The Pennsylvania State University

The Graduate School

Intercollege Graduate Program in Genetics

UNDERSTANDING GENE EXPRESSION AND GENETIC RECOMBINATION BY NEXT GENERATION SEQUENCING

A Dissertation in

Genetics

by

Xinwei Han

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2012

The dissertation of Xinwei Han was reviewed and approved* by the following:

Stephen W. Schaeffer
Professor of Biology
Chair of Committee

Hong Ma
Distinguished Professor of Biology
Dissertation Advisor

Naomi S. Altman
Professor of Statistics

Gong Chen
Associate Professor of Biology

Anton Nekrutenko
Associate Professor of Biochemistry and Molecular Biology

Robert F. Paulson
Associate Professor of Veterinary and Biomedical Sciences
Chair of Intercollege Graduate Degree Program in Genetics

*Signatures are on file in the Graduate School

ABSTRACT

The introduction of next-generation sequencing technologies has been changing the landscape of biological research. The plummeting cost of massive sequencing not only leads to the flourishing of various genome projects, but also opens up many opportunities in previously uncharted areas, two remarkable examples of which are sequencing different individuals of the same species and sequencing transcriptomes. With the availability of multiple genome sequences from a population, it is possible to systematically catalog natural variations and, more importantly, investigate their genome-wide distribution to deduce functional elements through conservation or selection. Moreover, as genomic changes reflect evolution at work, the comprehensive map of natural variations serves as a basis for studying the properties and effects of evolutionary forces. For transcriptomics, sequencing has become a revolutionary way to characterize gene expression. It not only offers high replicability and unparalleled accuracy, but also requires less input material, enabling transcriptome study in tiny structures, and no prior knowledge of gene structure, allowing detection of unknown transcripts. Here, three sequencing-based studies are presented. One study explored natural variations between two Arabidopsis ecotypes, Col and L*er*, and then investigated one crucial evolutionary force to shape variations, meiotic recombination. The other two studies investigated the small RNA transcriptome in Arabidopsis meiocytes and mRNA transcriptomes in developing mouse cortex.

In the first study, the sequencing and comparison between Col and L*er* uncovered 249,171 SNPs, 58,085 small and 2,315 large indels, with highly correlated genome-wide distributions of SNPs and small indels. Disease resistance genes contain significantly more variations, suggesting adaptation to specific environmental niches. These variations were then used as markers to investigate meiotic recombinations, crossovers and gene conversions, in two tetrads, detecting 18 crossovers, 6 of which had an associated gene conversion event, and 4

independent gene conversions. The number and length of identified recombination events suggest that Arabidopsis gene conversions are likely fewer and with shorter tracts than those in yeast. In addition, the analysis of variations in offspring plants showed meiosis provided a rapid mechanism to generate copy number variations (CNVs) by reshuffling existing variations.

In the second study, a recently developed method was applied to collect Arabidopsis meiocytes, a limited number of cells undergoing meiosis, and then small RNAs were profiled using SOLiD sequencing. 97 of 266 known miRNAs show expression in meiocytes. Interestingly, five miRNAs were found to account for more than half of the total miRNA expression in meiocytes, among which miR158a takes up about one third. The target genes of these five miRNAs have little or low expression in meiocytes. One putative novel miRNA was identified, which shows conservation with rice and maize. Analysis of longer reads provided clues for possible long ncRNAs in meiocytes.

The mouse transcritpome study uncovered 3,758 differentially expressed genes between two critical stages of cortex development, embryonic day 18 (E18) and postnatal day (P7). Neurogenesis-related genes, such as *Sox4* and *Sox11*, were more highly expressed at E18 than at P7. In contrast, the genes encoding synaptic proteins were up-regulated from E18 to P7, suggesting cortex development changes focus from neuron generation to synapse formation. In addition, approximately 500 genes with unknown function show dramatic change in expression level, serving as a blueprint for further experimental studies. Thousands of novel splice variants from 2,930 genes were identified, providing clues about another layer of dynamic expression.

These sequencing-based studies pushed the limit of previous work. The characterization of meiotic recombination reached single base-pair resolution. The mouse transcriptome work is one of several early studies utilizing RNA-seq. The small RNA transcriptome in meiocytes, single type cells at microscopic scale, was profiled for the first time. All these studies provided unprecedented information about complicated biological processes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Hong Ma, for his great support and guidance for the past five years. His insight, patience, enthusiasm, and broad knowledge have been motivating me all the time of research and writing of this thesis. I will never forget the time and efforts he devoted to helping me understand research projects, improving my writing and guiding me in my next step of career. He always makes lab feel like home. He arranged lab party for everyone's birthday and invited us to his home on every Thanksgiving. As an international student, this means a lot to me.

I also would like to thank my thesis committee members, Dr. Stephen Schaeffer, Dr. Naomi Altman, Dr. Gong Chen and Dr. Anton Nekrutenko for their insightful comments and suggestions. Their expertise and knowledge are invaluable to my research. It is so fortunate to have these outstanding scientists in my committee.

As most of my research is interdisciplinary, I really appreciate the assistance I got from many colleagues. Dr. Pingli Lu collected samples for both the meiotic recombination and meiocyte small RNA project, performed verification experiments and gave much advice on analysis. Dr. Qi Ji contributed a lot to the analysis of indels in L*er* and meiotic recombination. Dr. Wen-Yu Chung devised the strategy to detect novel splice variants in mouse cortex. Dr. Xia Wu dissected the mouse for sample collection. Dr. Tao Li did the initial analysis of Arabidopsis data and introduced Perl and Linux shell programming to me. Dr. Xiaofan Zhou taught me how to make phylogenetic trees. And many others helped to get these research projects moving smoothly.

Members in Dr. Hong Ma's lab are very helpful and good friends in life. In addition to those I collaborated with, Liye Zhang, Dihong Lu, Xuan Ma, Yazhou Sun, Zhao Su, Yi Hu, Bin Guo, Liyana Sukiran and Johnson McGovern gave me plenty of help. I learned a lot from discussing with them.

I want to thank Dr. Richard Ordway, Dr. Robert Paulson and other people in the genetics program. Dr. Richard Ordway recruited me and was always there when I needed his help.

My friends, Chicheng Sun, Xin Tang, Ning Li, Qiuying Shen, Zhenfeng Liu, Xia Wu, Xiaozhe Hu, He Xie and others, gave me a lot of encouragement. With their company, five years' life in graduate school becomes an exhilarating journey.

Finally, I would like to thank my parents and grandparents. Their love and care seed every progress in my life.

CHAPTER 1

INTRODUCTION

1.1 Overview of Next Generation Sequencing

Sequencing technologies have made astonishing progress in recent years. The cost of sequencing plummeted even faster than Moore's Law, which describes the exponential growth of computer performance. On the one hand, new sequencing technologies enable large-scale studies on previously unapproachable problems. On the other hand, the sheer volume and distinct characteristics of data from these technologies present substantial challenges to computational analysis.

1.1.1 Brief history of DNA sequencing

After the discovery of DNA being genetic material and the double helix structure of DNA molecule, sequencing DNA had been one of the holy grails in modern genetics. As the composition of DNA is fairly simple, only including four types of nucleotides, the sequence of these components was assumed to be the physical carrier of genetic information, which provides blueprint for the appearance and development of organisms. Thus determining the sequence of nucleotides chemically is the first step in decoding genetic information.

Besides some early attempts, the groundbreaking invention of chain termination methods by Frederick Sanger et al. in 1975-1977 (1) inaugurated the era of DNA sequencing. Since then the development of sequencing technologies has been through three generations. The first-generation sequencing includes Sanger method and Maxam-Gilbert method in 1977 (2). Maxam-Gilbert method sequences DNA by specific cleavage at modified nucleotides. Despite its popularity in the first several years, the Maxam-Gilbert method was not widely used due to its complexity of interpreting multiple lanes for the base call. In contrast to the Maxam-Gilbert method that is purely chemical reaction, Sanger sequencing mimics the DNA replication process

in the organism. By using dideoxynucleotide triphosphates (ddNTPs) that lack 3'-hydroxyl group and thus cannot form phosphodiester bond with the subsequent nucleotide, the elongation of newly synthesized DNA strand stops at the position where ddNTP is incorporated. Then DNA fragments of different length are separated in gel and DNA sequence can be deduced by reading these bands. Due to the relatively streamlined process, Sanger sequencing was the method of choice from then on.

Sanger sequencing kept being improved by companies like Applied Biosystem and Beckman, the most important progress of which is the automation of sequencing process by dye-terminator and capillary electrophoresis (3). The automation increased the read length and reduced the sequencing cost, which made sequencing the whole genome possible. The completions of human genome working draft at the turn of the 21st century (4, 5) and the first genome of an individual human, Craig Ventor, in 2007 (6) are two historical examples of the great success of Sanger sequencing.

The second-generation sequencing, also known as next generation sequencing, is characterized by "massively parallel" sequencing processes, including 454, Solexa/Illumina and SOLiD (7). Except for preprocessing of the sample DNA, all three technologies contain two major steps: amplification and sequencing. 454 sequencing, whose first instrument was commercially launched in 2005, is based on two techniques: emulsion PCR and pyrosequencing. Emulsion PCR isolates and then amplifies DNA in an aqueous droplet surrounded by an oil environment. Pyrosequencing is a method based on the detection of released pyrophosphate from the incorporation of dNTPs. Illumina sequencing, commercially released in 2006 and formerly named Solexa sequencing, relies on bridge PCR and reversible dye-terminator. Bridge PCR amplifies a DNA template on a solid surface by pairing primers fixed on the surface with the bent template. Reversible dye-terminator used in the sequencing step carries fluorescent label, similar to the terminator used in Sanger sequencing, but is reversible so that the end of DNA fragment

can be unblocked and the sequencing process continue on the same fragment. SOLiD sequencing, introduced in 2007, also employs emulsion PCR, but is different from 454 and Illumina that are both based on sequencing by synthesis, is sequencing by ligation. Fluorescently labeled oligonucleotides are ligated to elongating DNA strand, during which each di-base is interrogated. There is also another type of next generation sequencing technology developed by Complete Genomics, but because of its unique business model it has not been widely used in academia.

Compared to Sanger sequencing, the second-generation sequencing produces significantly more data. In other words, sequencing costs have dropped substantially. As shown in Figure 1-1, the cost per megabase decreased exponentially before the introduction of Illumina and SOLiD, but it decreased even faster after. Although due to shorter read length more data are needed to assemble the genome, the cost per human genome also showed dramatic reduction.

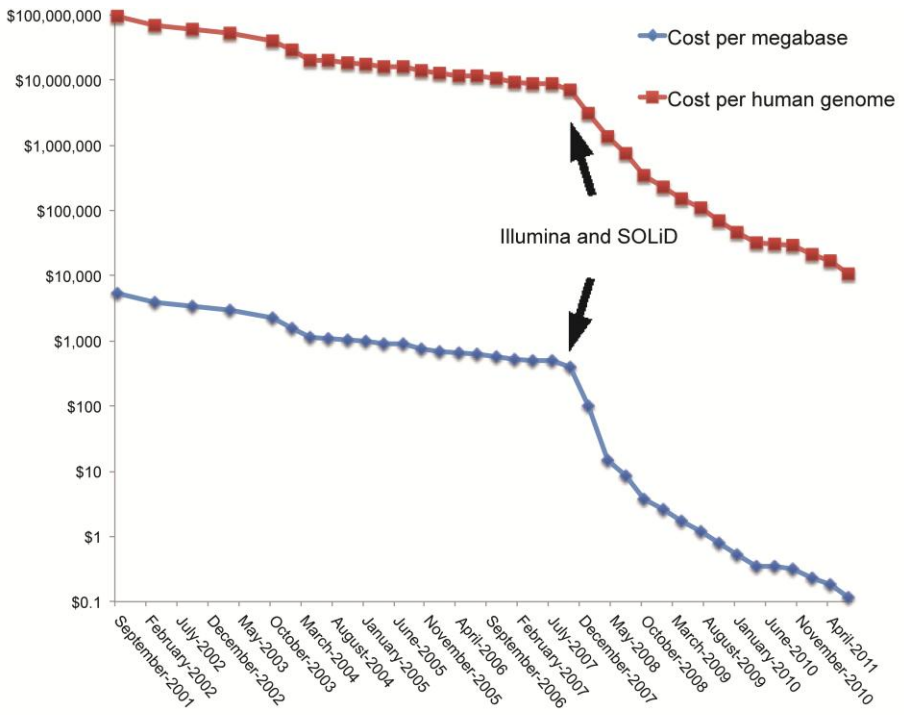


Figure 1-1. The decrease of sequencing cost per megabase and per human genome. Data are from National Human Genome Research Institute.

In addition to huge volume, data from next generation sequencing technologies have other distinct characteristics. All of these technologies produced shorter reads with lower sequencing quality compared to Sanger data. Currently Illumina generates reads of 100-150bp and SOLiD produces 50-75bp each read; 454 has much longer reads, around 700bp. Due to the property of pyrosequencing, in homopolymer region 454 determines the number of nucleotide according to the intensity of light signal, which becomes less linear when the number of nucleotides is large, and thus has relatively high error rate in these regions. But in other regions, 454 has comparable sequencing quality to the Sanger method. Both Illumina and SOLiD have inferior sequencing quality at the first several nucleotides or near the end of each read, although they do not suffer from the homopolymer problem. Another property of pyrosequencing used in 454 is that it only produces transient luminescence and requires constant monitoring by camera, so the throughput of 454 is limited by camera sensor (8). Currently 454 (up to 700Mb) has much less output than Illumina (up to 600Gb) and SOLiD (up to 180Gb) per sequencing run. In contrast to 454 and Illumina, which report DNA sequences directly, data from SOLiD are symbolized as fluorescent colors with each color coding for two nucleotides, which may be problematic in certain applications.

Now the third-generation sequencing is emerging, while second-generation technologies keep improving at a fast pace. There is still not much consensus regarding what defines third-generation sequencing. Some technologies, like Helicos and PacBio (Pacific Biosciences), are characterized by single molecule sequencing, which do not need amplification before sequencing and thus avoid bias in PCR. Other technologies, like Ion Torrent, still need amplification but have very different feature from second-generation sequencing, including “real time” sequencing, simple bench-top instrument at low cost (about one tenth of second-generation instrument) and short turnaround time of each sequencing run (less than 2 hours). Third-generation sequencing offers several advantages over the second generation: (1) longer reads. The average read length of

PacBio exceeds 1000bp, with instances of >10kb to facilitate genome assembly. Ion Torrent produces 200bp reads now and is expected to reach 400bp by the end of 2012. (2) Results from single molecule sequencing are closer to real copy numbers of molecules existent in the sample, which would be especially useful in RNA-seq or CNV (Copy Number Variation) detection. (3) Short turnaround time is particularly fit for diagnostics or other clinical use. On the other hand, third-generation sequencing is still in its infancy and has critical weakness for now: (1) lower throughput compared to the second generation. (2) PacBio reads have a much higher error rate, with raw read accuracy being just 85%.

1.1.2 Application of next generation sequencing

With the plummeting cost of sequencing, various genome projects have sprouted. For organisms having a reference genome sequence, now it is possible to sequence multiple individuals of the same species so that variation in a population can be investigated. Liti et al. sequenced 38 strains of the baker's yeast, *Saccharomyces cerevisiae*, and 32 strains of its closet relative, *Saccharomyces paradoxus*, using a mix of Sanger and Illumina technologies (9). *S. paradoxus* showed much more variation than *S. cerevisiae*, the world-wide differentiation of which only amounts to a single population of *S. paradoxus*. Moreover, in addition to geographically isolated lineages, some *S. cerevisiae* strains are mosaics of existing isolates, suggesting cross breeding due to human influence. For flies, Freeze 1 release of Drosophila Genetic Reference Panel (http://www.hgsc.bcm.tmc.edu/project-species-i-Drosophila_genRefPanel.hgsc) provides genome sequence of 162 inbred *Drosophila melanogaster* strains, each of which has at least 8.5X coverage. Combined with extensive phenotyping efforts in these lines, the genome sequence presents an invaluable resource for gene mapping of complex traits (10). For plants, the 1001 genome project (http://1001genomes.org/)

aims to sequence 1001 strains of *Arabidopsis thaliana*. The initial analysis of 80 strains cataloged the majority of common SNPs, indels and sequences missing in the reference assembly (11). Similar to fly, the next phase of 1001 project for Arabidopsis will relate genomic variation to gene activity and finally investigate how different strains adapt to specific environments.

More genome resequencing efforts have been devoted to human. The objective of the 1000 Genome Project (http://www.1000genomes.org/) is to deliver a comprehensive catalog of human genetic variation by sequencing thousands of individuals (12). The International Cancer Genome Consortium (http://www.icgc.org/) plans to detect driver mutations that underlie tumor genesis by sequencing and comparing cancer tissue and normal tissue, the challenge of which is that there are a huge number of accompanying mutations in cancer samples. A more ambitious undertaking is the Human Variome Project (http://www.humanvariomeproject.org/), which aims to establish a system not only for collecting but also storing and sharing variation data related to all kinds of diseases. Personal Genome Project (http://www.personalgenomes.org/) is a long-term study that sequences complete genomes, collects medical records and also traces future health condition of participants in order to advance personalized medicine. Besides these whole genome projects, another type of human resequencing efforts are exome sequencing, which have proven to be very successful in identifying causal mutation of Mendelian diseases (13). There are estimated to be 6000-8000 Mendelian diseases in total, about 3000 of which have a known causal variants (14). The goal of the International Rare Disease Research Consortium (http://www.geneticalliance.org/irdirc) is to find causal mutation for most of Mendelian diseases by 2020. Moreover, after knowing the causal variants sequencing can serve as a quick and decisive diagnostic method that is lacking for the majority of rare diseases for now.

In addition to resequencing, new sequencing technologies have also accelerated de novo genome assembly of more organisms. Despite the challenge in assembling short reads, the giant panda genome sequence marks the feasibility of using next generation sequencing to assemble

mammalian genome (15). Now almost every ongoing *de novo* genome project incorporates second-generation technologies to reduce cost, including sheep, tomato and wheat (the most challenging genome assembly so far due to its 16Gb size and being a hexaploid). The fast dropping cost even motivates coordinated efforts to fill in gaps in major branches of the tree of life, like Genomic Encyclopedia of Bacteria and Archaea (http://jgi.doe.gov/programs/GEBA/), Genomic Encyclopedia of Fungi (http://genome.jgi-psf.org/programs/fungi/about-program.jsf), Insect 5000 Genomes initiative (http://arthropodgenomes.org/wiki/i5K), 1000 Plant and Animal Reference Genomes Project (http://ldl.genomics.org.cn/page/pa-research.jsp) and Genome 10K for vertebrates (http://www.genome10k.org/).

Another type of genome study is metagenomics that investigates the composition of microorganisms in specific environment. Previous studies showed that metagenomes are closely related to crop disease (16), human health (17) and biofuel (18). With the aid of next generation sequencing, Human Microbiome Project (http://www.hmpdacc.org/) and Earth Microbiome Project (http://www.earthmicrobiome.org/) are now in full swing to characterize microorganism diversity in various environmental niches.

Second-generation technologies not only enable quick collection of genomic sequences, but also provide unparalleled way to measure transcriptome activity. A transcriptome is the complete set of expressed transcripts in certain cells or tissues at a specific developmental stage or particular physiological condition. Because gene expression change is a fundamental biological process and underlying all biological phenomena, transcriptome profiling technologies have been in active development since 1990s. Two types of approaches have been devised ever since: hybridization-based method and tag-based method (19). Hybridization-based method is exemplified by microarray technology that was introduced in 1995 and based on hybridization between transcripts and probes on a solid surface. The hybridization ignites fluorescence, the intensity of which represents the abundance of corresponding transcript. Due to its relatively low

cost and simple sample preparation, microarray has been widely used to detect differentially expressed genes and infer genetic network. However, microarray has critical limitations: (1) For highly expressed genes, due to probe saturation the fluorescence signal stagnates and cannot reflect the real expression level; (2) For extremely lowly expressed genes, it is difficult to tell between real signal and noise from nonspecific binding; (3) The design of probes on microarray need a reference genome. Tag-based methods include SAGE (serial analysis of gene expression), CAGE (cap analysis of gene expression) and MPSS (massively parallel signature sequencing). Because these methods rely on Sanger sequencing and thus have relatively high cost, they are less widely used than microarray. RNA-seq is an extension of current tag-based methods by incorporating next generation sequencing (19). The principle of RNA-seq is simple. After reverse transcription of mRNAs, cDNAs are fragmented and sequenced using Illumina or SOLiD technology, each sequencing run of which generates millions of reads. Each sequencing read is a tag representing one mRNA molecule. Due to the huge number of reads, a deep sampling of existent RNA molecules is achieved. Several studies showed RNA-seq is not only highly replicable but has less background noise and much more linear representation of highly expressed genes compared to microarray (19-22). To detect unannotated genes, although microarray can still be used by designing tilling probes, RNA-seq does not require prior knowledge and thus has much less cost. RNA-seq can even be applied to organisms without a reference genome by mapping reads to a close relative or assembling reads into contigs. Due to these advantages, RNA-seq has been widely adopted in areas where microarray prevailed and beyond (23).

Besides genome sequencing and RNA-seq, with creative design sequencing has become a general-purpose research tool. ChIP-seq, combining chromatin immunoprecipitation (ChIP) with next generation sequencing, maps genome-wide distribution of nucleosomes or transcription factors (24, 25). Especially, with the use of specific antibodies to histone modification, the genome-wide map of “histone code” can be identified. By selectively enriching and sequencing

the double-stranded portion of RNAs, sequencing results elucidate the folding of RNA molecules (26). By capturing specific chromatin structure, sequencing can be used to investigate 3D structure of the genome due to physical interaction of genomic elements (27). With further combination of sequencing and molecular biology techniques, more and more aspects of biological processes will be explored.

1.1.3 The development of read mapping tools

Read mapping is a key step in computational analysis of data from genome resequencing and RNA-seq. Alignment of nucleotide sequences has been a classical problem in bioinformatics. However, due to the short read length and sheer volume of next generation sequencing data, it takes hundreds of CPU hours to map just a few million reads by conventional methods like BLAST or BLAT. To accommodate the unique characteristics of these data, new algorithms for mapping tools have been actively developed (28).

To achieve fast lookup, either sequencing reads or the reference genome is indexed. As indexing genome requires large memory occupancy, early mapping tools, including RMAP, Eland and MAQ, indexed sequencing reads instead, which typically use “seed sequence”, a fragment of a read. The first conceptual progress in seed indexing is the use of spaced seeds (29), which tolerates more sequencing errors since next generation sequencing has inferior sequencing quality. Each mapping tool has slightly different implementation of spaced seeds. Take Maq, a widely used tool, as an example (30). Maq uses the first 28bp of each read for seeding, as bases towards the end have higher error rate in both Illumina and SOLiD data. The 28bp is divided into four 7bp seeds. If there is one sequencing error or SNP in the 28bp, there are still 3 seeds that have exact match with the genome sequence. Similarly, if there are two errors or SNPs, 2 seeds will have exact match. By searching the genome for exact matches with all possible combination

of 2 seeds (6 possible pairs in total), the candidate locations of that read are pinpointed. Thus by using spaced seeds, 2 errors or SNPs are tolerated in the first 28bp without much sacrifice of mapping speed. The second major advance in the seeding approach lies in the seed extension part, which deals with bases after the first 28 in the Maq case. The use of CPU SIMD instructions in seed extension facilitates parallelization of alignment (29). Dynamic programming, the key algorithmic step in BLAST, is also used to increase the speed.

Another critical development is the introduction of Burrows-Wheeler transformation (BWT). Using this transformation, indexing large mammalian genome, like human genome (3Gb), only occupies less than 2 gigabyte of memory, similar to index reads. Later tools, including Bowtie, BWA and NovoAlign, usually utilize BWT to transform and index the genome. In addition, for repeated mapping, only a single BWT is needed for the same organism.

Other development in mapping tools includes taking quality score into consideration during mapping (Maq and many others), performing spliced alignment to accommodate reads spanning exon-exon junctions (Tophat), complex seed sequence design to tolerate even more errors (BFAST) and incorporation of detailed statistical model in mapping (Stampy) (30-33). With novel algorithm and new pieces of mapping software keeping emerging, benchmark studies have been conducted to compare speed, sensitivity (the percentage of reads mapped) and accuracy (the percentage of reads mapped correctly) of different tools (29, 34).

1.2 Arabidopsis Natural Variation and Meiotic Recombination as One Crucial Shaping Force

The emergence of high throughput technologies has had a profound impact on the study of population genomics, the study of the amount, causes and dynamics of genome variation in natural populations (35). Topics in this field have been studied since Darwin and due to the scarcity of data, early studies mainly centered on mathematical modeling and theoretical derivation. After the availability of genome-wide technologies, especially sequencing, it has been possible to examine the presumption and projection of mathematical models by empirical evidence.

Compared to animals, plants cannot migrate by themselves to circumvent adverse environment. Thus in addition to neutrally evolved variants, natural variation among different stains of the same plant conveys specific adaptation to their geographic niches. The start of systematic study on genome variation in Arabidopsis, a model organism for plants, is marked by the publication of its reference genome sequence in 2000 (36), which provides a benchmark for subsequent research. Columbia (Col) is the strain of choice in this reference genome project. Around 2002, the Monsanto Company offered 2X coverage sequencing and initial assembly of another widely used Arabidopsis strain, Landsberg *erecta* (L*er*), providing useful markers for hundreds of gene mapping studies later on. Nordborg et al resequenced 876 short fragments in 96 Arabidopsis strains and found linkage disequilibrium decays rapidly, within 50kb, suggesting Arabidopsis is particularly appropriate for GWAS (genome wide association studies) to search for roots of complex traits (37). Another microarray-based study interrogated 20 Arabidopsis accessions and detected regional variation of polymorphism and selective sweeps in some locations (38). Since the availability of next generation sequencing, not only SNPs but also indels and large variations can be explored. Ossowski et al sequenced two strains, Bur-0 and Tsu-1, and tested the feasibility of using read mapping and assembly to catalog genome variations (39).

Based on this, the 1001 genome project aims to provide a comprehensive list of genome variations in Arabidopsis and examine the evolutionary forces shaping these variants.

Besides selection and genetic drift, two types of meiotic recombination, crossover (CO) and noncrossover (NCO), are important evolutionary forces and reshuffle genome variation every generation. Crossing over happens when two non-sister chromatids exchange genetic material during meiosis. In addition to physically bringing together two chromosomes during synapsis, crossover breaks down linkage, creating more combination of existing variants. The genome-wide distribution of crossovers per meiosis is not random, because (1) there should be at least one CO per chromosome per meiosis to make homologous chromosomes close to each other; (2) since the formation of CO begins with a double strand break of DNA, too many COs are detrimental and there is tight control of the total number of COs per meiosis, which is achieved by interference between neighboring COs. In Arabidopsis, a map of crossover rate in chromosome 4, the smallest chromosome and having distinct structure features, was constructed by observing the distribution of genome variations in the F2 progeny of crossing between Col and L*er* (40). The map clearly showed hotspots of crossover across the chromosome and the rate of crossover is inversely related with GC content.

In contrast to CO which leads to large-scale exchanges among chromosomes, NCO only unidirectionally copy kilobase(s) or less of DNA from one chromatid to the other. The generation of both CO and NCO begins with a double strand break of DNA and then continues with the formation of D-loop by invasion of DNA strand from the broken chromosome. Several pathways are proposed to produce NCO from D-loop (41). The first model is synthesis-dependent strand-annealing pathway (SDSA), in which the invasion strand is displaced and the repair of this strand leads to NCO. The second model requires the formation of double Holliday junctions from D-loop by DNA repair and ligation. Depending on how to resolve the double Holliday junctions,

either CO or NCO can be formed. The third possible mechanism is the dissolution of double Holliday junctions that can also result in NCO.

Although NCO tract is short, it can affect evolution in many ways (41). (1) NCO can also break down linkage between genomic variants. (2) NCO leads to non-Mendelian ratio of alleles. In the heterozygous loci where NCO takes place, the 2:2 ratio changes to be 3:1. (3) Since the formation of NCO needs DNA mismatch repair pathways at work, in cases where A:C or G:T mismatch is present GC are favored to be used as templates and AT are corrected, which is called biased gene conversion. Thus NCO has the tendency to increase GC content. (4) Meiotic NCO usually happens between allelic sequences, but occasionally it can also occur between repeated sequences in the same chromosome or even in non-homologous chromosomes. Because NCO copies sequence from one location to the other, it homogenizes repeated sequences in the end. Since gene duplication is the main mechanism to generate new genes during evolution, NCO inside genic sequence will affect the fate of duplicated genes. To quantitatively estimate or model these effects of NCO on genome variation after many generations, knowing the frequency of NCO per meiosis is essential.

However, due to the short length of NCO, the frequency per meiosis was difficult to measure by traditional approaches. Next generation sequencing makes it possible to directly observe NCO events and calculate their frequency by comparing sequencing results. Moreover, Arabidopsis is particularly suited for the study of NCO, because it has relatively simple genome and the *qrt* mutant of Arabidopsis allows collection of sequences from all four progenies from a single meiosis.

1.3 Brief Overview of Small RNAs Studies in Arabidopsis

The study of small RNAs in plants has a long history, which dates back to work on transgenic plants in 1980s. Agrobacterium-mediated transformation technology was developed then and it was soon applied to breeding research. Some interesting phenomena were reported in tobacco transformation experiments (42). After introducing the coat protein (CP) gene of tobacco mosaic virus (TMV) into tobacco, the transgenic plants showed less severe or delayed symptoms compared to wild-type ones when infected with TMV. Initially, it was assumed that the exogenous CP activated immune response in host plants. However, after examining the expression level of CP, the resistance of transgenic plant did not increase with the activity of CP. Some plants with low-level CP, unexpectedly, have higher resistance. Even more surprisingly, introduction of a mutated CP gene, e.g. loss of start codon or truncated coding sequence, leads to similar resistance. In another line of research, multiple copies of a key gene in petunia pigment biosynthetic pathway, Chalcone synthase (CHS), were introduced into wild-type petunia. Completely unexpected was the results that the transgenic plant with more copies of CHS has paler color than wild type. Further study utilizing nuclear run-off assay (also known as nuclear run-on), which marks nascent mRNAs by radioactive NTP and then hybridizes them with CHS probe to tell whether CHS is being expressed, found the transcription of CHS gene is unaffected, but most CHS transcripts are degraded afterwards. Later, all these phenomena were found to be due to posttranscriptional gene silencing (PTGS). In 1999, Hamilton and Baulcombe detected 25nt long RNAs in plants showing PTGS by hybridization and blotting, which was the first discovered instances of small interfering RNAs (siRNAs) (43).

In contrast to early studies that identified several siRNAs by low throughput approaches, after the completion of Arabidopsis genome assembly and especially the accessibility of high throughput sequencing, a huge number of siRNAs, including 3 major types, have been discovered

at accelerated pace (44, 45). Heterochromatic siRNAs (hc-siRNA), predominantly 24nt long in Arabidopsis, are the most abundant category, which usually account for more than 80% of reads in sequencing-based studies. hc-siRNA mainly originated from transposable elements, heterochromatic DNA and other repeated sequences like centromere regions. Interestingly, they functioned in cis, repressing the activity of the genomic regions where they originated. The second type of siRNA is trans-acting siRNA (ta-siRNA), which is 21nt long and degrades transcripts from regions different from its origin. Although the trans-acting mode of function is similar between ta-siRNA and microRNA (miRNA), ta-siRNA can migrate to neighboring cells and then exert its function, while miRNA is only discovered to function autonomously. Another major type of siRNA is natural antisense transcript siRNA (nat-siRNA) that is of heterogeneous length and mainly produced in plants under stress. nat-siRNA also regulates the level of its originating RNA by cleavage and degradation.

Different from siRNA, the first discovery of miRNA was in the worm, *Caenorhabditis elegans*. lin-4, the first identified miRNA gene in 1993 by genetic screen, produces 21nt long mature miRNA which regulates juvenile-to-adult developmental transition by incomplete base-paring with lin-14. However, there is no homolog of lin-4 in other animals. Thus initially miRNA was regarded as unique regulatory mechanism in worm and did not gain wide appreciation (44). In 2000, the second miRNA gene, let-7, was discovered in *C.elegans* (46). let-7 is a conserved gene regulating developmental timing and has homologs in all other animals with bilateral symmetry, which aroused widespread interest and motivated the search for miRNAs in other organisms, including Arabidopsis. In 2002, three groups independently reported miRNAs in Arabidopsis by cloning and Sanger sequencing (47-49). In 2003, Palatnik et al presented the first direct evidence that miRNA could regulate a specific aspect of plant morphogenesis, leaf development for *JAW* locus particularly (50). Similar to siRNA, the adaptation of new sequencing technologies greatly accelerated the discovery of novel miRNAs. Now we know that there are

266 miRNA genes in Arabidopsis, regulating diverse biological processes from differentiation to development. Some of these miRNA genes were identified by homology to miRNA genes in other organisms. Others, especially rare and less conserved miRNAs, were uncovered after the application of next generation sequencing.

As sequencing-based detection does not require prior knowledge, other types of small RNAs can also be investigated and have gained attention. Small nuclear RNA (snRNA) and small nucleolar RNA (snoRNAs) are present in several deep-sequencing studies (45). Although these RNA species were mainly generated from RNA degradation, recent studies showed snoRNA also participate in gene regulation. Moreover, deep sequencing enables the study of previously unappreciated long noncoding RNAs (lncRNA) (51). Another advantage of next generation sequencing is that compared to microarrays, it requires less starting material, which makes it possible to study small RNAs in tiny structures or tissues like Arabidopsis meiocytes. Meiocytes are cells undergoing meiosis. Due to the limited number of meiocytes and also the difficulty in isolation, the functions of small RNAs in meiocytes are still largely a mystery.

1.4 Transcriptome Studies in Mouse and Brief Introduction of Early Cortex Development

The mouse transcriptome has been actively studied since genome-wide gene expression profiling technologies were available in 1990s, because (1) mouse is a model organism for human biology and disease; (2) mouse has a published and reliable reference genome sequence; and (3) The rich prior knowledge about mouse anatomy, development and related techniques ensure precise and replicable sample collection. Early studies include microarray-based research on gene expression in 45 mouse tissues (52) and EST (expressed sequence tag) frequency based research on the transcriptome of mouse stem cells and early embryos (53). In addition to individual studies on different aspects of the transcriptome, there are also coordinated efforts to investigate mouse transcriptome in key tissues or at major developmental stages. The Mouse Transcriptome Project at NIH (www.ncbi.nlm.nih.gov/projects/geo/info/mouse-trans.html) used MPSS (massively parallel signature sequencing) to portray transcriptome in various mouse tissues or organs, including embryonic stem cells, liver, kidney and brain. FANTOM project, initiated at RIKEN of Japan, is long-term effort to catalog expression in mouse (and now expand to human). FANTOM project has gone through 4 stages. FANTOM 1 and 2 focused on cloning 103,000 cDNAs, which provided solid basis for gene annotation of mouse genome (54, 55). FANTOM 3 and 4 developed CAGE (cap analysis of gene expression) method, systematically explored transcription start sites and composed an atlas of transcription factor combinatorial network (56-58). Allen Brain Altas (http://www.brain-map.org/) is specifically devoted to brain transcriptome, which utilizes automated *in situ* hybridization to show gene activity in developing mouse brain. Other large-scale project/repository of mouse gene expression include BioGPS (http://biogps.org/), GenePaint (http://www.genepaint.org/) and Genevestigator (https://www.genevestigator.com/). Compared to other organisms, these studies collectively provided impressively rich resource of expression data in mouse.

However, despite increasing data in other stages or tissues, the study on transcriptome changes of the mouse cortex during the first week after birth is still lacking. Postnatal cortex development is an essential part of the maturation of mouse brain, during which the first seven days are particularly important and marked by glia generation. The production of glia is related to the activity and property of neural stem cells (59). In the embryonic stage, neural stem cells mainly divide and produce neurons within the germinal ventricular zone (VZ). On around 18 day after oocyte implantation (E18), neural stem cells shift from VZ to subventricular zone (SVZ) and developing neurons start to send out axons and dendrites. Neural stems cells change their characteristics, like response to growth factors, and mainly produce glial progenitor cells. After birth, active production of astrocytes, one type of glial and assumed to help synapse formation, is accompanying the continuing development of neurons. By postnatal day 7 (P7), the elongated axons and dendrites of many neurons start to establish synaptic connection with other neurons. So the period between E18 and P7, one week after birth, is a critical development stage for the progression of synaptogenesis. By profiling the change of the transcriptome during this stage, many genes crucial for synapse formation are expected to be identified.

Moreover, transcriptome profiling could lead to unexpected discoveries. Previous studies have shown RAG2 (recombination activation gene 2), a component of V(D)J site-specific recombination machinery and essential for antibody diversity, is expressed in embryonic and postnatal neurons (60). RAG1, however, does not have similar expression in neurons. The role of RAG2 in the developing cortex is still mystery. Profiling the transcriptomes of E18 and P7 may discover more recombination related genes or other genes with unexpected functions for further study.

CHAPTER 2

ANALYSIS OF ARABIDOPSIS GENOME-WIDE VARIATIONS BEFORE AND AFTER MEIOSIS AND MEIOTIC RECOMBINATION BY RE-SEQUENCING LANDSBERG *ERECTA* AND ALL FOUR PRODUCTS OF A SINGLE MEIOSIS

The work described in this chapter has been published in Pingli Lu*, Xinwei Han*, Ji Qi*, Jiange Yang, Asela J. Wijeratne, Tao Li and Hong Ma, Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by re-sequencing Landsberg *eracta* and all four products of a single meiosis, *Genome Research*, 2012, 22(3): 508-518 (* equal contribution).

2.1 Summary

Meiotic recombination, including crossovers (COs) and gene conversions (GCs), impacts natural variation and is an important evolutionary force for sexually reproducing organisms. COs increase genetic diversity by redistributing existing variation, whereas GCs can alter allelic frequency. Here we sequenced *Arabidopsis* Landsberg *erecta* (L*er*) and two sets of all four meiotic products from a Columbia (Col)/L*er* hybrid, to investigate genome-wide variation and meiotic recombination at nucleotide resolution. Comparing L*er* and Col sequences uncovered 349,171 Single Nucleotide Polymorphisms (SNPs), 58,085 small and 2,315 large insertions/deletions (indels), with highly correlated genome-wide distributions of SNPs and small indels. 443 genes have at least 10 nonsynonymous substitutions in protein coding regions, with enrichment for disease resistance genes. Another 316 genes are affected by large indels, including 130 genes with complete deletion of coding regions in L*er*. Using the *Arabidopsis qrt* mutant, two sets of four meiotic products were generated and analyzed by sequencing for meiotic recombination, representing the first tetrad analysis in a nonfungal species. We detected 18 COs, 6 of which had an associated GC event, and 4 GCs without COs (NCOs), and revealed that *Arabidopsis* GCs are likely fewer and with shorter tracts than those in yeast. Meiotic recombination and chromosome assortment events dramatically re-distributed genome variation in meiotic products, contributing to population diversity. In particular, meiosis provides a rapid mechanism to generate copy number variation (CNV) of sequences that have different chromosomal positions in Col and L*er*.

2.2 Introduction

Natural genomic variations, including SNPs, insertions, deletions, and CNVs, are prevalent in many species and can generate new alleles/genes, reshape gene structures, alter gene dosage, and change gene expression level (61, 62). In human and animals, genome variations are associated with severe genetic diseases, such as Parkinson and Alzheimer (63, 64). In plants, genome variations contribute to adaptive fitness by affecting traits such as flowering time, disease resistance and seed dormancy (65-68). Recently, genome-wide studies using microarray or next-generation sequencing (NGS) indicate that natural variations in humans, mouse and flies are more abundant than previous thought (69-71).

Genome variations are shaped by meiotic recombination and chromosome assortment, which reshuffles the genome in every generation. Because chromosome assortment has limited possibilities whereas meiotic recombination, either as crossover (CO) or non-crossover (NCO, or GC without exchange of flanking regions), can occur at many sites along the chromosome, recombination has a greater potential to increase genetic diversity. Furthermore, in COs, recombinations result in large-scale (megabases) reciprocal exchanges of genetic materials between homologous chromosomes, whereas in GCs associated with COs or NCOs, kilobase(s) or less of DNA sequences are unidirectionally copied from one homolog to the other, thereby altering frequency of natural variations (72, 73). COs and NCOs can be inferred by analysis of haplotype markers in population studies (74, 75), which can only detect fixed recombination events, not changes per meiosis.

In fungi such as the budding yeast *Saccharomyces cerevisiae*, meiotic products are kept together as spores in an ascus, forming a tetrad. This allows direct examination of the consequence of meiotic recombination in parallel cultures derived separately from the four spores, using “tetrad analysis” (72). Tetrad analysis has contributed significantly to the

understanding of the molecular basis of meiotic recombination, including strong support for the steps in the double strand break repair model (DSBR) (72, 73, 76, 77) and used in yeast to examine the frequency and genome-wide distribution of meiotic recombination (78, 79). Here, we combined next generation sequencing and tetrad analysis to investigate natural variations and meiotic recombination in the flowering plant *Arabidopsis*.

Arabidopsis thaliana is native to Europe and central Asia (67), with many genetic (geographical) variants (derivative lines are called accessions) adapted to different environments. The Columbia (Col) accession is widely used for molecular genetic studies and was sequenced by the *Arabidopsis* Genome Initiative as the genomic reference (80). Similar to Col, the Landsberg *erecta* (L*er*) accession is also widely used for functional studies (81). The Col and L*er* lines have extensive DNA polymorphisms (37) (82), even though they were both derived by George Redei at University of Missouri, Columbia in 1950s, from a heterogeneous population named Landsberg collected by Laibach: Col was selected from a group of nonirradiated Landsberg plants, whereas the L*er* line was derived from X-ray mutagenesis experiments with Landsberg plants (http://arabidopsis.info/protocols/ler.html). Polymorphisms among many diverse accessions were analyzed by sequencing hundreds of short fragments or using oligonucleotide arrays (37) (38). Moreover, genome variations among Col-0, Bur-0, and Tsu-1 were recently analyzed by Illumina sequencing (39). Because natural variations can have profound effects on gene function (67, 83-85), a genome-wide examination of *Arabidopsis* natural variation, such as that between Col and L*er*, will greatly facilitate the understanding of how they adapt to specific environmental niches. Genes related to certain functions or participating in specific pathways may contain exceedingly more variations so that these functions or pathways are purposely altered between these two accessions to accommodate environmental changes. Some altered genes may show strong signal of positive natural selection, as variations in these genes represent functional innovation.

Moreover, because the *Arabidopsis qrt* mutant meioses produce all four meiotic products as attached spores, which then develop into attached functional pollen grains, *Arabidopsis* offers a unique opportunity among plants and animals to carry out tetrad analysis, as was done using hundreds of DNA markers (86, 87). Here, we sequenced the L*er* genome using high-throughput sequencing to uncover over 400,000 genome variations between L*er* and Col. In order to distinguish variations with functional implications from neutrally evolved ones or mutations due to radiation when selecting L*er*, gene ontology enrichment and dN/dS analysis was applied. We then utilized SNP markers to examine meiotic recombination in two meioses, by sequencing all of their products, allowing the detection of CO and NCO/GC events and the characterization of the GC tracts at nucleotide resolution. The sequencing data also displayed the re-distribution of natural variation sites following a single meiotic generation.

2.3 Material and Methods

2.3.1 Plant material and growth conditions

Arabidopsis thaliana "Columbia-*qrt*" and "Landsberg *erecta-qrt*" from Dinesh Kumar's lab at Yale University were crossed to obtain F1 hybrids, whose tetrads of pollen were used to pollinate the *Arabidopsis* Columbia accession (Fig. 4A; see Supplemental Information for details). The resulting four seeds were named MPP-A, MPP-B, MPP-C, and MPP-D and allowed to mature. All of the plants were grown under long-day conditions (16 h day and 8 h night) in a growth chamber at 18–22 ℃.

2.3.2 DNA isolation, genotyping, and genome re-sequencing

DNAs were extracted from each of MPPs and L*er* leaves using the QIAGEN DNeasy Plant Mini kit (Cat#: 69104) and genotyped using SSLP markers, including NGA126, CIW4, NGA1126, and NGA63. Genomic DNA from L*er* and meiosis progeny plants (MPPs) were subjected to Illumina sequencing. The genomic DNA of L*er* were sequenced in 6 lanes of Genome Analyzer II (26,420,944 single-end reads of 36 bp) at FASTERIS SA, Switzerland and 1 lane of Genome Analyzer IIX (17,591,335 paired-end reads of 75 bp) at Beijing Genomics Institute (BGI), Shenzhen, China. Each MPP of the 1st meiosis was sequenced in two or three lanes of Genome Analyzer II (paired-end, 40 bp) at National Center for Gene Research, Shanghai, China (see Supplemental Table 1 for read counts). Each MPP of the 2nd meiosis was sequenced in one lane of Genome Analyzer I (single-end 36 bp) and one lane of Genome Analyzer II (single-end, 35 bp) at FASTERIS SA and two lanes of Genome Analyzer IIX (single-end, 55~75 bp) at BGI (see Appendix Table 2-1 for read counts).

2.3.3 Calling of small polymorphisms

The Tair9 assembly of *Arabidopsis* Columbia ecotype genome was downloaded from the TAIR (ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes) website. Sequenced reads of L*er* were then mapped against the Col genome and SNPs were called using Maq 0.7.1 (http://maq.sourceforge.net/) (30), whereas small indels were predicted using inGAP (88). The mapping and SNP prediction procedure follows the online Maq instruction from FASTQ format transformation to build consensus sequences. Only uniquely mapped reads with mapping quality score equal or greater than 20 were used in subsequent analyses. We further removed pseudo-SNPs due to repetitive sequences or amplification errors with Perl scripts written for the analysis here (available upon request). Finally, the sequencing reads of meiosis products were used to rule out errors in Col genome sequences and confirm predicted SNPs, which requires the first and second most abundant bases in SNP loci to be Col-specific base and L*er*-specific bases or vice versa. Otherwise the SNP loci were considered to be unreliable.

2.3.4 Identification of larger indels

All L*er* paired-end Illumina reads were assembled using Velvet (89) and the output contigs mapped to the Col genome by Mummer (90), with redundant contigs removed before the prediction of large indels using custom scripts (available upon request). To minimize false-positives, we implemented a step of read mapping depth estimate in the pipeline, because indels affect mapping pattern of paired ends in their flanking regions, and removed those predicted indels that lacked “gapped region” in the landscape of read mapping. Finally, homology searches with coding regions in indels against Col genes was performed by BlastN search (identity>80%).

Very recently, Schneeberger *et al.* reported assembly of L*er* sequences based on high-throughput sequencing reads (91). By using these contigs as reference and importing more sequences from 8 meiotic datasets and the Monsanto L*er* contigs, we built a new assembly with the longest contig of ~253 Kb and N50 of ~26 Kb (more than twice as long as the ones from our reads only, respectively). The newly assembled contigs can be accessed along with the *de novo* assembly from reads in this study. When selecting L*er* reads from 8 meiotic datasets, we screened the alignment results of paired-end reads against the Col genome and collected 15,791,682 reads with at least one end un-mapped. The insertion sizes and their standard deviation are estimated automatically by Velvet.

Genes in L*er* inserted sequences were predicted using geneid. Reciprocal best BLAST hit and syntenic map between Col and *A. lyarta* from SynMap were used to identify the *A. lyrata* ortholog of Col specific genes. Orthologs of L*er* unique genes were identified by reciprocal best BLAST hit only.

2.3.5 Further bioinformatic analysis of SNP/indel affected genes

Gene Ontology analysis was performed in agriGO with default parameters (http://bioinfo.cau.edu.cn/agriGO/) (92). dN/dS analysis was conducted using PAML with all four codon models (http://abacus.gene.ucl.ac.uk/software/paml.html) (93). To estimate the *Arabidopsis* branch-specific dN/dS values, we used Poplar and *A. lyrata* orthologous genes downloaded from Phytozome (http://www.phytozome.net/) in tree-guided analysis. Since similar results were attained with all codon models, we reported results from the simplest codon model with CodonFreq set to 0. Normalization of microarray data of affected genes was done in SNOMAD (http://pevsnerlab.kennedykrieger.org/ snomadinput.html) before comparing

expression in Col and L*er* (94). The enrichment of SNP/indel affected genes in multigene families was based on clustering of all *Arabidopsis* genes by MCL (95).

2.3.6 Identification of crossovers and non-crossovers

Sequenced reads of meiosis progeny were mapped to the Col genome with Maq 0.7.1. At each SNP locus, the read counts of all present bases were recorded. Crossovers were identified from the genome-wide distribution of the L*er* allele at SNP loci. L*er*-specific alleles flanked by Col markers were noted as potential GC events. To minimize noise from sequencing errors, we required high-quality calling of a L*er* allele in at least three reads to support a converted SNP. Converted SNPs less than 1 Kb apart were grouped into one gene conversion event. To further reduce false GCs due to repetitive sequence, the 35 bp flanking sequences of each converted SNPs were used as queries for BLAST searches against the Col genome. GCs with half or more converted SNPs in repetitive regions were ignored. The minimal length of CO/NCO was the length between two farthest converted SNPs and maximal length was the length between two closest unconverted SNPs. Midpoint length was the arithmetic average of minimal and maximal length.

To verify the detected COs and NCOs/GCs events, we used primers away from the SNPs supporting the recombination event to perform PCR (40 cycles of denaturing at 95 ℃ for 30s, annealing at 54 ℃ for 30 s, and extension at 72 ℃ for 1min, and an additional step of 72 ℃ for 10 min to allow complete extension). The PCR products were mixed with the corresponding primers and sequenced at the Nucleic Acid Facility at the Pennsylvania State University.

2.3.7 Meiosis simulation and statistical analysis

Custom scripts (available upon requests) were used to simulate 10,000 meioses by assigning crossovers to each chromosome according to the probability of recombination estimated based on genetic and physical maps, with one crossover in each chromosome arm, so that the total number of crossovers per meiosis was 10. The integration of genetic and physical map was according a previous study (96). The number of SNP/indel carried by each gamete was calculated based on the location of crossovers. Different iterations of simulation were tried and because results from 5,000 simulated meioses were similar to 10,000 meioses, final results were based on the 10,000 meioses simulation.

Enrichment of GO groups was analyzed using Fisher's exact test with Yekutieli FDR multi-test correction (92). The relation between multi-gene family members and genes with 10 or more nonsynonymous SNPs was analyzed by the χ^2 test. Enrichment of unknown genes was also analyzed by the χ^2 test.

2.3.8 Data access

The high throughput sequencing data from this study are available at NCBI under the accession number: SRP007172 and SRP008819. SNPs, indels, large DNA polymorphisms and L*er* contigs are available on the website: http://www.personal.psu.edu/hxm16/suppdatafile.zip.

2.4 Results

2.4.1 Sequencing of the L*er* genome uncovered numerous SNPs with functional Implications

We sequenced L*er* to obtain over 61.6 million reads, ~62.7% of which were uniquely mapped to the Col reference genome, providing ~18.7x coverage (Appendix Table 2-1). We identified a total of 349,171 SNPs between the two genomes, offering a map of molecular markers with an average distance of 340 bp (median distance of 118 bp) between adjacent markers (Appendix Figure 2-1A and Appendix Figure 2-1A and C). PCR amplification and Sanger sequencing of 23 randomly selected regions with 62 predicted SNPs on Chromosome 1 confirmed all of them and also detected a few additional SNPs (data not shown), indicating that SNP identification was conservative and accurate, providing a valuable resource for functional studies.

To assess the possible functional implications of genome variations, we analyzed their distribution and effect on protein sequences. Of the SNPs, 76,649 (~22%) are in protein-coding sequences (CDS) and 194,911 (~56%) in intergenic regions. Among the SNPs in CDS, nearly half of them (35,798) caused nonsynonymous changes, affecting 13,158 of the 27,169 annotated protein-coding genes. Most of the affected genes carry only a few nonsynonymous SNPs, but 443 genes have 10 or more nonsynonymous substitutions (Figure 2-2A; referred to as the 443 genes hereafter). In addition, 357 SNPs lead to premature stop codons in 319 genes (referred to as the 319 genes) in L*er* relative to Col, suggesting possible defects in the L*er* alleles. In contrast, 89 genes were found to have a sense codon in L*er* corresponding to the stop codon in Col, suggesting longer coding regions in L*er* and might be defective in Col.

To obtain clues about the potentially functional effect of the coding differences, all genes were classified into three annotation types: “known”, “unknown” with EST or cDNA

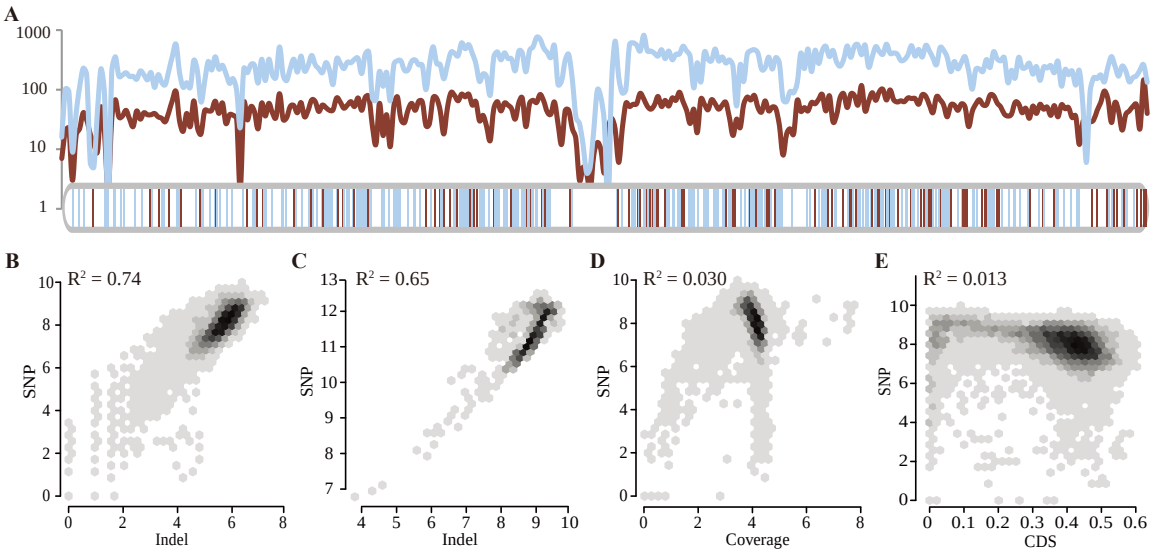


Figure 2-1. The correlation of SNP and small indel densities. (A) Parallel change of SNP and small indel density on Chr1. The density was defined to be the number of SNPs/indels per 100 Kb. The blue curve represents SNP density and the red curve for small indels. Blue and red vertical bars below show the location of large deletions and insertions, respectively. (B, C) Logistic regression of SNP and indel densities in 100 Kb (B) or 1 Mb (C) sliding windows. (D, E) Logistic regression of SNP density with read coverage (D) or CDS fraction (E) in 100 Kb sliding window. A near zero R^2 value suggests read coverage or CDS fraction does not contribute much to the correlation.

support, and "unknown" without expression information. Surprisingly, smaller fractions of the 443 genes were found in either of the "unknown" categories than whole-genome frequencies (χ^2 test, p=0.0019), suggesting that the genes with amino acid differences are not more likely to be pseudogenes than the genome average. The 319 genes have more than expected unknown genes without expression information, but still the majority of them have annotated functions. 68.2% of the 443 genes are members of multi-gene families (χ^2 test, p=1.27e-10), compared with 53% of all annotated protein-coding genes in multi-gene families, suggesting that nonsynonymous mutations might be more tolerated in multi-gene family members. In contrast, only 57.1% of the 319 genes belong to multi-gene families (χ^2 test, p=0.15). Strikingly, all members are altered

(missense or nonsense) in L*er* in six gene families, two of which are involved in resistance to biotrophic oomycetes and other pathogens (97).

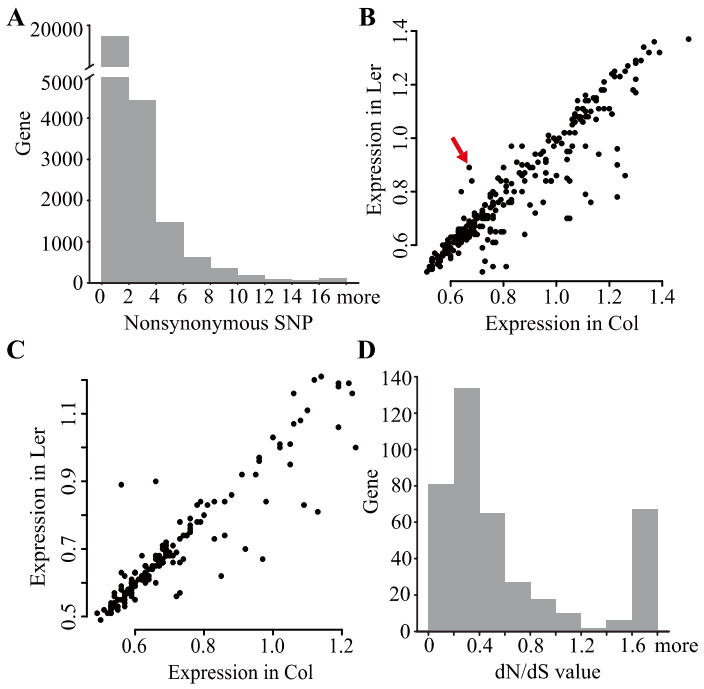


Figure 2-2. Nonsynonymous SNPs and affected genes. (A) The number of nonsynonymous SNPs per gene. Although about half of the genes contain nonsynonymous SNP, a much smaller set of genes has 10 or more nonsynonymous SNPs. (B) Expression levels in Col (x-axis) and L*er* (y-axis) of genes with 10 or more nonsynonymous SNPs. Red arrow points to AT5G58120 which has higher expression in L*er* and encodes a disease resistance protein. (C) Expression levels in Col (x-axis) and L*er* (y-axis) of genes with a L*er*-premature stop codon. (D) *Arabidopsis* branch specific dN/dS value of genes affected by 10 or more nonsynonymous SNPs with regard to *A. lyrata*. Only a few genes show neutral evolution, but most are under either positive or negative selection.

We further examined Gene Ontology (GO) for possible enrichment of specific categories among the genes affected by genome variations (Figure 2-3). Among the 443 genes, those related to defense response, apoptosis, transmembrane receptor and ATP binding are enriched, supporting the idea that Col and L*er* differ in defense response. We also examined the expression profile using the information from the Plant Ontology (PO) database and found that

most enriched PO groups are reproductive tissues and stages (Appendix Table 2-2). According to microarray analyses (98), most of mutated genes have similar or lower expression levels in L*er* than that in Col (Figure 2-2 B and C). However, a few genes, such as the AT5G58120 gene encoding a disease resistance protein, are expressed at higher levels in L*er* than in Col (Figure 2-2B). In short, changes in gene sequence and expression might impact functions, such as pathogen response.

To investigate whether these genes have been under selection, we estimated the dN/dS ratio of the *Arabidopsis thaliana* branch after divergence from a close relative, *Arabidopsis lyrata*. dN is the rate (observed over possible changes) of nonsynonymous substitutions, whereas dS is the rate of synonymous substitutions. Neutrally evolved genes tend to have dN/dS values close to 1. In contrast, genes under negative or positive selection tend to have dN/dS values close to zero or larger than 1, respectively. Most of the 443 genes have dN/dS values of 0~0.4 (negative selection indicative of highly conserved functions), but 50 genes have much larger dN/dS values (≥ 1.6) using all four amino acid frequency models, suggesting positive selection for new functions (Figure 2-2D). Interestingly, a comparison of GO results with dN/dS analysis revealed that 21 of those 50 genes belong to enriched GO groups (Figure 2-3), further supporting the idea of altered functions.

2.4.2 Numerous small indels with similar distribution patterns to those of SNPs

Small indels of several nucleotides were previously found to be more prevalent between *Arabidopsis* variants than among humans (39). We uncovered 58,085 indels of 1 to 4 bp, with the median distance between indels being 919 bp, and noticed that distribution of SNPs and small indels showed parallel patterns across the entire genome (Figure 2-1A and Appendix Figure 1 B and C). To investigate the potential correlation between SNPs and small

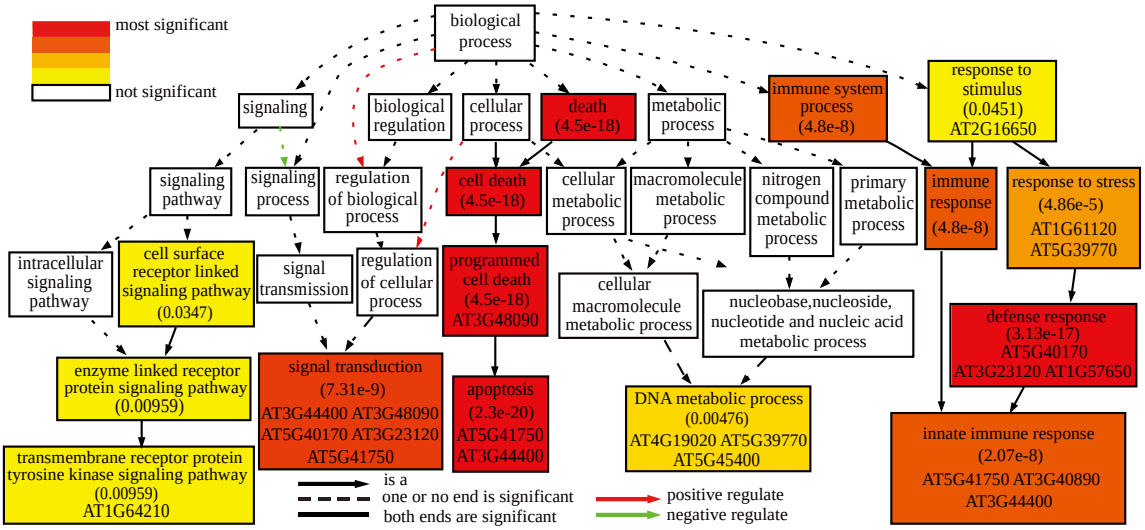


Figure 2-3. Gene Ontology groups enriched among genes with 10 or more nonsynonymous SNPs. Statistical significance is color coded, with Yekutieli FDR adjusted P value shown in each significant group. Most enriched GO groups contain some genes with large dN/dS value (≥1.6), as shown by TAIR gene IDs in the box.

indels frequencies, logistic regression was applied in 10 Kb, 100 Kb or 1 Mb sliding genomic windows. SNPs and indels are highly correlated in 100 Kb and 1 Mb windows, with R^2 values of 0.74 and 0.65, respectively (Figure 2-1 B and C), more than that in 1 Kb windows (R^2 = 0.52; Appendix Figure 2-1D). To exclude the possibility that more SNPs and indels might be detected in regions with high read coverage, the density of SNPs and indels were examined with regard to read coverage in each 100 Kb genomic window, but no correlation was observed (Figure 2-1D and Appendix Figure 2-1E). Moreover, the fraction of CDS in each window cannot explain the change in frequencies of SNP/indel (Figure 2-1E and Appendix Figure 2-1F). This strong correlation between frequencies of SNPs and small indels is similar to the mosaic pattern observed in mouse (99-101), which is probably due to variation in divergence time among different genomic regions. Since Col and L*er* were derived from same natural population, some

genomic regions might have inherited the same haplotype, whereas in other regions Col and L*er* might have had different haplotypes with longer divergence time and more polymorphisms.

Further, 1, 674 of the small indels are inside CDS, causing frameshift in 844 genes and non-frameshift changes in 461 genes. The number of affected genes is very large, considering that Col and L*er* have diverged only approximately 200,000 years ago (82). Similar to the SNPs-impacted genes discussed in the previous section, most genes affected by indels have very low dN/dS values in comparison with *A. lyrata*, indicating these genes have been under purifying selection during the 10 million years of divergence between *A. thaliana* and *A. lyrata* (102), but have changed more recently between Col and L*er*. Nevertheless, some of the genes affected by small indels have dN/dS ratios of 1.6 or higher , suggesting that they might have been under positive selection. In addition, many genes with frameshift mutations have lower expression in L*er*, but fewer non-frameshift genes do so. GO categories of transmembrane receptors and ATP binding are overrepresented among genes affected by frameshift mutations, whereas genes with non-frameshift mutations are enriched for transcription factors, suggesting functional differences in these categories.

2.4.3 Detection of large indels and CNVs

Because mapping of the short reads could not identify large indels, to detect large genomic variations, we assembled paired-end reads from L*er* into 30,217 contigs that ranged from 100 bp to 119 Kb (N50=11 Kb), representing ~78% of the Col reference genome, and then aligned them with the Col genome. 16,560 contigs were mapped to unique sites, 7,503 had their segments mapped orderly, 994 had rearrangements for mapping positions, and the remaining were not mapped. From the contigs mapped to unique sites, we identified 1,658 large deletions with a median size of 730 bp and 700 large insertions with a median size of 266 bp, spanning

cumulatively 2,841 Kb and 372 Kb, respectively. To evaluate the indels, we compared them with the Monsanto L*er* contigs (81,306) (downloaded from the Monsanto Co. database), which matched to ~60% of the Col genome. About 78% of the deletions we detected were also uncovered by the Monsanto contigs with ~99% of them consistent in both data sets. Moreover, ~71% of the insertions we detected were confirmed by the Monsanto contigs with ~96% of them in agreement. In addition, 28 of the indels we identified were tested by PCR and 20 of them displayed different fragment sizes between Col and L*er*.

1,759 of 2,315 large indels (75.9%) are located in intergenic regions, while the others contribute to the gain/loss of exons/introns/untranslated regions (UTRs) or even the entire genes (Figure 2-1A, Appendi Figure 2-1C and Table 2-1). 130 single copy genes were absent in the L*er* genome, with one example (At1g51430.1) confirmed by PCR. Of these 130 genes, 25 were found to have an ortholog in *A. lyrata*. In addition, 107 putative genes were predicted from L*er*-specific sequences; 9 of the 107 were detected in *A. lyrata*. Furthermore, 186 genes with exons/UTRs affected by indels could have changed/disrupted expression or functions. F-box genes encode subunits of E3 ubiquitin ligases that are involved in physiological and environmental responses and differ dramatically in gene number among *A. thaliana*, poplar and rice (103). We found that eight F-box genes are absent from L*er* and sixteen other F-box genes are partially affected, suggesting that these rapidly evolving genes are highly unstable even within the *Arabidopsis* species. In addition, L*er* lacked four disease resistance genes and part of six others. As large deletions affecting genes can cause phenotypic variations among different accessions (104), it was surprising to see that most genes affected by large indels show low dN/dS value in comparison with *A. lyrata*, suggesting that they have been conserved and under purifying selection since the separation of the two species. However, many of these genes have lower expression levels in L*er* than those in Col, possibly due to nonsense-mediated decay. Furthermore, large indels also lead

Table 2-1. The number of genes/non-coding segments affected by large deletions/insertion

Deletion/insertion position		Deletions			Insertions		
		NOT TE-mediated	TE-mediated	Sum	NOT TE-mediated	TE-mediated	Sum
Exon	5'-UTR[a]	11	4	15	0	0	0
	3'-UTR[a]	22	18	40	16	1	17
	5'-CDS portion[b]	7	1	8			
	3'-CDS portion[b]	12	3	15			
	Middle portion of CDS[b]	6	0	6	20	0	20
	Full exon[c]	22	12	34			
	Multiple-exons	21	13	34			
Complete gene		52	78	130	-	-	-
Intron		86	25	111	121	3	124
Intergenic		387	841	1228	377	154	531
Pseudogene		22	2	24	8	0	8
ncRNA/miRNA[d]		8	5	13	-	-	-
Total		656	1002	1658	542	158	700

[a] Indels were detected within the UTR region of a 5' or 3' terminal exon;
[b] A portion of a coding region within an exon was deleted from the L*er* genome when compared with Col;
[abc] For cases when a single exon is affected;
[d] ncRNA: non-protein-coding RNA; miRNA: microRNA

to reciprocal loss of genes in Col and L*er* for 22 homologous gene pairs, providing possible examples of gene loss following duplication.

Large indels might include copy numbers variations (CNVs) between Col and L*er*. To test this, we used the sequences affected by these indels to search against the Col genome. We defined CNVs as indels of one or more copies of similar sequences. Using a criterion of 80% identity over 80% of the query, we identified 614 deletions (~38%) and 20 insertions (~3%) affecting sequences similar to other copies in Col genome and 85 of these affected genes. Some CNVs occurred in tandemly duplicated genes; for example, in a cluster of genes encoding carbohydrate-binding X8 domain proteins with over 96% identity in amino acid sequence, Col has five copies (AT4G09462.1, AT4G09464.1, AT4G09465.1, AT4G09466.1, and AT4G09467.1), but L*er* had a deletion of AT4G09467.1. A major type of CNVs affected copy

number of transposable elements (TEs) or gain/loss of adjacent genes. For example, among the members of the ATREP1, ATREP2, and ATREP3 TE families present in Col, 13, 13, and 20, respectively, were absent in L*er*. In our study, 997 of 1758 (56.7%) gain/loss of DNA segments in intergenic regions and 149 of 557 (26.8%) in genes contained a segment with high sequence similarity to known TEs in Col, suggesting a role of TEs in generating CNVs.

2.4.4 Generating and sequencing "tetrads" of meiotic progeny plants

To observe meiotic recombination using tetrad analysis (Appendix Figure 2-2) with homologous chromosomes with numerous polymorphic markers, we constructed a hybrid between Col and L*er*, each mutated for the *QRT* gene (Figure 2-4A). A tetrad of four attached pollen grains from the Col/L*er qrt/qrt* F1 hybrid was used to pollinate a single pistil of an emasculated Col flower, producing four seeds. Here we named the plants grown up from the four seeds as meiotic progeny plant (MPP)-A, -B, -C and -D (Figure 2-4A), each containing the paternal genome with a mixture of Col and L*er* DNA and the maternal genome of 100% Col DNA. We sequenced 8 MPP genomes from two independent meioses, named as the 1st meiosis and the 2nd meiosis, yielding sequence information for each plant with about 8.2~16.6X coverage, matching 94~97.1% of the Col genome (Appendix Table 2-1).

2.4.5 Single-base resolution analysis of COs and NCOs/GCs

Meiotic CO events at single base resolution were investigated by analyzing sequencing reads from different MPPs (Appendix Figure 2-3), such as the CO on Chr2 as revealed by mapped reads from MMP-C and MMP-D (Figure 2-4B). We detected a total of 18 COs (Figure 2-5), which were verified by PCR and conventional sequencing (data not shown). All COs were

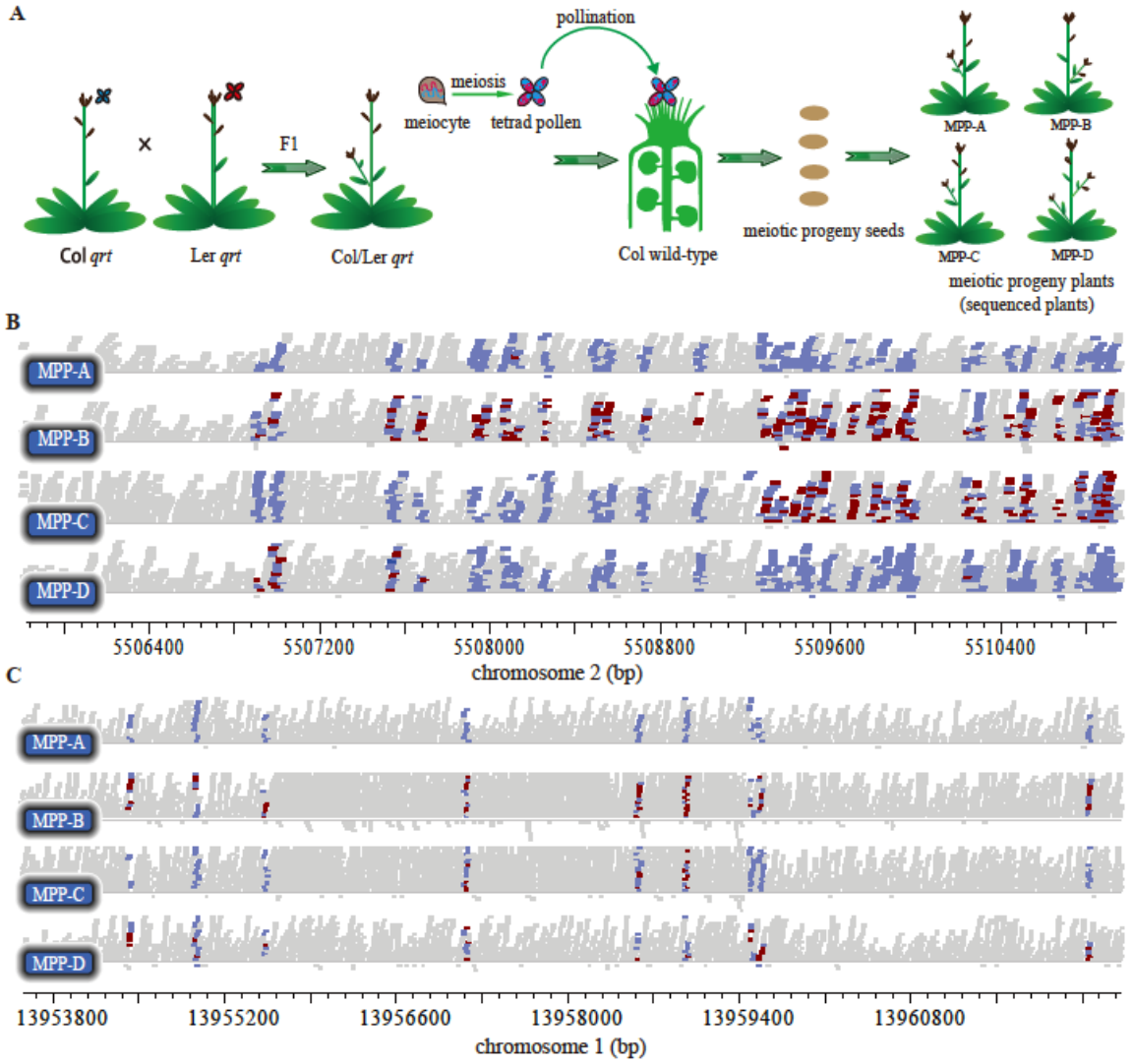


Figure 2-4. Tetrad analysis detecting meiotic COs and NCOs tracts using genomic sequencing in Arabidopsis. (A) A schematic illustration for generation of meiotic progeny plants (MPPs) to detect meiotic recombination using high throughput sequencing. Blue: Col genotype. Red: L*er* genotype. (B) An example of detected CO on Chr2. MPP-A has pure Col genotype, except one red bar possibly due to sequencing error. MPP-B has equal representation of blue and red bars, carrying the L*er* paternal genotype and the Col maternal genotype. Sequence exchange between MPP-C and MPP-D shows a CO event. One red bar in MPP-D to the right was likely due to sequencing error. The CO tract in the middle contains a 3:1 gene conversion, indicating repair of DSB in the L*er* chromatid by using the homologous Col chromatid as template. (C) An example of detected NCO in Chr1 from the 1st meiosis. Conversion occurred in chromatid inherited by MPP-C from Col to L*er* genotype, leading to the 3:1 ratio in this region. Blue horizontal bars represent mapped reads with a Col specific SNP and red bars represent reads with a L*er*-specific SNP. Grey bars represent reads without a SNP. Each MPP plant contains one set of chromosomes from a Col/Col mother and another set of chromosome from a Col/L*er* hybrid.

located in the intergenic regions and each chromosome experienced at least one CO (Figure 2-5), consistent with its role in holding homologs together for accurate segregation. A total of 9 COs were found in each meiosis (Figure 2-6A and Appendix Table 2-3), in remarkable agreement with previous estimate of 9.24 COs from cytological and molecular genetic analyses (105, 106).

Similar to the budding yeast, *Arabidopsis* uses the interference-sensitive pathway for the formation of a large majority of COs and the interference-insensitive pathways for a clearly detectable minority of COs (106-109) (110) (111). From physical distance between two COs, 16 of the 18 COs could be derived via the interference-sensitive pathway. In contrast, two COs were closely located on chromosome 3 in the 2nd meiosis (* in the Figure 2-5), being only about 244 Kb (<1 cM) apart, much smaller than the average interval (10.6 Mb) between other adjacent COs, suggesting that one or both of these COs could be generated by the interference-insensitive pathway.

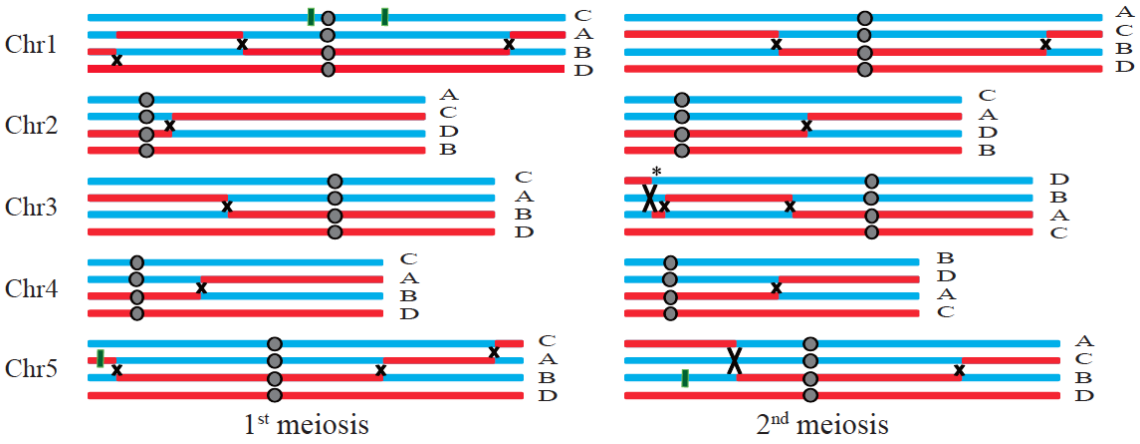


Figure 2-5. The distribution of COs and NCOs in the 1st and 2nd meioses. Either meiosis has 9 COs. The Col genotype is shown in cyan and the L*er* genotype in red. "×" represents the location of CO. A, B, C and D represent four meiotic progeny plants, respectively. Green vertical bars show the detected NCOs positions. One or both of two closely spaced COs (* marked) in Chr3 from the 2nd meiosis could be from interference-insensitive pathway.

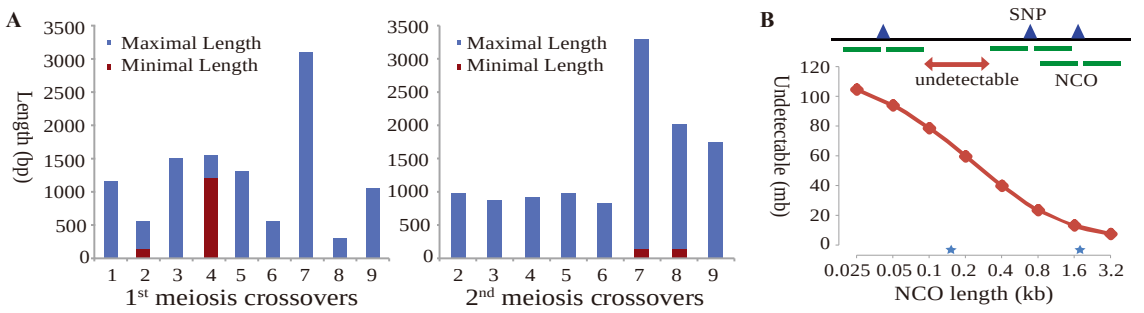


Figure 2-6. Properties of COs and NCOs in Arabidopsis. (A) The minimal (red) and maximal (blue) lengths of detected crossover tracts from the 1st meiosis (left) and the 2nd meiosis (right). The maximal length is the distance between two closest SNPs in unchanged regions flanking CO. The minimal length refers to the region having multiple 3:1 converted SNPs (one of the chromatid having a converted allele). In either the 1st or the 2nd meiosis, there are 2 COs containing multiple converted SNPs, with maximal lengths ranging from ~500 to ~3,000 bp. In the 2nd meiosis, CO-1 is not displayed here due to a large region without SNPs, making the position of CO-1 uncertain. (B) Predicted cumulative length of uncovered regions changes with possible length of NCO. NCO can only be detected if it covers at least one SNP, but are invisible between two adjacent SNPs. The predicted cumulative length of uncovered regions increases significantly when the length of NCO diminishes. Two blue stars show previously reported median NCO tract length from yeast (1.8 Kb) and human (156 bp).

Meiotic COs are known to distribute unevenly along the chromosomes, with recombination hot and cold spots (106, 112). In *Arabidopsis*, putative hot spots have been reported in short regions (a few kilobases) on Chr4 (113). Here, one CO we detected on Chr4 was in one of these regions. In addition, we also found that two COs, one in each meiosis analyzed here, around 25.6 Mb on Chr1 were located with a distance of only about 2 Kb (Figure 2-5), possibly representing a hot spot, which is within the size range of 1 to 10 Kb for mammalian meiotic recombination hot spots (114).

Tetrad analysis of all four meiotic products with single-base resolution allowed us to estimate the maximum (using flanking SNPs) and minimum (using converted SNPs) sizes of CO-associated conversion tracts (COCTs) for the first time in a multicellular organism (Appendix Figure 2-2; see Methods). One CO was located in a 129,507 bp region (Chr1: 8,733,517 to 8,863,024 bp) without SNPs, whereas the remaining 17 COs had maximum lengths of COCTs

ranging from 306 to 3,288 bp (Figure 2-6A), with a median of maximum sizes of COCT tracts of 1,115 bp, shorter than the median maximum estimate (2,643 bp) of the budding yeast COCTs (79). Among the 17 COCTs described here, 47% had maximum lengths of less than 1 Kb and 35% had maximum lengths of 1-2 Kb. In another study of yeast CO utilizing microarrays (78), the arithmetic average of minimal and maximal estimates of COCTs was defined as the midpoint length. The median of midpoint length of 18 COCTs detected here in *Arabidopsis* was 558 bp, significantly shorter than the 2 Kb value in yeast (78) (Wilcoxon rank-sum test, p = 5.14e-05).

According to the double strand break repair model (DSBR) for meiotic recombination (72, 76), the gap generated following the DSB is repaired using homologous sequences, leading to GC if there is sequence polymorphism. In the budding yeast, genome-wide analyses showed that most COCTs have a simple 3:1 GC pattern, but a small fraction of COCT regions had complex patterns (78, 79). Our analysis showed that all 6 COs with internal SNPs for GC detection were associated with 3:1 type GCs, including two with a single SNP, 3 with 140-150 bp COCTs, and a large one with a COCT of 1,208 bp. It is possible that the size of the initial DSB gap in *Arabidopsis* could be ~150 bp or shorter, whereas possible expansion of the dHJs could lead to longer conversion tracks. Interestingly, one CO (the CO on Chr2 in the 1st meiosis), spanning an 86-bp deletion in L*er*, resulted in the removal of the deletion in one daughter cell via GC.

In addition to COs, we also detected 3 and 1 NCO/GC events in the 1st and the 2nd meioses, respectively, as confirmed by PCR and sequencing. All NCOs/GCs were located in intergenic regions. The longest NCO/GC tract, in the 1st meiosis, showed conversion of three SNPs spanning 1,799 bp from Col to L*er* in the MMP-C plant (Figure 2-4C). The other NCO/GC tracts had minimum sizes (the region between the converted SNPs) of 1 bp. The estimated maximum NCO/GC tracts (the distance between the closest SNPs unaffected by the NCO) ranged from 3,078-6,696 bp (Appendix Table 2-4). Because each MPP contains the maternal Col

genome, we could only detect NCO/GC events with sequence change from the Col allele to the Ler allele. Assuming equal frequency in both directions of conversion, *Arabidopsis* could have ~6 NCO/GC events per meiosis.

2.4.6 Redistribution of genome variations after meiosis

The sequences from the 8 MPPs provided a unique opportunity to investigate the newly generated genetic architectures following meiosis. Figure 2-7A and Appendix Table 2-5 show the patterns of redistributed single nucleotide variants and indels in the two sets of MPPs, in comparison with Col. For the 1st meiosis, the MPPs had 42,669 to 255,360 L*er*-specific single nucleotide variants and 7,632 to 53,149 indels. The MPPs from the 2nd meiosis had 113,563 to 179,602 L*er*-specific single nucleotide variants and 19,288 to 40,607 indels, quantitatively demonstrating the reshuffling of genetic variations due to meiotic recombination and chromosome assortment. To further investigate the re-distribution of genetic variations, we simulated 10,000 meioses, producing 40,000 meiotic products, by comparing genetic map and physical map and assigning COs according to genetic distance. The simulated COs were used to predict the number of SNPs and indels in meiotic products. Strikingly, 2 meiotic products had very small or large number of variations, respectively, in the tails of simulated distribution of number of SNPs and indels, with one highly similar to Col and the other to L*er* (Figure 2-7 A and B). The generation of two extreme products might be due to two factors in this meiosis (1st meiosis): (1) preferential occurrence of COs between the same two chromatids, such as the 3 COs for Chr1 and 2 COs for Chr5; (2) non-recombinant chromatids tended to be assorted into the same two products (MPP-C or MPP-D), whereas recombinant chromosomes tended to be the other two products (MPP-A and MPP-B) (Figure 2-5).

A

Genotype		Total SNPs
L*er*		349171
1st meiosis	Plant A	91913
	Plant B	255360
	Plant C	42669*
	Plant D	308386*
2nd meiosis	Plant A	113563
	Plant B	179602
	Plant C	172329
	Plant D	232842

B

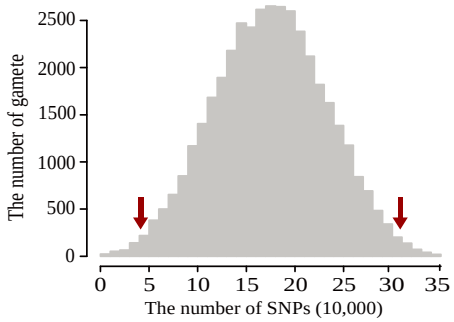


Figure 2-7. The distribution of single nucleotide variants in meiotic products. (A) The number of single nucleotide variants in L*er* and 8 experimented meiotic products. As Col sequence was used as reference, pure Col region had zero variant and only bases different from Col were counted. Two extreme gametes were marked by red asterisk, with one highly similar to Col and the other very similar to L*er*. (B) The distribution of single nucleotide variants in 40,000 simulated gametes. Simulation was performed according to genetic and physical maps of *Arabidopsis*. The unit of x-axis is 10,000. Two red arrows indicate the location of two extreme gametes in the simulated distribution.

2.4.7 CNVs due to meiotic reshuffling of structural variants

CNVs have been shown to affect various biological processes (115). Estimating the rate of CNV generation is critical to understanding their effects on genome evolution and gene functions. The rates of *de novo* CNV in human have been estimated; for example, one study found that the rate for total genome-wide new large CNVs (>100 Kb) is about 1.2×10^{-2} per genome per transmission (116) and another study reported that most of over 4000 CNVs analyzed had individual rates of ~10^{-5} per generation (117). Our limited analysis revealed that meiosis can rapidly generate CNVs among siblings, producing 21 and 32 CNVs in the two sets of four meiotic products, respectively. Further examination of L*er* reads with PCR verification showed that these CNVs were due to reshuffling of existing highly similar sequences that map to different locations (Appendix Figure 2-4). These non-allelic similar sequences could be on the same chromosome, and a CO between them can lead to CNVs in the meiotic products (Appendix Figure 2-4B). When the similar sequences are on different chromosomes, only the assortment of

the Col and L*er* chromosomes is needed to cause CNVs in the meiotic products (Appendix Figure 2-4C). *Arabidopsis* can outcross 3% of the time in environments with natural populations (118), generating hybrids in which CNVs can be generated from reshuffling much more frequently in comparison to *de novo* mutation.

2.5 Discussion

2.5.1 Genetic variation and phenotypic variation

Natural variations are shaped by integrated forces of mutation, recombination and selection, causing phenotypic differences and affecting individual adaptation to local environments. We have identified >400,000 SNPs, indels and CNVs between Col and L*er*, two accessions that are thought to have diverged approximately 200,000 years ago (82), even though there probably have been more recent gene flow between them (see Introduction). We have analyzed the Col/L*er* genome variations and found evidence for functional divergence between alleles in Col and L*er*, including multiple amino acid substitutions, premature stop codons or extension of reading frame in L*er*, small indels causing frameshifts, and large indels of part or all of coding regions. Some of these genes have predicted functions that could influence the adaptive fitness of Col or L*er*, potentially impacting traits such disease resistance and flowering time for plant health and reproductive success, respectively.

In *Arabidopsis*, genetic variations from SNPs to CNVs could affect gene functions; for example, SNPs can change amino acid in phytochrome A (PHYA) and B (PHYB), affecting light responses and flowering time (119, 120), and create a new splicing site with defective gene function (84). Small and large indels in the *RPS2* and *MAM2* genes, respectively, caused pathogen sensitivity (104, 121). We found genes responsible for biotic stress responses are enriched among genes specifically altered between Col and L*er*, including those encoding F-box proteins, LRR (Leucine Rich Repeat)-RLK (Receptor-Like Kinase), RLP (Receptor-Like Protein), and NBS (Nucleotide Biding Site)-LRR proteins (Appendix Table 2-6).

2.5.2 Possible relationship between frequency of CO, genome size and length of synaptonemal complex

Human and mouse, as well as other animals and plants, have very different genome sizes, yet all have 1-3 COs per homolog pair (122, 123), similar to *Arabidopsis* but unlike the 2-11 COs per chromosome in the budding yeast (78, 79). Recent studies indicated that, between individuals of the same species for human and others, the genetic distance (CO number) is positively correlated with the length of the synaptonemal complex (SC) but not the length of DNA (124, 125). However, this correlation does not seem to hold between species; for example, human, *Arabidopsis*, and the budding yeast have SC lengths per chromosomes of approximate 10-25, 2-3, and 1-2 microns, yet the CO numbers per chromosomes are 1-3, 1-3, and 2-11, respectively (126) (111, 123). Strikingly, the ratios of genome size to SC length are very similar between human (~10-12 Mb/micron) and *Arabidopsis* (~12 Mb/micron), but much smaller in the budding yeast (~0.5 Mb/micron). Because similar number of chromatin loops are packed into the same SC length (72), the sizes of chromatin loops associated with SC are likely similar between human and *Arabidopsis* and are ~20 times larger than that in yeast, providing a possible explanation for the difference in number of CO per chromosome. Our results also suggest that *Arabidopsis* has shorter COCTs than that in yeast (78, 79). The ability to conduct tetrad analysis in *Arabidopsis* offers great opportunities to gain further insights into the molecular control of meiosis in multicellular organisms.

2.5.3 The low frequency of detected NCOs and possible short GC tracts

We detected only ~15 (9 COs + 6 NCOs) recombination events per meiosis; however, fluorescence immunolocalization studies in *Arabidopsis* detected about 120 AtRAD51 foci per meiotic cell (127), suggesting that there are over 100 DSB sites in a single meiosis. It is possible

that there are many recombination events, but most are not detected because the sizes of DSB gaps and GC tracts are very small. An analysis of genomic regions relative to SNP distribution indicated that near 80% of the genome is at least 100 bp from any SNP (Figure 2-6B). If the GC tract length is 100 bp or shorter, ~80% or more of the NCO recombination events would be undetectable; therefore, our results could suggest that GC tracts in *Arabidopsis* are very short. Alternatively, in a fraction of the DBS repair events, the repair of meiotic heteroduplex DNA could also result in restoration of the parental genotypes in regions flanking the initial DSBs, as observed in yeast (128). A third possibility is that some DSBs might be repaired using the sister chromatids as templates, resulting in no GC. Recent studies showed that about one-third of the breaks could be repaired using the sister chromatid (129); however, this could not fully explain the difference between the numbers of detected CO and NCO events and the observed AtRAD51 foci. Therefore, short GC tracts of approximately 100 bp or less are at least one of the explanations for our results. On the other hand, if we assume the length of NCO in *Arabidopsis* is between those of yeast and human, the frequency of NCO is estimated to be 4~8 per meiosis (Figure 2-6B). Another way to estimate the frequency of NCO per meiosis is to use the fraction of COs with detected GCs in COs, because all COs should have a conversion tract if there are SNPs. Among the 18 COs we observed, 6 had detectable GCs but 12 did not (Figure 2-6). If the same fraction of NCO events were not detected due to the lack of SNPs, then there should be another 12 GCs/NCOs per meiosis, in addition to the 6 we estimated.

2.5.4 The redistribution of natural variations and generation of new CNVs

Although *Arabidopsis* is a predominantly self-pollinating plant, 3% outcrossing still allows gene transfer among different accessions (118). When the population faces the challenges of external environment changes, some hybrids that possess newly generated genotypes might

confer better adaptation and out-compete others. Our data of four meiotic descendants from either of two meioses showed two interesting outcomes of outcrossing: (1) meiosis indeed dramatically alters genetic variations, distributing the alleles from the two parents, creating new strains with new combinations of genes that are vastly different from either parent and (2) reshuffling of existing structural variants can generate new CNVs in a rapid manner. These results provide a direct view of the landscapes of genetic variations at whole genome scale, revealing how a single round of meiotic recombination and chromosome assortment can serve to reshape natural variations.

CHAPTER 3

GENOME-WIDE ANALYSIS OF SMALL RNAS IN ARABIDOPSIS MEIOCYTES BY sRNA-SEQ

3.1 Summary

Small RNAs are key regulators of gene expression in plants, contributing to development, disease, stress response and many other biological processes. The discovery of small RNAs has been greatly accelerated with the adaptation of next generation sequencing technologies. However, the profile of small RNAs in meiocytes, the cell population undergoing meiosis, is still lacking, due to the difficulty of isolating meiocytes from tiny Arabidopsis anthers and thus the low amount of RNA that could be retrieved. Here a recently developed method was applied to collect meiocytes and small RNAs were profiled using SOLiD sequencing. 97 of 266 known miRNAs show expression in meiocytes. Interestingly, five miRNAs were found to account for more than half of the total miRNA expression in meiocytes, among which miR158a takes up about one third. The target genes of these five miRNAs have little or low expression in meiocytes. One putative novel miRNA was identified, which shows conservation with rice and maize. Analysis of longer reads provided clues for possible long ncRNAs in meiocytes. Our small RNA transcriptome study uncovered miRNAs and other type of small RNAs possibly crucial for normal meiosis progression.

3.2 Introduction

Since the first discovery of miRNA in worm (130) and siRNAs in plant (43), small RNAs, 20-40 nucleotides in length, have been shown to be key regulators of diverse biological processes, like development and diseases, by various mechanisms, including mRNA degradation, translation inhibition and chromatin remodeling (131) (44). Arabidopsis encompasses a wide set of small RNAs, the most studied categories of which are miRNAs, hc-siRNAs, tasiRNAs and nat-siRNAs. miRNAs, predominantly 21 nucleotides long in Arabidopsis, fine-tune the abundance of specific mRNAs. hc-siRNAs, tasiRNAs and nat-siRNAs are three major types of endogenous siRNAs in Arabidopsis and have different length and diverse function (44).

Despite of their distinct characteristics, the biogenesis and function of these small RNAs involve the same protein families: DICER-LIKE (DCL) enzymes and ARGONAUTE (AGO) effectors. The generation of small RNAs begins with double stranded RNA (dsRNA) precursors, with miRNAs produced from the dsRNA portion of hairpin structure and siRNA mainly from dsRNA made of two strands, although recent studies reported some siRNAs could also be generated from long hairpin structures (45). Relatively long dsRNA precursors are then cut by DCL into small RNAs. Different from animals, after DCL processing all plant small RNAs will be modified by the S-adenosyl-L-methionine-dependent dsRNA methyltransferase (MTase) HUA ENHANCER1 (HEN1) to bear 2'-O-methylation at 3'-terminal ribose so that small RNAs are protected from degradation. Finally, modified small RNAs are loaded into AGO effectors, which keep the effective strand and remove the other, to exert function by base pairing. Arabidopsis has 4 DCL and 10 AGO proteins with specific, but somewhat redundant, function in the generation of distinct types of small RNAs.

Early studies identified novel small RNAs using cloning approach, the basic process of which follows RNA extraction, size selection on gel, ligation of 5' and 3' adaptors, RT-PCR

(Reverse Transcription-Polymerase Chain Reaction) and finally Sanger sequencing. This approach successfully identified small RNAs of high abundance in normally growing Arabidopsis (47) (49) and also stressed individuals (132). Later studies utilized high-throughput sequencing technologies, which not only allows rapid discovery, but also enables the detection of low-abundance small RNA species. Lu et al adapted MPSS technology to explore the landscape of small RNAs in Arabidopsis seedling and inflorescence, which detected 77 of 92 known miRNAs at that time and uncovered tens of novel miRNAs depending on the selection criteria used in addition to thousands of siRNAs (133). Sequencing small RNA population in RNA-Dependent RNA Polymerase 2 (*rdr2*) mutant by MPSS and 454 technologies identified 13 new miRNAs, all of which are of low abundance, and a vast repertoire of ta-siRNAs (134). As small RNAs are shorter than sequencing reads of Illumina or SOLiD, small RNA detection is not troubled by short read length in these technologies. On the contrary, the extraordinarily high throughput and low cost of these two sequencing technologies facilitated the investigation of small RNAs in various tissues or organs, including pollen, root, leaf and others (135-137). The combination of high sequencing depth and tissue specificity discovered even more small RNAs. For example, the profiling of small RNA population in root identified 66 novel miRNA genes and 15 novel miRNAs from known miRNA genes (136).

However, the study of small RNAs in Arabidopsis meiocytes is still lacking. Unlike animals of which germline forms during embryo development, germline in plant develops from somatic cells later in the adult stage through the differentiation of floral meristem and meiosis. Meiocytes, part of the germline, are cells undergoing meiosis, producing four haploid cells with genome variation due to crossover and gene conversion. As meiosis is a highly complicated process marked by not only distinct cell division stages but also synapsis and meiotic recombination, the potential role of small RNAs in gene regulation and chromatin remodeling during meiosis is particularly interesting. Moreover, recent study reported that siRNAs repressed

the activity of transposable elements in pollen to maintain genome integrity (137) and transposable elements were shown to be active in Arabidopsis male meiocytes (138, 139). siRNA may exert similar function in male meiocytes to suppress active transposition. To search for novel RNA species in meiocytes and characterize the function of small RNAs in meiosis, we collected Arabiodopsis male meiocytes and sequenced small RNA populations using SOLiD technology.

3.3 Material and Methods

3.3.1 Male meiocytes collection and sequencing sample preparation

The Columbia (Col) ecotype of Arabidopsis thaliana was grown in Metro-Mix 200 soil (Greenhouse & Nursery Supplies, http://www.griffins.com/) under 16h light and 8h dark in a growth chamber at 18-22 degree (Celsius). To isolate male meiocytes, young floral buds were harvested, and stage 5–7 anthers were separated from other floral organs on a glass slide under a Nikon stereoscopic microscope. The dissected anthers were immediately transferred to a microchamber with liquid meiocyte medium. Subsequently, two tiny needles were used to gently dissect anthers to release the meiocytes under the dissecting microscope. The microchamber with dissected samples was moved onto a Zeiss inverted microscope with a micromanipulation platform. A glass pipette was mounted onto the micromanipulator, and used to absorb meiocytes by gentle mouth pipetting. We refer to this newly developed method as "micro-collection of male meiocytes". The preparation had little contamination, with an estimated purity of over 95%. Total RNA was extracted from isolated male meiocytes using Trizol reagent from Invitrogen. Then standard small RNA purification procedure was followed to prepare sequencing sample using SOLiD small RNA expression kit (part number 4397682), which includes adding adaptors, reverse transcription, size selection on the gel and gel purification.

3.3.2 Read mapping and expression profiling known miRNAs

Adaptor sequence was trimmed from SOLiD reads by searching for the first 6 bases of the 3' adaptor in each read and trimmed reads with exactly the same sequence were merged using custom Perl script. Then trimmed reads of 17-26nt were mapped onto TAIR9 genome assembly of Arabidopsis by SHRiMP version 2.1.1b (-n 1 -U -o 100000 -v 50% -h 50%). As the mapping is

done by local alignment, the aligned read was extended in the form of global-local alignment, which requires the whole read to be aligned, by self-developed Perl scripts. Then the mapping position of each read was compared to the genomic location of known miRNAs. Reads in known miRNAs (+/- 2bp of either end to allow for imprecise cutting from pre-miRNA) were tallied to reflect the expression level. All Perl scripts are available upon request.

3.3.3 Identifying novel miRNA

Two 150bp genomic windows beside each mapped read were taken based on the mapping position and genomic sequence. Then the whole region, including the mapped read and two 150bp windows, was used as input for miRcheck. Only regions with valid hairpin structure were kept for further analysis. After removing genomic regions spanning known miRNAs, reads close to each other, <=5bp, were clustered. Candidate miRNAs were selected based on 25% rule, which requires at least 25% of total reads are from the exactly the same position on the same strand. Conservation among different species was checked in VISTA genome browser (http://pipeline.lbl.gov/cgi-bin/gateway2).

3.3.4 Longer read analysis

Reads without adaptors or containing small RNA longer than 26nt were mapped onto the genome using SHRiMP. Then mapped reads were clustered based on proximity and read number (distance between reads <= 5bp; each cluster has at least one base with coverage equal to or greater than 5). Then these clusters were compared to the genomic location of genes.

For reads from larger bands in gel, as most reads are from known genes in the initial analysis, all reads in exons were removed after mapping. Then the remaining reads were used for clustering.

3.4 Results and Discussion

3.4.1 Isolation of arabidopsis male meiocytes and read mapping

Arabidopsis male meiocytes develop within the anther, a tiny floral organ at the top of stamen and containing several somatic tissues in addition to germline cells. Since anther is at millimeter scale and meiocytes just constitute about 1% of all anther cells, the isolation of meiocytes from surrounding tissues poses a great challenge. Our lab developed a new extraction method named "micro-collection of male meiocytes", in which meiocytes were released from dissected anther and absorbed into a glass pipette under an inverted microscope (138). Collected meiocytes were then used for small RNA preparation and SOLiD sequencing.

Since meiocytes are difficult to retrieve, three sequencing datasets were obtained from the same small RNA library (Table 3-1). Sequencing reads are 36nt long, including both small RNA and adaptor sequences. As most Arabidopsis small RNAs range from 20 to 25nt, the adaptor sequences are expected to make up the last 11-16nt of each read. To remove adaptor sequences of different length, the exact match of the first six colors, comparable to six nucleotides due to the unique encoding scheme in SOLiD output, of the 3' adaptor was required to locate the boundary between small RNA and adaptor, for six bases balanced between capturing real adaptors and avoiding artifact due to random composition (Appendix Figure 3-1).

After the removal of adaptor sequences, trimmed reads of the exactly same composition were merged and then categorized based on their length. As the shortest miRNA known is 17nt, to maximize the possibility of identifying novel miRNAs trimmed reads of 17~26nt were used for miRNA and siRNA analysis, while longer reads were kept to detect lncRNAs (140). All these reads were then mapped to the TAIR9 assembly of Arabidopsis genome using SHRiMP (141).

Table 3-1. Read count in each step of analysis

	Replicate 1	Replicate 2	Replicate 3
Original dataset*	2,059,845	1,796,270	469,902
Without adaptor	1,295,489	1,307,295	385,518
With adaptor	764,356	488,975	84,384
After unification	396,079	264,905	54,078
Length >26	11,689	4,820	1,716
Length <17	165,836	122,828	18,894
Length 17~26	218,554	137,257	33,468
Uniquely mapped	95,364	57,998	12,976
Multiply mapped	51,553	31,931	8,815

* The difference of read number among replicates is due to the fluctuation of sequencer output.

3.4.2 The expression level of known miRNAs and their target genes

We defined the expression level of known miRNAs as the number of mapped reads in their genomic loci. Since different loci can produce the same miRNA, multiply mapped reads were also taken into consideration when calculating expression level. Each uniquely mapped read was counted once and each multiply mapped read contributed an equal fraction to every mapped locus. In Arabidopsis, there are 266 mature miRNAs documented in miRBase (142). 97 of them have mapped reads in at least two replicates and 75 of them show expression in all three replicates. Interestingly, a single miRNA, miR158a, accounted for about one third of total expression in each replicate (Appendix Table 3-1). The top five miRNAs, miR158a, miR167a, miR167b, miR172a, miR172b, took up more than half of the total expression. None of these miRNAs have been reported to be involved in meiosis. The function of miR158a is still unknown. miR167a and miR167 regulate auxin response that coordinates many aspects of plant

development. miR172a and miR172b target *AP2* (*APETALA2*) domain containing genes, including *AP2* which is an essential transcription factor in determining floral organ identity and serves as an A function gene in the ABC model of floral development. The high expression of miR172a and miR172b in male meiocytes suggest that AP2 and its close homologs are repressed during meiosis, consistent with the fact that A function genes are suppressed in stamen.

Unlike miRNAs in animals that typically have an extensive suite, up to hundreds, of target genes, plant miRNAs only target a few genes. ASRP (http://asrp.cgrb.oregonstate.edu/) (135) and MPSS plus (http://mpss.udel.edu/at/mpss_index.php) databases list experimentally confirmed and predicted targets of known Arabidopsis miRNAs. Using mRNA-seq data from the study by Yang et al. (101) on gene expression in male meiocytes, the expression level of target genes was matched to related miRNAs (Figure 3-1). Generally, highly expressed miRNAs correspond to low expression of target genes and vice versa. The expression level of four miR158a target genes is very low. Each has about 10 mapped reads, in addition to one target

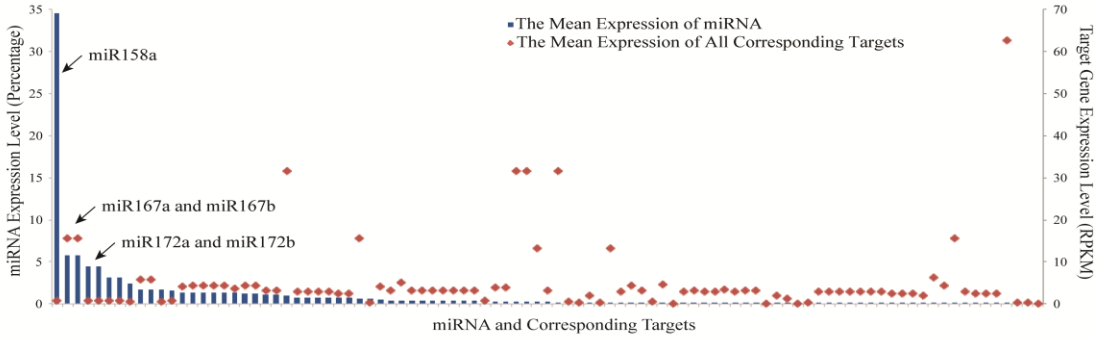


Figure 3-1. The mean expression level of miRNA and corresponding targets. miRNA expression level is shown in percentage of the total expression. Target gene expression is shown in RPKM (Reads per Kilobase per Million Reads).

gene, *FUCOSYLTRANSFERASE 2* (*FUT2*), with no read at all. Similarly, targets genes of miR172a and miR172b have no or very low expression. Only *TOE2* (*Target of Early Activation*

Tagged 2) and *AP2* has 10~20 or so reads, while others, including *TOE1*, *SCHNARCHZAPFEN* (*SNZ*) and *SMZ*, have no read mapped. Two target genes of miR167a and miR167b, *ARF6* (*Auxin Response Factor*) and *ARF8*, have relatively high expression, with about 50 reads.

Recent study by Todesco et al. constructed a series of target mimics, in which genes containing pseudo-targets of known miRNAs were transferred into wild type Arabidopsis, to systematically study the function of miRNAs (143). The transgenic line expressing target mimics for miR158, MIM158, does not have observable defects in morphology or during development, but phenotyping is limited to major organs, like leaf and flower, and the phenotype of meiocytes has not been investigated. MIM167 plants show delayed flowering and twisted leafs. Especially, anther in MIM167 does not mature completely and produces less pollen, suggesting there might be defects during meiosis. MIM172 plants also show late flowering, due to the negative regulation of *SCHLAFMÜTZE* (*SMZ*), a potent repressor of flowering, by miR172 (144). Flowers in MIM172 are normal, but meiocytes were not examined, either.

3.4.3 One putative novel miRNA

As sequencing data include reads from various RNA species, to distinguish possible novel miRNA from siRNA and other types of small RNAs, one definitive rule agreed upon in the community is more than 25% of reads derived from the same location in a hairpin structure (145). Following this standard and other criteria (See 4.2 Materials and Methods), one 17bp region in Chr3 is selected to be putative novel miRNA, despite of its short length. Interestingly, this region shows conservation between *Arabidopsis thaliana* and *Arabidopsis lyrata*, rice or maize (Figure 3-2). Further experimental examination is needed to conclude this RNA as novel miRNA.

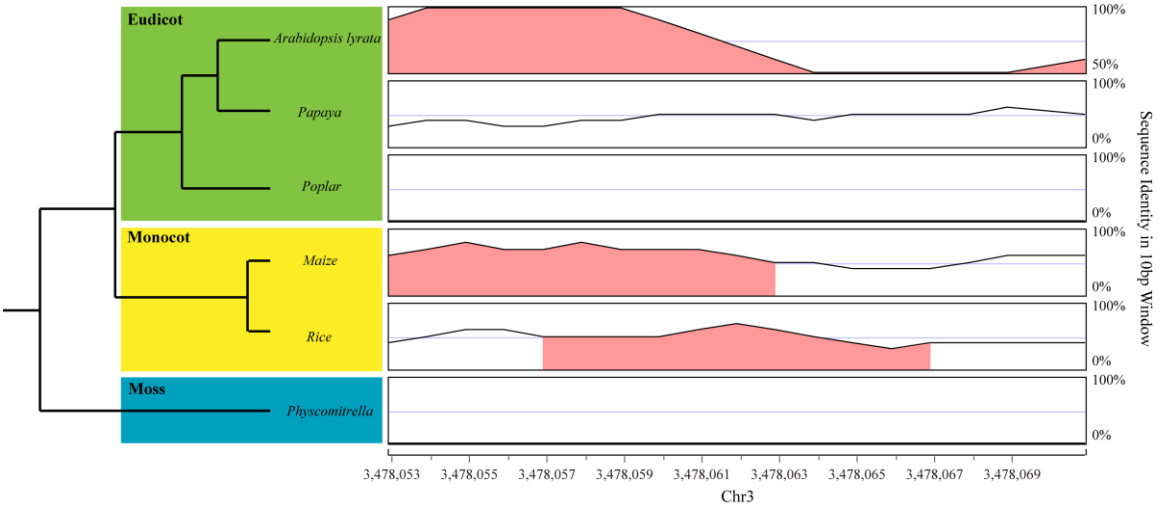


Figure 3-2. The sequence conservation between *Arabidopsis thaliana* and *Arabidopsis lyrata*, Poplar, Maize, Rice and Physcomitrella. A phylogenetic tree of these organisms is shown on the left. The region selected to be putative miRNA is shown on the right. Black curve stands for sequence identity in each 10bp window. Conserved region, at least 70% identity in 10 or more base pairs, is shaded in pink.

3.4.4 Longer ncRNAs

The majority of reads in the datasets do not contain adaptor sequence toward the end (Table 3-1), some of which may come from longer ncRNAs than miRNAs and siRNAs. In addition, even for reads with adaptor, a large portion of them has small RNA part longer than 26bp (Table 3-1). To investigate the existence of long ncRNAs in meiocytes, these reads were mapped to the Arabidopsis genome and clustered based on mapping positions. One possible source of these reads is the degradation product of mRNA. To explore this possibility, the genome-wide distribution of read clusters was compared to Arabidopsis gene density in individual genomic regions (Figure 3-3). However, the location of mapped reads shows little correlation with genes, suggesting degradation products only constitute a minor part. Interestingly, all identified read clusters, 55 in total, are located in intergenic regions.

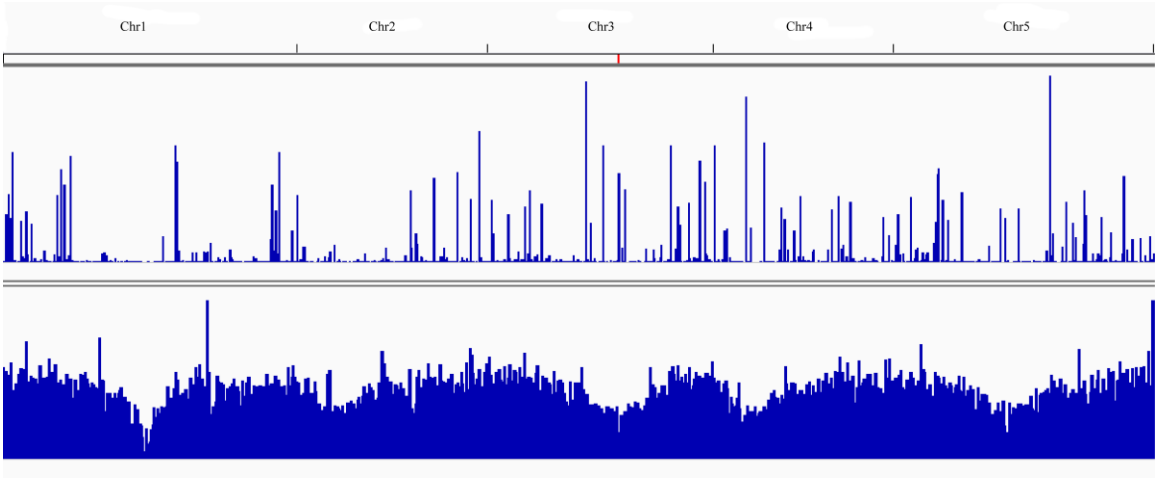


Figure 3-3. The distribution of mapped reads (top) and genes (bottom) across the Arabidopsis genome.

When performing small RNA sample preparation, only bands near 100bp in the gel were collected and purified, based on the length of typical small RNAs and sequencing adaptors used. Long ncRNA may be contained in larger bands. To further study these long ncRNAs, another sequencing dataset was obtained by sequencing the RNA population retrieved from larger bands in the gel. After read mapping and clustering, most of read clusters are located in exons, suggesting these reads might be from mRNA degradation. However, there are still 1137 clusters in intergenic regions and 205 in introns. Moreover, in addition to clusters in protein coding genes, some clusters correspond to snRNAs and snoRNAs, the function of which in meiosis have not been well characterized.

In summary, the analysis of reads possibly from longer ncRNAs provides an extensive list of potential candidates, awaiting further verification by experiments.

CHAPTER 4

TRANSCRIPTOME OF EMBRYONIC AND NEONATAL MOUSE CORTEX BY HIGH-THROUGHPUT RNA SEQUENCING

The work described in this chapter has been published in Han et al, *Proc. Natl. Acad. Sci. USA*, 2009 Aug 4,106(31): 12741-6.

4.1 Summary

Brain structure and function experience dramatic changes from embryonic to postnatal development. Microarray analyses have detected differential gene expression at different stages and in disease models, but gene expression information during early brain development is limited. We have generated over 27 million reads to identify mRNAs from the mouse cortex for over 16,000 genes at either embryonic day 18 (E18) or postnatal day 7 (P7), a period of significant synaptogenesis for neural circuit formation. In addition, we devised novel strategies to detect alternative splice forms and uncovered new splice variants. We observed differential expression of 3,758 genes between the two stages, many with known functions or predicted to be important for neural development. Neurogenesis-related genes, such as those encoding Sox4 and Sox11 and zinc-finger proteins, were more highly expressed at E18 than at P7. In contrast, the genes encoding synaptic proteins such as synaptotagmin, complexin 2, and syntaxin were upregulated from E18 to P7. We also found that several neurological disorder-related genes were highly expressed at E18. Our transcriptome analysis may serve as a blueprint for gene expression pattern and provide functional clues of previously unknown genes and disease-related genes during early brain development.

4.2 Introduction

Mammalian brain development can be largely divided into two periods: embryonic and postnatal. Embryonic mouse brain development starts around 10-11 days after gestation (E10-E11) with massive neuronal production from neural stem cells. The development of rodent cerebral cortex is a well-studied model system, where the initial neurons form the subplate layer, while subsequent neurons migrate in an inside-out pattern to form the multilayer cortical structure (146, 147). By embryonic day 18 (E18), neurons start to send out axons and dendrites to be poised for synaptic connections. After birth, the first week of postnatal brain development is characterized by elevated production of astrocytes, which are crucial for neuronal synaptogenesis (148, 149). By postnatal day 7, many neurons start to establish synaptic connections with other neurons, forming a primitive neural circuit.

Early brain development is precisely controlled by transcription factors, cell adhesion molecules, receptors and channels, synaptic proteins, and other effectors. A single misstep might result in a severe deformation of the brain circuit. For example, loss of Otx2 function results in the absence of early brain development (150). Since many psychiatric disorders such as autism and mental retardation are closely associated with early brain development, understanding the gene expression profile will facilitate the search for an optimal treatment for these disorders. Previous surveys of early brain development have focused on a small number of genes. More recently, microarray studies and others have revealed differential expression of groups of genes in specific brain regions or associated with brain disorders (151-159). Specifically, gene expression has been examined using whole brains or brain tissues from young and old, as well as diseased mice (152, 153, 157, 160). In one microarray study of 11,000 genes and ESTs, the expression level of 1,926 genes was found to change significantly during hippocampal development from E16 through P30 (156). Also, 366 ProbeSets were found to be differentially expressed during

postnatal 2-10 weeks (158). However, microarray technology has several limitations: (1) the number of genes is fixed; (2) the sensitivity is limited by background hybridization; and (3) the inaccuracy of mRNA levels due to difference in hybridization among probes.

Recent technologies have allowed massive amounts of sequencing at relatively low costs (161-163) and analysis of gene expression by sequencing is highly reproducible and more sensitive than microarrays (21). In particular, the Solexa/Illumina sequencing technology has advantages in high coverage and relatively low cost (161-163). It has been used to interrogate transcriptomes of yeast, mouse, and human tissues (20, 164-166). However, previous studies have not analyzed the change of transcriptome during early brain development. As the change of expression level could provide clues to gene function, large-scale sequencing to detect gene expression has great potential in finding important genes for development. We have conducted Solexa/Illumina sequencing of cDNAs from E18 and P7 mouse brain cortices, and report here the detection of over 16,000 genes, with 3,758 being differentially expressed between these two stages. In addition, we report on the discovery of novel putative splices. Our results pave the way for further functional analysis of a large number of genes in the early developing brain.

4.3 Materials and Methods

4.3.1 Mouse brain dissection, RNA extraction, cDNA synthesis and sequencing

The animal protocol was approved by the Penn State University IACUC committee. A total of 4 pregnant female mice (C57BL/6J, 10-12 weeks old), purchased at 2 different time points as biological replicates, were from Charles River Laboratories (Wilmington, MA) at 14-days pregnancy. The mice were fed with Purina laboratory rodent diet 5001. For each replicate of two female mice, one was sacrificed with CO_2 at 18-days pregnancy to collect the E18 embryos, and the other was allowed to give birth to collect pups at P7. 6 to 8 of the E18 embryos and 5 to 6 P7 pups were decapitated, and cortical hemispheres were collected by removing the brain stem, cerebellum, and midbrain. The cortices of all embryos (or all pups) from each litter were dissected, immediately frozen in liquid nitrogen, and stored at -80oC before RNA extraction. Total RNA was isolated from ~260 and ~750 mg of E18 and P7 cortical tissues, respectively, using an RNA isolation kit from Ambion Inc. (Austin, TX) according to manufacturer's protocol, yielding 1.2 and 5.4 mg of RNA, respectively. The RNA samples were treated with DNase I (Invitrogen, Carlsbad, CA), then sent to Fasteris SA (Plan-les-Ouates, Switzerland) for mRNA purification and cDNA library construction for sequencing using the Illumina/Solexa technology.

4.3.2 Sequencing quality, reads mapping and sequence analyses

The quality of sequencing result was summarized and plotted using Galaxy (http://galaxy.psu.edu). The mouse genome sequence of July 2007 assembly was downloaded from UCSC genome informatics portal (http://genome.ucsc.edu/). RMAP (http://rulai.cshl.edu/rmap/) reports the uniquely mappable and ambiguous reads into two separate files. To simplify subsequent analysis, the chromosome sequences were concatenated into one

pseudo-sequence by self-developed scripts (these and others are available upon request) and then single-end and paired-end reads were mapped onto the concatenated sequence. We only used the uniquely mapped reads in the subsequent analysis. For paired-end reads, we required both ends to be uniquely mapped.

The exon coordinates were downloaded from the UCSC table browser and analyzed using scripts developed here. The number of reads that matched each exon and UCSC transcript cluster was calculated. Fisher's exact test was applied with R (http://www.r-project.org/) and "sagenhaft" library (http://tagcalling.mbgproject.org) in Bioconductor to identify statistically significant differentially expressed clusters between P7 and E18. Also, for sufficiently highly expressed genes, read counts were converted to percent of total mapped reads, and a supplementary analysis of the log2 (percentages) was done as a randomized complete block design using the "LIMMA" library (41) in Bioconductor. The reads mapped to transcript- specific sequence were selected (see Supplemental Methods for the selection criteria). Using read mapping information, transcripts were compared for splicing events between P7 and E18. In single-end analysis, 20 bps on either side of all possible junctions between known exons were connected into one pseudo sequence. In paired-end analysis, the 20 bps flanking possible junctions and the full-length exonic sequences were connected into one pseudo sequence (see supplemental method). We then mapped our reads to this concatenated junction sequence using RMAP, separately from the genomic mapping. Reads that were already mapped in the genomic analyses were excluded (167).

4.4 Results and Discussion

4.4.1 Isolation of RNAs from dissected mouse brain cortex tissues and library construction

We compared transcriptomes of two important stages of developing mouse brain, E18 and P7, characterized by the production of the majority of neuronal cells and significant synaptic contacts between neurons, respectively. For each of two biological replicates, cortices were dissected from 5-8 E18 embryos and P7 pups, and used to isolate total RNAs and mRNAs, which were used to generate cDNA libraries without vectors. A small portion of each library was cloned into a plasmid vector and several clones were sequenced to determine the quality of the cDNAs. The cDNA libraries were then used for high-throughput sequencing.

4.4.2 High throughput sequencing and mapping of the reads

High-throughput sequencing of the cDNA libraries was carried out using the Solexa/Illumina technology. As this technology was still in its early stages of application, the initial sequencing was done using single-end cDNA libraries from one biological replicate of mouse brain at the E18 and P7 stages. After analysis of the single-end cDNAs, cDNAs from the same two mRNA samples were used to construct libraries for paired-end sequencing. From these experiments, the single-end sequencing produced 2,956,444 reads from the E18 and 3,619,970 from P7 stages, respectively. The paired-end round generated 4,536,964 reads from the E18 library and 4,019,273 reads from the P7 library. A second biological replicate of mouse brain cortices at the E18 and P7 stages was subjected to paired-end sequencing, resulting in 5,501,311 reads at E18 and 6,042,658 reads at P7 (Table 4-1). For both single and paired-end sequences we examined the quality of the base calls using Galaxy (http://galaxy.psu.edu) (168, 169). The distribution of quality score [Phred-equivalent metric (170, 171)] at each base of the read

(Appendix Figure 4-1) indicated that bases before 28 generally had quality scores of 20 or higher; therefore, only the first 27 high-quality bps were used in the subsequent analysis to maximize the sequence reliability, while retaining sufficient length for analyses.

To identify genes expression in the mouse brain cortex, we mapped the reads against the July 2007 assembly of the mouse genome using the RMAP (64, 71, 171) tool specifically designed for Illumina (Solexa) data. As shown in Table 4-1, 60.9% of single-end reads were uniquely mapped and there were 51.8% of paired-end reads that were mapped uniquely at both ends.

We compared the genomic coordinates of reads against the gene locations of UCSC (University of California at Santa Cruz) Known Gene collection (172, 173) (Referred to as "Known Genes" in the remainder of this study). Because the UCSC Genes dataset is highly redundant due to multiple splice variants, we aggregated the reads for each gene cluster – a collection of transcripts representing a single gene. Interestingly, besides the large number of reads that mapped to annotated exons, additional large number of reads mapped to

Table 4-1. Summary of read number

	Biological Replicate 1				**Biological Replicate 2**	
Read Category	**Single-end**		**Paired-end***		**Paired-end**	
	E18	**P7**	**E18**	**P7**	**E18**	**P7**
Total Reads	2,956,444	3,619,970	4,536,964	4,019,273	5,501,311	6,402,658
Uniquely Mapped	1,886,668	2,117,328	2,329,560	2,238,480	2,784,748	3,220,724
In Exons	1,099,869	1,404,798	1,376,421	1,577,198	1,948,700	2,448,499
In Novel Transcripts	2,302	2,689	621	746	794	1,084

* For paired-end data, one pair of sequencing results was counted as one read. Only if both ends were uniquely mapped, we count this pair as the uniquely mapped read.

introns and intergenic regions (Table 4-1), consistent with the previous discovery of pervasive transcription of the human genome (174). Finally the number of reads matching

each gene was calculated for all the Known Genes. An example of how reads were related to exons and genes is shown in Figure 4-1. The per- gene distribution of reads was highly consistent between the two paired-end data, with Pearson's correlation coefficient being 0.97 and 0.96 for E18 and P7, respectively (Figure 4-2 A and B). The single-end data and paired-end data were also highly consistent in which Pearson's correlation coefficient was 0.96 and 0.95 for E18 and P7 stages, indicating that both the single-end and paired-end sequencing yielded similar results (Appendix Figure 4-2 A and B). This consistency was in agreement with the high reproducibility reported recently (20, 21, 164-166) and permitted the detection of differential expression and analysis of novel transcripts. Interestingly, the distributions of the number of genes according to read counts were very similar between the two replicates for each stage, and even between the two stages (Figure 4-2C).

To minimize false positives for expressed genes, we required at least 2 uniquely mapped reads as detectable expression of a given gene. The single-end and paired-end sequences yielded similar numbers of detected genes (Appendix Figure 4-2 C and D). Our single-end reads revealed the expression for 13,463 and 14,243 genes at E18 and P7, respectively. From the first paired-end analysis, we identified reads for 12,590 genes from the E18 cortex and 12,991 genes from P7. From the second paired-end dataset, 13,642 and 14,223 genes showed expression at E18 and P7, respectively. Altogether, we detected the expression of 14,787 genes at E18 and 15,423 genes at P7, with a total of 16,083 genes, representing 58.7% of mouse Known Genes (Figure 4-3A). In addition to the large overlap between the number of expressed genes in E18 and P7, there were also a substantial number of genes that are preferentially expressed at a particular stage (660 at E18; 1,296 at P7). If we regarded the top 5% genes among total clusters of UCSC Known Genes (27,389) as highly expressed, we found that besides a large number of genes (974) expressed at both stages, an appreciable number of highly expressed genes were also uniquely expressed at each stage (396 at E18; 399 at P7) (Figure 4-3B, Appendix Figure 4-2E).

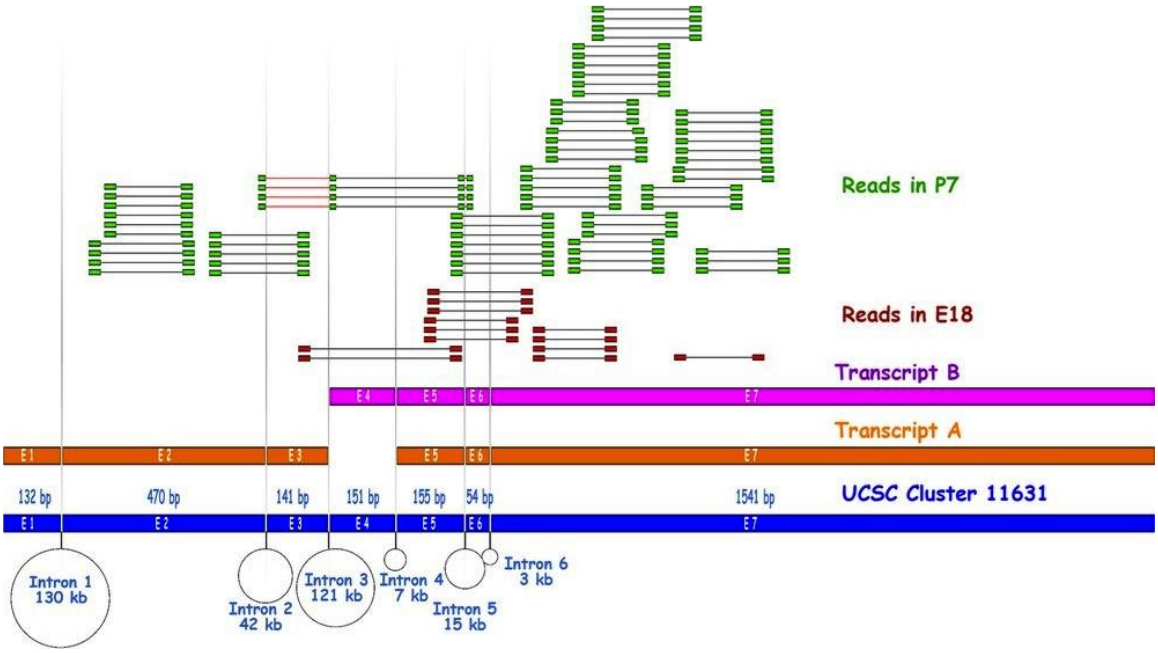


Figure 4-1. An example of mapping reads to a gene. The UCSC gene/cluster 11631 has two transcripts, A and B. Transcript A has exon 1, 2, 3, 5, 6 and 7. Transcript B has exon 4,5,6 and 7. 13 paired-end reads (red) mapped to this gene at E18 and 65 paired-end reads (green) mapped to this gene at P7. At P7 stage, some reads mapped to transcript-specific exon 1 and exon 2, providing evidence for the expression of transcript A. Four reads mapped to a novel junction (the red line) between exon 2 and 4, indicating a possible transcript at P7 stage. As the cDNA synthesis was poly-T primed, there were more reads in 3' region than 5'.

4.4.3 Strong evidence for differential gene expression

The large numbers of sequencing reads represent a deep sampling of the transcriptome and can be an excellent digital measure of the relative abundance of transcripts (164). To obtain statistical support for the differences between the two stages, we applied Fisher's exact test on the read count of each gene at E18 and P7. We performed the test in two different ways to obtain relatively conservative estimates. To balance false positive and false negative results, we summed reads from our two paired-end biological replicates and then performed the test. 7,688 genes were found to be differentially expressed between E18 and P7 (at 0.01 significance level). In a more

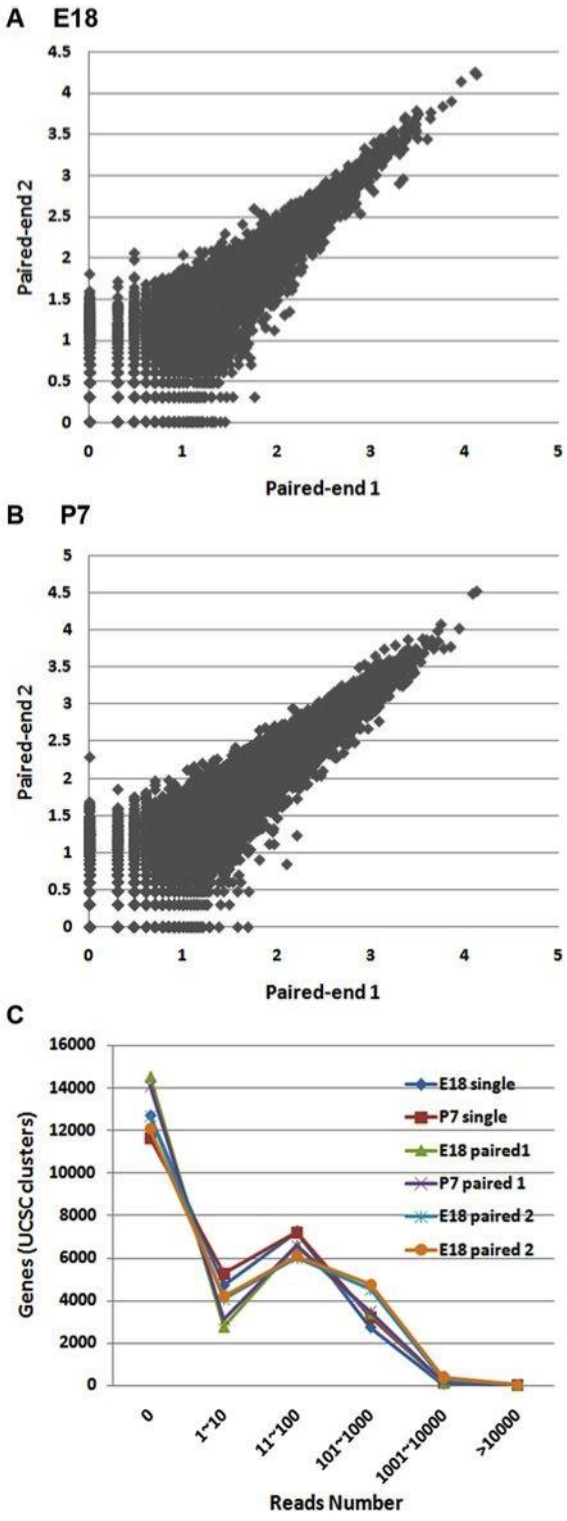


Figure 4-2. A comparison between two biological replicates and among all datasets. (A) The comparison of reads per gene between the first and second paired-end data in E18. Since the read number per gene ranges from 0 to over 10,000, the read numbers adding 1were transformed by log10. There is a good correlation between the first and second paired (R = 0.97). (B) The comparison of reads per gene between the first and second paired-end data in P7 (R = 0.96). (C) The line chart showing the distribution of genes with different reads. The y-axis is the number of genes. The x-axis is different intervals of read number. The number of genes at E18 and P7 showed parallel changes in both single-end and paired-end analysis.

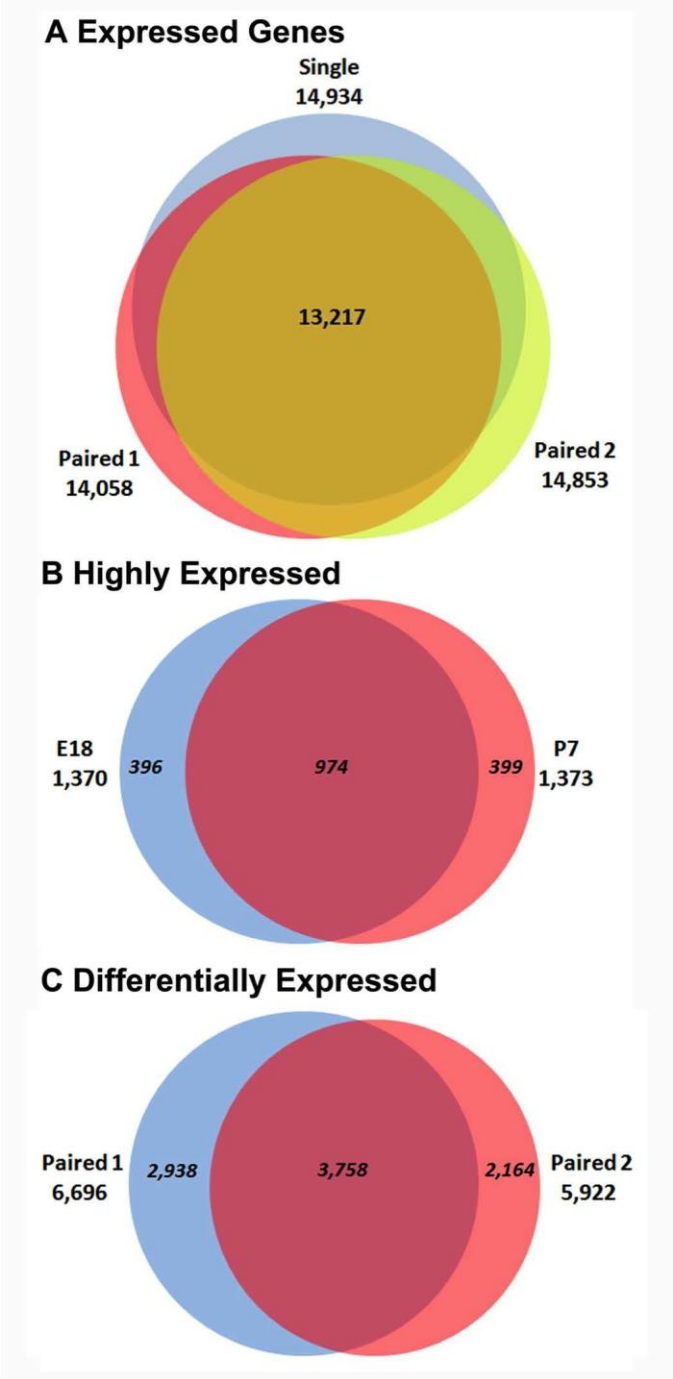


Figure 4-3. Venn diagrams showing the number of expressed genes. (A) The number of expressed genes in three analyses, with a total of 16,083. (B) The number of highly expressed genes. The top 5% genes were regarded as highly expressed ones. The two paired-end data were summed. (C) The number of differentially expressed genes between E18 and P7. There were a substantial number of differentially expressed genes (3,758) supported by both paired-end analyses.

stringent analysis, we applied the test on each paired-end dataset separately and then found the intersection between the two replicates that was significant and concordant in the direction of differential expression in both replicates. There were still 3,758 differentially expressed genes with significance level of 0.01 or less in both replicates (Figure 4-3C). We also performed an analysis that combined the data from both biological replicates using LIMMA (see Materials and Methods). Although LIMMA is less powerful than Fisher's exact test in detecting differential expression, LIMMA allows a combined analysis of the replicates. Our LIMMA analysis showed that 5811 genes had a q value between 0.10 and 0.15, a moderate support for differential expression. These 5811 genes included 3,337 (88.8%) of the above-mentioned 3,758 differentially expressed genes from the Fisher's exact test with data of both replicates. Therefore, these two types of analyses gave generally consistent results, and represent statistically well-supported findings of differential gene expression. Many of the differentially expressed genes were known to play important roles in neuronal network formation. Strikingly, approximately 500 genes showed dramatic changes in expression level but did not have functional annotation in the UCSC Known Genes.

4.4.4 Differentially expressed and unreported splice variants

Many mammalian genes are known to have alternatively spliced transcripts, which are usually not specifically detected in microarray experiments. Specific splice variant(s) can be detected by the reads that map to exon(s) or exon combination(s) that are unique to a transcript. Such exonic sequences are markers allowing the comparison of splice variants between two samples (Figure 4-1). In addition to reads mapped to unique exon-exon junctions, the paired-end reads can identify splicing variants if the two ends map to two exons in a transcript-specific combination. Our single-end reads allowed the detection of alternatively spliced transcript(s) for

4,131 and 4,169 genes at E18 and P7, respectively. Furthermore, alternative spliced transcripts for 2,112 genes were detected in only one of these stages. Similarly, paired-end reads of the first replicate supported alternatively spliced transcript(s) for 4,062 and 4032 genes at E18 and P7, respectively. Again, at least one alternatively spliced transcript for each of 2,557 genes was detected in only one of the two stages. The second paired-end replicate uncovered alternative splicing of 4,354 and 4,496 genes, respectively, at E18 and P7, among which 2,055 genes had alternatively spliced transcript(s) detected in only one of the stages. Among the several thousand genes detected to have alternative splice variants, 320 genes had splicing differences that were supported by all three sequencing datasets. A recent study also employing Solexa sequencing found that 92-94% of human genes had alternative splicing among 15 types of tissues (175). Our results show that there is also substantial splicing variation between two developmental stages of the same tissue.

To detect novel exon-exon junctions, we searched for reads that mapped on novel junctions. In addition, a pair of ends that mapped in two exons could support novel junctions if they did not belong to the same known transcripts (Appendix Figure 4-3). We modified a previous procedure (164) to incorporate both single-end and paired-end reads (see Materials and Methods). Single- and paired-end junction analyses differed in three ways: (1) the target pseudo-sequences were different; (2) paired-end data allowed the detection of novel splicing forms when both ends mapped two exons in a novel combination; (3) paired-end analysis required both ends mapping onto the same transcript. In single-end analysis, we found that 2,302 and 2,689 reads mapped to novel junctions in 1,367 and 1,596 genes at E18 and P7, respectively. In the first paired-end analysis, we found 621 reads as evidence of novel transcripts in 143 genes at E18. At P7, we found 746 reads supporting novel transcripts in 172 genes. From the second paired-end analysis, we identified 794 and 1084 reads for novel transcripts in 448 and 536 genes at E18 and P7, respectively (Table 4-1 and Appendix Table 4-1). In total, we identified novel splicing forms

in 2,930 genes from the single end and two paired-end samples, with 974 genes shown to have novel transcripts supported by two or more datasets at either stages. Among them were genes that play important roles in the brain, such as *FBXO2* (maintaining postmitotic neurons), *App* (amyloid beta (A4) precursor protein, involved in plasticity and Alzheimer's), *Aplp1* (synaptic maturation), and *BPTF* (gene regulation, related to Alzheimer's disease).

4.4.5 The most expressed genes during early brain development

Our analysis showed that genes with highest early brain expression level were responsible for ATP production in mitochondria or encoding microtubule proteins. Among the top 10 most expressed genes at E18 and P7 from single-end analysis (most reads), 7 genes were the same: NADH dehydrogenase 1, cytochrome b, NADH dehydrogenase 4, NADH-ubiquinone oxidoreductase chain 2, cytochrome c oxidase subunit I, tubulin β5, and microtubule-associated protein tau. Paired-end analysis also showed that these genes were among most expressed ones except for minor changes in the ranking. The results are consistent with the need for mitochondrial biogenesis and energy during active cell proliferation and growth in early brain development. Microtubules are crucial both for cell division and differentiation generally and for neuronal axon and dendritic growth specifically. At P7, an unclassified gene AK157178 in mouse (ortholog of human *WASF2/SCAR2/WAVE2*) is also among the top 10 expressed genes, suggesting an important function during early brain development and warrants further studies.

4.4.6 Developmental regulation of genes encoding synaptic proteins and receptors

Consistent with a significant increase in synaptogenesis during neonatal brain development, we found that many synaptic protein genes and receptor genes were substantially up-regulated from E18 to P7. Genes for synaptic proteins were already expressed at modest to high level at E18 and were further increased at P7. The highest among these at P7 encoded synaptophysin, complexin 2, syntaxin 1A, and synucleins (paired-end analysis). The observation that many genes coding for synaptic proteins were almost uniformly up-regulated by 4- to 5-fold from E18 to P7 suggests a potentially common mechanism for transcriptional regulation. Similarly, genes for neurotransmitter receptors were also up-regulated in general, but different subunits showed some distinct developmental changes. Among genes encoding glutamate receptors, which mediate the majority of neuronal excitation, *GluR2* for the AMPA receptor subunit 2 was highly expressed, whereas *GluR1, 3, 4* were modestly expressed at E18. Interestingly, from E18 to P7, *GluR2* was down-regulated but *GluR1* was up-regulated. Among genes for NMDA receptor subunits, *NR1A* and *NR2B* were significantly expressed and up-regulated from E18 to P7, but *NR2A*, *NR2C*, and *NR2D* had very low levels at both stages. *NR2A* expression will likely be up-regulated during later brain development (176). Among metabotropic glutamate receptor genes, *mGluR5* was most highly expressed at E18 and further increased at P7. GABA receptors mediate the major inhibition in the brain. Interestingly, genes for $GABA_A$ receptor α subunits were expressed at low or modest ($\alpha 5$) level in E18 and P7, whereas those for $GABA_B$ receptors were highly expressed at E18 and further increased at P7, suggesting an important role in the early brain. In contrast, most genes encoding receptors for glycine, acetylcholine, dopamine, and serotonin were detected at low levels.

Among voltage-dependent channels, genes for the majority of Na^+, K^+, and Ca^{2+} channels were expressed at low levels in E18 but generally up-regulated by P7. For voltage- gated Ca^{2+}

channels, genes for low-voltage sensitive T-type channel α subunits were expressed more than those for L-, N-, and P/Q-type channels at E18, but the latter ones were significantly up-regulated from E18 to P7. For K^+ channels, *Kcnab2, Kcnc4, Kcnd2, Kcnf1, Kcnh3, Kcnj4,* and *Kcnq2* were the most highly expressed genes. For Na^+ channels, *Scn2α1, Scn8α, Scn1β, Scn2β,* and *Scn3β* were all highly expressed, and up-regulated from E18 to P7.

4.4.7 Expression of cell signaling genes

Among cell signaling molecules, calmodulin 3, adenylate cyclase 1, and calmodulin kinase-like vesicle-associated protein were encoded by the top expressing genes that were highly upregulated from E18 to P7. Other genes encoding protein kinases, such as PKC β1 and γ, CaMKII, and protein tyrosine kinase 2β and 3, together with those for protein phosphatase 1 and protein tyrosine phosphatase N, were also up-regulated by more than 4-fold from E18 to P7.

Wnt, BMP (bone morphogenetic protein), and hedgehog are known to be involved in early brain development (177, 178). Among all Wnts detected, only *Wnt7b* was significantly expressed and *Wnt7a* modestly expressed at E18, and both slightly down-regulated at P7. *Wnt4* was up-regulated from E18 to P7, at modest levels. For the BMP group, only *BMP1* was modestly expressed at similar levels in the E18 and P7 mouse cortices. Consistently, we found that only BMP1a and 1b receptors were modestly expressed. Other BMPs were generally at very low expression level. Similarly, the *sonic hedgehog* gene, which is crucial for early embryogenesis, was substantially reduced to almost undetectable level at P7.

4.4.8 Detecting a large number of genes encoding transcription factors

Transcription factors (TFs) perform important regulatory functions by controlling a variety of cellular processes. In the mouse genome, 1445 genes were identified to encode TFs and 983 were expressed in the brain (179). Our sequence data uncovered expression of at least 1,024 genes encoding TFs at the E18 stage and 1039 genes at the P7 stage, totaling 1,079 genes. Among these, 559 and 504 genes were detected with 50 reads from combined single-end and paired-end datasets at E18 and P7, respectively. Interestingly, many of the most highly expressed TF genes at both E18 and P7 stages are the same, including *Thra, Tcf4, Pbx1, Nr2f1, Sox11, Sox4,* and *Bcl11b*. Among the 349 differentially expressed TF genes supported by both replicates, a large majority showed higher levels at E18 than P7, suggesting that transcriptional regulation is important for active neuronal cell division and differentiation, more so than the later neuronal maturation stage. Many of the genes that showed highest ratios of read numbers at E18 to P7 encoded zinc finger proteins; among 117 differentially expressed zinc-finger genes, 56 were more highly expressed at E18, suggesting that they play crucial roles in embryonic neuronal development. In addition, the expression of *Otx2*, encoding a TF required for initial forebrain development (150), was substantially reduced at P7. A recent report found that *Otx2* is expressed again at relatively high levels later in the visual cortex during the critical period of P28-P30 (180).

4.4.9 Detection of genes for autophagy and apoptosis

Although autophagy has been reported to participate in many biological processes, its exact role in vertebrate development, especially neurodevelopment, is far from fully characterized (181). A recent study found that *Ambra1* regulates autophagy and is critical for neural tube development (182). Our result showed that *Ambra1* has appreciable expression levels in both E18

and P7. Besides *Ambra1*, there are 17 other autophagy-related genes. Nine of these had similar or higher expression levels than *Ambra1*. Especially, *Atg9a* showed significant increase in expression level from E18 to P7. Apoptosis is important for development and is also associated with neurodegenerative diseases (183). Among genes related to apoptosis, the gene for Cugbp2/Napor-1, an apoptosis related RNA-binding protein, had extraordinarily high expression levels at both E18 and P7. In addition, the Bcl family genes *Bcl11a*, *Bcl11b*, and *Bcl7a* were highly expressed. These highly expressed apoptosis-related genes were often down-regulated from E18 to P7, suggesting that programmed cell death is more active at E18 than P7. In addition, many other genes annotated to participate in apoptosis also had substantial expression level, suggesting the important role of apoptosis in embryonic neuronal development.

4.4.10 Up- and down-regulated genes and previously unknown genes

Many transporter genes showed greatly increased expression from E18 to P7. For example, Na^+/K^+-ATPases were most abundantly expressed at E18 already and further increased by 4-5 folds at P7, consistent with their critical functions in maintenance of resting membrane potential and cell volume. We have also identified a group of genes that are substantially down-regulated from E18 to P7. As noted above, two genes encoding the transcription factors Sox4 and Sox11 were among the mostly highly expressed transcription factor genes at both stages. These genes are known to function in cell fate determination and were more highly expressed at E18 than at P7 (6-10 folds down-regulation), supporting their greater roles in early neural differentiation than neuronal maturation. A group of genes with unknown functions were also identified that display substantial up- or down-regulation from E18 to P7.

4.4.11 Genes related to neurological disorders

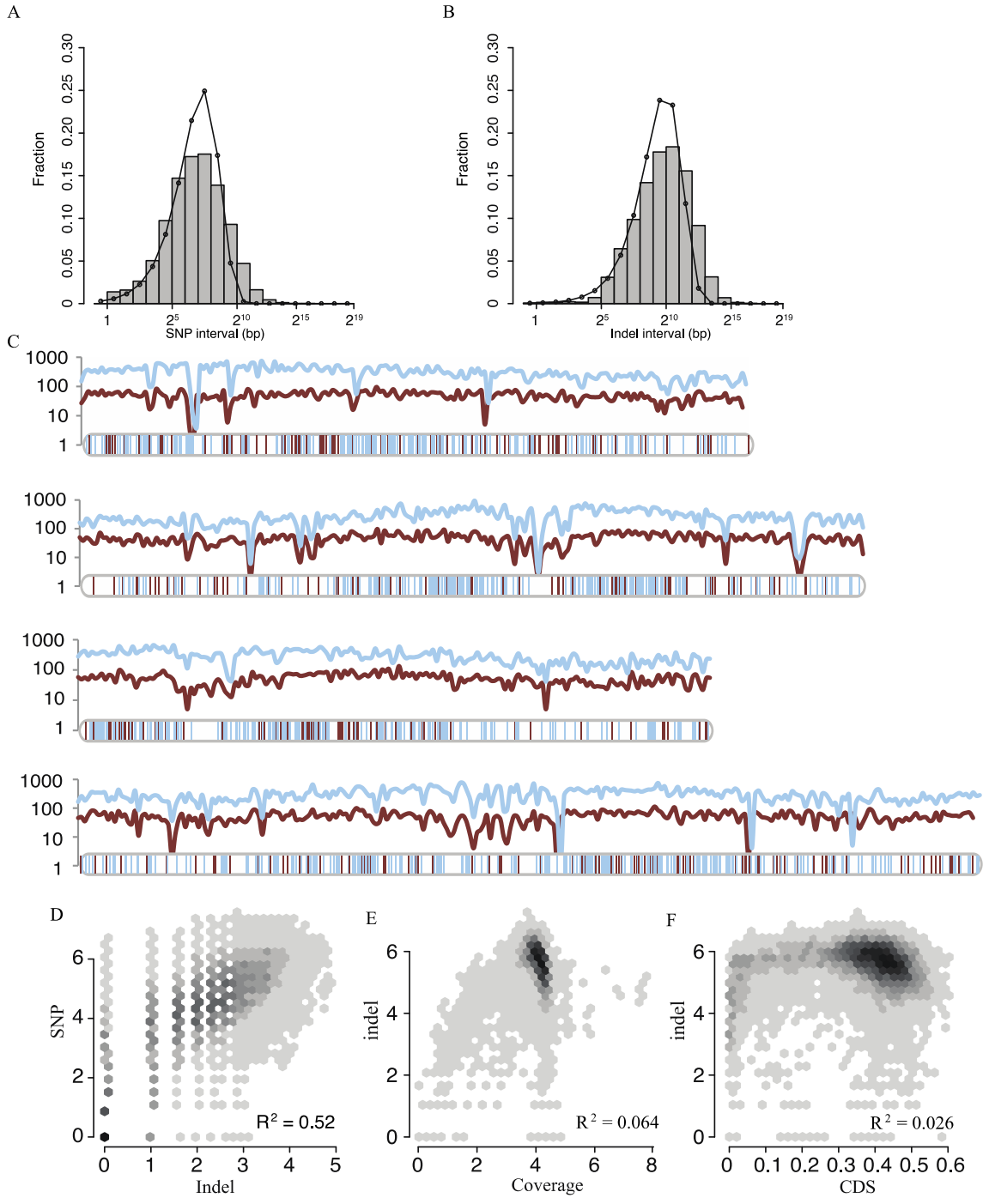
Among the highly expressed genes during early brain development, we further identified genes associated with neurological diseases. The *APP* gene (amyloid beta precursor protein) was already highly expressed at E18, and further up- regulated at P7. Cleavage of APP by β and γ secretases results in amyloid β peptide deposits in senile brains, especially among Alzheimer's patients, but the high expression of APP in E18 and P7 cortex suggests an uncharacterized function in brain development. Some mental retardation (*Atrx*) and seizure-related genes (*Sez6*) were highly expressed from E18 to P7, suggesting a possible link to infant brain defects or infant seizures if these genes do not function properly. Interestingly, autism related gene (*Auts2*) was highly expressed at E18 but substantially down- regulated by P7, raising the possibility that continuous expression of this gene might be associated with autism. Some prion protein genes (*Prnp* and *Prnpip1*) were also highly expressed at E18 and further up-regulated by P7, suggesting a potential function of these genes in neuronal maturation or synapse formation and plasticity.

4.5 Conclusions

Our high-throughput sequence analysis of the transcriptome of the mouse developing cortices detected the expression of over 16,000 genes and uncovered 3,758 genes that were differentially expressed between the E18 and P7 stages. The methods we used and the comparison between single-end and paired-end analyses will provide foundation for further development of bioinformatic tools to analyze next-generation sequencing data. The expression information suggested important functions for a number of regulatory genes, and provided strong evidence for a highly dynamic transcriptome during mouse brain development. Previous microarray work in postnatal cortical development found an increase of *sox11* but decrease of *sox4* expression from postnatal week 2 to week 3 (158). However, our results revealed a substantial decrease of both *sox11* (10 fold) and *sox4* (5 fold) from E18 to P7, suggesting that transcriptional regulation may differ significantly before and after birth. The detection by our deep sequencing data of the expression of a majority of the known transcription factor genes during E18-P7 stages further suggested that the transcriptional regulation plays a prominent role during a critical window of brain circuit formation.

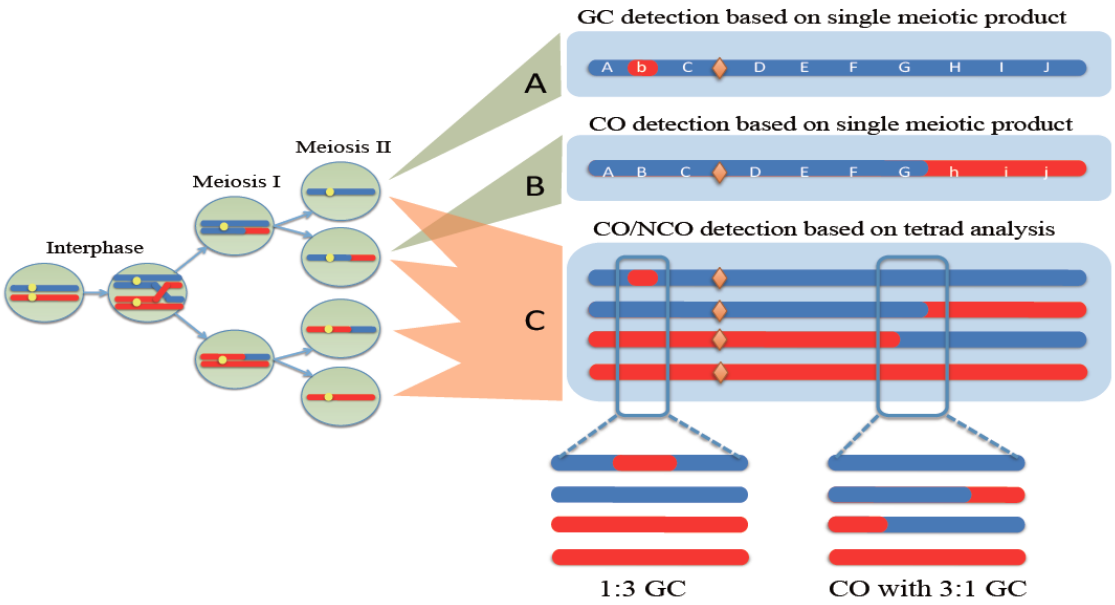
APPENDIX

SUPPLEMENTAL FIGURES AND TABLES

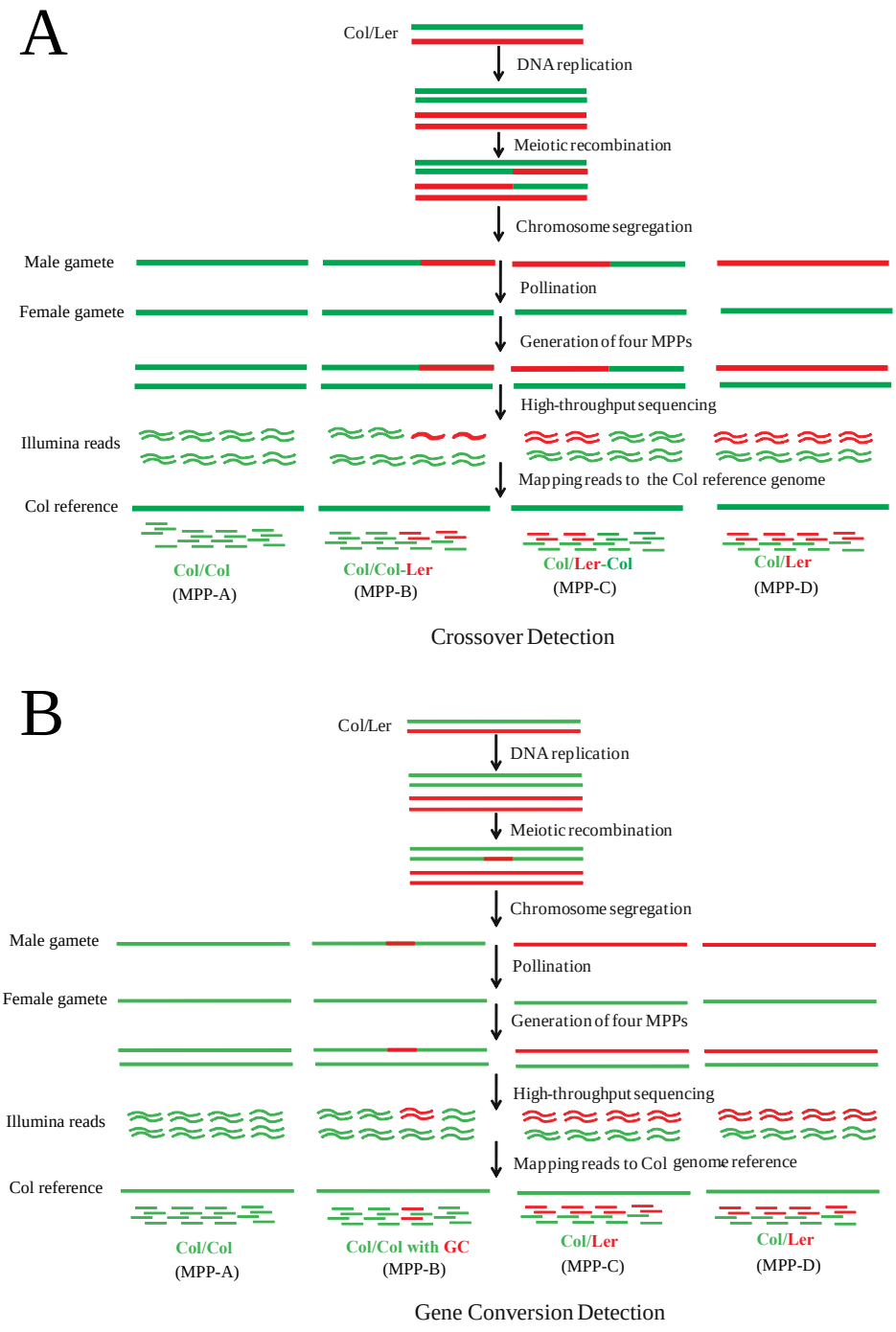


Appendix Figure 2-1. The distribution of intervals between SNPs/indels and correlation of SNP and small indel densities. (A, B) Histogram of interval between SNPs (A) or indels (B). The fraction of intervals in each length range was plotted. The curve shows the fraction from exponential distribution, as the length of intervals between SNPs/indels follows exponential distribution if SNPs/indels are randomly distributed across the genome. The actual distribution of

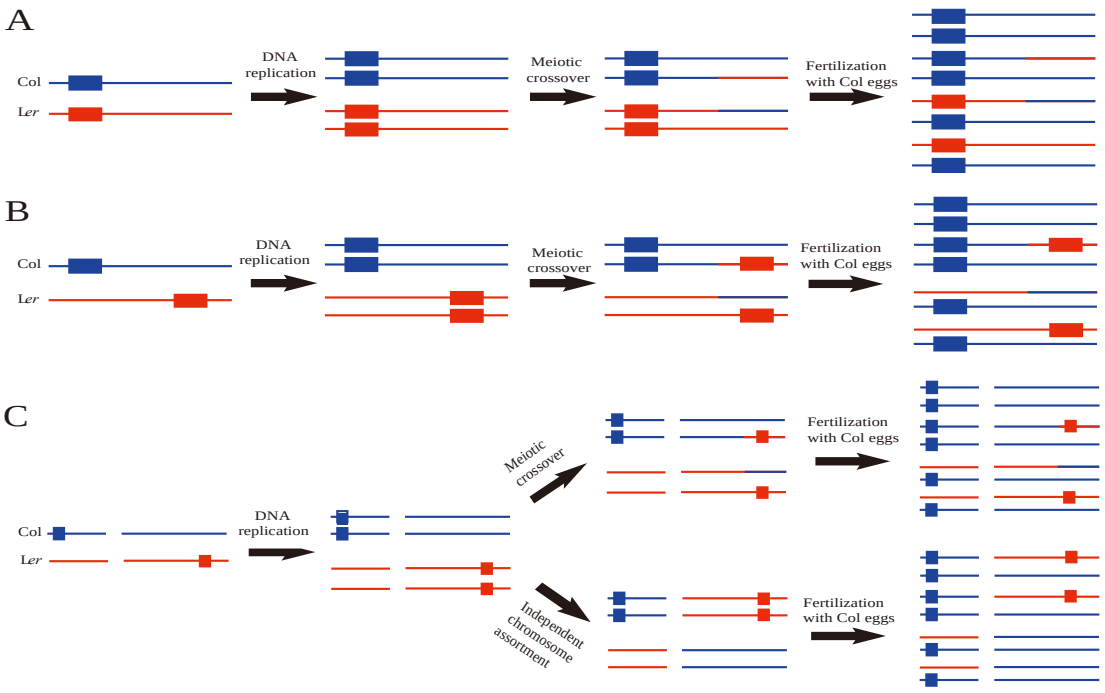
interval length has heavier tail than exponential distribution. For either SNP o indel, the difference between actual distribution and exponential distribution is statistical different (Kolmogorov-Smirnov test, $p < 10e-10$) (C) Parallel change of SNP and small indel density on Chr2-5. The density was defined to be the number of SNPs/indels per 100 Kb. The blue curves represent SNP densities and the red curves for small indels. Blue and red vertical bars below show the locations of large deletions and insertions, respectively. (D) Logistic regression of SNP and indel densities in 10 Kb sliding windows. (E, F) Logistic regression of small indel densities with read coverage (E) or CDS fraction (F) in 100 Kb sliding window. Near zero R^2 value suggests read coverage or CDS fraction does not contribute much to the correlation.



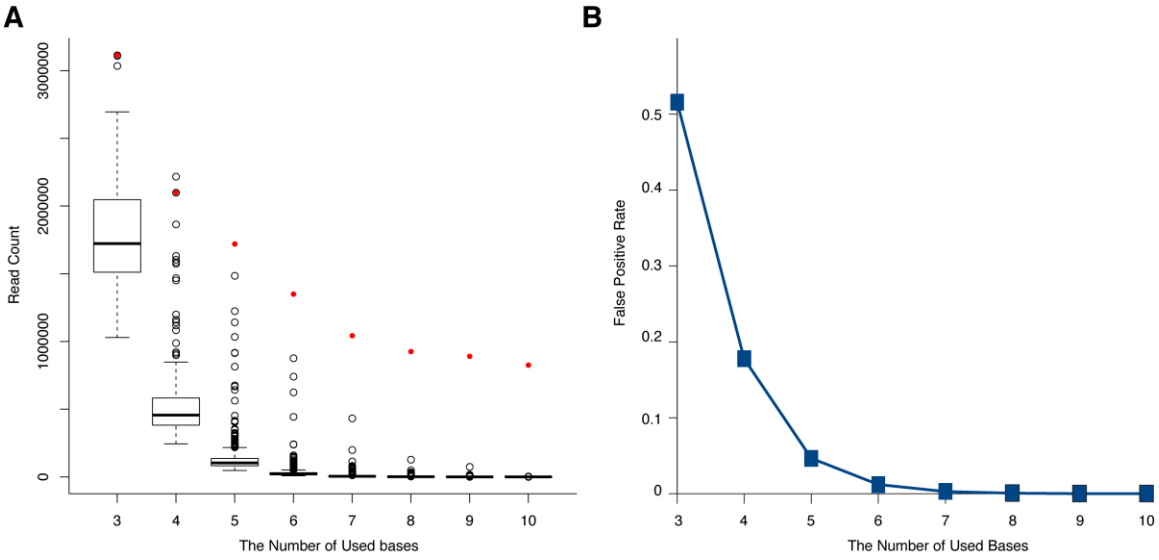
Appendix Figure 2-2. Tetrad analysis for identification of meiotic CO and NCO. The comparison between analysis of single meiotic product and tetrad analysis. (A, B) GC and CO detection based on only single meiotic products. Without all four gametes, the tract length of CO cannot be determined. (C) GC and CO detection using tetrad analysis. With all four gametes included in the tetrad, we can find the length of CO tract and further distinguish GC within CO and GC independent of CO (NCO).



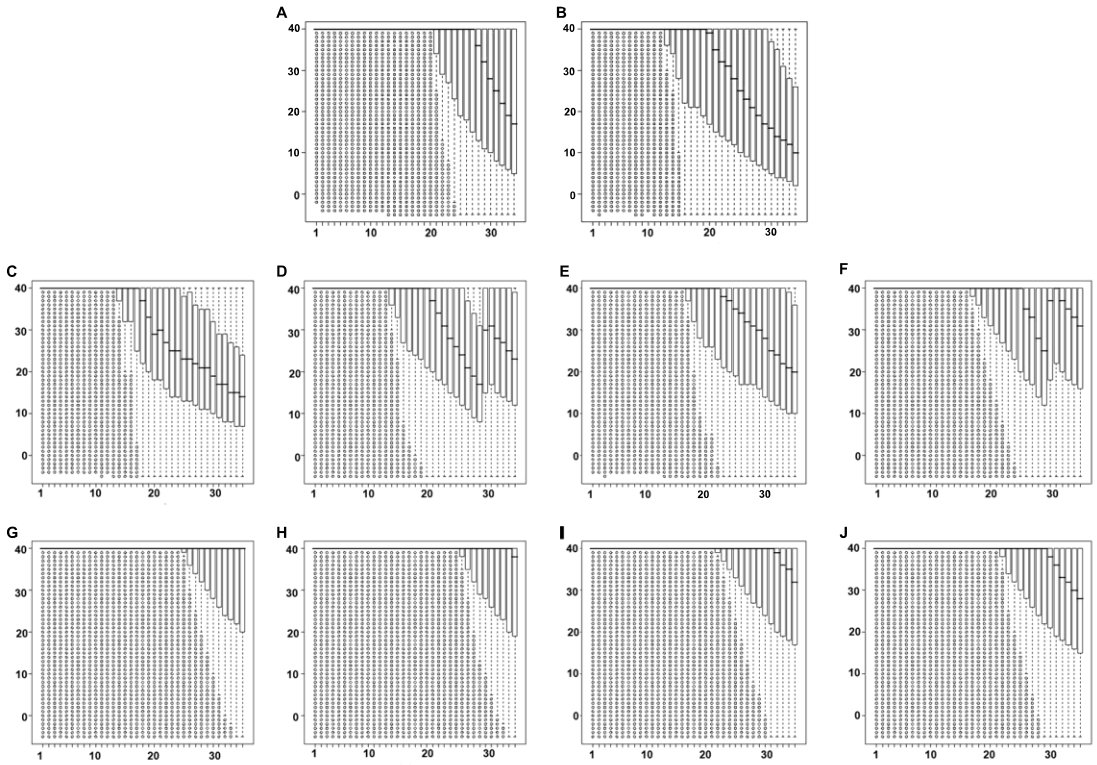
Appendix Figure 2-3. A schematic illustration for detecting meiotic recombination using Illumina sequencing reads. Green bars represent Col genome, red bars show L*er* genome. (A) Meiotic crossover was revealed by the reads from MPP-B and MPP-C plants. (B) Meiotic gene conversion was detected by the L*er* reads in the MPP-B plant.



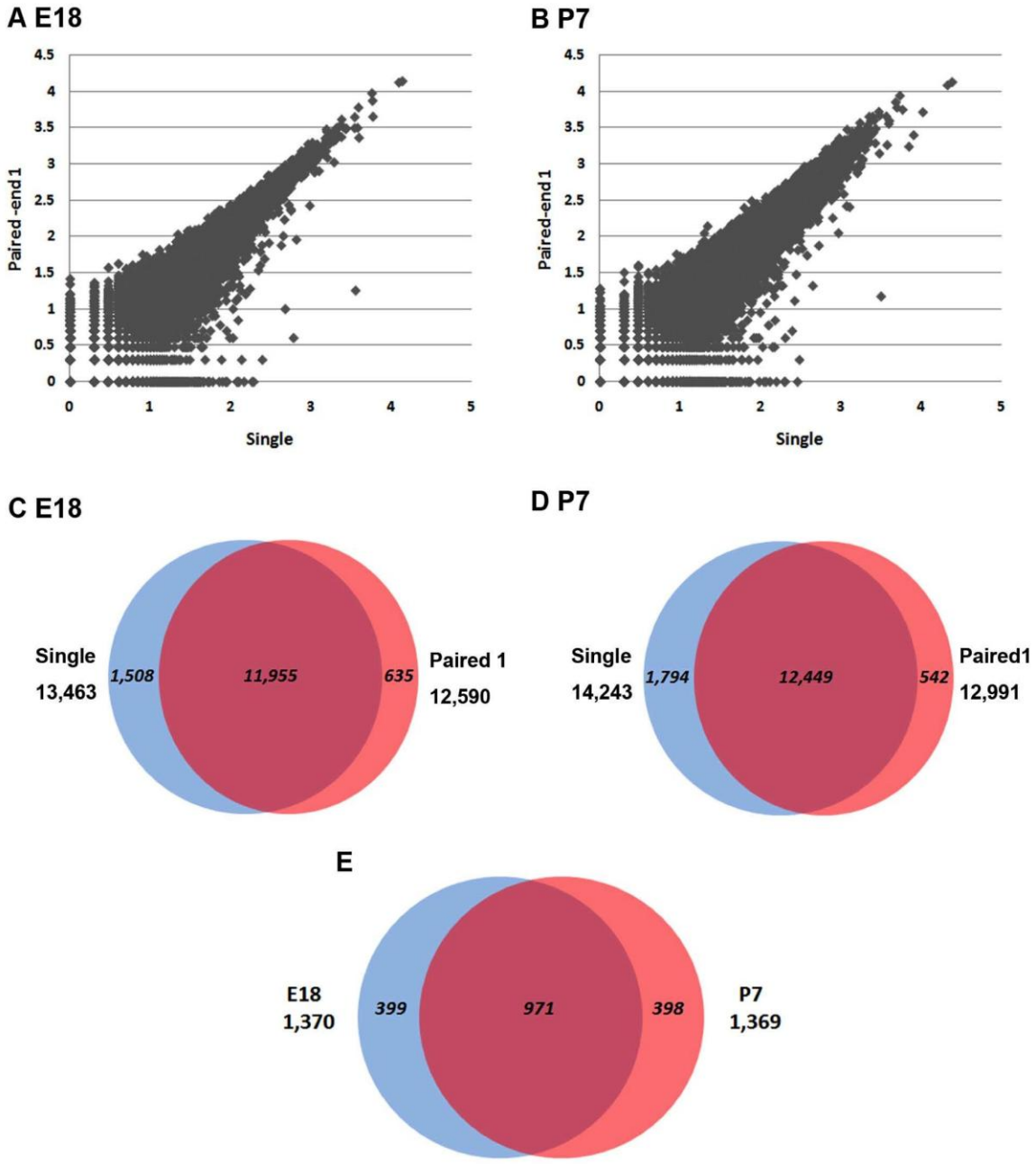
Appendix Figure 2-4. Meiosis turns genomic transposition into copy number variations. (A) Blue (Col) and red (L*er*) rectangles are represented two allelic regions. Without translocations, meiotic products have the same number of copies as parents. (B) In this case, blue and red rectangles are shown homologous regions with different locations on the same chromosome. After meiotic crossover, the second meiotic product obtained two copies, while the third meiotic product has none. (C) Blue and red rectangles are located in different chromosomes. Then either meiotic crossover and/or subsequently independent chromosome assortment can lead to copy number variation.



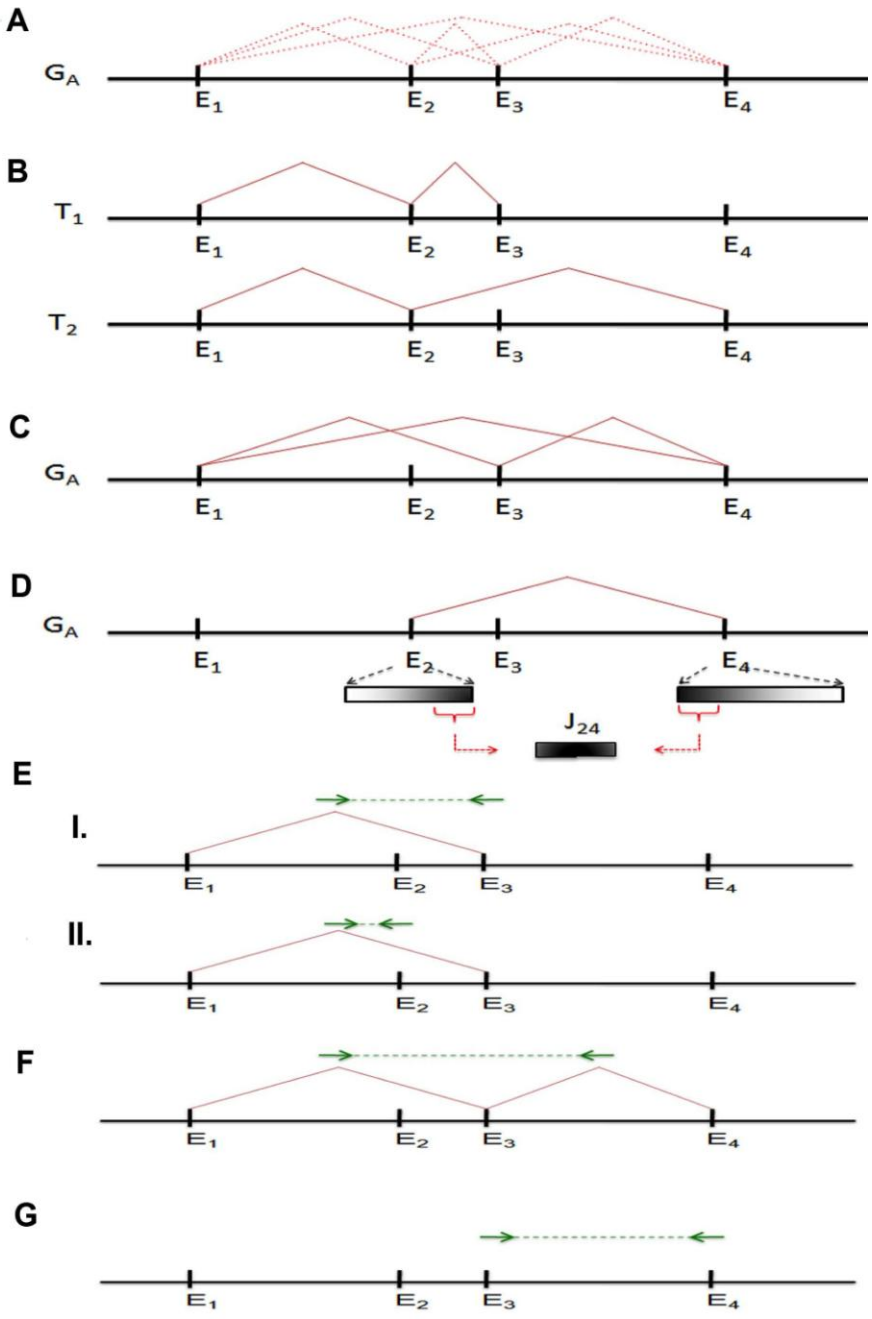
Appendix Figure 3-1. The selection of the number of bases to be used in adaptor trimming. (A) The number of reads containing these bases decreases with the number of bases used to locate the boundary between small RNA and the adaptor sequence. Red dots show the read count when using the real adaptor sequence. Black dots and box plots show read counts when using1000 random sequences of corresponding length, which represent possible artifacts during adaptor trimming. (B) The false positive rate decrease with the number of bases used. False positive rate is calculated by dividing the median read count from random sequences (horizontal bar in box plot) by the read count from the real adaptor sequence (red dot). When 6 or more bases are used, the false positive rate is negligible.



Appendix Figure 4-1. The box plot of quality score at each base. The y-axis was the quality score (Phred-equivalent metric). The x-axis showed each base position in a read. The short bold horizontal line was the median of quality score at that base and the medians above 40 of the first several bases were not shown. At each base the box plot of quality score was shown. Small circles stand for outliers. (A) Single-end reads from E18. (B) Single-end reads from P7. (C) The first end of the first paired-end reads from E18. (D) The second end of the first paired-end reads from E18. (E) The first end of the first paired-end reads from P7. (F) The second end of the first paired-end reads from P7. (G) The first end of the second paired-end reads from E18. (H) The second end of the second paired-end reads from E18. (I) The first end of the second paired-end reads from P7. (G) The second end of the second paired-end reads from P7.In all the ten plots the first 27 bases have quality scores above 20 (less than 1% error rate).



Appendix Figure 4-2. Single-end analysis results and comparison with the first paired-end data (from the same cDNA sample) (A and B) The comparison of reads per gene between single-end analysis and the first paired-end analysis. Because the read number per gene ranges from 0 to over 10,000, the read numbers adding 1 were transformed by log10. (A) E18 R=0.96 (B) P7 R=0.95 (C and D) The comparison of the number of expressed genes between single-end analysis and the first paired-end analysis. Genes having two or more reads were regarded as expressing at certain stage. (C) E18 (D) P7 (E) The number of highly expressed genes in single-end analysis. The top 5% genes were regarded as highly expressed ones.



Appendix Figure 4-3. A hypothetical example of the strategy to detect novel transcripts. (A) Gene A had four exons, E1, E2, E3 and E4. (B) Two transcripts, T1 and T2, were alternatively splicing variants from gene A. T1 had E1, E2 and E3. T2 had E1, E2 and E4. (C) For gene A, junctions between E1 and E3, E1 and E4, E3 and E4 were novel junctions, which were not present in known transcripts. (D) 20 bps on either side of every possible junction were taken and attached to form junction sequences computationally. For gene A, there were 6 possible junctions in total, in

which 3 were known junctions and 3 were novel junctions. In single-end analysis, all junction sequences from UCSC Known Gene database were concatenated into one long pseudosequence (because Rmap only tolerated 1 sequence as target) and reads were mapped to the concatenated sequence with Rmap. Reads mapping to novel junctions were selected and considered as evidence for novel transcripts. In paired-end analysis, all junction sequences and exon sequences were concatenated together to accommodate both ends. One or both ends mapping to novel junctions were selected and considered as evidence for novel transcripts. In addition, paired-end reads mapped to novel exon-exon combination were selected and regarded as evidence for novel transcripts as well. (E to G) Detailed illustration of paired-end reads. Paired-end reads can not only detect novel transcripts by novel junctions (E and F), but also through novel exon-exon combinations (G). (E) One end mapped on a novel junction; the other end mapped in an exon. (i) Valid mapping: one end of a paired-end read mapped on a novel junction (J13); the other end mapped in E3. This indicates a novel splicing form of E1 and E3. (ii) Invalid mapping: one end of a paired-end read mapped at J13; the other end mapped at E2. It is not reasonable to have a splicing form of this kind. (F) Both ends were mapped on junctions. One end of a paired-end read mapped on a novel junction (J13); the other end also mapped at a novel junction (J34). This indicated a novel splicing form of E1, E3 and E4. (G) Both ends were mapped in exons. One end mapped to E3 and the other end mapped to E4. A novel transcript with E3 and E4 was indicated.

Appendix Table 2-1. Summary of read number and coverage

		Total reads [a]	Mapped reads	Mapping quality ≥ 20	Coverage [b]
L*er* ecotype		61,603,614	58,102,231	38,674,710	18.7
1st meiosis	Plant-1A	36,590,794	33,241,502	24,541,695	8.2
	Plant-1B	54,404,688	53,474,928	39,851,797	13.4
	Plant-1C	53,810,544	53,442,224	39,538,948	13.3
	Plant-1D	43,117,198	42,265,828	31,753,318	10.7
2nd meiosis	Plant-2A	48,220,664	45,774,118	34,705,911	16.5
	Plant-2B	50,370,054	47,877,914	33,159,674	16.6
	Plant-2C	45,102,640	43,347,370	32,485,254	13.8
	Plant-2D	39,898,699	38,375,746	27,887,166	10.0

a. The number of total reads includes both single-end reads and paired-end reads. The two ends from paired-end technology are counted as two reads.

b. Only take reads with mapping quality score ≥20 into account in calculating coverage.

Appendix Table 2-2. Expression of genes with 10 or more nonsynonymous mutation and genes with nonsense mutation based on plant ontology*

Mutation type	Plant ontology categories		Genes with mutation	Genes without mutation	P value
10 or more nonsynonymous mutation	Growth stage	L mature pollen stage	22	3878	0.00022
		M germinated pollen stage	32	4557	0.0028
		petal differentiation and expansion stage	223	16298	0.014
		Seedling growth	3	884	0.017
	Structure component	Egg cell	6	166	0.016
		Guard cell	7	1811	0.00063
		Male gametophyte	111	12551	0.0026
		Pollen tube	31	4809	0.00036
Nonsense mutation	Growth stage	petal differentiation and expansion stage	78	16443	0.017
	Structure component	flower	79	16611	0.021
		Sperm cell	35	5380	0.0015

*PO categories with less than 3 mutated genes are not included.

Appendix Table 2-3. The list of meiotic crossovers detected in the two meioses

Meiosis Sample	Crossover	Chromosome	Offspring	Maximal Length	Minimal Length	Containing GC
1st meiosis	CO-1	Chr1	A,B	1166	0	No
	CO-2	Chr1	A,B	562	143	Yes
	CO-3	Chr1	A,B	1514	1	Yes
	CO-4	Chr2	C,D	1549	1209	Yes
	CO-5	Chr3	A,B	1305	0	No
	CO-6	Chr4	A,B	568	0	No
	CO-7	Chr5	A,B	3099	0	No
	CO-8	Chr5	A,B	306	0	No
	CO-9	Chr5	A,C	1064	1	Yes
2nd meiosis	CO-1	Chr1	B,C	129508	0	No
	CO-2	Chr1	B,C	980	0	No
	CO-3	Chr2	A,D	873	0	No
	CO-4	Chr3	A,D	926	0	No
	CO-5	Chr3	A,B	977	0	No
	CO-6	Chr3	A,B	829	0	No
	CO-7	Chr4	A,D	3288	141	Yes
	CO-8	Chr5	A,B	2026	150	Yes
	CO-9	Chr5	B,C	1749	0	No

Appendix Table 2-4. The list of gene conversions detected in the two meioses

Meiosis Sample	Chromosome	Min Start	Min End	Min Length	Max Start	Max End	Max Length	Supporting SNPs
1st meiosis	Chr1	13957178	13958976	1799	13955540	13959511	3972	3
	Chr1	17235790	17235790	1	17232765	17236663	3899	1
	Chr5	121436	121436	1	120397	123474	3078	1
2nd meiosis	Chr5	4114437	4114437	1	4109948	4116643	6696	1

Appendix Table 2-5. Redistribution of SNPs and indels by crossing over meiotic crossovers and independent chromosome assortment *

Genotype			Chr1	Chr2	Chr3	Chr4	Chr5	Total
SNP	L*er*		82392	63075	66040	55579	82085	***349171***
	Meiosis I	Plant A	26422	0	15184	27630	22677	***91913***
		Plant B	55967	63075	50855	27949	57514	***255360***
		Plant C	0	40776	0	0	1893	***42669***
		Plant D	82392	22290	66040	55579	82085	***308386***
	Meiosis II	Plant A	0	17266	42417	36510	17370	***113563***
		Plant B	54391	63075	20134	0	42002	***179602***
		Plant C	28001	0	66040	55579	22709	***172329***
		Plant D	82392	45809	3489	19067	82085	***232842***
Small Indel	L*er*		14399	9536	10772	9648	13730	***58085***
	Meiosis I	Plant A	5534	0	3239	5955	4768	***19496***
		Plant B	8864	9536	7532	3693	8584	***38209***
		Plant C	0	6974	0	0	377	***7351***
		Plant D	14399	2559	10772	9648	13730	***51108***
	Meiosis II	Plant A	0	3421	6159	5583	3393	***18556***
		Plant B	8524	9536	3865	0	5752	***27677***
		Plant C	5872	0	10772	9648	4582	***30874***
		Plant D	14399	6115	748	4064	13730	***39056***
Large Indel	L*er*		588	396	402	397	532	***2315***
	Meiosis I	Plant A	170	0	80	175	131	***556***
		Plant B	417	396	322	222	394	***1751***
		Plant C	0	274	0	0	7	***281***
		Plant D	588	122	402	397	532	***2041***
	Meiosis II	Plant A	0	87	277	287	81	***732***
		Plant B	409	396	113	0	313	***1231***
		Plant C	179	0	402	397	138	***1116***
		Plant D	588	309	12	110	532	***1551***

*All SNPs and indels were using the Col sequence as reference. Pure Col sequence has 0 in the table.

Appendix Table 2-6. Gene families enriched for genomic variations

	Gene family	10 or more nonsynonymous SNPs (443/27206)	Nonsense SNPs (319/27206)	Frameshift indels (845/27206)	Nonframeshift indels (462/27206)	Large indels (316/27206)	All variations (2171/27206)
Not/moderately enriched	MADS	-	-	-	-	-	-
	Myb	-	-	-	0.01115 (6/132)	-	-
	bHLH	-	-	-	-	-	-
	AP2/EREBP	-	-	-	0.01572 (6/138)	-	-
	WD repeat	-	-	-	-	-	-
Strongly enriched	WRKY	-	-	-	1.57E-05 (6/73)	-	-
	F-box	-	<2.2e-16 (36/656)	<2.2e-16 (63/656)	-	1.18E-07 (22/656)	<2.2e-16 (134/656)
	LRR-RLK	3.61E-08 (14/223)	-	-	-	-	-
	LRR-RLP	<2.2e-16 (37/57)	-	-	-	-	2.10E-08 (16/57)
	NBS-LRR	<2.2e-16 (37/147)	-	6.55E-16 (15/147)	-	1.19E-06 (8/147)	<2.2e-16 (58/147)
	Receptor kinase-like	3.97E-13 (21/307)	-	-	-	-	8.89E-05 (43/307)

Appendix Table 3-1. The expression level of 97 mature miRNAs having mapped reads in at least two replicates

miRNA ID	Replicate 1	Percentage	Replicate 2	Percentage	Replicate 3	Percentage
ath-miR156a	115.6281746032	1.9	63.6095238095	1.7	5.5488095238	0.8
ath-miR156b	118.501984127	1.94	63.9761904762	1.71	6.1321428571	0.88
ath-miR156c	115.6281746032	1.9	63.6095238095	1.7	5.5488095238	0.8
ath-miR156d	121.6281746032	1.99	65.4428571429	1.75	5.5488095238	0.8
ath-miR156e	118.0448412698	1.93	66.2761904762	1.77	6.1321428571	0.88
ath-miR156f	118.0448412698	1.93	66.2761904762	1.77	6.1321428571	0.88
ath-miR156g	15.6785714286	0.26	9.9464285714	0.27	0.5714285714	0.08
ath-miR156h	0.25	0	1	0.03	0	0
ath-miR157a	24.1916666667	0.4	12.45	0.33	0.9166666667	0.13
ath-miR157b	23.6916666667	0.39	12.1166666667	0.32	0.9166666667	0.13
ath-miR157c	21.025	0.34	14.45	0.39	0.5833333333	0.08
ath-miR157d	89.275	1.46	53.45	1.43	7	1.01
ath-miR158a	2182.5166666667	35.77	1446.2666666667	38.6	203	29.2
ath-miR158b	82.8333333333	1.36	57.3666666667	1.53	8	1.15
ath-miR159a	69.5166666667	1.14	30.85	0.82	29.45	4.24
ath-miR159b	34.5166666667	0.57	14.1	0.38	15.45	2.22
ath-miR159c	10.1833333333	0.17	1.5166666667	0.04	3.45	0.5
ath-miR160a	51.25	0.84	30.95	0.83	3.5833333333	0.52
ath-miR160b	49.75	0.82	32.45	0.87	4.0833333333	0.59
ath-miR160c	49.75	0.82	32.45	0.87	4.0833333333	0.59
ath-miR161.1	22	0.36	14	0.37	2	0.29
ath-miR161.2	4	0.07	3	0.08	0	0
ath-miR162a	105	1.72	77	2.06	14.5	2.09
ath-miR162b	105	1.72	77	2.06	14.5	2.09
ath-miR163	2	0.03	2.3333333333	0.06	0	0
ath-miR164a	16.2666666667	0.27	6.4761904762	0.17	2	0.29
ath-miR164b	14.2666666667	0.23	6.4761904762	0.17	2	0.29
ath-miR164c	1.2666666667	0.02	2.4761904762	0.07	0	0
ath-miR165a	46.6111111111	0.76	17	0.45	11.9444444444	1.72
ath-miR165b	52.6111111111	0.86	13.25	0.35	12.9444444444	1.86
ath-miR166a	17.0158730159	0.28	10.0238095238	0.27	3.9801587302	0.57
ath-miR166b	14.0158730159	0.23	9.0238095238	0.24	3.9801587302	0.57
ath-miR166c	14.3492063492	0.24	9.0238095238	0.24	3.9801587302	0.57
ath-miR166d	14.0158730159	0.23	9.0238095238	0.24	3.9801587302	0.57
ath-miR166e	12.0158730159	0.2	8.7738095238	0.23	3.7301587302	0.54
ath-miR166f	12.0158730159	0.2	8.9404761905	0.24	3.7301587302	0.54
ath-miR166g	12.0158730159	0.2	8.9404761905	0.24	3.7301587302	0.54
ath-miR167a	231.75	3.8	178.2833333333	4.76	44.8333333333	6.45
ath-miR167b	230.75	3.78	178.2833333333	4.76	44.8333333333	6.45
ath-miR167c	1.5	0.02	1.2	0.03	0	0
ath-miR167d	42.4166666667	0.7	40.45	1.08	4.3333333333	0.62
ath-miR168a	17.5	0.29	15	0.4	1	0.14
ath-miR168b	16.5	0.27	15	0.4	0	0
ath-miR169a	133.1666666667	2.18	27.8666666667	0.74	2.0714285714	0.3
ath-miR169b	21.6666666667	0.36	11.3666666667	0.3	0.0714285714	0.01
ath-miR169c	12.6666666667	0.21	2.1666666667	0.06	0.0714285714	0.01
ath-miR169d	4.25	0.07	1.2	0.03	0.3214285714	0.05
ath-miR169e	4.25	0.07	1.2	0.03	0.3214285714	0.05
ath-miR169f	1.25	0.02	0.5	0.01	0.3214285714	0.05
ath-miR169g	3.25	0.05	0.5	0.01	0.3214285714	0.05
ath-miR169h	0.2857142857	0	0.3428571429	0.01	0.0714285714	0.01
ath-miR169i	1.7857142857	0.03	1.5928571429	0.04	0.0714285714	0.01
ath-miR169j	1.7857142857	0.03	0.5928571429	0.02	0.0714285714	0.01
ath-miR169k	2.2857142857	0.04	0.3428571429	0.01	0.0714285714	0.01
ath-miR169l	1.7857142857	0.03	0.5928571429	0.02	0.0714285714	0.01
ath-miR169m	1.2857142857	0.02	1.1428571429	0.03	0.0714285714	0.01
ath-miR169n	1.7857142857	0.03	0.5928571429	0.02	0.0714285714	0.01
ath-miR170	9	0.15	2	0.05	0	0
ath-miR171a	7.25	0.12	2.5	0.07	0.25	0.04
ath-miR171b	2.5	0.04	0.5	0.01	0.5	0.07
ath-miR171c	2.5	0.04	0.5	0.01	0.5	0.07
ath-miR172a	293.3428571429	4.81	159.2666666667	4.25	35.1666666667	5.06
ath-miR172b	305.0095238095	5	156.2666666667	4.17	34.6666666667	4.99
ath-miR172c	177.6761904762	2.91	110.1	2.94	23.0833333333	3.32
ath-miR172d	177.3428571429	2.91	110.4333333333	2.95	23.5833333333	3.39
ath-miR172e	30.15	0.49	21.2333333333	0.57	0.6666666667	0.1
ath-miR173	7	0.11	7	0.19	3	0.43

ath-miR319a	7.5	0.12	9.1666666667	0.24	3.1666666667	0.46
ath-miR319b	37	0.61	15.6666666667	0.42	2.1666666667	0.31
ath-miR319c	5.5	0.09	4.1666666667	0.11	1.6666666667	0.24
ath-miR390a	38.1666666667	0.63	22.5833333333	0.6	3.3333333333	0.48
ath-miR390b	38.1666666667	0.63	22.5833333333	0.6	3.3333333333	0.48
ath-miR391	3	0.05	2	0.05	0	0
ath-miR394a	9	0.15	9.5	0.25	4.5	0.65
ath-miR394b	9	0.15	9.5	0.25	4.5	0.65
ath-miR395a	0.3333333333	0.01	1.3333333333	0.04	0	0
ath-miR395b	0.3333333333	0.01	0	0	0.3333333333	0.05
ath-miR395c	0.3333333333	0.01	0	0	0.3333333333	0.05
ath-miR395d	0.3333333333	0.01	1.3333333333	0.04	0	0
ath-miR395e	0.3333333333	0.01	1.3333333333	0.04	0	0
ath-miR395f	0.3333333333	0.01	0	0	0.3333333333	0.05
ath-miR396a	55.5	0.91	22	0.59	4.5	0.65
ath-miR396b	81.5	1.34	57	1.52	18.5	2.66
ath-miR399b	2	0.03	2	0.05	0	0
ath-miR399c	3	0.05	2	0.05	0	0
ath-miR403	8.5	0.14	3.8333333333	0.1	1	0.14
ath-miR408	3	0.05	0	0	2	0.29
ath-miR5017	1	0.02	1	0.03	0	0
ath-miR775	14	0.23	15	0.4	2	0.29
ath-miR780.1	9	0.15	9.5	0.25	1	0.14
ath-miR780.2	2	0.03	2	0.05	0	0
ath-miR823	8	0.13	3	0.08	1	0.14
ath-miR824	88	1.44	29	0.77	4.5	0.65
ath-miR830	1	0.02	2	0.05	0	0
ath-miR839	45	0.74	26.5	0.71	10	1.44
ath-miR845a	2	0.03	1	0.03	0	0
ath-miR858	2.5	0.04	0	0	0.5	0.07

Appendix Table 4-1. Summary of novel transcript analysis

Summary of reads, junctions and genes	**Single-end 1**		**Paired-end 1**		**Paired-end 2**	
	E18	P7	E18	P7	E18	P7
Reads	2,302	2,689	621	746	794	1,084
Junctions	1,800	2,177	151	189	504	631
Genes*	1,367	1,596	143	172	448	536
Total genes	**2,396**		**295**		**811**	

* Numbers in this row showed how many genes the detected novel junctions belonged to.

REFERENCES

1. Sanger F, Nicklen S, & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74(12):5463-5467.
2. Maxam AM & Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 74(2):560-564.
3. Drossman H, Luckey JA, Kostichka AJ, D'Cunha J, & Smith LM (1990) High-speed separations of DNA sequencing reactions by capillary electrophoresis. *Analytical Chemistry* 62(9):900-903.
4. Lander ES, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.
5. Venter JC, *et al.* (2001) The sequence of the human genome. *Science* 291(5507):1304-1351.
6. Levy S, *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biology* 5(10):e254.
7. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, & Barron AE (2011) Landscape of next-generation sequencing technologies. *Analytical Chemistry* 83(12):4327-4341.
8. Sims PA, Greenleaf WJ, Duan H, & Xie XS (2011) Fluorogenic DNA sequencing in PDMS microreactors. *Nature Methods* 8(7):575-580.
9. Liti G, *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature* 458(7236):337-341.
10. Ledford H (2008) Population genomics for fruitflies. *Nature* 453(7199):1154-1155.
11. Cao J, *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics* 43(10):956-963.
12. Anonymous (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061-1073.
13. Biesecker LG, Shianna KV, & Mullikin JC (2011) Exome sequencing: the expert view. *Genome Biology* 12(9):128.
14. Green ED & Guyer MS (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature* 470(7333):204-213.
15. Li R, *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463(7279):311-317.
16. Mendes R, *et al.* (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* 332(6033):1097-1100.
17. Faith JJ, McNulty NP, Rey FE, & Gordon JI (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* 333(6038):101-104.
18. Hess M, *et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331(6016):463-467.
19. Wang Z, Gerstein M, & Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1):57-63.
20. Nagalakshmi U, *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881):1344-1349.
21. Marioni JC, Mason CE, Mane SM, Stephens M, & Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 18(9):1509-1517.

22. Malone JH & Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology* 9:34.
23. Ozsolak F & Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics* 12(2):87-98.
24. Zhang Z & Pugh BF (2011) High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* 144(2):175-186.
25. MacQuarrie KL, Fong AP, Morse RH, & Tapscott SJ (2011) Genome-wide transcription factor binding: beyond direct target regulation. *Trends In Genetics* 27(4):141-148.
26. Zheng Q, *et al.* (2010) Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genetics* 6(9).
27. Dostie J, *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research* 16(10):1299-1309.
28. Trapnell C & Salzberg SL (2009) How to map billions of short reads onto genomes. *Nature Biotechnology* 27(5):455-457.
29. Bao S, *et al.* (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics* 56(6):406-414.
30. Li H, Ruan J, & Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18(11):1851-1858.
31. Lunter G & Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21(6):936-939.
32. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.
33. Homer N, Merriman B, & Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. *PloS One* 4(11):e7767.
34. Holtgrewe M, Emde AK, Weese D, & Reinert K (2011) A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics* 12:210.
35. Charlesworth B (2010) Molecular population genomics: a short history. *Genetics Research* 92(5-6):397-411.
36. Anonymous (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796-815.
37. Nordborg M, *et al.* (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology* 3(7):e196.
38. Clark RM, *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 317(5836):338-342.
39. Ossowski S, *et al.* (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Research* 18(12):2024-2033.
40. Giraut L, *et al.* (2011) Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS Genetics* 7(11):e1002354.
41. Chen JM, Cooper DN, Chuzhanova N, Ferec C, & Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics* 8(10):762-775.
42. Watanabe Y (2011) Overview of plant RNAi. *Methods Mol Biol* 744:1-11.
43. Hamilton AJ & Baulcombe DC (1999) A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* 286(5441):950-952.
44. Chen X (2010) Small RNAs - secrets and surprises of the genome. *The Plant Journal* 61(6):941-958.
45. Vazquez F, Legrand S, & Windels D (2010) The biosynthetic pathways and biological scopes of plant small RNAs. *Trends in Plant Science* 15(6):337-345.

46. Reinhart BJ, *et al.* (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403(6772):901-906.
47. Llave C, Kasschau KD, Rector MA, & Carrington JC (2002) Endogenous and silencing-associated small RNAs in plants. *The Plant Cell* 14(7):1605-1619.
48. Park W, Li J, Song R, Messing J, & Chen X (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Current Biology* 12(17):1484-1495.
49. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, & Bartel DP (2002) MicroRNAs in plants. *Genes & Development* 16(13):1616-1626.
50. Palatnik JF, *et al.* (2003) Control of leaf morphogenesis by microRNAs. *Nature* 425(6955):257-263.
51. Kim ED & Sung S (2011) Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends in Plant Science*.
52. Su AI, *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 99(7):4465-4470.
53. Sharov AA, *et al.* (2003) Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biology* 1(3):E74.
54. Kawai J, *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature* 409(6821):685-690.
55. Okazaki Y, *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420(6915):563-573.
56. Ravasi T, *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140(5):744-752.
57. Carninci P, *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309(5740):1559-1563.
58. Carninci P, *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* 38(6):626-635.
59. Temple S (2001) The development of neural stem cells. *Nature* 414(6859):112-117.
60. Chun JJ, Schatz DG, Oettinger MA, Jaenisch R, & Baltimore D (1991) The recombination activating gene-1 (RAG-1) transcript is present in the murine central nervous system. *Cell* 64(1):189-200.
61. Long M, Betran E, Thornton K, & Wang W (2003) The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics* 4(11):865-875.
62. Mitchell-Olds T & Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* 441(7096):947-952.
63. Stankiewicz P & Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* 18(2):74-82.
64. Hurles ME, Dermitzakis ET, & Tyler-Smith C (2008) The functional impact of structural variation in humans. *Trends in Genetics* 24(5):238-245.
65. Johanson U, *et al.* (2000) Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science* 290(5490):344-347.
66. Michaels SD, He Y, Scortecci KC, & Amasino RM (2003) Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* 100(17):10102-10107.
67. Koornneef M, Alonso-Blanco C, & Vreugdenhil D (2004) Naturally occurring genetic variation in Arabidopsis thaliana. *Annual Review of Plant Biology* 55:141-172.

68. Krieger U, Lippman ZB, & Zamir D (2010) The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nature Genetics* 42(5):459-463.
69. Graubert TA, *et al.* (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genetics* 3(1):e3.
70. Emerson JJ, Cardoso-Moreira M, Borevitz JO, & Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* 320(5883):1629-1631.
71. Kidd JM, *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453(7191):56-64.
72. Zickler D & Kleckner N (1999) Meiotic chromosomes: integrating structure and function. *Annual Review of Genetics* 33:603-754.
73. Ma H (2006) A molecular portrait of Arabidopsis meiosis. *The Arabidopsis Book*: 1-39.
74. Hurst DD, Fogel S, & Mortimer RK (1972) Conversion-associated recombination in yeast (hybrids-meiosis-tetrads-marker loci-models). *Proceedings of the National Academy of Sciences of the United States of America* 69(1):101-105.
75. Haubold B, Kroymann J, Ratzka A, Mitchell-Olds T, & Wiehe T (2002) Recombination and gene conversion in a 170-kb genomic region of Arabidopsis thaliana. *Genetics* 161(3):1269-1278.
76. Keeney S (2001) Mechanism and control of meiotic recombination initiation. *Current Topics in Developmental Biology* 52:1-53.
77. Mezard C, Vignard J, Drouaud J, & Mercier R (2007) The road to crossovers: plants have their say. *Trends in Genetics* 23(2):91-99.
78. Mancera E, Bourgon R, Brozzi A, Huber W, & Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479-485.
79. Qi J, *et al.* (2009) Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC genomics* 10:475.
80. Initiative TAG (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796-815.
81. Meyerowitz E & Ma H (1994) *Genetic variations of Arabidopsis thaliana*.
82. Ziolkowski PA, Koczyk G, Galganski L, & Sadowski J (2009) Genome sequence comparison of Col and Ler lines reveals the dynamic nature of Arabidopsis chromosomes. *Nucleic Acids Research* 37(10):3189-3201.
83. Bentsink L, *et al.* (2010) Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways. *Proceedings of the National Academy of Sciences of the United States of America* 107(9):4264-4269.
84. Guyon-Debast A, Lecureuil A, Bonhomme S, Guerche P, & Gallois JL (2010) A SNP associated with alternative splicing of RPT5b causes unequal redundancy between RPT5a and RPT5b among Arabidopsis thaliana natural variation. *BMC Plant Biology* 10:158.
85. Todesco M, *et al.* (2010) Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature* 465(7298):632-636.
86. Preuss D, Rhee SY, & Davis RW (1994) Tetrad analysis possible in Arabidopsis with mutation of the QUARTET (QRT) genes. *Science* 264(5164):1458-1460.
87. Francis KE, *et al.* (2007) Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* 104(10):3913-3918.
88. Qi J, Zhao F, Buboltz A, & Schuster SC (2010) inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics* 26(1):127-129.
89. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18(5):821-829.

90. Delcher AL, *et al.* (1999) Alignment of whole genomes. *Nucleic Acids Research* 27(11):2369-2376.
91. Schneeberger K, *et al.* (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences of the United States of America* 108(25):10249-10254.
92. Du Z, Zhou X, Ling Y, Zhang Z, & Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research* 38(Web Server issue):W64-70.
93. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24(8):1586-1591.
94. Colantuoni C, Henry G, Zeger S, & Pevsner J (2002) SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics* 18(11):1540-1541.
95. Jiao Y, *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97-100.
96. Meinke D, Sweeney C, & Muralla R (2009) Integrating the genetic and physical maps of Arabidopsis thaliana: identification of mapped alleles of cloned essential (EMB) genes. *PloS One* 4(10):e7386.
97. Rentel MC, Leonelli L, Dahlbeck D, Zhao B, & Staskawicz BJ (2008) Recognition of the Hyaloperonospora parasitica effector ATR13 triggers resistance against oomycete, bacterial, and viral pathogens. *Proceedings of the National Academy of Sciences of the United States of America* 105(3):1091-1096.
98. Schmid M, *et al.* (2003) Dissection of floral induction pathways using global expression analysis. *Development* 130(24):6001-6012.
99. Sakai T, *et al.* (2005) Origins of mouse inbred strains deduced from whole-genome scanning by polymorphic microsatellite loci. *Mammalian genome* 16(1):11-19.
100. Tsang S, *et al.* (2005) A comprehensive SNP-based genetic analysis of inbred mouse strains. *Mammalian genome* 16(7):476-480.
101. Yang H, *et al.* (2011) Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics* 43(7):648-655.
102. Hu TT, *et al.* (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics* 43(5):476-481.
103. Xu G, Ma H, Nei M, & Kong H (2009) Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proceedings of the National Academy of Sciences of the United States of America* 106(3):835-840.
104. Kroymann J, Donnerhacke S, Schnabelrauch D, & Mitchell-Olds T (2003) Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proceedings of the National Academy of Sciences of the United States of America* 100 Suppl 2:14587-14592.
105. Sanchez-Moran E, Santos JL, Jones GH, & Franklin FC (2007) ASY1 mediates AtDMC1-dependent interhomolog recombination during meiosis in Arabidopsis. *Genes & Development* 21(17):2220-2233.
106. Copenhaver GP, Housworth EA, & Stahl FW (2002) Crossover interference in Arabidopsis. *Genetics* 160(4):1631-1639.
107. de los Santos T, *et al.* (2003) The Mus81/Mms4 endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics* 164(1):81-94.
108. Higgins JD, Armstrong SJ, Franklin FC, & Jones GH (2004) The Arabidopsis MutS homolog AtMSH4 functions at an early step in recombination: evidence for two classes of recombination in Arabidopsis. *Genes & Development* 18(20):2557-2570.

109. Hollingsworth NM & Brill SJ (2004) The Mus81 solution to resolution: generating meiotic crossovers without Holliday junctions. *Genes & Development* 18(2):117-125.
110. Chen C, Zhang W, Timofejeva L, Gerardin Y, & Ma H (2005) The Arabidopsis ROCK-N-ROLLERS gene encodes a homolog of the yeast ATP-dependent DNA helicase MER3 and is required for normal meiotic crossover formation. *The Plant Journal* 43(3):321-334.
111. Wijeratne AJ, Chen C, Zhang W, Timofejeva L, & Ma H (2006) The Arabidopsis thaliana PARTING DANCERS gene encoding a novel protein is required for normal meiotic homologous recombination. *Molecular Biology of the Cell* 17(3):1331-1343.
112. Baudat F, *et al.* (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967):836-840.
113. Drouaud J, *et al.* (2006) Variation in crossing-over rates across chromosome 4 of Arabidopsis thaliana reveals the presence of meiotic recombination "hot spots". *Genome Research* 16(1):106-114.
114. Kauppi L, Jeffreys AJ, & Keeney S (2004) Where the crossovers are: recombination distributions in mammals. *Nature Reviews Genetics* 5(6):413-424.
115. Zhang F, Gu W, Hurles ME, & Lupski JR (2009) Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* 10:451-481.
116. Itsara A, *et al.* (2010) De novo rates and selection of large copy number variation. *Genome Research* 20(11):1469-1481.
117. Fu W, Zhang F, Wang Y, Gu X, & Jin L (2010) Identification of copy number variation hotspots in human populations. *American Journal of Human Genetics* 87(4):494-504.
118. Platt A, *et al.* (2010) The scale of population structure in Arabidopsis thaliana. *PLoS Genetics* 6(2):e1000843.
119. Maloof JN, *et al.* (2001) Natural variation in light sensitivity of Arabidopsis. *Nature Genetics* 29(4):441-446.
120. Filiault DL, *et al.* (2008) Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light. *Proceedings of the National Academy of Sciences of the United States of America* 105(8):3157-3162.
121. Mindrinos M, Katagiri F, Yu GL, & Ausubel FM (1994) The A. thaliana disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell* 78(6):1089-1099.
122. Baker SM, *et al.* (1995) Male mice defective in the DNA mismatch repair gene PMS2 exhibit abnormal chromosome synapsis in meiosis. *Cell* 82(2):309-319.
123. Barlow AL & Hulten MA (1998) Crossing over analysis at pachytene in man. *European Journal of Human Genetics* 6(4):350-358.
124. Lynn A, *et al.* (2002) Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science* 296(5576):2222-2225.
125. Kleckner N, Storlazzi A, & Zickler D (2003) Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. *Trends in Genetics* 19(11):623-628.
126. Dresser ME & Giroux CN (1988) Meiotic chromosome behavior in spread preparations of yeast. *The Journal of Cell Biology* 106(3):567-573.
127. Sanchez Moran E, Armstrong SJ, Santos JL, Franklin FC, & Jones GH (2001) Chiasma formation in Arabidopsis thaliana accession Wassileskija and in two meiotic mutants. *Chromosome Research* 9(2):121-128.
128. Borts RH, Chambers SR, & Abdullah MF (2000) The many faces of mismatch repair in meiosis. *Mutation Research* 451(1-2):129-150.

129. Baarends WM & Mercier R (2010) Sisters dancing in meiosis. *EMBO Reports* 11(2):76-78.
130. Lee RC, Feinbaum RL, & Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75(5):843-854.
131. Zamore PD & Haley B (2005) Ribo-gnome: the big world of small RNAs. *Science* 309(5740):1519-1524.
132. Sunkar R & Zhu JK (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *The Plant Cell* 16(8):2001-2019.
133. Lu C, *et al.* (2005) Elucidation of the small RNA component of the transcriptome. *Science* 309(5740):1567-1569.
134. Lu C, *et al.* (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Research* 16(10):1276-1288.
135. Backman TW, *et al.* (2008) Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic Acids Research* 36(Database issue):D982-985.
136. Breakfield NW, *et al.* (2012) High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. *Genome Research* 22(1):163-176.
137. Slotkin RK, *et al.* (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136(3):461-472.
138. Yang H, Lu P, Wang Y, & Ma H (2011) The transcriptome landscape of Arabidopsis male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *The Plant Journal* 65(4):503-516.
139. Chen C, *et al.* (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated Arabidopsis male meiocytes. *BMC Plant Biology* 10:280.
140. Hackenberg M, Rodriguez-Ezpeleta N, & Aransay AM (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research* 39(Web Server issue):W132-138.
141. David M, Dzamba M, Lister D, Ilie L, & Brudno M (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* 27(7):1011-1012.
142. Kozomara A & Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* 39(Database issue):D152-157.
143. Todesco M, Rubio-Somoza I, Paz-Ares J, & Weigel D (2010) A collection of target mimics for comprehensive analysis of microRNA function in Arabidopsis thaliana. *PLoS Genetics* 6(7):e1001031.
144. Mathieu J, Yant LJ, Murdter F, Kuttner F, & Schmid M (2009) Repression of flowering by the miR172 target SMZ. *PLoS Biology* 7(7):e1000148.
145. Meyers BC, *et al.* (2008) Criteria for annotation of plant MicroRNAs. *The Plant Cell* 20(12):3186-3190.
146. Allendoerfer KL & Shatz CJ (1994) The subplate, a transient neocortical structure: its role in the development of connections between thalamus and cortex. *Annual Review of Neuroscience* 17:185-218.
147. Rakic P (2006) A century of progress in corticoneurogenesis: from silver impregnation to genetic engineering. *Cereb Cortex* 16 Suppl 1:i3-17.
148. Christopherson KS, *et al.* (2005) Thrombospondins are astrocyte-secreted proteins that promote CNS synaptogenesis. *Cell* 120(3):421-433.
149. Ullian EM, Christopherson KS, & Barres BA (2004) Role for glia in synaptogenesis. *Glia* 47(3):209-216.
150. Matsuo I, Kuratani S, Kimura C, Takeda N, & Aizawa S (1995) Mouse Otx2 functions in the formation and patterning of rostral head. *Genes & Development* 9(21):2646-2658.

151. Funatsu N, Inoue T, & Nakamura S (2004) Gene expression analysis of the late embryonic mouse cerebral cortex using DNA microarray: identification of several region- and layer-specific genes. *Cereb Cortex* 14(9):1031-1044.
152. Jiang CH, Tsien JZ, Schultz PG, & Hu Y (2001) The effects of aging on gene expression in the hypothalamus and cortex of mice. *Proceedings of the National Academy of Sciences of the United States of America* 98(4):1930-1934.
153. Lee CK, Weindruch R, & Prolla TA (2000) Gene-expression profile of the ageing brain in mice. *Nature Genetics* 25(3):294-297.
154. Miki R, *et al.* (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proceedings of the National Academy of Sciences of the United States of America* 98(5):2199-2204.
155. Mirnics ZK, *et al.* (2003) DNA microarray profiling of developing PS1-deficient mouse brain reveals complex and coregulated expression changes. *Molecular Psychiatry* 8(10):863-878.
156. Mody M, *et al.* (2001) Genome-wide gene expression profiles of the developing mouse hippocampus. *Proceedings of the National Academy of Sciences of the United States of America* 98(15):8862-8867.
157. Pavlidis P & Noble WS (2001) Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology* 2(10):RESEARCH0042.
158. Semeralul MO, *et al.* (2006) Microarray analysis of the developing cortex. *Journal of Neurobiology* 66(14):1646-1658.
159. Zapala MA, *et al.* (2005) Adult mouse brain gene expression patterns bear an embryologic imprint. *Proceedings of the National Academy of Sciences of the United States of America* 102(29):10357-10362.
160. Cahoy JD, *et al.* (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *The Journal of Neuroscience* 28(1):264-278.
161. Holt RA & Jones SJ (2008) The new paradigm of flow cell sequencing. *Genome Research* 18(6):839-846.
162. von Bubnoff A (2008) Next-generation sequencing: the race is on. *Cell* 132(5):721-723.
163. Strausberg RL, Levy S, & Rogers YH (2008) Emerging DNA sequencing technologies for human genomic medicine. *Drug Discovery Today* 13(13-14):569-577.
164. Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5(7):621-628.
165. Sultan M, *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321(5891):956-960.
166. Wilhelm BT, *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453(7199):1239-1243.
167. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3:Article3.
168. Blankenberg D, *et al.* (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Research* 17(6):960-964.
169. Giardine B, *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* 15(10):1451-1455.
170. Ewing B & Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8(3):186-194.

171. Bentley DR, *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53-59.
172. Kent WJ, *et al.* (2002) The human genome browser at UCSC. *Genome Research* 12(6):996-1006.
173. Hsu F, *et al.* (2006) The UCSC Known Genes. *Bioinformatics* 22(9):1036-1046.
174. Birney E, *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799-816.
175. Wang ET, *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470-476.
176. Cull-Candy S, Brickley S, & Farrant M (2001) NMDA receptor subunits: diversity, development and disease. *Current Opinion in Neurobiology* 11(3):327-335.
177. Salinas PC & Zou Y (2008) Wnt signaling in neural circuit assembly. *Annual Review of Neuroscience* 31:339-358.
178. Charron F & Tessier-Lavigne M (2005) Novel brain wiring functions for classical morphogens: a role as graded positional cues in axon guidance. *Development* 132(10):2251-2262.
179. Gray PA, *et al.* (2004) Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* 306(5705):2255-2257.
180. Sugiyama S, *et al.* (2008) Experience-dependent transfer of Otx2 homeoprotein into the visual cortex activates postnatal plasticity. *Cell* 134(3):508-520.
181. Cecconi F, *et al.* (2007) A novel role for autophagy in neurodevelopment. *Autophagy* 3(5):506-508.
182. Fimia GM, *et al.* (2007) Ambra1 regulates autophagy and development of the nervous system. *Nature* 447(7148):1121-1125.
183. Yeo W & Gautier J (2004) Early neural cell death: dying to become neurons. *Developmental Biology* 274(2):233-244.

VITA

Xinwei Han

Education:

2007 – Present Ph.D. candidate in Genetics Program (minor in Statistics),
The Pennsylvania State University

2003 – 2007 B.S., Biological Sciences, Fudan University, China

Publication:

Lu P*, **Han X***, Qi J*, Yang J, Wijeratne AJ, Li T, Ma H, Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by re-sequencing Landsberg erecta and all four products of a single meiosis, Genome Research, 2012, 22(3),508-18. (* equal contribution)

Han X, Wu X, Chung WY, Li T, Nekrutenko A, Altman NS, Chen G, Ma H, Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing, PNAS, 2009, 106(31), 12741-6.

Fu W, Chu L, **Han X**, Liu X, Ren D, Synergistic antitumoral effects of human telomerase reverse transcriptase-mediated dual-apoptosis-related gene vector delivered by orally attenuated Salmonella enterica Serovar Typhimurium in murine tumor models, Journal of Gene Medicine, 2008, 10(6), 690-701.

Fu W, Lan H, Li S, **Han X**, Gao T, Ren D, Synergistic antitumor efficacy of suicide/ePNP gene and 6-methylpurine 2'-deoxyriboside via Salmonella against murine tumors, Cancer Gene Therapy, 2008, 15(7), 474-84.

Awards and Fellowships:

2010 Braddock research award, The Pennsylvania State University
2007 Top award, The Pennsylvania State University
2003-2007 People Scholarship, Fudan University

Conference Presentations:

2011 18th Biennial Penn State Plant Biology Symposium, plenary talk:
"Analysis of Arabidopsis Natural Variation before and After Meiotic Recombination"

2010 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), poster: "Transcriptome of Embryonic and Neonatal Mouse Cortex by RNA-seq"

Teaching Experience:

Fall semesters of each of 2008-2011, Teaching Assistant of Bio110 Biology: Basic Concepts and Biodiversity