wang et al. 2005). "Quartet" comparisons of the two resulting paralogous (homoeologous) chromosomal regions in rice and sorghum show that 99% of post-duplication gene losses are orthologous (Paterson et al. 2009), consistent with the ρ event predating the diversification of major grass lineages (Paterson et al. 2004; Salse et al. 2008). This suggests that rice-sorghum gene arrangements are probably representative of those of most grass genomes, albeit in some lineages modified by additional cycles of duplication and gene loss. The ρ duplication is extensive, involving all modern chromosomes of rice and sorghum and covering much of the euchromatin (Bowers et al. 2005; Paterson et al. 2009). Even one duplicated block previously thought to be recent and segmental

appears to also result from ρ with subsequent concerted evolution (Paterson et al. 2009; Wang et al. 2009).

While several studies (Jaillon et al. 2007; Salse et al. 2008; Zhang et al. 2005) have hinted that additional monocot duplications may have predated ρ, the extent of such earlier duplications has not yet been elucidated. Inferences of more ancient polyploidy based on inspection of amino acid differences between duplicate genes (d_A) (Zhang et al. 2005) are affected by varying substitution rates among different gene families (Bowers et al. 2003b). A recent study identified 29 duplications in the rice genome including 19 minor blocks that overlap with 10 major blocks (Salse et al. 2008), but did not systematically study these segments in a hierarchical context to reflect their evolutionary history.

3.3.2 Quartet alignments among rice and sorghum gene orders (ρ-blocks)

To facilitate WGD analysis, we first compiled a list of syntenic gene quartets from rice and sorghum, showing both orthologous and ρ-paralogous matches. A total of 9 large segmental duplications attributed to the ρ-genome duplication were analyzed using previously described block identifiers (Paterson et al. 2004). The boundaries of ρ blocks are highlighted in a rice intra-genomic dot plot (**Figure 3.5**). Indeed, these 9 ρ-blocks correspond to the 9 blocks identified in (Paterson et al. 2004) and agree with 9 of 10 major blocks described in (Salse et al. 2008). We consider one block involving chromosomes 4-10 in Salse et al. (Salse et al. 2008) to overlap with both ρ2 and ρ5, indicating an origin more ancient than ρ.

Each ρ-block merges two regions of rice and two regions of sorghum into a single gene order that approximates the genome composition prior to the ρ duplication. In summary, the ρ-order collapses 15640 rice genes and 15636 sorghum genes into 13308 ρ-nodes (~50% of the rice and sorghum transcriptomes), excluding tandemly duplicated genes. The incorporation of sorghum gene orders into the ρ-blocks validate the previously identified blocks in rice while offering more resolution of a few duplicated regions that are reciprocally silenced in rice or

sorghum. This reconstruction of pre-ρ gene order is intended to computationally reverse post-ρ gene loss, increasing the sensitivity of subsequent analysis. We emphasize that this order is only an approximation, since the ancestral positions of the intervening singleton genes between consecutive pairs of ρ-paralogs cannot be precisely determined. Nonetheless, we show below that this intermediate order is useful to mask post-ρ events and infer the structure of more ancient blocks.

3.3.3 Pre-ρ duplications in the cereal lineage (σ-blocks)

The **σ**-blocks (involved in duplication events prior to ρ) were identified through further bottom-up reconstruction, similar to the procedure in (Bowers et al. 2003b). Reconstructed ρ-orders of 13308 ρ-nodes from the previous step were compared among themselves, revealing collinear patterns of correspondence that involve all major ρ-blocks (**Figure 3.5**). Some collinear patterns between pairs of ρ-blocks are one-to-one; while others (σ2, σ4 and σ5) involve more than two ρ-blocks, suggesting that additional duplications have been identified. In this step, we curated a second list of 8 large σ-blocks that have retained collinearity following σ. These blocks contain a total of 4168 σ-nodes, covering 5747 rice genes and 5738 sorghum genes (~20% of the rice and sorghum transcriptomes). It is difficult to exhaustively enumerate all patterns of σ collinearity, since some duplicated regions become highly degenerate during post-WGD diploidization, creating gene orders that are largely reciprocal or sometimes complementary (Thomas et al. 2006; Van de Peer 2004).

The bottom-up approach, starting from the modern gene order to deduce ρ- and σ-orders, offers inherent hierarchical structures that reflect the relationships among chromosome segments. Collinearity is well retained and anchored gene pairs, including rice-sorghum orthologs, ρ-paralogs and σ-paralogs often retain consistent transcriptional orientations. Nonetheless, gene losses (due to fractionation) are extensive, particularly across the σ duplication (between the two ρ-blocks) where there are the fewest corresponding genes.

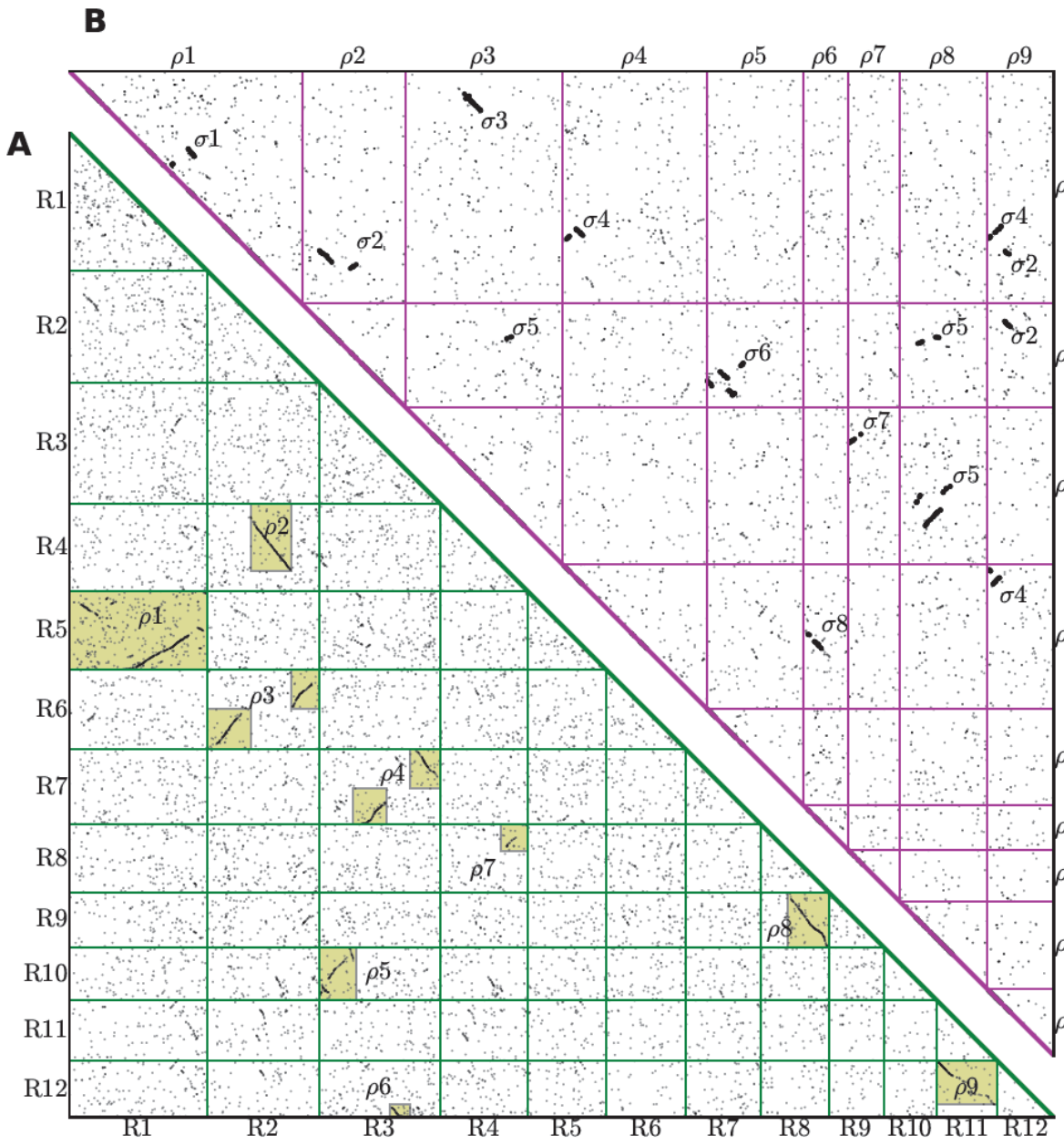

Figure 3.5: Illustration of bottom-up reconstruction of ρ-blocks and σ-blocks. (A) Classifications of ρ duplicated blocks are visualized in the lower left triangle. (B) During the second iteration, the paired hits are each converted into ρ-nodes and then plotted in the upper right triangle. Gene positions are in their rank orders along the chromosome (A) or the reconstructed ρ-order (B).

3.3.4 Genetic distances of the gene pairs

Synonymous nucleotide substitutions per synonymous site (*Ks*) for the groups of orthologs and paralogs from different events (ρ and σ) were observed to be well separated (**Figure 3.6**). However, variations in GC-content of cereal genes can impact *Ks* calculations, with different algorithms generating differing estimates of *Ks* values for pairs involving genes with high third codon position GC content (GC_3) (Shi et al. 2006). In light of such complications, we focus on gene pairs with average GC_3 less than 75%.

Rice-sorghum orthologs show a sharp *Ks* peak (median 0.62) consistent with previous estimates (Paterson et al. 2009). The population of ρ-paralogs from both rice and sorghum show a major peak at *Ks* 0.94, along with a small peak at *Ks* ~0.15 resulting from concerted evolution of the terminal part of ρ9 (Wang et al. 2007; Wang et al. 2009).

Paralogs derived from the more ancient σ duplication(s) show a well defined peak around much older *Ks* (median 1.72), and with a larger variance than that of other groups. Based on a commonly used molecular clock estimate of 6.5×10^{-9} synonymous substitutions per synonymous site per year (Gaut et al. 1996), the σ duplications are estimated to have occurred approximately 130 million years ago. Since the *Ks* values for many σ-paralogous pairs are almost saturated and there are substantial uncertainties in the calibration of the molecular clock (Hedges et al. 2004; Vicentini et al. 2008), this date can only be considered a rough estimate.

Decomposition of mixed *Ks* distribution into different event groups explains why previous studies were not able to identify the σ event (and to some extent also the ρ event) based solely on the *Ks* distribution of ESTs (Blanc et al. 2004). Several analyses relied on curve-fitting methods to find multiple duplication events based on *Ks* distributions (Cui et al. 2006; Tang et al. 2008b). The combined set of ρ and σ paralogs show a distribution with the mixed peak extending from 1.0 to 2.0, which becomes easily separable into distinct components using our synteny-based classifications. Synteny-based classifications of gene pairs also remove the *L*-

shaped component resulting from recent single gene duplication events in the *Ks* plot (Tang et al. 2008b).

Judging from the *Ks* distribution, both distances between ρ and σ duplicates appear bounded between rice-sorghum orthologs and grape-cereal orthologs (**Figure 3.6**), suggesting that the relative timing of these WGDs might be between cereal diversification and monocot-eudicot divergence. Indeed, the distances between grape-cereal orthologs (median *Ks* 1.95) are higher than those between the cereal paralogs from σ duplications (P=4.8×10^{-24}, student's *t*-test). However, differences in lineage-specific mutation rate between grass and grape confound interpretation of *Ks* values and we re-emphasize that our divergence time estimates must be considered rough approximations. Initial interpretation of the *Arabidopsis* β WGD duplication provides a cautionary example – *Ks* analyses of duplicated genes suggested that the β-duplication predated the divergence of *Arabidopsis* and *Carica* but analyses of blocks of genomic sequence indicated that the β-duplication occurred after the divergence of lineages leading to these two species (see previous section in this chapter).

We again caution the interpretations based on *Ks* distributions, as already noted in (Tang et al. 2008b; Van de Peer et al. 2009a). Large *Ks* estimates have large variances, therefore although σ (1.72) and cereal-grape (1.95) differ by 0.2 *Ks*, the peaks still look overlapping (**Figure 3.6**). We consider this difference real, for two reasons: 1) for cereal-grape ortholog *Ks*, the estimate is conservative (we calculated reciprocal best matches only); 2) the alternative (i.e. the sigma duplicates have same age or older than cereal-grape) implies that sigma is in the grape lineage, which we consider unlikely; otherwise we should already see this in the later PAR analysis.

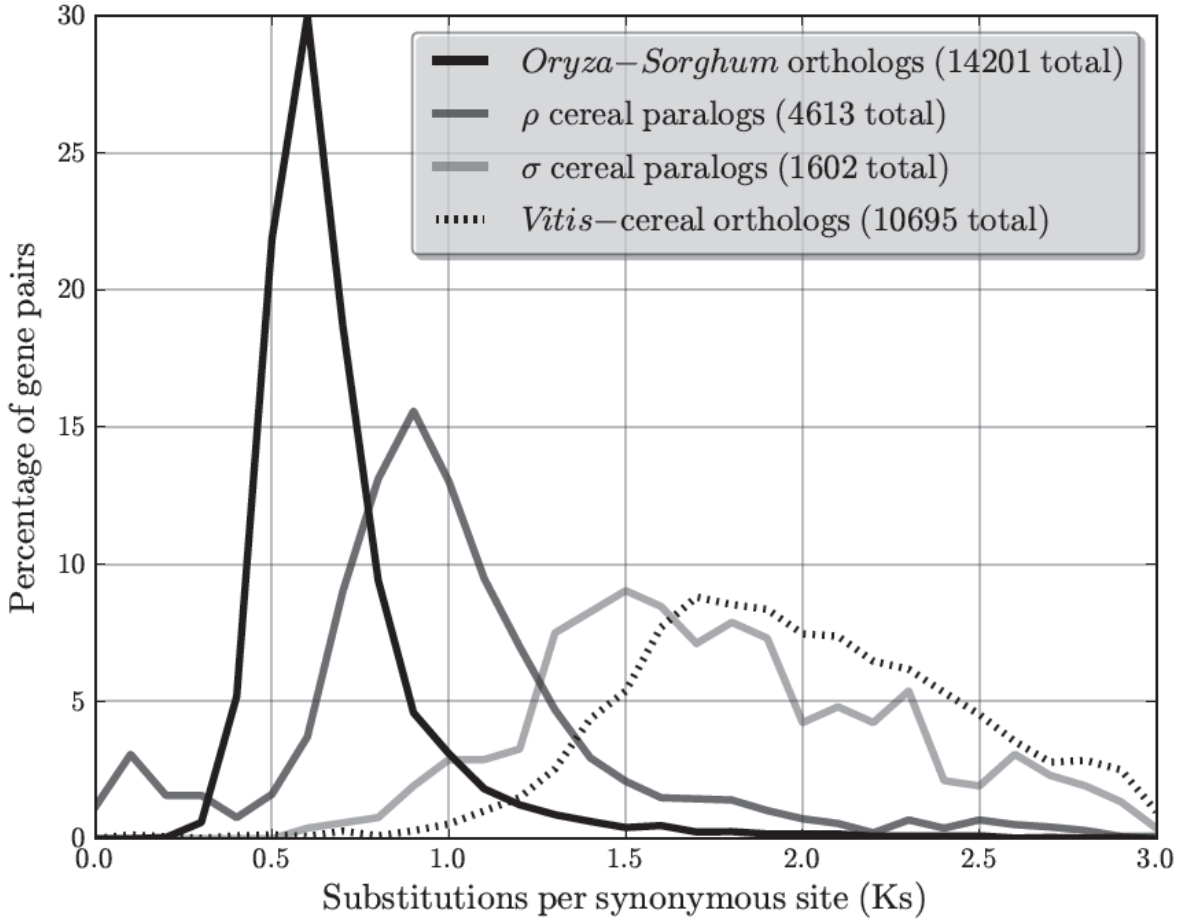

Figure 3.6: *Ks* distributions for rice-sorghum orthologs, cereal WGD paralogs (ρ and σ paralogs) and grape-cereal orthologs. The recent ρ paralog pairs (rice-rice, sorghum-sorghum pairs) are readily derived from the ρ-nodes. However, σ-nodes contain several possible paralog pairs. To calculate *Ks* for σ paralogs, we include all paralog pairs within rice and sorghum (but exclude the ρ pairs). Cereal-grape orthologs are inferred from reciprocal best hits in rice-grape BLAST or sorghum-grape BLAST.

3.4 Comparisons between eudicot and monocot genomes

3.4.1 A novel hierarchical clustering method for deep synteny inference

Similarities between monocot and eudicot genomes resulting from common ancestry have been obscured by many rounds of paleo-polyploidy and numerous genome rearrangements (Jaillon et al. 2007; Liu et al. 2001). To compare monocot and eudicot genomes, we apply a hierarchical clustering approach that partially circumvents such difficulties to identify synteny across grape

and rice. Briefly, the chromosomes were first cut into small segments and comparisons were made between every pair of rice and grape segments. For example, assume we have rice segments *O1* and *O2*, grape segment *V1*, and comparisons *O1-V1* and *O2-V1* show a significant number of homologs. Based on this information, *O1* and *O2* can be clustered together, because they both match the same grape region(s). In this approach, only the "dense" (syntenic) portions of the whole-genome dot plot are clustered, assembled and interpreted; the "sparse" (non-syntenic) portions are ignored from further analysis (**Figure 3.7**). Part of the method is inspired by the methodology used in the analysis of sea anemone and amphioxus genome (Putnam et al. 2008; Putnam et al. 2007). The whole analysis, streamlined in a set of computer programs, follows three major steps as detailed below.

Filtering of the matching set. We first scanned for tandem gene families, defined as clusters of genes within 10 intervening genes from one another, and kept the longest peptide. Next, we used *c*-value filtering to exclude weak peptide matches. The *c*-value is defined as $c(x,y) = b(x,y)/max(b(x,z) \text{ for } z \text{ in } Y \text{ or } b(w,y) \text{ for } w \text{ in } X)$, for each BLAST hit between peptide *x* in genome *X* and peptide *y* in genome *Y*. The *c*-value generalizes the concept of mutual best hit, as the mutual best hit would have a value of 1 (Putnam et al. 2008). We used *c*-value cutoff of 0.7, which implies that we excluded matches that are less than 70% similar to the best match in either genome. The filtered BLASTP results contain 35386 matches between 14982 grape genes and 15395 rice genes. The genes were re-indexed according to the rank order on each chromosome.

Segmentation of Chromosomes and scaffolds. BLASTP matches within 40 Manhattan distance units were clustered as first-pass evaluation of syntenic blocks, and as before we kept the blocks with more than 10 gene pairs. The start and stop boundaries of the first-pass syntenic blocks were used as indications of the breakpoints which disrupt otherwise even distributions of homologues. The chromosomes or scaffolds in both genomes were segmented into "atomic" intervals using a sweeping line algorithm that identifies the breakpoints. A total of 180 and 266

"atomic" segments were identified in grape and rice, respectively, including the breaks created by chromosomal or scaffold ends. Such segments are less affected by genome rearrangements and suitable for defining simple synteny patterns.

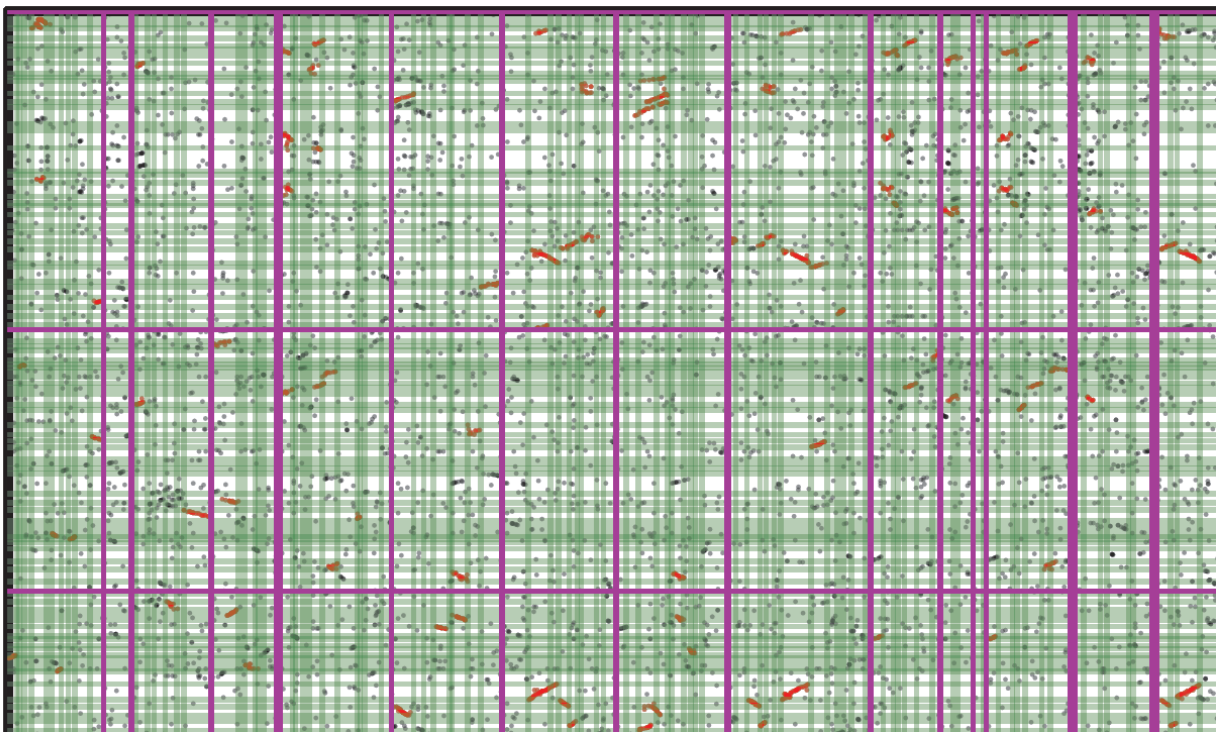

Figure 3.7: Example of chromosomal segmentation. We start with a raw dot plot between two genomes, with the magenta lines as the chromosomal ends and dots are the matching gene pairs. We identify the diagonal patterns and then compute the breakpoints and segment the genome into disjoint regions that are less affected by genomic rearrangements.

Clustering of segments free of rearrangements into PARs. The segments from grape and rice identified above were compared in a pairwise manner and homologue concentration score (Putnam et al. 2008) were calculated using $-log(p)$, where p is the probability of observed number of homolog pairs as modeled by a Poisson distribution. The log-likelihood scores for each segment against segments in another genome were used as the homologue distribution profile. With Pearson correlation coefficient between the profiles as the distance metric, we

hierarchically clustered the segments using average linkage method. The final clusters were defined at cutoff of r=0.3, as selected by visually inspecting the resulting clusters (**Figure 3.8**).

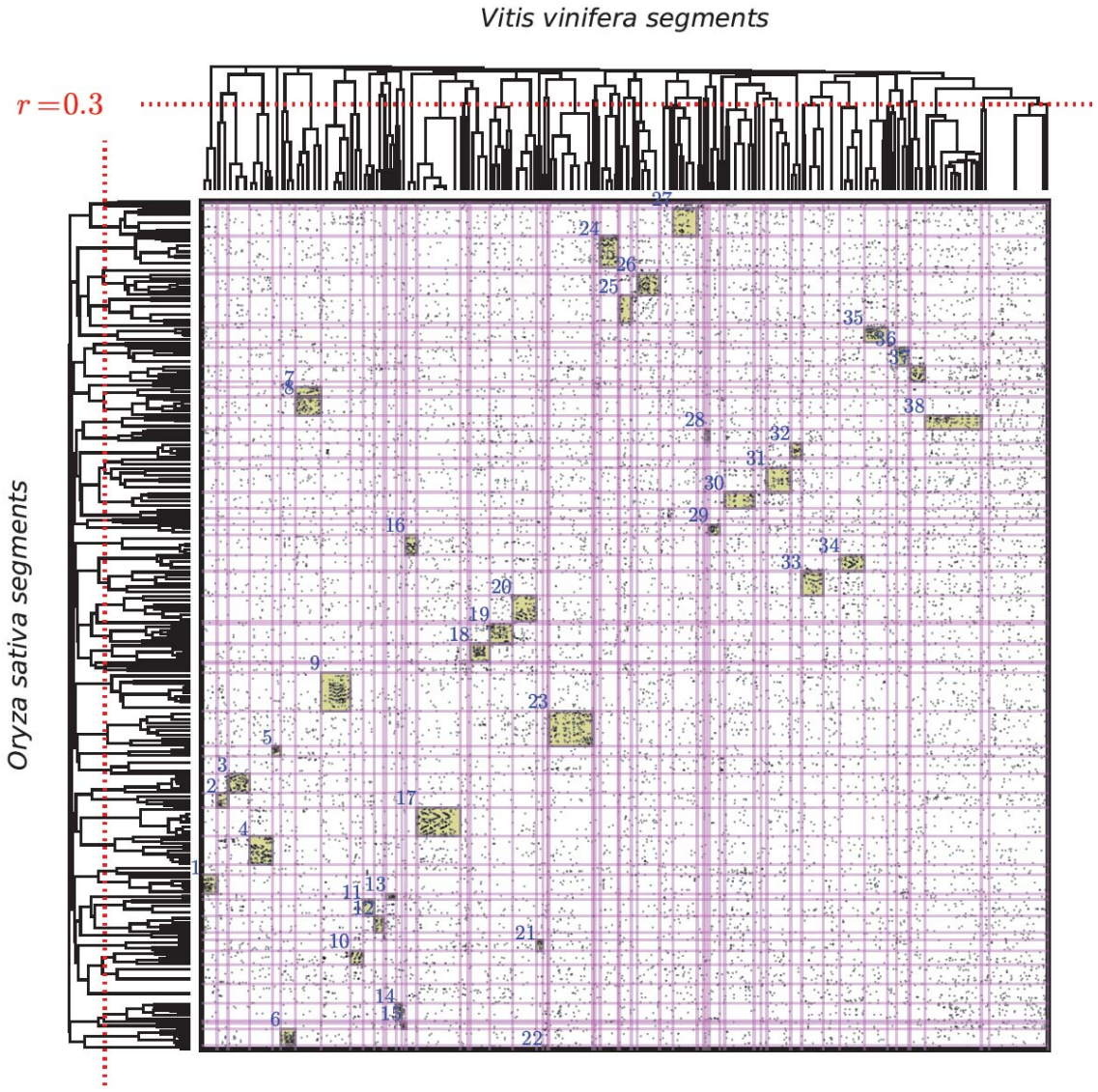

Figure 3.8: Hierarchical clustering method for constructing putative ancestral regions (PARs). Dots represent pairs of homologous genes between grape and rice. The chromosomal segments in the two genomes are reordered by the hierarchical clusters and concatenated. The trees on the top and left of the plot represent average correlation coefficient (r) among clustered grape and rice segments. Horizontal and vertical lines separate clusters of grape and rice segments, as defined by having an average correlation coefficient of distribution of hits to the other genome greater than 0.3 (i.e. r=0.3 as the cutoff of the trees). The squares highlighted in yellow are the 38 PARs that show high density of gene pairs between grape and rice clusters, with the PAR identifier shown in the upper-left corner of each highlighted block.

We finally get down to 56 and 56 reconstructed regions in the grape and rice genomes, respectively. Significant synteny between the reconstructed regions was finally evaluated using the Poisson distribution by summing the likelihoods of observing as many or more gene pairs under the null hypothesis of these pairs occurring randomly. For all pairwise comparisons in grape and rice, we kept 38 blocks that were significantly enriched for homologs ($P<1\times10^{-10}$), employing this stringent cut-off to limit consideration to particularly strong synteny. These 38 blocks were finally referred to as "putative ancestral blocks" (PARs), with a unique PAR identifier assigned to each.

3.4.2 Effective comparisons between cereal and eudicot genomes through PARs

Based on our unique clustering approach, duplicated segments retained in grape following the eudicot γ hexaploidy event (Jaillon et al. 2007), and homologous segments retained in rice following at least two rounds of duplication (ρ and σ), contain 38 "putative ancestral regions" (PARs). Each PAR consists of regions that show high density of homologs ($P<1\times10^{-10}$). The PARs collectively explain 19.1% of all observed homolog pairs and 31.0% of reciprocal best hits between grape and rice genes, although by chance they should only explain 2.1% for both categories (the 38 PARs, as highlighted in **Figure 3.8**, occupy only 2.1% of the total area on the dot plot), achieving a ~10-fold enrichment. The PARs interleave multiple grape and rice genomic regions collectively covering around 70% of each genome. By consolidating much of the redundancy in each genome, the PARs create syntenic blocks with much less ambiguity and in most cases show association between one γ block and one σ block (i.e. we did not find any PAR that is simultaneously mapping to two different γ or σ blocks).

When a particular PAR is scrutinized, syntenic relationships among the clustered regions are more informative than analyzing any individual pair of syntenic segments that contribute to the PAR. For example, in PAR17 (**Figure 3.9**), three grape regions resulting from the γ triplication (γ6) (Jaillon et al. 2007; Tang et al. 2008a) correspond to several regions in rice

matching each other, which can be partially explained by σ1 (and additional duplications unobserved in intra-genomic comparisons).

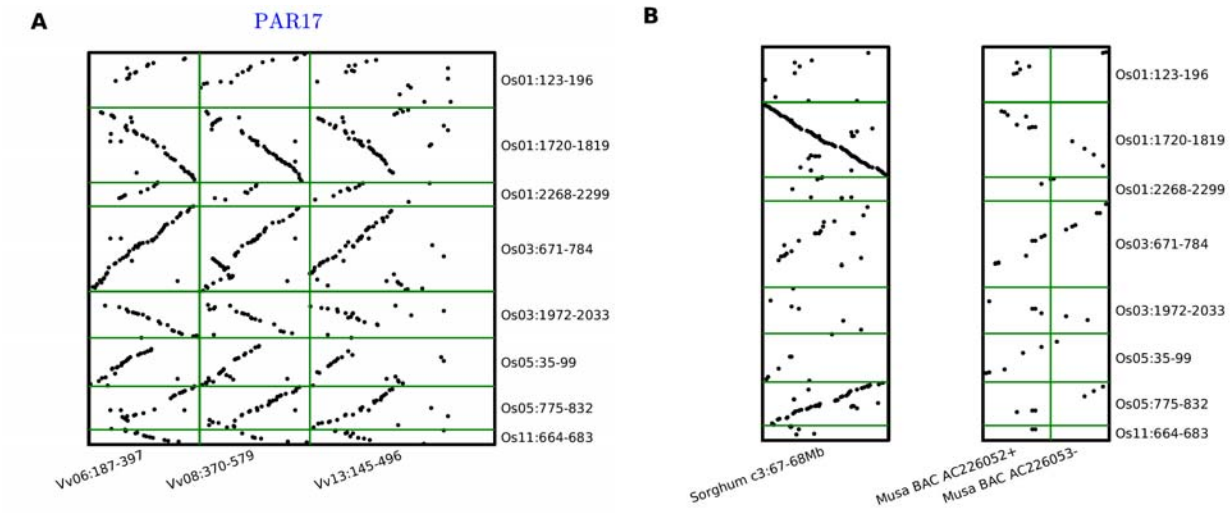

Figure 3.9: Synteny comparisons with putative ancestral regions (PARs). (A) Zoom-in view of one exemplary PAR17 consisting of corresponding regions from grape and rice. The segment labels on the right and below the graph has the format species (*Vv* for grape, *Os* for rice) followed by "chromosome:start-stop", where start and stop are the re-indexed gene rank after removal of tandem genes and singletons (see Methods). (B) Synteny between one sorghum genomic region and two contiguous *Musa* BACs (+/- indicating flipping of order) to the rice duplicated regions identified in PAR17.

3.4.3 Implications for comparisons between cereal and basal genomes

The high level of synteny and collinearity among cereal genomes has long been clear, but parallels to other monocots such as banana (Lescot et al. 2008), onion and asparagus (Jakse et al. 2006) have been more difficult to discern. The generally low synteny found in these prior studies may improve after accounting for redundancies within cereal and other genomes.

The duplicated regions we identified in rice are also evident in comparisons to banana, a non-grass monocot (Lescot et al. 2008). Our synteny search in limited outgroup sequences revealed that two banana BACs (AC226052.1 and AC226053.1) match the set of rice regions in PAR17, which was used as an example in the rice-grape PARs (**Figure 3.9A**). A sorghum genomic region (c3:67-68*Mb*) was selected as a cereal reference (**Figure 3.9B**). Sorghum

shows very strong synteny corresponding to the orthologous rice region (*Os*01:1720-1819), then lesser but still easily discernible synteny to one matching ρ-block (*Os*05:775-832), while σ-blocks (the remaining six regions) only show a few homologs. In contrast, banana-rice homolog concentrations in each duplicated regions are comparable to one another, suggesting that the banana-rice divergence may have predated both ρ and σ duplications. Limited amounts of banana sequence data prevent us from falsifying the alternative hypothesis that this lesser stratification of synteny patterns simply reflects greater divergence between banana-rice.

3.4.4 The number of WGD events in the monocot lineages

In many lineages, the existence and the numbers of WGD events have been contentious. Whether the vertebrate lineage had experienced two (2R) or three (3R) WGDs was long debated, and only recently resolved through careful analysis of synteny patterns of WGD paralogs (Dehal et al. 2005). Similarly, various studies offered conflicting estimates of the number of WGDs in *Arabidopsis* (Bowers et al. 2003b; Vision et al. 2000). Different sources of evidence might favor different models – in particular, estimates based on the distribution of genetic distances of paralogs or topologies of gene trees alone are now known to be complicated by unequal evolutionary rates between gene families and lineages (Fares et al. 2006; Tang et al. 2008a). So far, analyses based on synteny patterns provide the most accurate inferences of WGD events (Tang et al. 2008a).

Our unique approach to synteny analysis provides new insight into the number(s) of WGD events experienced by modern cereal genomes. The pattern exemplified by the one PAR we had space to show with fine resolution (**Figure 3.9A**), with usually 3-fold redundancies on the grape axis and at least 4-fold redundancies on the rice axis, is representative of all 38 PAR patterns. In 22 of the 38 PARs, grapevine-rice collinearity is clear, which allows us to evaluate the level of redundancies in both genomes. These redundancies reflect the number of genome duplication events observable in both lineages. Among the 22 PARs, 12 are 3-fold redundant in

grapevine, consistent with hexaploidy (3). The level of redundancy in rice is less clear, ranging from as little as 2-fold (1 PAR) to 7-fold (3 PARs) and 8-fold (5 PARs). In line with the intra-genomic evidence from our bottom-up analysis, these high redundancies suggest that the rice lineage experienced more than two, perhaps three, rounds of WGD.

3.5 Conclusion

Using the improved methodology to identify genomic synteny patterns, I discovered evidence of ancient whole genome duplications that occurred earlier in the eudicot (γ) and monocot lineage (σ), respectively. I further use the knowledge of these duplications, to infer "scattered" synteny patterns between eudicot and monocot, through a novel clustering and sorting approach (PAR), thus bridging the comparison between the two distantly separated plant lineages. Such correspondences between eudicots and monocots were suspected long before (Paterson et al. 1996), but previous efforts were not successful on a genome-scale (Liu et al. 2001; Salse et al. 2009a). This study presents compelling evidence of large regions of good synteny conservation between a eudicot and a monocot genome.

CHAPTER 4 STUDY OF DOMESTICATION IN THE POST-GENOMICS ERA

4.1 Introduction

Pre-historic people were able to transform the wild plant species into the crops that are more amenable for human utility. Compared to their wild relatives, the domesticated crops typically show synchronization of flowering time, enlargement of grain size, loss of seed dispersal, increased apical dominance, among many other characteristics collectively known as the "domestication syndrome" (Hammer 1984). Additional valued traits include modified plant stature, grain yield, tolerance of biotic and abiotic stress, as well as better nutritional value and taste. Most of these desirable traits were driven to high frequency or eventually fixed in the cultivars during domestication.

In this review chapter, I discuss both the genomic and population changes underlying the domestication process, and then reiterate some common strategies for dissecting and quantifying these changes. Different trait mapping strategies, including linkage mapping and association mapping, are discussed in more details, to provide a foundation for Chapter 5.

4.2 Genomic and population changes associated with the domestication

Several genes that were targets of domestication or crop improvement have been identified (Doebley et al. 2006). Specific mutations are linked to shattering (Li et al. 2006a), tillering (Wang et al. 1999), fruit size (Frary et al. 2000) and shape (Xiao et al. 2008), and seed color (Sweeney et al. 2006) etc. , and more crop related genes are reviewed in (Doebley et al. 2006; Izawa et al. 2009). The mutations most likely occurred in the progenitor population; humans simply selected for these mutations and later spread them to the cultivars. The mutated alleles of most domestication-related genes are often thought to confer negative reproductive fitness to the wild individuals bearing the alleles. For example, the non-functional (domesticated) allele of

the rice shattering gene *sh4* is also found in some individuals of the progenitor species *O. rufipogon* (Lin et al. 2007). It is unclear whether the same mutation is somehow maintained in low frequency in the wild population, or instead went extinct in the wild but later introgressed from cultivated individuals.

Identification of several key domestication genes reveals an array of genomic changes that are associated with the transitions from wild to domesticated plant. The form and nature of the genetic mutations is highly variable. Common types of changes associated with crop related genes include: 1) amino acid substitutions, e.g. rice *sh4* gene (Li et al. 2006a); 2) deletions and truncation, e.g. rice *rc* gene (Sweeney et al. 2006); 3) insertional mutation caused by transposon activity, e.g. maize *sh2* gene (Bhave et al. 1990); 4) DNA mutation in the regulatory elements, e.g. maize *tb1* gene (Wang et al. 1999); 5) splice site mutation causing alternative splicing, e.g. rice *waxy* gene (Wang et al. 1995); 6) gene duplication creating a new genomic context or dosage change for particular genes, e.g. tomato *SUN* gene (Xiao et al. 2008).

Some types of mutations disrupt the normal coding of the protein product therefore are considered non-functional "knock-outs" in the domesticated species. In this class, genes have mutations that induced frameshifts and early stop codon and have become pseudogenes. By contrast, some types of the mutations are not in the coding sequence but instead are mutations in the regulatory elements that modifies expression levels or spatio-temporal expression patterns (Doebley et al. 2006). It is important to note that one type of change – amino acid substitutions likely involve no change of the gene expression at all but disrupt the interaction of the protein with the downstream targets, as in the case of *sh4* (Li et al. 2006a).

The mutations in crop-related genes provided novel genetic variants but still need time to spread to the population; therefore we need to understand the population changes as well. This is typically studied by collecting genetic information from a diverse sampling of both domesticated and wild plant varieties. One common feature of the domesticated genomes is the reduction of genetic diversity in crops relative to the wild progenitors (Li et al. 2006a). This

reduction has resulted from two major forces. First, domestication is typically thought to have involved initial population of small size that has constrained *genome-wide* genetic diversity known as "bottleneck effect". The second factor is the directional selection for *local* genomic regions that distinguish crops from their ancestors. Both forces can be tested for deviations from the neutral Wright-Fisher model, which assumes constant population size and no selection (Yamasaki et al. 2007).

The development of neutral markers has decoupled the above two factors so that it becomes more tractable to first study the demographic changes associated with domestication (Londo et al. 2006). Although the conventional wisdom was that the cultivated crops are usually inbred and expected to lose a significant portion of ancestral diversity, recent sequence data suggest otherwise. For example, the diversity in domesticated maize has only reduced to about 60-80% of the diversity in its progenitor teosinte (Tenaillon et al. 2004; Wright et al. 2005). Surprisingly, this estimate is typical of several crops, including einkorn wheat (70%-100%) (Kilian et al. 2007), sorghum (~80%) (Casa et al. 2005) and chile peppers (~90%) (Aguilar-Melendez et al. 2009). However, such estimates are likely over-estimates, given the potential selection bias and possible genetic erosion in the wild population. It is also likely that some crops have partially restored the diversity through recent gene flow from wild population after the initial domestication (Glemin et al. 2009).

Population bottlenecks are usually quantified by two factors – the bottleneck population size (*Nb*) and duration of the bottleneck (*d*). The severity of the bottleneck is given by coefficient $k = Nb/d$ (Wright et al. 2005). Results from earlier analysis suggested that most domesticates form a monophyletic group, which was interpreted as a single, rapid localized domestication event. Recent archaeobotanical evidence and coalescent simulations instead favor a "protracted" model of domestication (Allaby et al. 2008; Brown et al. 2009). For example, the classical population model in maize suggested only one bottleneck with a single value of *k* (Wright et al. 2005), whereas the "protracted" model fits several values of *k*, representing multiple bottlenecks

of different strengths. The prolonged period of several bottlenecks might suggest additional events after the initial domestication, perhaps reflecting the process of dispersal of the cultivars and plant breeding (Chrispeels 2003). Frequency spectrums of allele variants also reveal unique demographic history of particular domesticated species. For example, studies of SNPs in domesticated rice show an excess of high-frequency alleles, supporting a rather complicated breeding history of rice (Caicedo et al. 2007).

Detailed analyses of the domestication genes reveal remarkable reduction of diversity that drove only a few haplotypes into fixation. Selection (both artificial and natural) is expected to reduce diversity at the domestication-related genes and also tightly linked loci ("selective sweep") as the favorable alleles are driven to high frequency, and such reduction is more striking compared to the bottleneck effect alone. The size and shape of the selective sweeps depend on the time and strength of the selection as well as local recombination rates in the genome. Several studies in maize reported particularly large sweep blocks (Palaisa et al. 2004; Tian et al. 2009). However, in some domesticated species like sorghum, the power to detect selection is weak when the levels of variations at neutral loci are already low, perhaps due to much more recent domestication (Hamblin et al. 2006).

The strong artificial selection unintentionally imposes genetic load on the crop genome in harboring deleterious mutations that are often quickly purged in freely recombining natural population, therefore potentially interferes with the natural selection. Recent genome comparisons of two rice cultivars (*japonica* and *indica*) show a high level of deleterious mutations, suggesting a genome-wide relaxation of selective constraint due to domestication (Lu et al. 2006), and is consistent with the findings in domesticated animals like dogs (Bjornerfeldt et al. 2006). This is also known as the genetic hitchhiking effect, again due to limited recombination in the crop genomes.

The history of domestication and breeding can also be revealed by tracing the distribution of major domestication genes in chronologically and geographically stratified

sampling of landraces and cultivars. For example, among the six rice domestication-related genes identified so far, spread of the mutations in *Rc* and *qSW5* were probably the most ancient since these mutations can be readily found in most heritage landraces, while the recruitment of *qSH1* was relatively recent and indeed only found in a few modern temperate *japonica* cultivars (Izawa et al. 2009; Konishi et al. 2008).

4.3 Methods for dissecting domestication traits

Quantitative trait locus (QTL) mapping is a powerful way to study the domestication-related genes and chromosomal regions, with the only requirement of the presence of both domesticated and non-domesticated alleles in the mapping population (Paterson 2002). Two popular experimental methods are available for QTL mapping.

One common strategy for QTL mapping is linkage mapping. For linkage studies, the researcher usually look for set of markers that are present in the individuals showing the phenotype, and absent in the individuals without the phenotype, therefore the markers are called segregating markers and the population is called segregating population. In practice, segregating populations in plants are generated through a number of experimental designs, including F2, backcross (BC), near isogenic lines (NILs), bulked segregant analysis (BSA) and recombinant inbred lines (RIL) and many others (Paterson 2002). QTLs with simple genetics, large phenotypic effect can be identified easily through linkage mapping studies, and indeed most of the identified domestication genes so far are from this category. However, caution is still warranted since many QTL studies are dependent on the environment and the parental lines, therefore the generality of QTL analysis in small populations are sometimes questionable.

Compared to linkage mapping, association mapping in plants started relatively recently, with one of the earliest applications described in (Thornsberry et al. 2001), which confirmed the association between maize *dwarf8* gene and flowering time. Association mapping is a powerful genetic tool to identify alleles or polymorphisms responsible for trait variations, and is gaining

popularity recently due to decreasing genotyping cost. In some applications, the target locus may be known *a priori* to be contained within a candidate chromosomal region (often by linkage mapping studies) or those genes purported by biochemical analysis (candidate genes). Association would then be used as a fine-scale approach to narrow down and identify the responsible gene. We can genotype candidate genetic loci within the target region, and test whether certain sites within those genes are strongly associated with the trait.

Another variant of the association mapping called "whole-genome association" (WGA) approach, also known as "genome scans", requires no prior knowledge of the locations of the QTLs. With available whole genome sequences, the genetic diversity can be sampled at well spaced intervals across the whole genome. For example, a recent study in *Arabidopsis* searched for genome-wide associations with flowering time and pathogen resistance in 95 individuals (Aranzana et al. 2005). Even though the result showed a high false-positive rate, they still detected a few validated genes, thus suggesting that genome-wide association mapping is feasible (Aranzana et al. 2005). In some studies, whole genome association mapping can also be combined with the candidate gene approach. For example, a more recent study focused on 51 known loci involved in flowering pathway, and tested the association between different alleles and flowering time in 275 *Arabidopsis* accessions (Ehrenreich et al. 2009).

In order to have better statistical power for association studies, panels of many naturally occurring individuals (or a wide collection of germplasms from seed banks) that show trait variation are often used in the association study. Ideally, the individuals should be unrelated and randomly selected to avoid the complications from population structure. The population structures or genetic relateness of individuals in the sample violates the assumption of sample independence in the linear model for the association test. Indeed, the strong structures for plant populations is one reason that limits the application of association mapping in plants (Zhu et al. 2008). One remedy for this, is a statistical approach called "structured association mapping", where complex familial relationships and population structures are included as covariates (and

thus accounted for) in the same general linear model (Yu et al. 2006). However, for plant species that show substantial phenotypic differentiation over different geographic areas due to local adaptations (like flowering time), structured association can suffer loss in statistical power (Yu et al. 2008).

There are both pros and cons to the traditional linkage mapping method and the relatively new association method, yet both methods are based on the simple notion that the trait difference can be explained by underlying DNA polymorphism. Linkage mapping often has good statistical power, since strong linkage disequilibrium (LD) affords high power to detect QTL. However, the strong LD is a double-edged sword for linkage mapping in that there are a limited number of meiotic crossing-over within a few generations thus a large number of individuals need to be genotyped to narrow down the interval. This is complemented by the association approach which investigates the patterns in mostly unrelated individuals from natural populations. The breeding pool of the natural population is much larger and the coalescence of the individuals can go back many generations so that there were many historical recombination events to break down LD. The shorter LD offers higher resolution than linkage mapping. However, the power of association mapping is usually weaker than a bi-parental cross. Additionally as stated above, association between marker and trait may have arisen from unknown population structure rather than a causative site (Zhu et al. 2008). It is important to note that even under the assumption of no population structure; the presence of an association is still not necessarily the result of a direct factor but rather due to a statistical correlation between the causal gene and other genes.

A different approach – "selection scan" takes advantage of the unusual polymorphism patterns of *domestication*-related genes (Chapman et al. 2008; Wright et al. 2005). This method looks for loci that show significant reduction of sequence diversity in the domesticated compared to wild samples. Selection scan is a relatively high-throughput method, often generating a large set of "candidate genes" compared to the traditional QTL mapping which

interrogates only a few loci at a time. It was originally thought that only a few key genes of "large effect" were responsible to transform the wild teosinte to the maize crop (Doebley 2004). However, a study based on a sample of 774 genes in maize extrapolates that 2-4% of the maize genome is under selection, indicating that domestication affected a large number of loci (Wright et al. 2005). Another study identified 36 out of 492 (7%) of the sunflower genes that show evidence of selection (Chapman et al. 2008). It is possible that the abundance of "selected" loci is variable among different domesticated species, because of differences in the domestication history and extent of recombination.

4.4 Test for convergent evolution of domesticated genomes

Earlier analysis of a few domestication-related QTLs suggested that they occur in corresponding map locations across different cereal species, more often than explained by chance (Paterson et al. 1995). It was also postulated that convergent phenotypic evolution of major cereal crops can be explained by independent selection of mutations in orthologous gene loci (Paterson et al. 1995). It now appears that various domestication traits have divergent patterns of genetic architecture (i.e. underlying genetic basis). For example, the flowering time QTLs in maize often show synteny conservation with rice (Chardon et al. 2004; Paterson et al. 1995). In contrast, major genetic loci controlling for seed shattering appear different (non-orthologous) in barley, maize, rice and sorghum, indicating multiple genetic pathways (Freeling et al. 2006). It was also suggested that even in the case when the genetic control of a particular trait is well conserved, the natural variations in corresponding genes are independent in different species and might not show similar level of contribution to each trait (van Leeuwen et al. 2007; Yamamoto et al. 2009). For other traits, related species might have different morphological or phylogenetic constraints and therefore the major genetic determinants vary. For example, *ramose1* controls the floral branching system in the panicoid (maize, *Miscanthus* and sorghum) but is missing in rice (Vollbrecht et al. 2005).

Many key genes for domestication transitions are known transcriptional regulators (Doebley et al. 2006), yet their downstream targets are still unknown. Targets of these transcription factors can be studied through genome-wide expression QTLs (eQTLs), which simultaneously queries the expression level of many genes (Hansen et al. 2008; van Leeuwen et al. 2007). Additionally, we can study the set of genes that are differentially expressed in the domesticated versus the wild individuals and see how these genes are related to one another in the context of a large regulatory network. Loss-of-function mutations (as opposed to only regulatory changes) can also be examined by comparing the genomes of different related crop species. Pseudogenes in the corresponding chromosomal locations that have simultaneously experienced loss-of-function mutations in rice and sorghum might reveal potential targets of domestication that are perhaps the result of recent convergent changes. For example, Shang et al. did a genome-wide analysis of the rice NBS-LRR gene family with a focus on the pseudogene members (Shang et al. 2009). They found that the *Pid3* locus became pseudogenized in the *japonica* varieties after *indica-japonica* split, thus conferring rice blast susceptibility only to the *japonica* genotypes. This shows how *in silico* comparative genomics (in this case screening for nonsense mutations in one subspecies versus another) can quickly identify candidate gene loci that may be responsible for varietal and species differences.

4.5 New avenues for studying domestication

Among the six plant genomes sequenced thus far, papaya, grape, rice and sorghum are all domesticated species, and many more crop genomes are partially sequenced or pending. Gene content and arrangements are often well conserved in related plant species (Tang et al. 2008a), making it efficient to test the homologues in the related species when a certain gene is isolated in one species. Recent analytical methods provide more accurate correspondences between genomes, both at the gene level and nucleotide level (Tang et al. 2008a).

Sequencing technology is becoming increasingly parallel and high-throughput while cost per base continues to plummet. High density tiling arrays and next generation sequencing provide efficient sampling of the genetic diversity. Using resequencing microarrays to map genome-wide SNP variations, a recent study revealed clear phylogenetic relationship, population structure and introgression history among 20 rice cultivars and landraces (McNally et al. 2009). Another proof-of-concept study used short-read sequencing technology and was able to map the "green revolution" gene *sd1* in a 160-individual recombination inbred population (Huang et al. 2009). Although the current read length of next generation sequencing is still not ideal for *de novo* sequencing of the more repetitive crop genomes, steady efforts are being made (Rounsley 2009).

Typical populations for genetic mapping of quantitative traits are based on two types of populations – naturally occurring lines and synthetic lines (Paterson 2002), with respective strengths and weaknesses (discussed in the previous section). Efforts that combine the relative strengths of the two types of populations were available earlier but limited by genotyping efficiency. Given new technologies, mapping individuals can now be assayed with efficiency and less ambiguity. Buckler and colleagues recently established a nested association mapping (NAM) population in maize, where 25 different maize lines were all crossed with the same parent B73, and for each of the 25 families 200 recombinant inbred lines were generated (Buckler et al. 2009). This composite population captured a significant fraction of the maize diversity and identified numerous QTLs that are shared among different families (Buckler et al. 2009). This is in contrast to the classical family of only two parents, in which only a subset of QTLs is detected. Similarly, development of "multi-parent advanced generation inter-cross" (MAGIC) offers a new experimental platform for analyzing gene-trait correlations (Cavanagh et al. 2008). Such mapping strategies with combined power and resolution will provide a clearer picture of the genetic architecture underlying many domestication-related traits.

The “diversity-based” mapping population also provides a basis for an integrated mapping approach combining both linkage and association mapping (Yu et al. 2008). As discussed above, linkage mapping requires less marker coverage while association mapping offers higher resolution. The integrated approach improves the map resolution without the need for dense markers, as well as increasing the scope for otherwise limited QTLs inferred from a single two-parent population (Yu et al. 2008).

4.6 Conclusion

Plant domestication, breeding and biotechnology have modified plant genomes to be tailored to the needs of humanity with increasing efficiency and precision. Understanding such processes, crop domestication in particular, is crucial today because of the rising demand for improving yield and quality of grain crops, as well as renewed interest in utilizing biomass species for energy production. Crops also form a particularly good system for the study of accelerated evolution. The study of domestication intersects both genomics and population genetics, and informs us about the nature of selective constraints. Further knowledge of the genomics underlying crop domestication facilitates advancement of evolutionary theory while offering a solid foundation for full-fledged crop engineering in the near future.

CHAPTER 5 ASSOCIATION MAPPING OF THE SORGHUM GRAIN SHATTERING GENE *SH1*

5.1 Introduction

Cultivated sorghum (*Sorghum bicolor*) is a leading cereal in agriculture, ranking fifth in importance among the worlds' grain crops (Doggett 1976). Sorghum is used for food, feed fodder, and the production of ethanol. Sorghum plants are more tolerant to drought and heat than most other grasses, making it an ideal staple food in arid African countries. Among the more than 20 species within the *Sorghum* genus, *S. halepense*, *S. almum* and hybrids of these to the cultivated *S. bicolor,* collectively known as "Johnson grass", are notorious weeds affecting crop yields (Draye et al. 2001).

Sorghum is in the Poaceae subfamily Panicoideae, the tribe Andropogoneae, and sub-tribe Saccharinae. Sorghum is phylogenetically closer to sugarcane and maize than to rice (**Figure 5.1**). Sorghum and sugarcane diverged from a common ancestor an estimated 8-9 million years ago (Jannoo et al. 2007) while the sorghum-maize divergence is about 12 million years ago (Zuzana Swigonova 2004).

The domestication of sorghum started in Africa and then was carried to Europe and Asia before North America. Wild species of sorghum are found as early as 8000 years ago in the Nilotic regions of southern Egypt and Sudan, but the location of its true domestication within East Africa is still speculative (Dahlberg 1995). Wild sorghums disperse by two major ways: vegetative reproduction through subterranean rhizomes (e.g. *S. propinquum* and *S. halepense*; *S. bicolor* is not rhizomatous) and seed dispersal by shattering. Although disadvantageous in the wild habitat, non-shattering sorghums are thought to have been selected during domestication because humans could more efficiently harvest grains that remained attached to the plant.

During plant development, the shattering of seeds involves the formation of an abscission layer and is considered a process of programmed senescence.

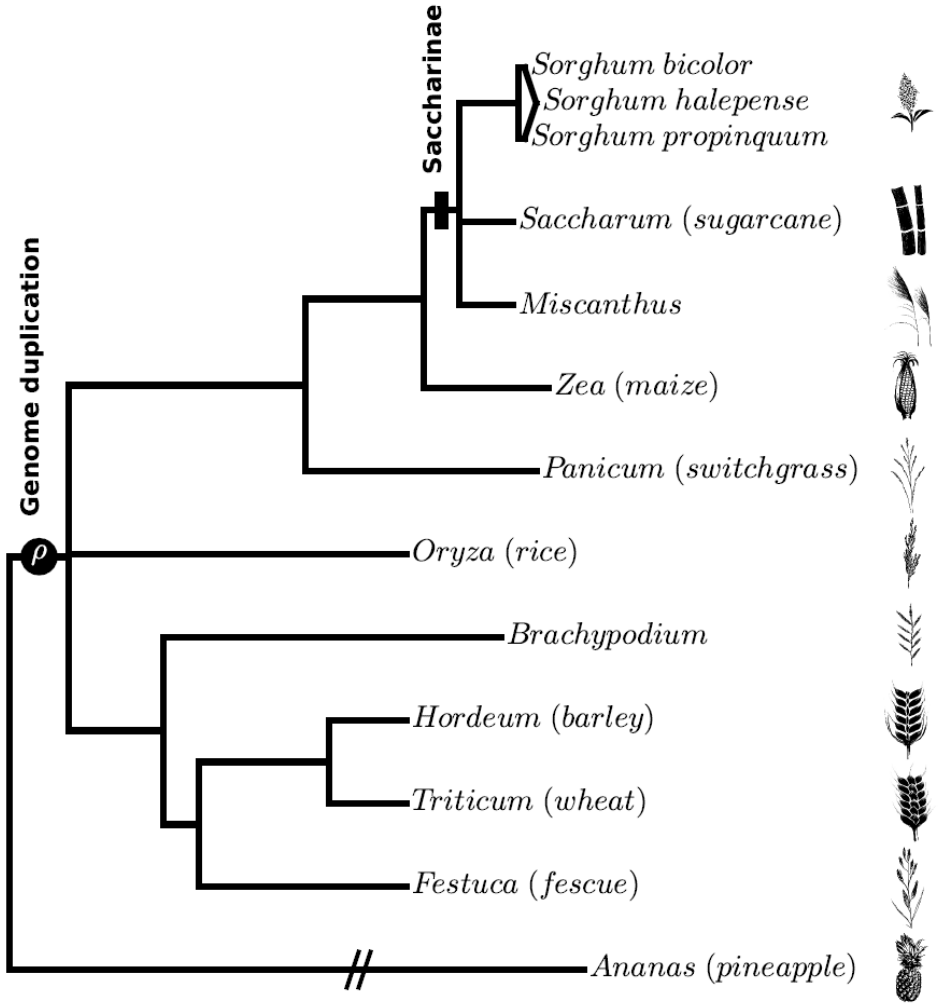

Figure 5.1: Phylogenetic relationships of sorghum with selected grasses.

The pathway involving the formation of the abscission layer is well characterized in some eudicot species. *SHATTERPROOF* genes *SHP1* and *SHP2* have been shown to specify valve margin cell identities in *Arabidopsis* (Liljegren et al. 2000). The expression of the *SHP* genes are reinforced through negative regulation from *FRUITFUL (FUL)* in valve development (Ferrandiz et al. 2000) and *REPLUMLESS (RPL)* in the replum (Roeder et al. 2003). However, the botanical origin of the abscission layer in *Arabidopsis* is clearly different from that of rice or

other cereals. The layer contributing to seed shattering studied in *Arabidopsis* is located at the valve-replum boundary and does not correspond to that of rice which is at the base of the pedicel. Therefore, it remains questionable whether orthologous genes are implicated in the seed dispersal mechanisms of dicots and cereals, respectively.

Two major genes that contribute to the shattering trait in rice (*Oryza sativa* ssp.) were identified in 2006 – *qSH1* and *sh4*, controlling 68% and 69% of the phenotypic variance in the studied crosses, respectively (Konishi et al. 2006; Li et al. 2006a). In both cases, the non-shattering phenotype is caused by the absence of the abscission layer (or dehiscence zone), though *sh4* shows a change of protein function while *qSH1* shows a change in expression pattern as a result of domestication (Konishi et al. 2006; Li et al. 2006a). The fixation of *sh4* occurred very early in rice domestication with the domesticated allele occurring in both *indica* and *japonica*, while *qSH1* is much more recent and is present only within temperate *japonica* individuals (Konishi et al. 2008; Zhang et al. 2009). In wheat, QTLs that are responsible for nonbrittle rachis are located in the homeologous regions of chromosome 3A (*Br2*), 3B (*Br3*) and 3D (*Br1*) (Nalam et al. 2007; Nalam et al. 2006). Comparative mapping hinted that this part of the chromosomal regions might correspond to the orthologous region in barley, controlled by two tightly linked loci, *Btr1* and *Btr2*, but do not appear to correspond to the region in other major cereals (Nalam et al. 2007; Nalam et al. 2006). Indeed, many of these genes in different cereal crops do not appear to be in corresponding (orthologous) chromosomal locations, therefore it is hypothesized that there are multiple pathways responsible for seed dispersal in the grasses (Li et al. 2006b). Steady progress in rice notwithstanding, many more rice genes that control shattering are known (Paterson et al. 1995) but have not yet been identified, therefore the above hypothesis remains to be tested. Additionally, since sorghum and maize are closer to one another than to rice, the shattering loci between the two panicoid species may still partially correspond (Paterson et al. 1995).

Sorghum appears to be a favorable species to investigate the genetic basis of shattering, since only one locus *Sh1*, explains 100% of the phenotypic variance in the cultivated × wild sorghum cross (Paterson et al. 1995). In the cross *S. bicolor* × *S. propinquum*, all F1 progenies shattered, indicating that *Sh1* was completely dominant (Paterson et al. 1995). The linkage mapping in 370 F2 individuals together with progeny testing of key recombinants, suggested that the region was defined by two flanking RFLP markers, with a genetic distance of 0.42cM (3 recombinants out of 740 gametes) between the two markers. However, due to the limited number of recombination events in the F2 individuals, the resolution of linkage mapping is quite coarse.

We can now attempt to fine map the shattering gene *Sh1* in sorghum, with the aid of genomic resources for the sorghum that have increased rapidly in recent years. Both a high-density genetic map (Bowers et al. 2003a) and physical maps of both *Sorghum bicolor* BTx623 and *Sorghum propinquum* (Bowers et al. 2005) are available. The *S. bicolor* genome size is approximately 730Mb, and has been sequenced to high quality with ~90% of its DNA and ~98% of the genes anchored to the chromosomes (Paterson et al. 2009). Comparative mapping and genomic data suggests that sorghum shows similar composition and high levels of synteny and micro-collinearity with maize and rice (Bowers et al. 2005), despite ~50 million years of divergence (Vicentini et al. 2008).

In this chapter, I first identify the target genome region that contains the previously mapped *Sh1* and compare the corresponding regions in *S. bicolor* and *S. propinquum*. Many gene loci differ at the DNA level between the two species. In order to find the DNA changes responsible for the loss of shattering in *S. bicolor*, I collected sequence data from more individuals to increase the statistical power of association mapping. I established a diversity panel with 24 sorghum genotypes. I validated the shattering/non-shattering phenotypes for these individuals, using both qualitative and quantitative methods. Extensive genotyping (resequencing) was done in the target chromosomal region, using a rather exhaustive approach.

Strong associations were detected in a small region following the analysis of the genotypes. The association mapping suggests promising candidates for further functional study.

5.2 Sequencing, assembly and annotation of *S. propinquum* BACs

An *S. propinquum* bacterial artificial chromosome (BAC) library with high coverage of the genome (Lin et al. 1999) was screened with the DNA markers closely linked to *Sh1*. BACs that hybridized to the two flanking genetic markers in the shattering region were fingerprinted via restriction enzyme digestion, and used to construct physical contigs (Soderlund 1997). One contig that spans the entire length between the two flanking markers was constructed. Several BACs forming a tiling path of the contig were selected. The DNA of the BACs was isolated, sheared, end-repaired into subclones and Sanger-sequenced.

Table 5.1: Assembly status of the *S. propinquum* BACs around the putative shattering region. I only counted contigs that are >1kb length.

BAC ID	# of scaffolds	# of contigs	Size	Total # of reads
00001	4	5	226kb	5898
00002	1	3	111kb	2118
00003	6	15	120kb	2304
00004	5	16	210kb	3355
00005	3	5	61kb	1772
00006	3	9	115kb	3840
00007	2	12	157kb	3137
00008	3	4	55kb	1536
00009	5	26	119kb	2304
00010	3	14	142kb	2131

Sequence assembly follows the PHRED/PHRAP/CONSED pipeline (Ewing et al. 1998). Alternative assemblies were also attempted with the TIGR and CELERA assemblers but we chose PHRAP because it shows the lowest error rate among the three programs. Thus far, we have draft assemblies for the 10 BACs containing un-finished contigs within each BAC (**Table 5.1**). Finally, the reads from the 10 overlapping BACs were pooled and assembled into 108 contigs, comprising a total size of 1.06Mb of the entire region in *S. propinquum*.

Gene structures in the *S. propinquum* shattering region were predicted using the similarity-based gene prediction software GENEWISE, using the *S. bicolor* predicted genes (Sbi version 1.4) as the reference sequences. GENEWISE predicted 95 *S. propinquum* gene models (with a median size of 906 base pairs), corresponding to 95 *S. bicolor* gene models. A total of 80 genes are within the boundary of the two flanking markers in the linkage mapping.

Comparative analyses between *S. bicolor* and *S. propinquum* orthologs show that they are similar at the DNA level. For the 95 loci, 9 loci show no protein changes between the two species. The median of synonymous substitution per synonymous site (*Ks*) is 0.0215 in the shattering region. This median *Ks* value corresponds to ~1.7 million years of divergence between *S. propinquum* and *S. bicolor*, using a rate estimate of 6.5×10^{-9} synonymous substitutions per year (Gaut et al. 1996). Median non-synonymous substitution value (*Ka*) is 0.0063 between the two species. Most genes show *Ka/Ks* ration less than 1, indicating purifying selection (Yang et al. 2000). Surprisingly, 10 genes among the 95 genes have a *Ka/Ks* ratio greater than 1 (**Figure 2.1**), which is often interpreted as evidence supporting positive selection (Yang et al. 2000). However, since all 10 genes with high *Ka/Ks* ratio only have putative function, it is possible that some genes or some parts of the genes might be results of mis-annotations.

Repeats within the shattering region of the two sorghum species were identified using REPEATMASKER version 3.2 (Huda et al. 2009). The physical positions of these elements in *S. bicolor* are shown in **Figure 5.3**. The overall repeat level is comparable between the two sorghum species in this region. There is a higher level of retroelements in *S. propinquum* (7.7%)

than in *S. bicolor* (4.9%). Previous study found that the entire sorghum genome contains 55% retrotransposons, with preferential insertions of these elements in the heterochromatic regions (Paterson et al. 2009). Therefore, the relatively low percentage of retroelements we observed in this region compared to the genome average is consistent with features of euchromatin. Contrary to the relative abundance of retroelements, there are slightly more DNA transposons in *S. bicolor* (8.5%) than in *S. propinquum* (7.3%). The most abundant retroelement and DNA transposon in this region are Gypsy/DIRS1 and Tourist/Harbinger, respectively.

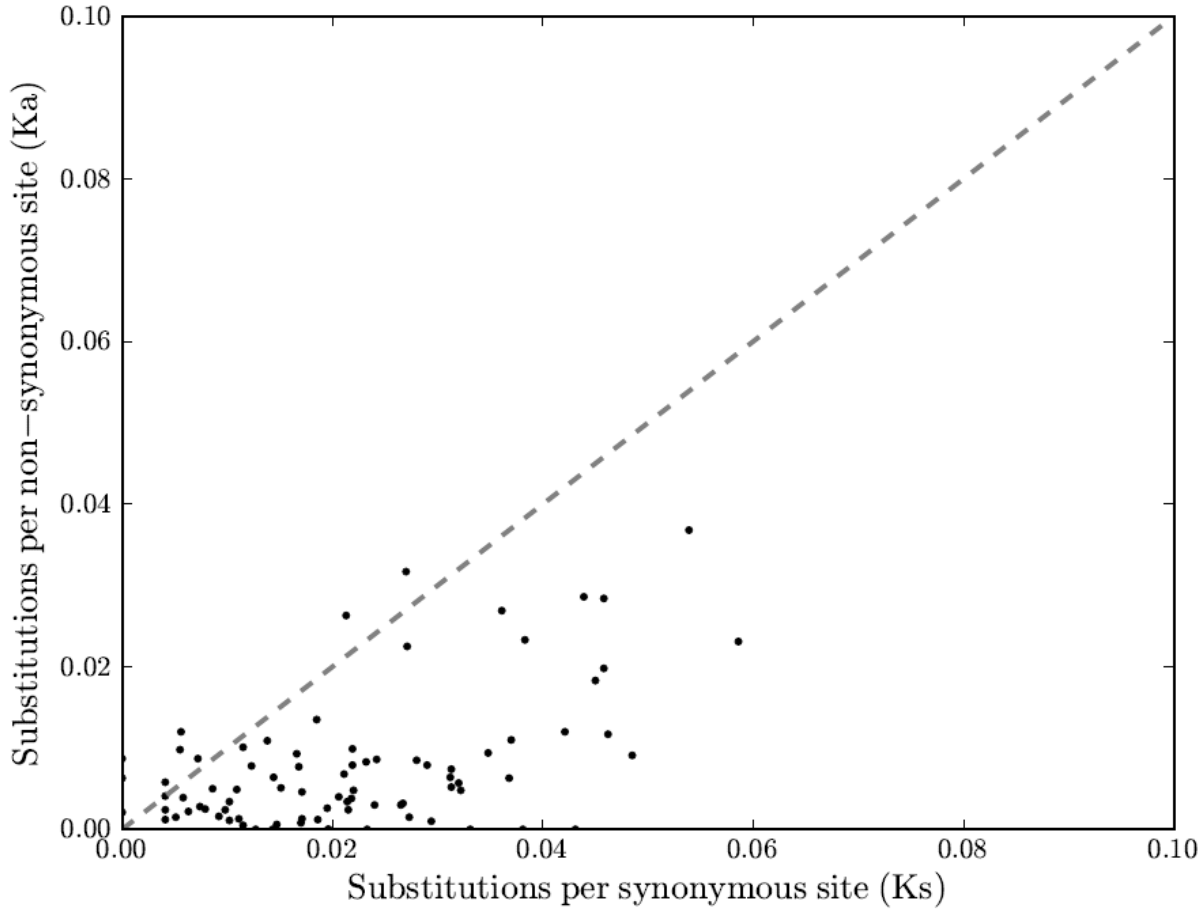

Figure 5.2: Synonymous and non-synonymous substitutions between pair of genes between *S. bicolor* and *S. propinquum*.

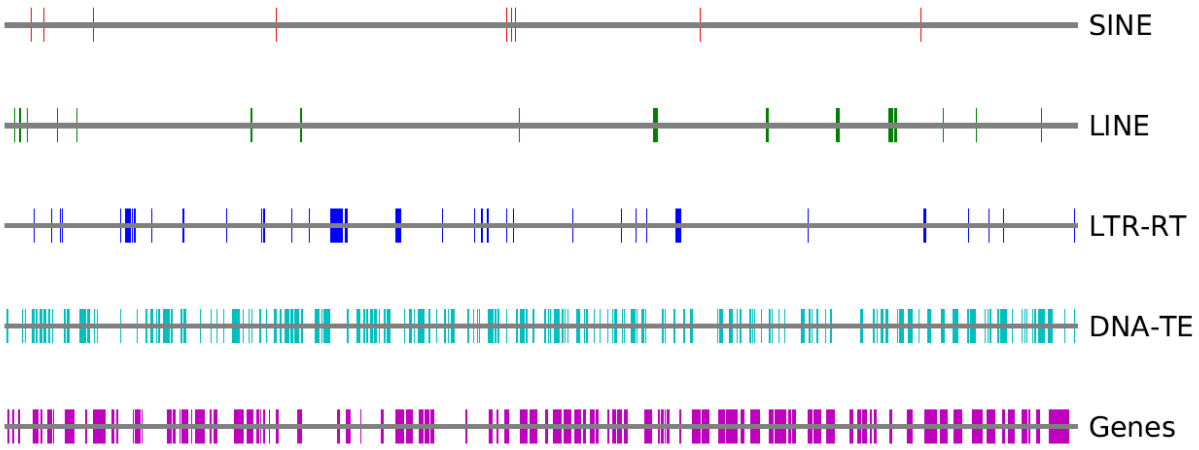

Figure 5.3: The distributions of repeats and genes in the shattering region of *S. bicolor*.

5.3 Alignment of the *S. propinquum* BACs to the orthologous *S. bicolor* region

The corresponding regions in *S. bicolor* and *S. propinquum* were aligned using MUMMER version 3.0 (Kurtz et al. 2004). The alignments show that the BAC sequences correspond to a ~1Mb region on *S. bicolor* chromosome 1 (**Figure 5.4**). Over 90% of this sequence is well aligned with *S. propinquum* contigs.

Genome alignments between *S. propinquum* BACs with the corresponding region in *S. bicolor* identified 127 sequences (>300bp) present in *S. bicolor* but not in *S. propinquum*. Some of these sequences are simple sequence repeats (SSRs) and known retrotransposons. This resource of genomic indels is useful for the discovery of novel transposon species. Because most sorghum helitrons lack structural features compared to other DNA transposons, helitron prediction software can use the indel differences between closely related species as a training set (Du et al. 2008). These indel sequences that are different between the two species of *Sorghum* were used to train the helitron prediction software used in describing the sorghum genome sequence (Paterson et al. 2009).

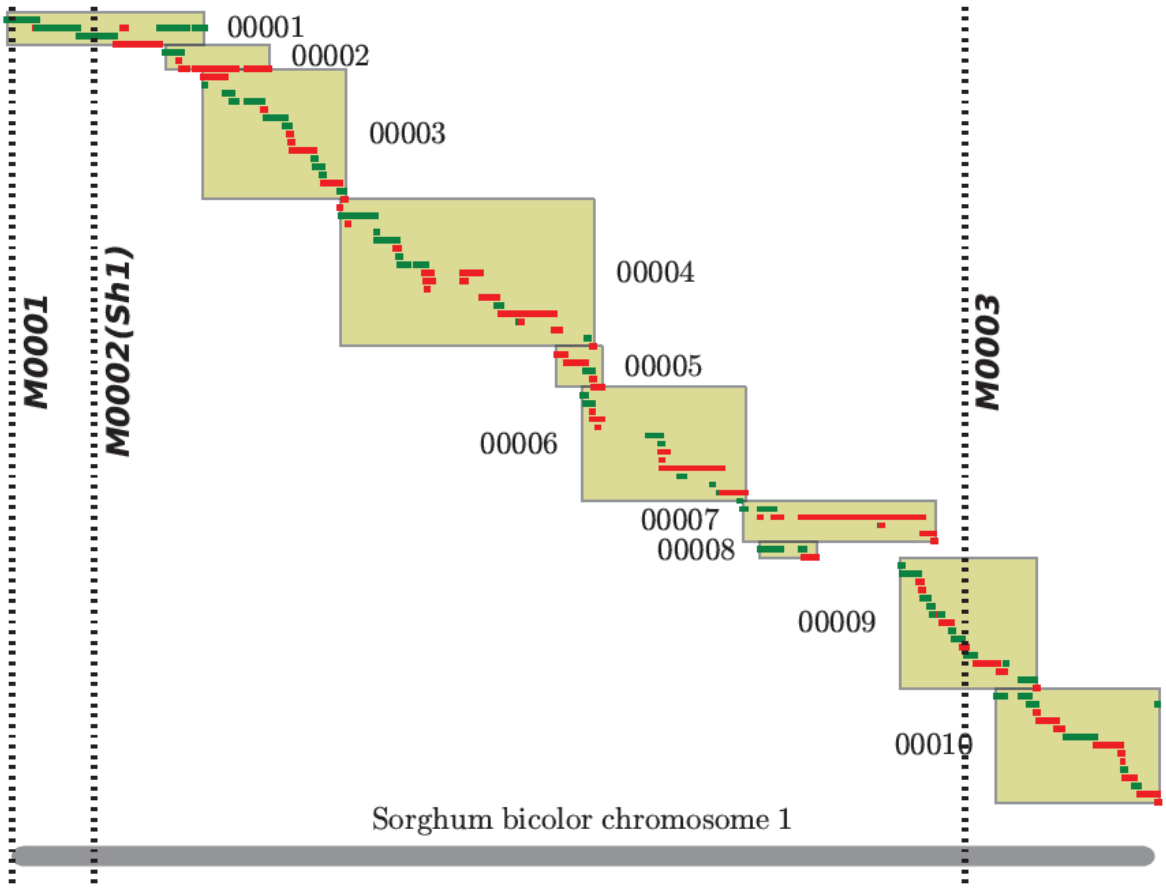

Figure 5.4: Aligned positions for *Sorghum propinquum* BACs. The line segments represent aligned contigs within each BAC, with green lines showing alignments with the same orientations, red lines showing alignments with the opposite orientations. The dotted lines represent the genetic markers flanking (M0001, M0003) or co-segregating (M0002) with *Sh1*.

I calculated the physical to genetic distance ratio, which appears non-uniform in this region. From marker M0001 to M0002 (~70kb, 2 recombinants), where most of BAC 00001 sits, the physical to genetic distance ratio is ~260kb/cM (kilobase/centimorgan), whereas between M0002 to M0003 (~790kb, 1 recombinant), the rest of the BACs, the physical to genetic distance ratio is ~5600kb/cM, suggesting that recombination is very limited in this part of the region. According to previous estimates, heterochromatic regions in sorghum showed a much lower recombination rate ~8700kb/cM compared to euchromatic regions ~250kb/cM (Kim et al.

2005). Therefore the drastic transition observed in our *Sh1* region from one side of the middle M0002 marker to the other side is comparable to the difference between euchromatin to heterochromatin, although the region generally appears to be euchromatic (Bowers et al. 2005). Such a precipitous transition is unlikely an artifacts due to sampling: assuming that the low-recombination part has an actual physical to genetic distance ratio of 260kb/cM, we should expect 22 recombinant gametes instead of only 1 observed (P=6×10^{-9}).

It is unclear what has caused the difference in recombination frequency in this region. The two parts appear to have similar repeat and gene density (**Figure 5.3**). One possibility is that there might be chromosomal inversion to suppress recombination between *S. bicolor* and *S. propinquum* in the right part of the region. However, due to the incompleteness of the *S. propinquum* assembly, I was not able to test this possibility.

5.4 Alignment of sorghum shattering region to homologous regions in other taxa

Gene content and collinearity is conserved across the sorghum shattering region, aligning well with a region on rice chromosome 3. Although the rice genome is smaller than sorghum (430Mb versus 730Mb), the corresponding region in rice appears to cover a larger physical distance than the sorghum region (1.4Mb versus 1.0Mb), although with a similar number of genes (98 versus 95). A total of 77 sorghum genes in the shattering region have syntenic rice orthologs with a median *Ks* value of 0.58, corresponding to ~44.6 million years of divergence.

Because of the most recent cereal polyploidy event, the shattering region is also syntenic to rice chromosome 12, as part of a duplication block ρ6 (Paterson et al. 2004). The region is also involved in a more ancient duplication block σ8 (consisting ρ4 and ρ6) (see Chapter 3).

Corresponding regions in a eudicot genome are less clear. Part of the sorghum shattering region is syntenic to several regions on grape chromosome 6 and chromosome 8 through PAR21 (see Chapter 3), but the synteny blocks are more degenerate, involving less than 10 gene pairs each.

5.5 A sorghum diversity panel for mapping the shattering trait

To test the gene-trait association and identify functional candidates in the region, I compiled a diversity panel of sorghum varieties that are suitable to study the shattering trait. These sorghum accessions were provided by S. Kresovich and M. Hamblin from Cornell University and from the USDA-ARS germplasm collection. Within the panel, the varieties were selected to represent a wide range of geographical locations including Africa and Asia (**Table 5.2**). Diverse varieties from wider geographical areas are chosen since in theory association mapping works better on unrelated individuals. Otherwise, if some individuals with similar genotypes are represented multiple times in our panel, this could create false positive associations.

Table 5.2: The sorghum accessions selected in the shattering diversity panel. There are three accessions that did not flower. In the "PGML index" column accessions with prefix (AL, AN, AP) are from Cornell and accessions with prefix BP are from USDA-ARS. "Race" information was taken from the accompanying documentations shipped with the samples.

Accession ID	PGML index	Race	Origin
Complete shatterers (11 varieties)			
PI 267436	BP03 (#5)	*bicolor*	India
PI 569834	BP10 (#6)	*bicolor*	Sudan
PI 521356	BP06 (#7)	*drummondii*	Kenya
PI 365024	BP05 (#8)	*verticilliflorum*	South Africa
L-WA 27	AL03 (#10)	*verticilliflorum*	Angola
L-WA 23	AL02 (#11)	*verticilliflorum*	Angola
L-WA 13	AL01 (#12)	*verticilliflorum*	Sudan
PI 155675	BP01 (#15)	*bicolor*	Malawi
S. propinquum	SP (#20)	*S. propinquum*	--

KFS (deciduous mutant)	KFS (#21)	*bicolor*	United States
PI 570917	BP11 (#22)	*bicolor*	Sudan
Non-shatterers (13 varieties)			
PI 221607	AP02 (#1)	*bicolor*	Nigeria
PI 302115	BP04 (#2)	*verticilliflorum*	Australia
PI 152702	AP01 (#3)	*bicolor*	Sudan
NSL 87902	AN07 (#4)	*bicolor*	Cameroon
NSL77217	AN05 (#9)	*bicolor*	Chad
NSL56003	AN03 (#13)	*bicolor*	Kenya
NSL56174	AN04 (#14)	*bicolor*	Ethiopia
PI 267408	AP03 (#16)	*bicolor*	Uganda
PI 563146	BP07 (#17)	*bicolor*	Sudan
PI 267539	AP04 (#18)	*bicolor*	India
PI 563474	BP09 (#19)	*bicolor*	United States
PI 591385	BP13 (#23)	*bicolor*	India
PI 584089	BP12 (#24)	*bicolor*	Uganda
Did not flower			
NSL 87666	AN06	*bicolor*	Cameroon
PI 585454	AP05	*bicolor*	Ghana
PI 156399	BP02	*bicolor*	Tanzania

5.6 Phenotyping and genotyping

5.6.1 Verification of shattering phenotypes

The shattering phenotype for each accession in the panel was carefully validated. A simple but subjective method is to classify the shattering phenotypes of the individuals into "shattering" and "non-shattering", through the hand tapping technique. The panicles were cut off from the plant and shaken vigorously, and the grains from the "shattering" varieties would usually fall off easily. Alternatively, breaking tensile strength (BTS) was used as a quantitative measurement for the degree of shattering (Konishi et al. 2006), using a digital force gauge (IMADA Inc. DPS-4) to clasp to the grain and measure the force required to break the pedicel when pulling the grain away (**Figure 5.5**). The BTS values were recorded at different developmental stages and stable values (after maturity of the grains) were used to distinguish the shattering/non-shattering phenotype for each variety (**Figure 5.5**). For each genotype, I recorded the BTS values for multiple panicles at roughly five-day intervals. Ideally, the sorghum accessions need to be measured at roughly equally spaced dates. However, since different sorghum accessions were flowering at different times, it is difficult to track each individual panicle and manage a well spaced sampling of measurements. Therefore, a few accessions were not sampled every five days.

The sorghum genotypes were first planted on June 10th, 2008, and the last measurements were taken on Nov 12th, 2008. In the span of five months, a total of 77 panicles were clipped from the planted sorghum individuals and measured in terms of degree of shattering at various stages (multiple panicles were measured for each genotype). On average, each panicle was tracked and measured around 4 times, with one case (AP03, panicle #8) measured 8 times to make sure that it is indeed non-shattering. The shattering varieties are often easier to distinguish since they are deciduous once the grains mature, while the non-shattering varieties need to be monitored for a longer period of time. I found that the breaking

force (BTS) for non-shattering varieties stabilize around 50g force after maturity, while the shattering varieties go to zero, i.e. capable of dispersal with little external force (**Figure 5.6**).

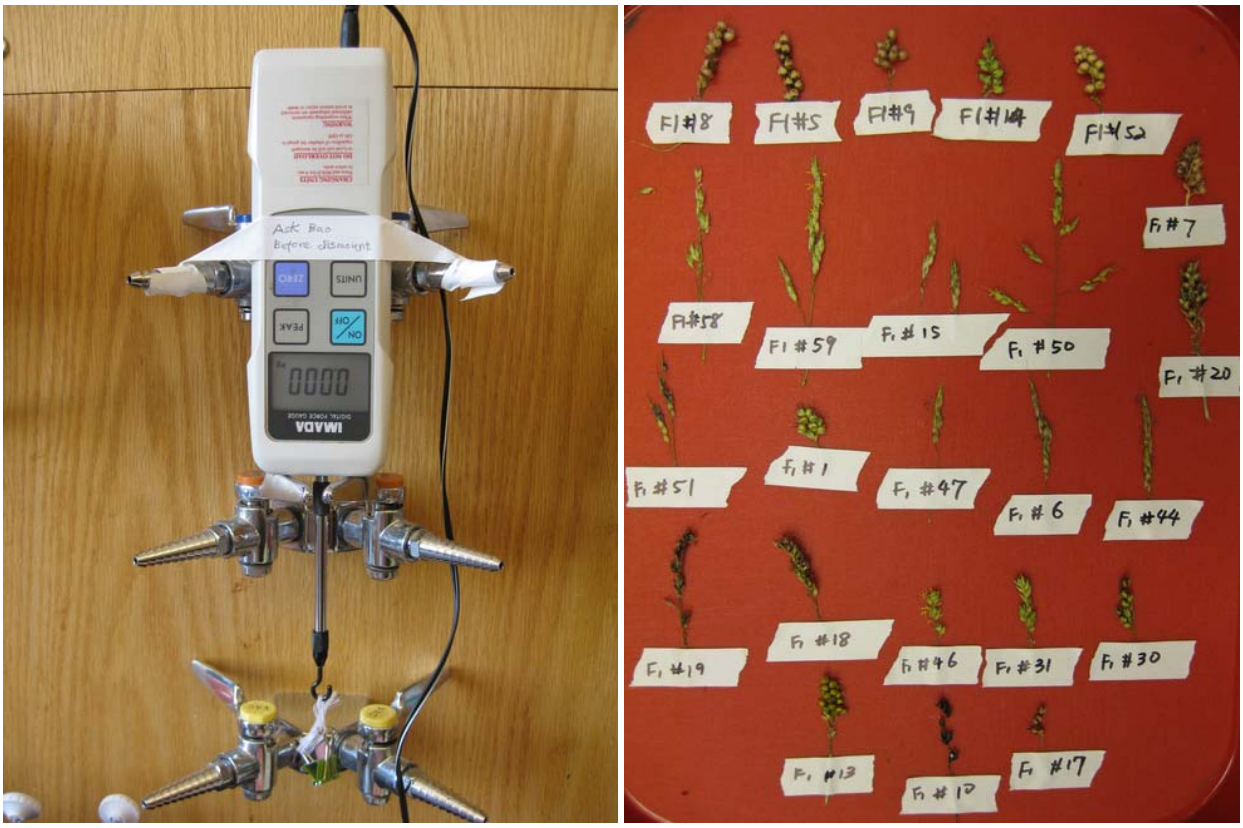

Figure 5.5: Force gauge device used to score the breaking strengths and the sample florets (with the panicle origin numbered and tracked) to illustrate the phenotyping procedure.

The final distributions of the mature BTS for the genotypes are therefore quite bimodal even without the quantitative measurements. I used 25g of mature BTS as a cutoff to distinguish the shattering/non-shattering genotypes, and 23 panicles (from 8 varieties) were scored as shattering and 52 panicles (from 13 varieties) were scored as non-shattering. These results are consistent with the qualitative hand tapping. One individual (BP06) did not flower in the five month period, so we moved the plant to the growth chamber to induce flowering. BP06, KFS and SP were not measured with force gauge but were verified as "shattering" varieties through hand tapping. The final phenotypes for the sorghum individuals are shown in **Table 5.2**.

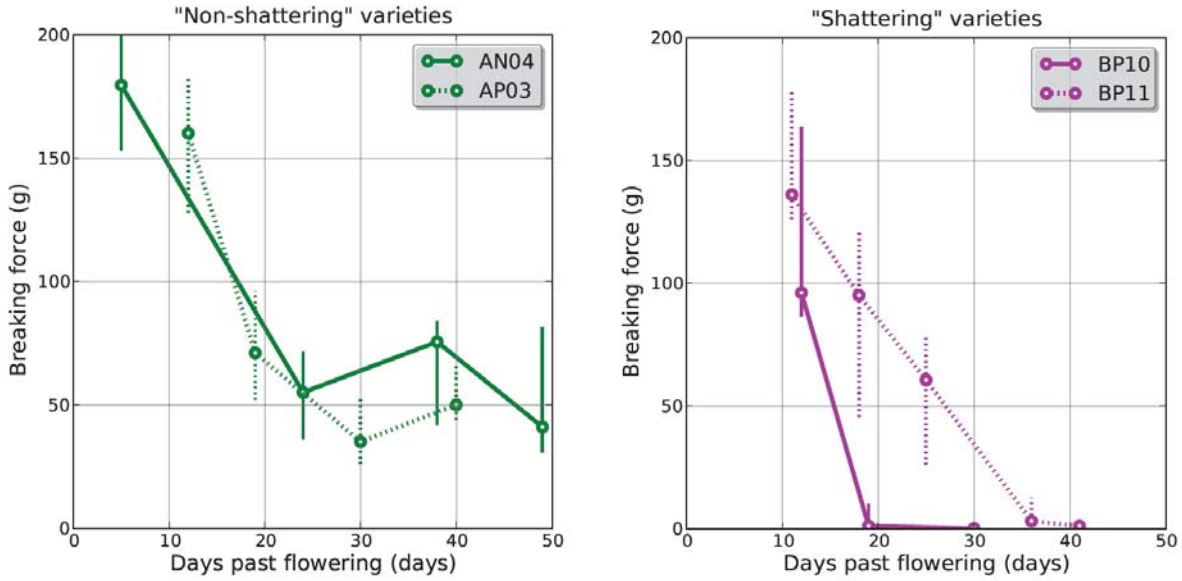

Figure 5.6: The progression of required breaking strengths for two example "non-shattering" varieties and two "shattering" varieties.

5.6.2 Amplification, resequencing and analyses of the genetic loci within the region

Among the predicted gene models within the region, 7 do not show differences at the protein level between non-shattering *S. bicolor* BTX623, and shattering *S. propinquum*. A total of 603 non-synonymous SNPs were found within the shattering region (on average 1.7nsSNP/kb), many of which are spatially clustered, making it possible to test several using single pairs of primers. Primers of 20-22bp that amplify between 700-1000bp amplicons were designed around the polymorphic sites of the candidate loci using PRIMER3 (Koressaar et al. 2007).

DNA was prepared from young leaves of individual plants. PCR reactions of 15µl per well were set up to amplify sampled regions using the following thermo-cycling program (ANN): 95˚C 30 sec, 58˚C 30 sec, 72˚C 1 min for a total of 36 cycles, 72˚C 10 min. The concentrations of the PCR amplicons were verified in 1% agarose gel and excessive primers and dNTPs in the PCR reactions were removed using exonuclease I and shrimp alkaline phosphatase enzymatic digestion. The amplicons were sequenced using BigDye 3.1 chemistry using the following

thermo-cycling program (BRISEQ): 96˚C 15 sec, 56˚C 30 sec, and 58.8˚C 1 min 30 sec for a total of 60 cycles. Excessive primers and dyes in the sequencing reactions were removed using Sephadex columns before the sequencing plates were loaded onto ABI3730 capillary sequencer.

The chromatograms were examined carefully using SEQUENCHER software (GENECODES Inc. version 4.1) and the polymorphisms were recorded in an EXCEL spreadsheet. From each PCR-amplicon sequence, I retained only the "informative" SNPs (tagging SNPs that are sufficient to reconstruct haplotype blocks), based on the observation that polymorphic sites within the same amplicon often show complete linkage disequilibrium (LD).

A total of 69 PCR fragments were sequenced with the DNA of 24 individuals in the compiled shattering/non-shattering panel. The public genome sequence of sorghum is from a non-shattering inbred cultivar *S. bicolor* BTX623 (Paterson et al. 2009), therefore a total of 25 different genotypes are available to be compared.

LD between multiple loci and the strength of marker-trait associations were analyzed using TASSEL (version 2.1) (Bradbury et al. 2007). I used r^2 as an indicator of linkage disequilibrium between pairwise SNP markers. Consider a pair of loci – alleles A/a in one and B/a in another, π_A, π_a, π_B, π_b are allele frequencies, π_{AB}, π_{aB}, π_{Ab}, π_{ab} are haplotype frequencies, then we have the following equation (Flint-Garcia et al. 2003),

$$r^2 = \frac{(\pi_{AB} - \pi_A \pi_B)^2}{\pi_A \pi_a \pi_B \pi_b}$$

For the association test, I used a generalized linear model (GLM) to evaluate the level of association between the shattering traits with the genotype data.

5.7 Results and discussion

5.7.1 Linkage disequilibrium in the *Sh1* region

A total of 58 informative sites were retained after removing a few sites with rare polymorphisms. The concatenated 58 sites comprise haplotype alignment among the individuals and were used

as input to the program TASSEL. Some sites are heterozygous for some individuals (e.g. plant #24 is heterozygous in least three sites). A total of 5 sites are indels (ranging from 3 to 11bp), but are treated similarly as SNP sites in the analysis.

Compared to maize, sorghum is a predominantly self-pollinating species with a range of outcrossing rates between 2% - 35%; Sorghum also has a smaller effective population size. Both factors can lead to higher levels of LD than maize (Hamblin et al. 2004). The strength of LD over the physical distance is shown in **Figure 5.7**. The LD in this region drops by half at a distance of ~500bp. This estimate of LD is largely consistent with a previous estimate of LD decay to 0.5 by 400bp (Hamblin et al. 2004).

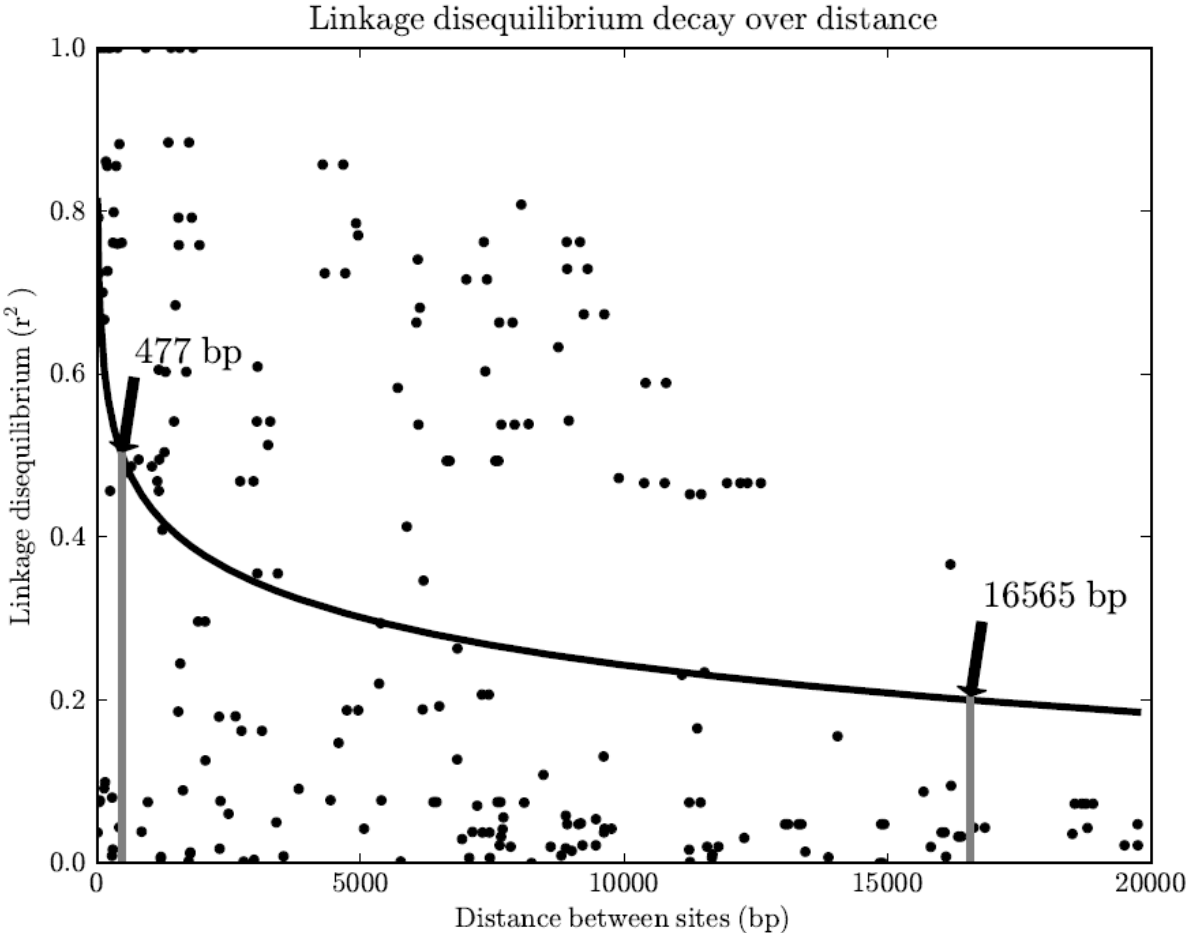

Figure 5.7: Strength of linkage disequilibrium over physical distance. The curve is the logarithmic fit of the data, and the distances at 477bp and 16565bp is shown as the distance where r^2 drops to 50% and 20%, respectively.

Pairwise LD values between the sampled sites were shown in **Figure 5.8**. Two relatively large LD blocks (with size ~48kb and ~44kb) are evident. Although the average estimate for our LD decay as calculated above is 477bp, in the two large LD blocks in **Figure 5.8**, sites that are separated by 40kb still show LD ~0.5.

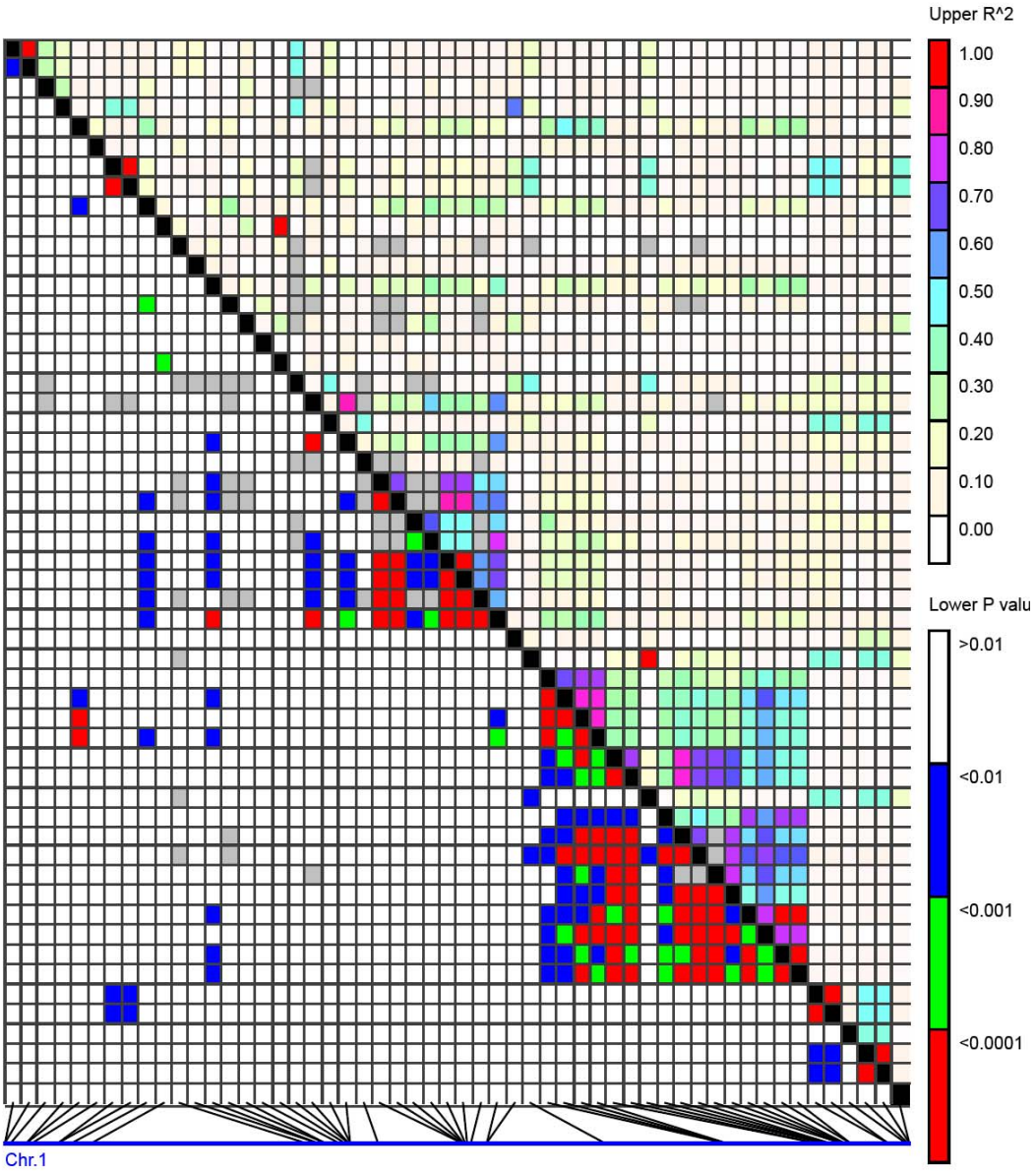

Figure 5.8: Pairwise LD matrix of the SNPs genotyped in this study. The markers are ordered according to their physical positions in the shattering region. The upper right matrix plots the pairwise r^2 score (ranging from 0 to 1, 1 means perfect LD). The lower left portion of the matrix plots the *P*-value from the Fisher's exact test (two-alleles) or test of independence (multiple alleles). Note that the SNP sites are not equally spaced.

There are also variations of LD in the region, as some regions do not show strong LD. This might be partially affected by the uneven sampling of polymorphic sites. Some LD occasionally persist over large distances and do not correspond to the tight linkage, as suggested in (Flint-Garcia et al. 2003).

5.7.2 Association analysis in the *Sh1* region

The general linear model (GLM) I used is a simple statistical model: $y = marker + e$, where y is the phenotype (0 for non-shattering, 1 for shattering). I chose not to include the population structure in the model, since I did not evaluate enough neutral markers to accurately estimate and to control for the population structure.

Among the 58 sites that I tested, I found 3 sites significantly associated with the shattering trait (amplicon P7E9, P3H11 and P4C3 in the shattering region) at significance level $P<0.001$ (**Figure 5.9; Figure 5.10**). The three sites are also in good LD. However, the intermediate sites between the two peaks are not significantly associated with the shattering trait.

The sites cover a region of ~50kb size and contain ~10 predicted gene models close to the three sites of high trait-association. Likely candidates are two transcription factors, but most predicted genes in this region have no good functional characterizations in sorghum or related plant species.

Additional PCR primers were designed to sample more sequences in the ~50kb region, in order to find the extent of the LD and also reveal sites that are even more associated with the shattering trait that might be the actual causal site or tightly linked sites. If we assume the causal locus *Sh1* indeed has perfect association with the shattering trait, the r^2 between P3H11 and *Sh1* is 0.48 – a relative tight linkage based on the LD decay trend in (**Figure 5.7**).

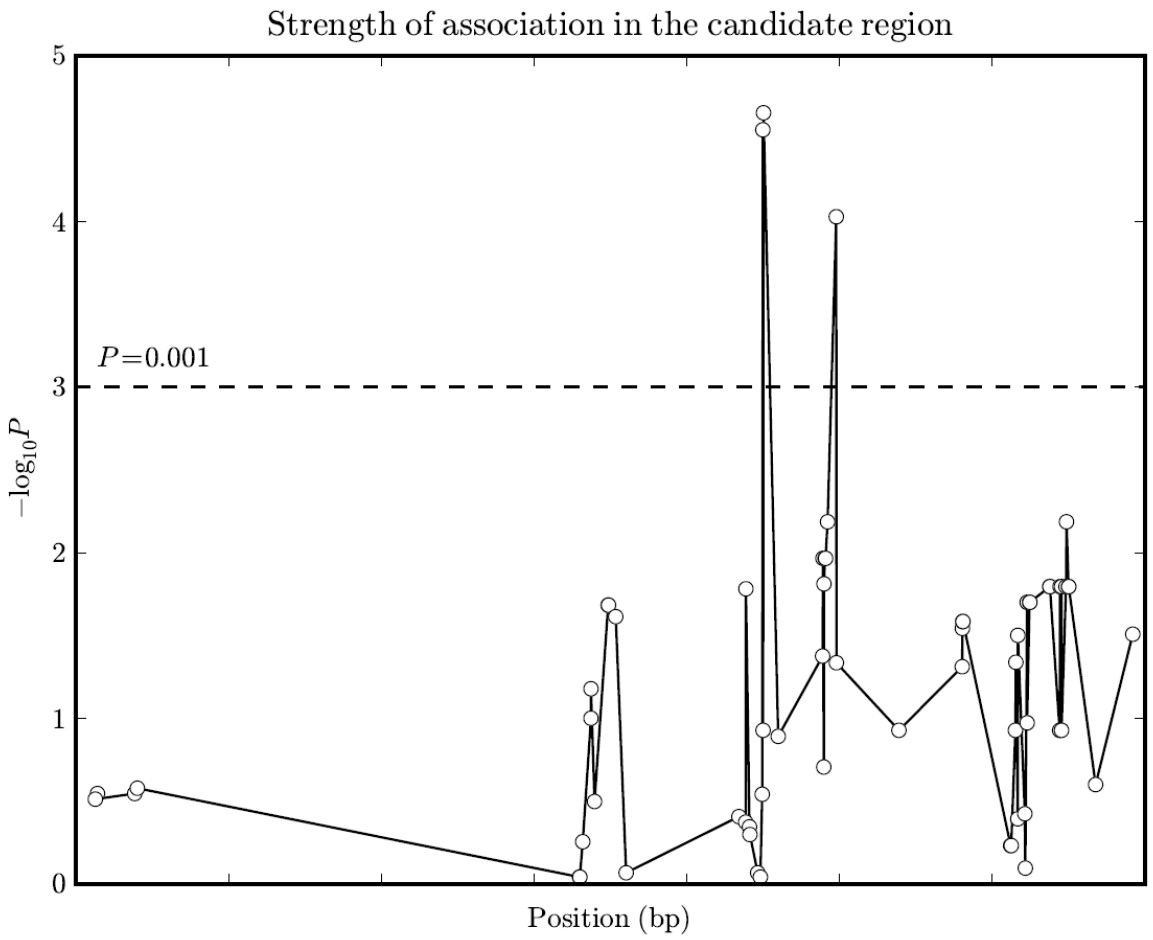

Figure 5.9: The strength of association at the tested SNP positions.

DNA	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
S/NS	N	N	N	N	N	S	S	S	S	N	S	S	S	N	N	S	N	N	N	N	S	S	S	N	N
P7E9	T	C	T	?	T	C	?	C	C	T	C	C	C	T	T	C	T	T	T/C	T	C	C	T	T	C
P3H11	G	C	G	G	G	C	C	C	C	G	C	C	C	G	G	C	G	G	C	G	C	C	G	G	C
P4C3	A	G	A	G	A	G	G	G	G	A	G	G	G	A	A	G	A	A	G	A	G	G	G	G	G

Figure 5.10: Two polymorphic sites with strong associations with the shattering trait (S/NS). Symbol "?" represents missing data (failed sequencing). The sites are listed in the order of increasing bp positions.

5.7.3 Relationship among the genotyped individuals

I also looked at the phylogenetic relationship among the haplotypes of the individuals. Visually, three sub-structures can be seen, note that #0 and #20 are the two parents used in the linkage mapping study (**Figure 5.11**). One clade contains *S. bicolor* BTX623 (#0) with four other non-shattering varieties, one clade contains *S. propinquum* (#20) and one other shattering variety, while the rest form the third clade with mixed shattering/non-shattering accessions. However, caution needs to be taken because of limited sampling of polymorphic sites (only 58 in total) in this study and potential bias induced by the selection (non-neutrality) at the candidate shattering locus.

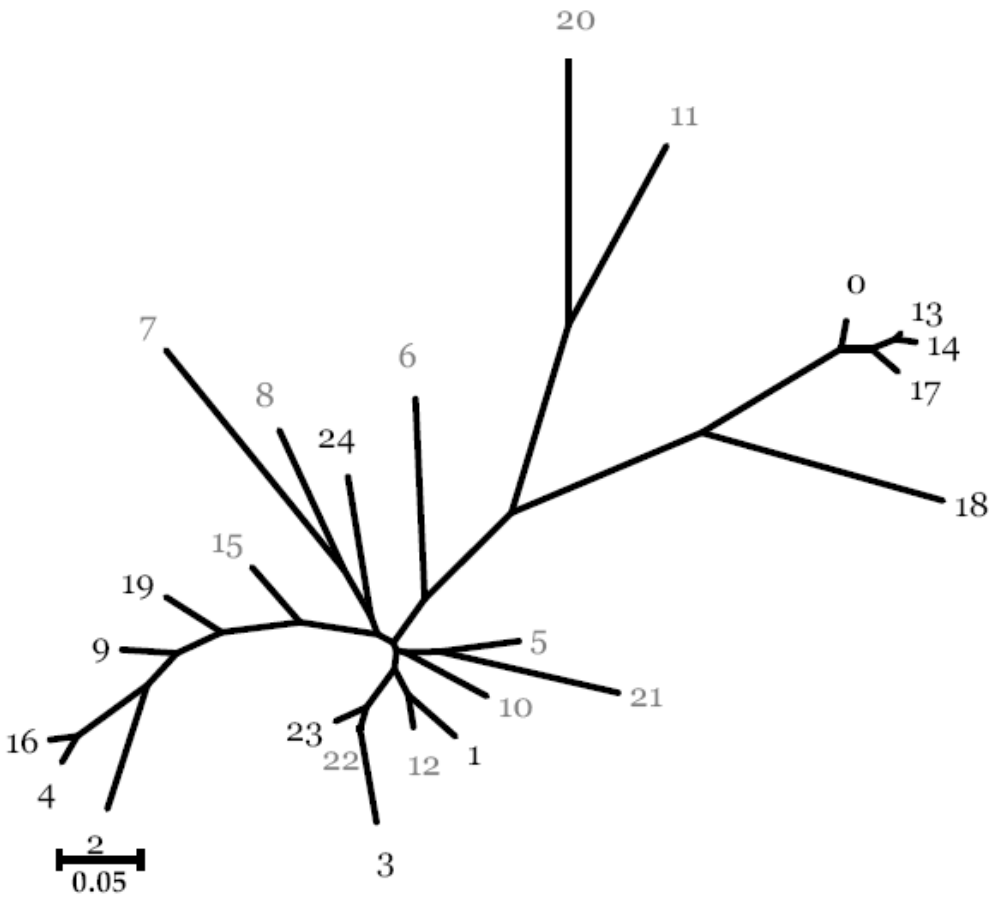

Figure 5.11: Neighbor-joining tree based on the 25 genotypes within the diversity panel. Gray labels are the accessions that shatter; black labels are the accessions that don't shatter. #0 is *S. bicolor* line BTX623, #20 is *S. propinquum*, the two parents used in the linkage mapping.

The rationale for the tree analysis is to see whether there is underlying population structure that accounts for the shattering/non-shattering varieties. If this were the case, then the associations that we identified might be false positives. We consider this unlikely, for two reasons. First, we can see that clade #3 in **Figure 5.11** includes both shattering/non-shattering individuals and therefore do not show significant partitions. Second, most sites in the region do not show significant association with the trait (except for the three sites shown in **Figure 5.10**).

5.8 Conclusion

As the result of the fine mapping effort, the sorghum *Sh1* gene is now mapped to a smaller chromosomal region than before, using an association-based approach based on a sorghum diversity panel. Linkage disequilibrium (LD) patterns are analyzed in the region surrounding the sites and the extent of LD decay over the physical distance is consistent with previous estimates. I further identified several sites that show significant associations with the shattering trait. Further resequencing and functional evaluations are needed to identify and confirm potential candidate loci in this region.

CHAPTER 6 CONCLUSIONS AND FUTURE PROSPECTS

6.1 Inference of synteny and collinearity

Identification of synteny and collinearity patterns across many genomes and subgenomes is a challenging computational task. Different assumptions, methods, criteria and models have been proposed, even with different optimizations depending on the organisms (Salse et al. 2009b). MCscan is my own software implementation to approach this problem. The algorithm in MCscan borrows heavily from previous theories in sequence alignments and extends to gene order alignments, providing a useful framework for analyzing the genome structure evolution in angiosperms.

The inference of synteny and collinearity is particularly important in studying molecular evolution in plants. Because of the relatively high variability in DNA substitution rates among plants, deviation from collinearity might be a more reliable phylogenetic character. DNA substitution rates can be highly variable among seed plant lineages (Smith et al. 2008), with extreme cases showing 100-fold variation within the same genus on the basis of a study of mitochondrial genes (Mower et al. 2007). Analysis of rare changes (when compared to DNA substitutions) in genomic structure such as specific rearrangements of gene order, insertions, or deletions — provides an informative and robust way to resolve relationships among many lineages (Rokas et al. 2000). In retrospect, early inferences on polyploidy in angiosperms and vertebrates were initially confused by gene phylogenies but later resolved with synteny (Dehal et al. 2005; Jaillon et al. 2007).

The emerging unified framework for comparative evolutionary analysis of angiosperm genes and genomes will improve in power and precision as more genomes are sequenced. Additional sequences from non-cereal genomes such as banana or pineapple, along with

sequences of basal eudicots such as California or opium poppy and columbine, and basal angiosperms such as *Amborella*, may further improve detection of collinearity and synteny across major angiosperm clades.

Improved synteny and collinearity alignments applied to multiple genomes and subgenomes are a potential foundation for reconstruction of the ancestral states of angiosperm genomes. Consensus gene orders within syntenic blocks can be approximated on the basis of top-down alignments. Ordering among the syntenic blocks themselves on the macro-level is more difficult; however, several combinatorial algorithms exist to reconstruct ancestral genomes under a most-parsimonious rearrangement scenario (Eichler et al. 2003). The resulting orders would reveal not only shared but also divergent genes inserted into novel locations, underlining lineage-specific changes. Additional genome sequences will improve power to resolve gene orders at the micro-level and also contribute to identifying functionally important DNA, such as the evolutionarily constrained elements among 28 vertebrate genomes (Miller et al. 2007).

6.2 Inference of paleopolyploidy

It is now widely accepted that ancient polyploidy events have affected many plant lineages since the evidence emerged from the analysis of the genome sequence from *Arabidopsis* 10 years ago (Paterson et al. 2000; Vision et al. 2000). Many analyses follow up with improved methodology for the identification of duplicated segments and phylogenetic dating of these polyploidy events, as reviewed in (Van de Peer et al. 2009a).

The analyses of deep or very ancient WGDs are of particular interest for plant researchers, for two reasons. First, because of the early occurrence of these events, they can potentially affect many plant lineages and impact the synteny correspondences between plants of divergent lineages. Second, understanding these ancient WGDs can help us determine the timing of the expansions of many plant gene families that might contribute to morphological diversifications and innovations of flowering plants.

The initial "bottom-up" (intra-genomic) analyses of the *Arabidopsis* (Bowers et al. 2003b) and rice genome (Paterson et al. 2004) yielded insightful results, yet limited by the number of plant genomes that can be compared at the time. The important addition to the available tools for uncovering these ancient duplication events is the development of "top-down" (inter-genomic) analysis (Tang et al. 2008a). The comparisons among many taxa complement the earlier intra-genomic analysis, and provide a much clearer picture of the duplication landscape.

The availability of the poplar, grape and papaya genome sequence offers much resolution to the WGDs in the eudicot lineage. This has produced the surprising result that the previous dating of the β and γ duplication events in (Bowers et al. 2003b) was erroneous. The β duplication was previously thought to be associated with the eudicot radiation but is now suggested to have occurred after the *Arabidopsis*-papaya divergence (Ming et al. 2008), while instead the γ duplication appears to be associated with the eudicot radiation. This incongruence is again caused by the large rate difference between plant lineages and has been extensively studied in (Tang et al. 2008b) and (Van de Peer et al. 2009a).

More knowledge has been gained from the comparisons within and between cereal genomes regarding the pan-cereal WGD (ρ). Some analyses hinting at duplication events that are more ancient than the ρ event were available but not quite conclusive (Salse et al. 2008; Zhang et al. 2005). Therefore, I did an in-depth analysis and suggested that the more ancient WGDs did indeed exist and perhaps involve two additional doublings, collectively called σ. The timing of σ is still unclear due to the scarce genome data from basal monocot species and uncertainties of molecular clock assumption (Vicentini et al. 2008), but is considered to have occurred only within the monocot lineage.

Full knowledge and better characterizations of these events in both monocot and eudicot lineages led to better analyses between the genomes from the two clades. Some earlier results suggested that the synteny between monocots and eudicots was poor, with only local synteny

stretching to a few genes (Liu et al. 2001; Salse et al. 2009a; Salse et al. 2002). The reason for the reported limited synteny is that WGDs in both lineages have "scattered" the synteny signals in many chromosomal regions in the modern genomes. The PAR analyses show that such comparisons across these two lineages are still possible, when the multiple-to-multiple synteny regions are pooled and interpreted simultaneously.

6.3 Association mapping of sorghum shattering gene

The mutation within the sorghum shattering locus *Sh1* is a key transition from wild to domesticated sorghum, and was perhaps utilized and spread by the humans. The original mapping study in a family of wild × domesticated sorghum cross, performed ~15 years ago, mapped the locus to a particular region on sorghum chromosome 1 (linkage group C) (Paterson et al. 1995). Progress has been slow since then, due to limited genomic resources for the sorghum species.

With the sorghum genetic map, physical map and genome sequence finished one after another, a new approach is now used to try to fine-map *Sh1*. The path that I took is still based on the principle of gene-trait association, but now exploits a diversity panel consisting of unrelated naturally occurring individuals rather than traditional synthetic F2 family. The use of diverse sampling of individuals with smaller linkage disequilibrium (LD) has improved the resolution of the mapping. Exploiting the genomic sequences for both parental species, I genotyped the diversity panel at different loci within the target region and found several sites that show strong association with the shattering trait. A second round of resequencing surrounding these sites is still under way, to accurately define the boundary of LD and also to search for sites with better association with the trait that might be more tightly linked to the *causal* site.

For future functional study as follow-up, we can transform non-shattering sorghums with the candidate shattering allele *Sh1* to complement the non-shattering phenotype. Following the transformation, we expect to see dispersal of mature grains in the transformants, using the

same method that we used to phenotype the diversity panel. Functional complementation offers the ultimate proof for the function of the candidate gene suggested in this study.

The nature of the mutation in the *Sh1* locus is still unclear. It was originally thought that the non-shattering trait was perhaps induced by a loss-of-function mutation in the *Sh1* locus. However, if *S. bicolor* and *S. propinquum* alleles are both functional variants (a change-of-function mutation), then we can also test the candidate genes in the sorghum TILLING population (Xin et al. 2008).

Once the locus is identified, we can explore the spatial and temporal expression of the *Sh1* locus, to suggest when and where *Sh1* participates in the development of grain abscission. It will be attractive if we can correlate the time-course gene expression with the trend of decreasing breaking force of pedicels. We can also take advantage of the data to test whether we can identify signals of domestication around *Sh1* locus. We expect to find reduced genetic diversity in domesticated compared to wild sorghums, as expected from intense artificial selection. More accessions from both cultivated and wild sorghums in various different global populations can be genotyped to determine the frequency of the mutation and also track the course of sorghum domestication.

REFERENCES

AGI. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Aguilar-Melendez, A., P.L. Morrell, M.L. Roose, and S.-C. Kim. 2009. Genetic diversity and structure in semiwild and domesticated chiles (Capsicum annuum; Solanaceae) from Mexico. *Am. J. Bot.* **96:** 1190-1202.

Allaby, R.G., D.Q. Fuller, and T.A. Brown. 2008. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc Natl Acad Sci U S A* **105:** 13982-13986.

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang et al. 2005. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* **1:** e60.

Aury, J.M., O. Jaillon, L. Duret, B. Noel, C. Jubin, B.M. Porcel, B. Segurens, V. Daubin, V. Anthouard, N. Aiach et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444:** 171-178.

Bennett, M.D. and J.B. Smith. 1991. Nuclear DNA Amounts in Angiosperms. *Philosophical Transactions of the Royal Society B: Biological Sciences* **334:** 309-345.

Bhave, M.R., S. Lawrence, C. Barton, and L.C. Hannah. 1990. Identification and molecular characterization of shrunken-2 cDNA clones of maize. *Plant Cell* **2:** 581-588.

Bikard, D., D. Patel, C. Le Mette, V. Giorgi, C. Camilleri, M.J. Bennett, and O. Loudet. 2009. Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. *Science* **323:** 623-626.

Bjornerfeldt, S., M.T. Webster, and C. Vila. 2006. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res* **16:** 990-994.

Blanc, G. and K.H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16:** 1667-1678.

Bowers, J.E., C. Abbey, S. Anderson, C. Chang, X. Draye, A.H. Hoppe, R. Jessup, C. Lemke, J. Lennington, Z. Li et al. 2003a. A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* **165:** 367-386.

Bowers, J.E., M.A. Arias, R. Asher, J.A. Avise, R.T. Ball, G.A. Brewer, R.W. Buss, A.H. Chen, T.M. Edwards, J.C. Estill et al. 2005. Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc Natl Acad Sci U S A* **102:** 13206-13211.

Bowers, J.E., B.A. Chapman, J. Rong, and A.H. Paterson. 2003b. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433-438.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23:** 2633-2635.

Brown, T.A., M.K. Jones, W. Powell, and R.G. Allaby. 2009. The complex origins of domesticated crops in the Fertile Crescent. *Trends in Ecology & Evolution* **24:** 103-109.

Buckler, E.S., J.B. Holland, P.J. Bradbury, C.B. Acharya, P.J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J.C. Glaubitz et al. 2009. The genetic architecture of maize flowering time. *Science* **325:** 714-718.

Caicedo, A.L., S.H. Williamson, R.D. Hernandez, A. Boyko, A. Fledel-Alon, T.L. York, N.R. Polato, K.M. Olsen, R. Nielsen, S.R. McCouch et al. 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3:** 1745-1756.

Calabrese, P.P., S. Chakravarty, and T.J. Vision. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19 Suppl 1:** i74-80.

Cannon, S.B., A. Kozik, B. Chan, R. Michelmore, and N.D. Young. 2003. DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol* **4:** R68.

Casa, A.M., S.E. Mitchell, M.T. Hamblin, H. Sun, J.E. Bowers, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2005. Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. *Theor Appl Genet* **111:** 23-30.

Cavanagh, C., M. Morell, I. Mackay, and W. Powell. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr Opin Plant Biol* **11:** 215-221.

Chapman, M.A., C.H. Pashley, J. Wenzler, J. Hvala, S. Tang, S.J. Knapp, and J.M. Burke. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (Helianthus annuus). *Plant Cell* **20:** 2931-2945.

Chardon, F., B. Virlon, L. Moreau, M. Falque, J. Joets, L. Decousset, A. Murigneux, and A. Charcosset. 2004. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics* **168:** 2169-2185.

Chrispeels, M., Sadava, D. 2003. *Plants, genes, and crop biotechnology*. Jones and Bartlett Publishers, Sudbury.

Coghlan, A., E.E. Eichler, S.G. Oliver, A.H. Paterson, and L. Stein. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet* **21:** 673-682.

Cui, L., P.K. Wall, J.H. Leebens-Mack, B.G. Lindsay, D.E. Soltis, J.J. Doyle, P.S. Soltis, J.E. Carlson, K. Arumuganathan, A. Barakat et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16:** 738-749.

Dahlberg, J. 1995. Dispersal of sorghum and the role of genetic drift. *African Crop Science Journal* **3:** 143-151.

Davies, T.J., T.G. Barraclough, M.W. Chase, P.S. Soltis, D.E. Soltis, and V. Savolainen. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc Natl Acad Sci U S A* **101:** 1904-1909.

Dehal, P. and J.L. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3:** e314.

Doebley, J. 2004. The genetics of maize evolution. *Annu Rev Genet* **38:** 37-59.

Doebley, J.F., B.S. Gaut, and B.D. Smith. 2006. The molecular genetics of crop domestication. *Cell* **127:** 1309-1321.

Doggett, H. 1976. Sorghum. In *Evolution of Crop Plants* (ed. N. Simmonds), pp. 112-117. Longman, Essex, UK.

Draye, X., Y.R. Lin, X.Y. Qian, J.E. Bowers, G.B. Burow, P.L. Morrell, D.G. Peterson, G.G. Presting, S.X. Ren, R.A. Wing et al. 2001. Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. *Plant Physiol* **125:** 1325-1341.

Du, C., J. Caronna, L. He, and H.K. Dooner. 2008. Computational prediction and molecular confirmation of Helitron transposons in the maize genome. *BMC Genomics* **9:** 51.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge.

Ehrenreich, I.M., Y. Hanzawa, L. Chou, J.L. Roe, P.X. Kover, and M.D. Purugganan. 2009. Candidate gene association mapping of Arabidopsis flowering time. *Genetics* **183:** 325-335.

Eichler, E.E. and D. Sankoff. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* **301:** 793-797.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8:** 175-185.

Fares, M.A., K.P. Byrne, and K.H. Wolfe. 2006. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Mol Biol Evol* **23:** 245-253.

Fawcett, J.A., S. Maere, and Y. Van de Peer. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A* **106:** 5737-5742.

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer, Sunderland, MA.

Ferguson-Smith, M.A. and V. Trifonov. 2007. Mammalian karyotype evolution. *Nat Rev Genet* **8:** 950-962.

Ferrandiz, C., S.J. Liljegren, and M.F. Yanofsky. 2000. Negative regulation of the SHATTERPROOF genes by FRUITFULL during Arabidopsis fruit development. *Science* **289:** 436-438.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S.t. Buckler. 2003. Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54:** 357-374.

Frary, A., T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K.B. Alpert, and S.D. Tanksley. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289:** 85-88.

Frazer, K.A., L. Pachter, A. Poliakov, E.M. Rubin, and I. Dubchak. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32:** W273-279.

Freeling, M. and B.C. Thomas. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* **16:** 805-814.

Gaut, B.S., B.R. Morton, B.C. McCaig, and M.T. Clegg. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci U S A* **93:** 10274-10279.

Glemin, S. and T. Bataillon. 2009. A comparative view of the evolution of grasses under domestication. *New Phytol* **183:** 273-290.

Haas, B.J., A.L. Delcher, J.R. Wortman, and S.L. Salzberg. 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20:** 3643-3646.

Hamblin, M.T., A.M. Casa, H. Sun, S.C. Murray, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: lessons from Sorghum bicolor. *Genetics* **173:** 953-964.

Hamblin, M.T., S.E. Mitchell, G.M. White, J. Gallego, R. Kukatla, R.A. Wing, A.H. Paterson, and S. Kresovich. 2004. Comparative population genetics of the panicoid grasses: sequence

polymorphism, linkage disequilibrium and selection in a diverse sample of sorghum bicolor. *Genetics* **167:** 471-483.

Hammer, K. 1984. Das Domestikationssyndrom. *Kulturpflanze***:** 11-34.

Hampson, S.E., B.S. Gaut, and P. Baldi. 2005. Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* **21:** 1339-1348.

Hansen, B.G., B.A. Halkier, and D.J. Kliebenstein. 2008. Identifying the molecular basis of QTLs: eQTLs add a new dimension. *Trends Plant Sci* **13:** 72-77.

Hedges, S.B. and S. Kumar. 2004. Precision of molecular time estimates. *Trends Genet* **20:** 242-247.

Hein, J., C. Wiuf, B. Knudsen, M.B. Moller, and G. Wibling. 2000. Statistical alignment: computational properties, homology testing and goodness-of-fit. *J Mol Biol* **302:** 265-279.

Hittinger, C.T. and S.B. Carroll. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449:** 677-681.

Holmes, I. and W.J. Bruno. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17:** 803-820.

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. Wang, J. Guan, D. Fan, Q. Weng, T. Huang et al. 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res* **19:** 1068-1076.

Huda, A. and I.K. Jordan. 2009. Analysis of transposable element sequences using CENSOR and RepeatMasker. *Methods Mol Biol* **537:** 323-336.

IRGSP. 2005. The map-based sequence of the rice genome. *Nature* **436:** 793-800.

Izawa, T., S. Konishi, A. Shomura, and M. Yano. 2009. DNA changes tell us about rice domestication. *Curr Opin Plant Biol* **12:** 185-192.

Jaillon, O., J.M. Aury, F. Brunet, J.L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431:** 946-957.

Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449:** 463-467.

Jakse, J., A. Telgmann, C. Jung, A. Khar, S. Melgar, F. Cheung, C.D. Town, and M.J. Havey. 2006. Comparative sequence and genetic analyses of asparagus BACs reveal no microsynteny with onion or rice. *Theor Appl Genet* **114:** 31-39.

Jannoo, N., L. Grivet, N. Chantret, O. Garsmeur, J.C. Glaszmann, P. Arruda, and A. D'Hont. 2007. Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J* **50:** 574-585.

Kellis, M., B.W. Birren, and E.S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428:** 617-624.

Kent, W.J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100:** 11484-11489.

Kilian, B., H. Ozkan, A. Walther, J. Kohl, T. Dagan, F. Salamini, and W. Martin. 2007. Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during Triticum monococcum (Einkorn) domestication: implications for the origin of agriculture. *Mol Biol Evol* **24:** 2657-2668.

Kim, J.S., M.N. Islam-Faridi, P.E. Klein, D.M. Stelly, H.J. Price, R.R. Klein, and J.E. Mullet. 2005. Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. *Genetics* **171:** 1963-1976.

Koch, M.A., B. Haubold, and T. Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). *Mol Biol Evol* **17:** 1483-1498.

Konishi, S., K. Ebana, and T. Izawa. 2008. Inference of the japonica rice domestication process from the distribution of six functional nucleotide polymorphisms of domestication-related genes in various landraces and modern cultivars. *Plant Cell Physiol* **49:** 1283-1293.

Konishi, S., T. Izawa, S.Y. Lin, K. Ebana, Y. Fukuta, T. Sasaki, and M. Yano. 2006. An SNP caused loss of seed shattering during rice domestication. *Science* **312:** 1392-1396.

Koressaar, T. and M. Remm. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23:** 1289-1291.

Ku, H.M., T. Vision, J. Liu, and S.D. Tanksley. 2000. Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci U S A* **97:** 9121-9126.

Kurtz, S., A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12.

Lee, C., C. Grasso, and M.F. Sharlow. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18:** 452-464.

Lescot, M., P. Piffanelli, A.Y. Ciampi, M. Ruiz, G. Blanc, J. Leebens-Mack, F.R. da Silva, C.M. Santos, A. D'Hont, O. Garsmeur et al. 2008. Insights into the Musa genome: syntenic relationships to rice and between Musa species. *BMC Genomics* **9:** 58.

Li, C., A. Zhou, and T. Sang. 2006a. Rice domestication by reducing shattering. *Science* **311:** 1936-1939.

Li, W. and B.S. Gill. 2006b. Multiple genetic pathways for seed shattering in the grasses. *Funct Integr Genomics* **6:** 300-309.

Liljegren, S.J., G.S. Ditta, Y. Eshed, B. Savidge, J.L. Bowman, and M.F. Yanofsky. 2000. SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. *Nature* **404:** 766-770.

Lin, Y.-R., L. Zhu, S. Ren, J. Yang, K.F. Schertz, and A.H. Paterson. 1999. A Sorghum propinquum BAC library, suitable for cloning genes associated with loss-of-function mutations during crop domestication. *Molecular Breeding* **5:** 511-520.

Lin, Z., M.E. Griffith, X. Li, Z. Zhu, L. Tan, Y. Fu, W. Zhang, X. Wang, D. Xie, and C. Sun. 2007. Origin of seed shattering in rice (Oryza sativa L.). *Planta* **226:** 11-20.

Liu, H., R. Sachidanandam, and L. Stein. 2001. Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order. *Genome Res* **11:** 2020-2026.

Londo, J.P., Y.C. Chiang, K.H. Hung, T.Y. Chiang, and B.A. Schaal. 2006. Phylogeography of Asian wild rice, Oryza rufipogon, reveals multiple independent domestications of cultivated rice, Oryza sativa. *Proc Natl Acad Sci U S A* **103:** 9578-9583.

Lu, J., T. Tang, H. Tang, J. Huang, S. Shi, and C.I. Wu. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet* **22:** 126-131.

Lynch, M. and A.G. Force. 2000. The origin of interspecific genomic incompatibility via gene duplication. *American Naturalist* **156:** 590-605.

Lyons, E. and M. Freeling. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* **53:** 661-673.

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102:** 5454-5459.

McNally, K.L., K.L. Childs, R. Bohnert, R.M. Davidson, K. Zhao, V.J. Ulat, G. Zeller, R.M. Clark, D.R. Hoen, T.E. Bureau et al. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci U S A* **106:** 12273-12278.

Miller, W., K. Rosenbloom, R.C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D.C. King, R. Baertsch, D. Blankenberg et al. 2007. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17:** 1797-1808.

Ming, R., S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J.H. Saw, P. Senin, W. Wang, B.V. Ly, K.L. Lewis et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452:** 991-996.

Mower, J.P., P. Touzet, J.S. Gummow, L.F. Delph, and J.D. Palmer. 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. *BMC Evol Biol* **7:** 135.

Nakatani, Y., H. Takeda, Y. Kohara, and S. Morishita. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17:** 1254-1265.

Nalam, V.J., M.I. Vales, C.J. Watson, E.B. Johnson, and O. Riera-Lizarazu. 2007. Map-based analysis of genetic loci on chromosome 2D that affect glume tenacity and threshability, components of the free-threshing habit in common wheat (Triticum aestivum L.). *Theor Appl Genet* **116:** 135-145.

Nalam, V.J., M.I. Vales, C.J. Watson, S.F. Kianian, and O. Riera-Lizarazu. 2006. Map-based analysis of genes affecting the brittle rachis character in tetraploid wheat (Triticum turgidum L.). *Theor Appl Genet* **112:** 373-381.

Palaisa, K., M. Morgante, S. Tingey, and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc Natl Acad Sci U S A* **101:** 9885-9890.

Paterson, A.H. 2002. What has QTL mapping taught us about plant domestication? *New Phytologist* **154:** 591-608.

Paterson, A.H., J.E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457:** 551-556.

Paterson, A.H., J.E. Bowers, M.D. Burow, X. Draye, C.G. Elsik, C.X. Jiang, C.S. Katsar, T.H. Lan, Y.R. Lin, R. Ming et al. 2000. Comparative genomics of plant chromosomes. *Plant Cell* **12:** 1523-1540.

Paterson, A.H., J.E. Bowers, and B.A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A* **101:** 9903-9908.

Paterson, A.H., T.H. Lan, K.P. Reischmann, C. Chang, Y.R. Lin, S.C. Liu, M.D. Burow, S.P. Kowalski, C.S. Katsar, T.A. DelMonte et al. 1996. Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nat Genet* **14:** 380-382.

Paterson, A.H., Y.R. Lin, Z.K. Li, K.F. Schertz, J.F. Doebley, S.R.M. Pinson, S.C. Liu, J.W. Stansel, and J.E. Irvine. 1995. Convergent Domestication of Cereal Crops by Independent Mutations at Corresponding Genetic-Loci. *Science* **269:** 1714-1718.

Putnam, N.H., T. Butts, D.E. Ferrier, R.F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. Terry, J.K. Yu et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453:** 1064-1071.

Putnam, N.H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro, E. Lindquist, V.V. Kapitonov et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317:** 86-94.

Rodelsperger, C. and C. Dieterich. 2008. Syntenator: Multiple gene order alignments with a gene-specific scoring function. *Algorithms Mol Biol* **3:** 14.

Roeder, A.H., C. Ferrandiz, and M.F. Yanofsky. 2003. The role of the REPLUMLESS homeodomain protein in patterning the Arabidopsis fruit. *Curr Biol* **13:** 1630-1635.

Rokas, A. and P.W. Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* **15:** 454-459.

Rounsley, S., Marri, P.R., Yu, Y., He, R., Sisneros, N., Goicoechea, J.L., Lee, S.J., Angelova, A., Kudrna, D., Luo, M., Affourtit, J., Desany, B., Knight, J., Niazi, F., Egholm, M., Wing, R.A. 2009. De Novo Next Generation Sequencing of Plant Genomes. *Rice***:** 35-43.

Salse, J., M. Abrouk, S. Bolot, N. Guilhot, E. Courcelle, T. Faraut, R. Waugh, T.J. Close, J. Messing, and C. Feuillet. 2009a. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci U S A* **106:** 14908-14913.

Salse, J., M. Abrouk, F. Murat, U.M. Quraishi, and C. Feuillet. 2009b. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief Bioinform*.

Salse, J., S. Bolot, M. Throude, V. Jouffe, B. Piegu, U.M. Quraishi, T. Calcagno, R. Cooke, M. Delseny, and C. Feuillet. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20:** 11-24.

Salse, J., B. Piegu, R. Cooke, and M. Delseny. 2002. Synteny between Arabidopsis thaliana and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res* **30:** 2316-2328.

Scannell, D.R., K.P. Byrne, J.L. Gordon, S. Wong, and K.H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440:** 341-345.

Scannell, D.R., A.C. Frank, G.C. Conant, K.P. Byrne, M. Woolfit, and K.H. Wolfe. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A* **104:** 8397-8402.

Shang, J., Y. Tao, X. Chen, Y. Zou, C. Lei, J. Wang, X. Li, X. Zhao, M. Zhang, Z. Lu 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:** 1303-1311.

Shi, X., X. Wang, Z. Li, Q. Zhu, W. Tang, S. Ge, and J. Luo. 2006. Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias. *Gene* **376:** 199-206.

Simillion, C., K. Janssens, L. Sterck, and Y. Van de Peer. 2008. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24:** 127-128.

Sjodin, A., N.R. Street, G. Sandberg, P. Gustafsson, and S. Jansson. 2009. The Populus Genome Integrative Explorer (PopGenIE): a new resource for exploring the Populus genome. *New Phytol.*

Smith, S.A. and M.J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* **322:** 86-89.

Soderlund, C., Longden, I. & Mott, R. 1997. FPC: A system for building contigs from restriction fingerprinted clones. *Cabios* **13:** 523-535.

Soltis, D.E., V.A. Albert, J. Leebens-Mack, C.D. Bell, A.H. Paterson, C. Zheng, D. Sankoff, C.W. dePamphilis, P.K. Wall, and P.S. Soltis. 2009. Polyploidy and angiosperm diversification. *Am. J. Bot.* **96:** 336-348.

Soltis, D.E., P.S. Soltis, P.K. Endress, and M.W. Chase. 2005. *Phylogeny and evolution of angiosperms.*

Spangler, R., B. Zaitchik, E. Russo, and E. Kellogg. 1999. Andropogoneae Evolution and Generic Limits in Sorghum (Poaceae) Using ndhF Sequences. *Systematic Botany* **24:** 267-281.

Spillane, C., K.J. Schmid, S. Laoueille-Duprat, S. Pien, J.M. Escobar-Restrepo, C. Baroux, V. Gagliardini, D.R. Page, K.H. Wolfe, and U. Grossniklaus. 2007. Positive darwinian selection at the imprinted MEDEA locus in plants. *Nature* **448:** 349-352.

Swarbreck, D., C. Wilks, P. Lamesch, T.Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz et al. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36:** D1009-1014.

Sweeney, M.T., M.J. Thomson, B.E. Pfeil, and S. McCouch. 2006. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* **18:** 283-294.

Swigonova, Z., J. Lai, J. Ma, W. Ramakrishna, V. Llaca, J.L. Bennetzen, and J. Messing. 2004. Close split of sorghum and maize genome progenitors. *Genome Res* **14:** 1916-1923.

Tang, H., J. Bowers, X. Wang, R. Ming, M. Alam, and A. Paterson. 2008a. Synteny and Collinearity in Plant Genomes. *Science* **320:** 486-488.

Tang, H., X. Wang, J.E. Bowers, R. Ming, M. Alam, and A.H. Paterson. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18:** 1944-1954.

Tenaillon, M.I., J. U'Ren, O. Tenaillon, and B.S. Gaut. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol Biol Evol* **21:** 1214-1225.

Thomas, B.C., B. Pedersen, and M. Freeling. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* **16:** 934-946.

Thorne, J.L., H. Kishino, and J. Felsenstein. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol* **33:** 114-124.

Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S.t. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* **28:** 286-289.

Tian, F., N.M. Stevens, and E.S.t. Buckler. 2009. Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proc Natl Acad Sci U S A* **106 Suppl 1:** 9979-9986.

Tuskan, G.A. S. Difazio S. Jansson J. Bohlmann I. Grigoriev U. Hellsten N. Putnam S. Ralph S. Rombauts A. Salamov et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313:** 1596-1604.

Van de Peer, Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* **5:** 752-763.

Van de Peer, Y., J.A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele. 2009a. The flowering world: a tale of duplications. *Trends Plant Sci.*

Van de Peer, Y., S. Maere, and A. Meyer. 2009b. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10:** 725-732.

van Leeuwen, H., D.J. Kliebenstein, M.A. West, K. Kim, R. van Poecke, F. Katagiri, R.W. Michelmore, R.W. Doerge, and D.A. St Clair. 2007. Natural variation among Arabidopsis thaliana accessions for transcriptome response to exogenous salicylic acid. *Plant Cell* **19:** 2099-2110.

Vandepoele, K., Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res* **12:** 1792-1801.

Vicentini, A., J.C. Barber, S.S. Aliscioni, L.M. Giussani, and E.A. Kellogg. 2008. The age of the grasses and clusters of origins of C₄ photosynthesis. *Global Change Biology* **14:** 2963-2977.

Vision, T.J., D.G. Brown, and S.D. Tanksley. 2000. The origins of genomic duplications in Arabidopsis. *Science* **290:** 2114-2117.

Vollbrecht, E., P.S. Springer, L. Goh, E.S.t. Buckler, and R. Martienssen. 2005. Architecture of floral branch systems in maize and related grasses. *Nature* **436:** 1119-1126.

Wang, R.L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. *Nature* **398:** 236-239.

Wang, X., X. Shi, B. Hao, S. Ge, and J. Luo. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol* **165:** 937-946.

Wang, X., X. Shi, Z. Li, Q. Zhu, L. Kong, W. Tang, S. Ge, and J. Luo. 2006. Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC Bioinformatics* **7:** 447.

Wang, X., H. Tang, J.E. Bowers, F.A. Feltus, and A.H. Paterson. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177:** 1753-1763.

Wang, X., H. Tang, J.E. Bowers, and A.H. Paterson. 2009. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res* **19:** 1026-1032.

Wang, Z.Y., F.Q. Zheng, G.Z. Shen, J.P. Gao, D.P. Snustad, M.G. Li, J.L. Zhang, and M.M. Hong. 1995. The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. *Plant J* **7:** 613-622.

Wright, S.I., I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen, and B.S. Gaut. 2005. The effects of artificial selection on the maize genome. *Science* **308:** 1310-1314.

Xiao, H., N. Jiang, E. Schaffner, E.J. Stockinger, and E. van der Knaap. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319:** 1527-1530.

Xin, Z., M.L. Wang, N.A. Barkley, G. Burow, C. Franks, G. Pederson, and J. Burke. 2008. Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. *BMC Plant Biol* **8:** 103.

Yamamoto, T., J. Yonemaru, and M. Yano. 2009. Towards the understanding of complex traits in rice: substantially or superficially? *DNA Res* **16:** 141-154.

Yamasaki, M., S.I. Wright, and M.D. McMullen. 2007. Genomic screening for artificial selection during domestication and improvement in maize. *Ann Bot (Lond)* **100:** 967-973.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.

Yang, Z. and J.P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15:** 496-503.

Yogeeswaran, K., A. Frary, T.L. York, A. Amenta, A.H. Lesser, J.B. Nasrallah, S.D. Tanksley, and M.E. Nasrallah. 2005. Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of A. thaliana. *Genome Res* **15:** 505-515.

Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178:** 539-551.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38:** 203-208.

Zhang, L.B., Q. Zhu, Z.Q. Wu, J. Ross-Ibarra, B.S. Gaut, S. Ge, and T. Sang. 2009. Selection on grain shattering genes and rates of rice domestication. *New Phytol.*

Zhang, Y., G.H. Xu, X.Y. Guo, and L.J. Fan. 2005. Two ancient rounds of polyploidy in rice genome. *J Zhejiang Univ Sci B* **6:** 87-90.

Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and Prospects of Association Mapping in Plants. *The Plant Genome* **1:** 5-20.

Zuzana Swigonova, J.L., Jianxin Ma, Wusirika Ramakrishna, Victor Llaca, Jeffrey L. Bennetzen, and Joachim Messing. 2004. Close Split of Sorghum and Maize Genome Progenitors. *Genome Research***:** 1916-1923.