1-2011

Comparison Genomics of Oryza rufipogon Chromosome Eight Centromere with Other Species Centromeres

Kalyani Ginjupalli
Montclair State University

MONTCLAIR STATE UNIVERSITY

COMPARISON GENOMICS OF *ORYZA RUFIPOGON* CHROMOSOME EIGHT CENTROMERE WITH OTHER SPECIES CENTROMERES

by

Kalyani Ginjupalli

A Master's Thesis Submitted to the Faculty of
Montclair State University
In Partial Fulfillment of the Requirements
For the Degree of

Masters in Molecular Biology

January 2011

College of Science and Mathematics

Department of Biology and Molecular Biology

Thesis Committee:

[redacted]

Dr. Chunguang Du, Ph.D.
Thesis sponsor

[redacted]

Robert S. Prezant, Ph.D
Dean

[redacted]

Dr. John J. Gaynor, Ph.D.
Committee member

January 2011

[redacted]

Dr. Ann Marie Dilorenzo
Committee member

[redacted]

Dr. Quinn C. Vega
Department chair

Abstract

Centromeres are playing a key role in proper segregation of sister chromatids during mitosis and meiosis. At the centromere region in the chromosome, spindle microtubules attached to the kinetochores that direct the chromosome segregation during mitosis and meiosis. Although centromeric function is conserved among different organisms, there is no conservation of centromeric DNA sequences ranging from budding yeast which has a consensus sequence of approximately 125 base pairs (bp) on each chromosome to *Caenorhabditis elegans* species which has a holocentric centromere that spans the entire chromosome. However, there are some common proteins that form the kinetochore, such as CENH3 and CENP-C.

In this study, wild rice (*Oryza rufipogon*) chromosome 8 centromere fragments were compared to all the domesticate rice (*Oryza sativa* ssp *japonica*) chromosome centromeres, *Arabidopsis thaliana* chromosome 5 centromere, *Lycopersicon esculentum* chromosome 12 centromere, and *Dictyostelium discoideum* chromosome centromere. Annotation results of all the genomes suggest that many of the structural aspects parallel in both the species of the rice centromeres. For example, both wild rice fragments and domesticated rice chromosome centromere sequences found to share their GC contents in between 40-45%. The *Arabidopsis thaliana* chromosome 5 centromere, the *Lycopersicon esculentum* chromosome 12 centromere, and the *Dictyostelium discoideum* chromosome centromere had less structural similarities with the *Oryza rufipogon* chromosome 8 centromere. Further, annotation of the wild rice fragments and the domesticated rice genome revealed that both the species had the cento satellite repeats and *Ty3/gypsy* retrotransposons.

COMPARISON GENOMICS OF *ORYZA RUFIPOGON* CHROMOSOME EIGHT CENTROMERE WITH OTHER SPECIES CENTROMERES

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science in Molecular Biology

by

KALYANI GINJUPALLI

Montclair State University

Montclair, NJ

2011

Table of Contents

List of Figures

List of Tables

Introduction

Centromeres play an important role in the chromosome segregation during cell division. Centromeres are the sites of kinetochore assembly and sister chromatid cohesion during cell division (Kitagawa *et al.*, 2001). Centromere function is conserved from plants to mammals and also centromeric chromatins have similar structural features in eukaryotes. Eucaryotic centromeres are characterized by having various types of elements with repetitive sequences in the center, including satellite DNA and by having transposons and retrotransposons in the flanking region except budding yeast (*Saccharomyces cerevisiae).* Centromeres are typically composed of repetitive satellite sequences that are evolving rapidly and are arranged in a tandem head to tail fashion. Cenrtomeric chromatin is characterized by the presence of specialized histone H3 called CENH3. Species specific CENH3 molecules have been identified in all eukaryotes investigated so far. Completely cloning, sequencing and assembling of the centromere sequences have been remained as challenging because of the heterochromatic nature of the centromeres (Jiang *et al.*, 2003; Ma *et al.*, 2007; kitagawa *et al.*, 2001; Talbert *et al.*, 2004).

This study focuses on the comparison of the centromeric region of the wild rice chromosome 8 centromere fragments with the domesicated rice centromeres , other plants (*Brassicaceae*, *Solanaceae*) centromeres, , and also with the *Dictyostelium discoideum* centromere.

So far, among the genomes analyzed, rice is the simplest monocotyledonous genome. The rice genome is composed of 12 chromosomes and has a total length of 430 mega bases (Mb). Because of its compact genome, synteny with other genomes, and its great economic value, rice

has been used as the model for plant genome analysis (Sharma *et al.*, 2008; Kurata *et al.*, 2002). Cytological studies have shown that the rice centromere is characterized by having seven repetitive elements. Tandem repeats are present in the core region and most of the centromere specific retrotransposons, and few genes are present in the peri centromeric region.

1. Tandem repeat family (CentO) with a 155 bp unit ranging from 65 kb to 2 Mb. The CentO is quantitatively variable among 12 rice centromeres (Cheng et al., 2002).
2. RCS2 is 168 bp short tandem repeat specific for *Oryza* species. This is present in only closely related species of rice compared to the remaining six repetitive elements which are present in all graminiae family (Kurata et al., 2002; Dong et al., 1998).
3. Centromere specific retrotransposon (CRR). These elements interrupt CentO elements irregularly (Cheng et al., 2002).
    a. Gypsy type retrotransposons are like RIRE3, RIRE7, RIRE8 and RCS1 are more common (Kurata et al., 2002).
    b. RCE1 is 1.9 kb unit, tandemly arrayed intervening sequences (Kurata et al., 2002).
4. Several putative genes, such as transforming growth facter (TGF)-beta receptor-interacting protein, putative endoplasmic reticulum retrieval protein (RER1), defective chloroplast, and leaf (DCL) protein are present in the rice centromeres (Cheng et al., 2002).

Ikuo *et al* identified CpG rich clusters in rice genomic sequences at the frequencies of one per 4.7 kb. The CpG sites are the region of DNA where the cytosine nucleotide is present next to the guanine nucleotide in a linear sequence along its length. These CpG containg sites are called CpG islands (2001).

Arabidopsis thaliana is a member of the *Brassicaceae* family and is so similar to most other plants in many aspects. It has a smaller genome compared to other species in this family. *Arabidopsis* chromosome 5 centomere is 0.1 Mb and the central domain consists of 180 base pair satellite repeats organized in tandem arrays that range from 0.4 to 1.4 Mb on different chromosomes and different athila retrotransposons; it forms the functional sequence of the *Arabidopsis* centromere (Ma *et al.*, 2001; Kumekawa *et al.*, 2001; Fang *et al.*, 2005). Ikuo *et al* identified CpG rich clusters in *Arabidopsis* genomic sequences at the frequencies of one per 4 kb. Plants with small genomes have these clusters associated with their genes. However, in plants with large genomes only few clusters are associated with genes (2001).

Lycopersicon esculentum (tomato) belongs to the *Solanaceae* species family. Among this family, the tomato has the smallest diploid genome. This family species show high conservation among each other, and thus, the tomato genome will serve as a "blue print" for other *Solanaceae*. The tomato chromosome 12 centromeric region is 1.9 Mb long. The *Lycopersicon* is characterized by having 162 bp tandem repeat TGR111 in the centromere. TGR111 is more predominate in the centromere. TGR1 and TGR11 are present only in lycopersicon species (Glockner *et al.*, 1998; Yang *et al.*, 2005).

Dictyostelium discoideum has a haploid genome. It consists of 6 chromosomes. It belongs to the group of social amoebas in the evolutionary branch of Amoebozoa. Each individual functional centromere is around 190 kilo bases (kb). All of the centromeres consist of 86% of highly repetitive elements. Also the centromere consists of tRNA gene-targeted retroelement (TRE) locate exclusively on the 3^1 or 5^1 end of tRNA genes (Andrew *et al.*, 2000). Previous studies showed that Dictyostelium Intermediate Repeat Sequence (DIRS) elements, Dictyostelium DNA Transposon (DDT) elements and the retrotransposon skipper contribute

30%-50%, 20% and 10%, respectively, to the overall length of the centromere. Inner centromere protein A is found in the centromere region (Glockner *et al.*, 2009).

Finally, from my study it was revealed the wild rice genome is closely related to the domesticated rice genome. Arabidopsis species is distantly related to the wild rice genome compared to the *Lycopersicon,* and the *Dictyostelium* species. If I had the full length centromeres of each species, I would have gotten clear idea of how many years ago these species are separated.

Materials and Methods

Centromere sequences:

The wild rice centromere fragments were obtained from Dr. Chungang Du's lab. The domesticated rice centromere sequences were obtained from "Rice genome annotation project website". These centromeres were identified using the CentO centromeric sequence, sequencing and flourescent in situ hybridization information. (http://rice.plantbiology.msu.edu/pseudomolecules/centromere.shtml). *Arabidopsis thaliana* chromosome 4 centromere sequences was obtained from NCBI and accession numbers is AB073166.1 and *Arabidopsis thaliana* chromosome 5 centromere sequence was obtained from EMBL nucleotide sequence database, accession number is AB046425. *Dictyostelium discoideum* chromosome 3 centromere genomic sequence were obtained from EMBL, accession number is FJ387222. *Lycopersicon esculentum* chromosome 12 centromeric sequence was obtained from NCBI, accession number is AY850394.1.

Annotation:

All these centromeres were further analyzed using several annotation softwares. First, RepeatMasker (Http://www.repeatmasker.org/cgi-bin/WEBRepeatmasker) (Smit and Green *et al*) was used to detect the known repeats (DNA transposons, simple repeats and retro elements such as SINEs, LINEs and LTR) present in the centromere sequences, using the appropriate repeat libraries for each species. Tandem repeats in these sequences are revealed by Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) (Benson *et al.*, 1999). GENSCAN

(http://genes.mit.edu/GENSCAN.html) (Burge *et al.*, 1997) was used to predict the genes in the repeat masked sequences. FGENESH (http://linux1.softberry.com/berry.phtml) was also used to predict the genes. Homology searches of the rice full-length cDNA sequences against the predicted genes were performed using BLASTN (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn) (Altschul et al., 1997) program was used to find the locations of genes.

Comparative tools:

Multiple sequence alignment tool MAFFT (http://align.bmr.kyushu-u.ac.jp/mafft/software/) (Katoh *et al.*, 2002) is used to align all the sequences. MultiPipMaker (http://pipmaker.bx.psu.edu/cgi-bin/multipipmaker) (schwartz *et al.*, 2000) and zPicture (http://zpicture.dcode.org/), (Ovachrenko et al, 2004) soft wares are user friendly based approaches for capturing, visualizing and disseminating comparative analysis of multi sequences. For all the comparisons, standard parameters of greater than 100 bp and homology greater than 70% were used to find the conserved non coding regions (CNS), genes and repeats. All these sequences were compared against the wild rice chromosome 8 centromere fragments.

RESULTS

Manual alignment of the wild rice chromosome 8 centromere fragments to the *Oryza sativa* chromosome 8 centromere:

I tried to align the wild rice chromosome 8 centromere fragments to the *Oryza sativa* chromosome 8 centromere because the *Oryza sativa* chromosome 8 centromere is the only one centromere sequence which has been sequenced fully so far. The wild rice E5P1 centromere fragment aligned at two regions and P2 cenrtomere fragment aligned at the one region. Centromere sequence fragments p2 and E5P1 are overlapped with some extent (Fig 1).

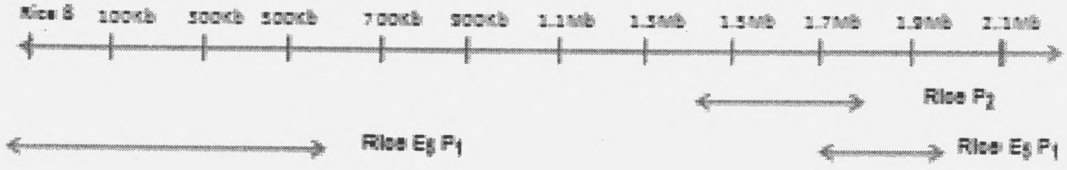

Figure 1: Manual alignment of the wild rice chromosome 8 centromere fragments with the domesticated rice chromosome 8 centromere .

This figure shows the alignment of wild rice chromosome 8 centromere fragments to the *Oryza sativa* chromosome 8 centromere. *Oryza sativa* chromosome 8 is fully sequenced centromere. The wild rice centromeres are partially sequenced.

Base composition, gene content and alignment with the wild rice chromosome 8 centromere fragments:

Once I have done alignment with the domesticated rice chromosome 8 centromere, I am curious to know what is in each centromere and which region of the each centromere is align with the wild rice chromosome 8 centromere fragments.

The *Oryza rufipogon* chromosome 8 centromere fragment (E5P1) has 58,516 base pairs (bp). It has a GC content of 49.84%. The percentage of bases masked is 61.64. It has 61.01% of class 1 retroelements, 0.59% of DNA transposons, and 0.04% of simple repeats. Ten genes were predicted by FGENESH. However, Blast program revealed chloroplast precursor protein (DCL) and chromatin organization modifier domain.

The *Oryza rufipogon* chromosome 8 centromere fragment (p2) is 15,751 bp long. The percentage of GC and bases masked were 45.64% and 79.44%, respectively. Class 1 retroelements and DNA transposons were 71.01% and 6.03% respectively. The FEGENESH predicted 3 genes. The Blast program revealed UDP-glucoronosyl and UDP-glucosyl transferase domain containing proteins, and cytochrome P450 enzyme.

The *Oryza sativa japonica* chromosome 1 centromere is 87,684 bp long. It has a 41.91% GC content. Four genes were predicted by FGENESH. However, blasting of this sequence in the Blastx program revealed chloroplast precursor protein. RepeatMasker analysis of this centromere revealed that 24.97% of the bases were masked. Class 1 retroelements occupied 24.97%. This sequence aligned to the wild rice E5P1 and P2 fragmenst up to 14,000 bp and 3,000 bp, respectively. However, there were lots of gaps in between.

The *Oryza sativa japonica* chromosome 2 is 68,673 bp long. No repetitive sequences were predicted by RepeatMasker and no genes were predicted by FGENESH. The wild rice

chromosome 8 centromere E5P1 fragment aligned with the domesticated rice chromosome 2 centromere. No alignment with the rice P2 centromere fragment.

The *Oryza sativa japonica* chromosome 3 centromere is 112, 886 bp long. It has a GC content of 45.28%. RepeatMasker analysis of this sequence revealed that it has 65.99% of class 1 retroelements, and 2.15% DNA transposons. FGENESH predicted 21 genes for chromosome 3. However, only tumor differentially expressed (TMS) protein was identified by Blast. MAFFT aligned this sequence with the wild rice E5P1 sequence from 0 to 40 kb and from 53 to 58 kb. However, there were lots of gaps found. P2 centromere fragment aligned here and there.

The *Oryza sativa japonica* chromosome 4 centromere is 143,702 bp long. It has a 43.18% GC content. It has 58.18% of class I retroelements. 20 genes were predicted by FGENESH. Only chloroplast precursor protein was revealed by Blast program. MAFFT revealed that this sequence aligned with the wild rice sequences from 0 to 58 kb. However there were lots of mismatches in between.

The *Oryza sativa japonica* chromosome 5 centromere is 62,350 bp long. RepeatMasker analysis of this sequence revealed that it has 9.4% of class 1 retroelements, and 1.44% DNA transposons. The Blast program identified 4 genes. They are core histone H2A/H2B/H3/H4 domain containing protein, kelch domain containing protein, jacalin-like lectin domain containing protein, and COBW domain containing proteins. The number of genes predicted by FGENESH was also 4. MAFFT revealed that rice 5 centromere is aligned from 0 to 5 kb and from 52 to 58 kb of the wild rice E5P1 centromere fragment. No alignment with the wild rice P2 centromere fragment.

The *Oryza sativa japonica* chromosome 6 centromere is 91,579 bp long. The GC content is 43.27%, and 47.18% of bases were masked. RepeatMasker analysis revealed that it has 38.57%, and 7.86% of the class 1 retroelements, and DNA transposons, respectively. It has 1 prefoldin protein, 1 protein of 1,4-alpha-glucan-branching enzyme, CGMC_MAPKCMGC_2.10 (CGMC includes CDA, MAPK, GSK3, and CLKC) kinases and autophagy related proteins were identified through blast. FGENESH predicted 13 genes. This sequence is aligned with the entire wild rice 8 centromere sequence with lots of gaps in between.

The *Oryza sativa japonica* chromosome 7 centromere is 65, 793 bp long. Masked bases and GC content is 1.07%, and 45.6%, respectively. 1.07% of the class 1 retroelements were identified. Only hypothetical proteins were identified through Blast and there were no genes predicted by FGENESH. This sequence is aligned with the entire wild rice 8 centromere sequence with lots of gaps in between.

The *Oryza sativa japonica* chromosome 8 centromere is 1, 972 kb long. It has a GC content of 45.22%. Analysis of rice 8 centromere with the Blastx software revealed that it has 3 genes. They are chloroplast precursor (DCL) protein, Rhamnogalacturonate lyase, and RER1. GENSCAN predicted 21 genes. RepeatMasker masked 37.54% bases in the entire genome. Among them, 35.44% were class 1 retroelements, 1.59% DNA transposons, and the remaining 0.51% was occupied by simple repeats, low complexity regions, and small RNAs. This sequence aligned with the wild rice chromosome 8 centromere E5P1 fragment from 0 to 10 kb and from 50 to 58 kb. This sequence aligned with the entire P2 centromere fragment with lots of gaps in between.

The *Oryza sativa japonica* chromosome 9 centromere is 32,982 bp long. No repetitive sequences were detected by RepeatMasker. The only gene identified by Blast is that of a oligo peptide transporter protein. Also, there were no predicted genes by FGENESH. This centromere aligned with wild rice chromosome 8 centromere E5P1 fragment from o to 7 kb and from 52 to 58 kb. No alignment was observed with the wild rice P2 centromere fragment.

The *Oryza sativa japonica* chromosome 10 centromere is 139,398 bp long. The GC content is 47.15%, and 74.65% of the bases were masked. Among them, 71.21% are class 1 retroelements, and DNA transposons occupy 2.70%, the remaining repeats are simple repeats, low complexity regions. 23 genes were predicted by FGENESH. However, the Blast program identified a CUE domain containg protein. MAFFT program showed that this sequence aligned with the wild rice chromosome 8 centromere from 0 to 58 kb. However there were lots of mismatches in between.

The *Oryza sativa japonica* chromosome 11 centromere is 44,813 long. The GC content is 41.33%, and 54.51% of bases were masked. RepeatMasker analysis revealed that this centromere has 42% ofclass 1 retroelements, and 12% of DNA transposons. The Blast program identified chloroplast precursor. FREGENSH predicted 7 genes. This sequence aligned with the entire wild rice E5P1 centromere fragment. However, there was lot of mismatches in between. No alignment was observed with the wild rice P2 centromere fragment.

The *Oryza sativa japonica* chromosome 12 centromere is 92,644 bp long. The GC% is 45.11%. RepeatMasker masked 41.39% bases. It consists of 37.22% of class 1 retroelements , and 3.79% of DNA transposons. Cytochrome p450 was identified by the Blast program.

FREGENSH predicted 5 genes. This sequence aligned with the entire wild rice sequences with gaps.

Arabidopsis chromosome 5 centromere is 123 kb long. It has 36.4% of GC. RepeatMasker analysis masked 0.78% bases. It has only simple repeats and low complexity regions. Galactinol synthetase, extensin like protein, and AP 2 domain proteins were identified by blast program. FGENESH predicted 21 genes. According to the MAFFT program, this sequence aligned here and there to the wild rice E5P1centromere fragment.

Lycopersicon chromosome 12 centromere is 1,972,546 bp long. Its GC content is 44.22%. RepeatMasker masked 4.18% bases. It has 2.23% of class 1 retro elements. Blast program identified cytochrome c biogenesis protein, and extension proteins. FGENESH predicted 49 genes. According to the MAFFT program, this sequence aligned here and there to the wild rice centromere fragments.

Dictyostelium centromere is 191,675 bp long. Percentage of GC is 30.54% and 7.02% bases were masked by RepeatMasker. Class 1 retroelements occupied 1.22%. FGENESH predicted 68 genes. No genes were identified by Blast.

Summary of the repeats and the genes:

Most of the rice centromeres that I have analyzed so far have repeats in between 40 to 80%. However, the domesticated rice 1, 2, 5 and 9 chromosomes centromeres have 22%, 0%, 11%, and 0% of the repeats, respectively. However, the percentage of repeats in other species like *Arabidopsis, Lycopersicon,* and *Dictyostelium* range from 0.5 to 10%. Most of the centromeres have *Gypsy LTR* retrotransposons rather than copia LTR retrotransposons. The

percent GC content in *Arabidopsis, Lycopersicon,* and *Dictyostelium* ranges between 30-45%. In contrast, all the rice centromeres have more than 45% GC content (Table 1).

Some centromeres like rice 6, 8, 10, 11, P2 and *Dictyostelium* have Penelope SINES and some centromeres like rice E5P1, 3, 6, 8, 10, 11, and 12 have DNA transposons. Total interspread repeats are lesser in *Arabidopsis, Lycopersicon, Dictyostelium* and rice 5, 7, 8 centromeres than other rice centromeres. There are more low complexity regions and simple repeats in *Lycopersicon* and *Dictyostelium* than in other species (Table 1).

Table 1: Retroelements detected in the centromeres

Name	Total Bases (bp)	GC Level(%)	Masked bases (%)	Non LTR retrotransposons	LTR retrotransposons	DNA transposons	Simple repeats	Low complexity regions
Rice 1	87684	41.91	24.97		10	1		
Rice 2								
Rice 3	112885	45.28	69.34		42	12	14	4
Rice 4	143702	43.18	58.19		30			1
Rice 5	62350	41.14	11.95		3		3	5
Rice 6	91579	43.27	47.18	3	47	34	4	7
Rice 7	65793	40.65	1.07		3			
Rice 8	1972546	45.22	.10	2	4	1		
Rice 9								
Rice 10	139398	47.1	74.65	3	32	17	13	4
Rice 11	44813	41.33	54.51	3	18	18	1	3
Rice 12	92644	45.11	41.39		16	4	4	5
Arabidopsis	123460	36.40	.78				6	9
Lycopersicon	1972546	45.22	4.8		67		255	377
Dictyostelium	191675	30.54	7.02	1	15		73	76
Rice P2	15751	45.64	79.44	1	2	4		1
Rice E5P1	58516	49.84	61.64		12	2	1	

Genes were predicted by FGENESH and GENSCAN software. The Blast program was used to find the proteins. Most of the centromeres have genes. However, there were no genes conserved in all the sequences. Most of the rice chromosome centromeres have chloroplast precursor proteins, and cytochrome p 450 proteins. The rice 2 chromosome centromere did not have genes. The rice chromosome 7 centromere and *Dictyostelium* centromere had only hypothetical proteins. The rice 6 chromosome centromere has a large number of proteins than other centromeres (Table 2).

Table 2: Genes detected in the centromeres

Sequence name	Blast identified genes
Rice 1	chloroplast precursor proteins
Rice 2	None
Rice 3	TMS protein/ tumor differentially expressed proteins
Rice 4	chloroplast precursor proteins
Rice 5	Core histone H2A/H2B/H3/H4 domain containing protein; kelch domain containing protein, jacalin-like lectin domain containing protein; COBW domain containing proteins.
Rice 6	Zinc knuckle containing protein; 1,4-alpha-glucan-branching enzyme; CGMC_MAPKCMGC_2.10 (CGMC includes CDA, MAPK, GSK3, and CLKC) kinases; prefoldin, lecithin:cholesterol acyltransferase, aminotransferase,
Rice 7	Hypothetical protein
Rice 9	Oligopeptide transporter -like protein
Rice 10	CUE domain containing protein
Rice 11	Adenylate kinase; sulfate transporter 4.1, chloroplast precursor; caleosin related protein; Aminotransferase
Rice 12	Cytochrome P450
Arabidopsis	Galactinol synthetage; Extensin like protein; AP 2 domain; Transcriptional Activator
Lycopersicon	Ccytochrome c biogenesis protein; extension enzyme; putative transcription regulator CPL1

Dictyostelium	Hypothetical proteins
Rice E5P1	Chloroplast precursor protein; Chromatin organization modifier
Rice P2	UDP-glucoronosyl and UDP-glucosyl transferase domain containing protein; cytochrome P450 enzyme

Distribution of the retroelements, genes and satellite sequences:

Further, I am interested to see not only where do the genes and repeats present in the centromeres but also look for location of the centromeric specific repeats, which may provide a way to identify the region of centromere that the partially sequenced chromosome came from. By doing so, it was revealed that the *Oryza sativa* chromosome 2, and 9 centromeres were entirely from the core region because they have only Cent0 satellite repeats. The Oryza *sativa* 6, and 11 chromosome centromeres, and the wild rice P2 centromere fragments are entirely from flanking regions because they have only retrotransposons. The remaining centromeres have both satellite repeats and retrotransposons (Fig 2).

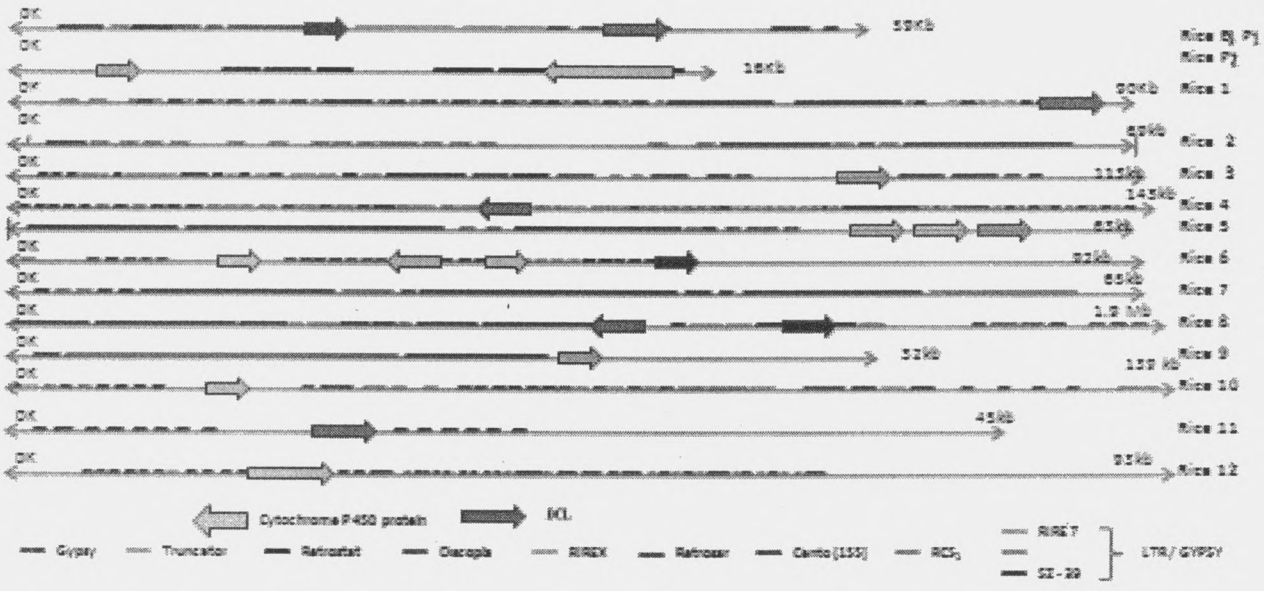

Figure 2: Annotation map of the centromeric regions of the rice chromosomes

This picture shows the location of the tandem repeats, retroelements and genes in each centromere. In this figure, RCS2 and CentO satellite repeats presence indicates that this region is from core region. Only transposable elements are seen in pericentromeric region. Genes are indicated by arrows. Direction of the arrow indicates the orientation of the gene.

Comparative analysis of all the centromeres against the wild rice chromosome 8 centromere:

After aligning each centromere with the wild rice chromosome 8 centromere fragments, I wanted to see if there is a repeat or a gene common in all the sequences. I submitted all the sequences to the zPicture browser in FASTA format. All the sequences were compared against the wild rice chromosome 8 centromere fragments. Outcome results indicated that there are no evolutionary conserved regions (ECR) between centromere fragment P2 and domesticated rice chromosome 2, 5, 9, 11 centromeres, *Arabidopsis* centromere and *Dictyostelium* centromere. ECRs are more in between P2 fragment and the domesticated rice 3, 6, 8, 10, 12 chromosome centromeres and Lycopersicon centromere than the domesticated rice chromosomes 1, 4, 5

centromeres (Fig 3b). The wild rice E5P1 centromere fragment showed more ECR between all the centromeres except with *Dictyostelium* centromere (Fig 3a). There are no conserved genes, exons, and UTRs in these sequences.

Figure 3: Figure showing evolutionary conserved regions (ECR)

zPicture analysis for the comparision of the all the centromeres against the wild rice chromosome 8 centromere fragments E5P1 (left) and P2 (right). In this figure, default parameters (>100 bp/>70% ID) were used to highlight intronic (pink) and intragenic (red) conserved elements. Black lines indicate the alignment. Length of the black line is proportional to the length of ungaped region. Height of the graph is proportional to the percent identity.

These results are conformed by MultiPipMaker software. All the centromeres are compared against the wild rice chromosome E5P1 fragment through MultiPipMaker. All light

orange boxes are conserved non coding regions which are present in all the sequences except *Arabidopsis*, *Lycopersicon* and *Dictyostelium* centromeres (Fig 4). MultiPipMaker also showed conserved CpG islands which are indicated by light dark gray boxes (CpG/CpC >0.75) and white boxes (CpG/CpC < 0.75). The CpG sites are the region of DNA where the cytosine nucleotide is present next to the guanine nucleotide in a linear sequence along its length. These CpG containg sites are called CpG islands. MultiPipMaker software gives outcome results only if all the given sequences are aligned with the comparing sequence. I did not get the outcome results when I submitted all the sequences against the wild rice P2 fragment because fragment p2 had no alignment with rice chromosomes 2, 5, 9, 11 centromeres , *Arabidopsis* centromere and *Dictyostelium* centromere.

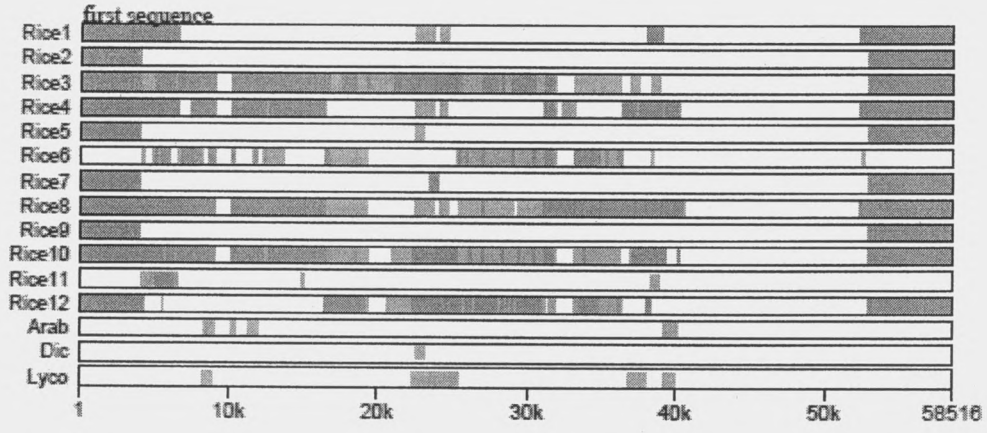

Figure 4: percentage identity plot (PIP) showing conserved non coding regions by comparing all the centromeres against the wild rice chromosome 8 centromere E5P1 fragment

In this figure, all conserved non coding regions are showed as light pink color. Light green region indicates the presence of genes.

PhylogeneticTree:

Next, I wanted to know the relationship between all the centromeres. Phylogenetic tree is a convenient way of representing evolutionary relationships among a group of organisms. I attempted to construct a phylogenetic tree using all the centromere sequences. All the rice centromere sequences which are aligned to the wild rice E5P1 centromere fragment from 101 bp to the 350 bp were used to construct the phylogenetic tree. The MAFFT create trees by using the simpler algorithms (i.e. those based on distance) of tree construction.

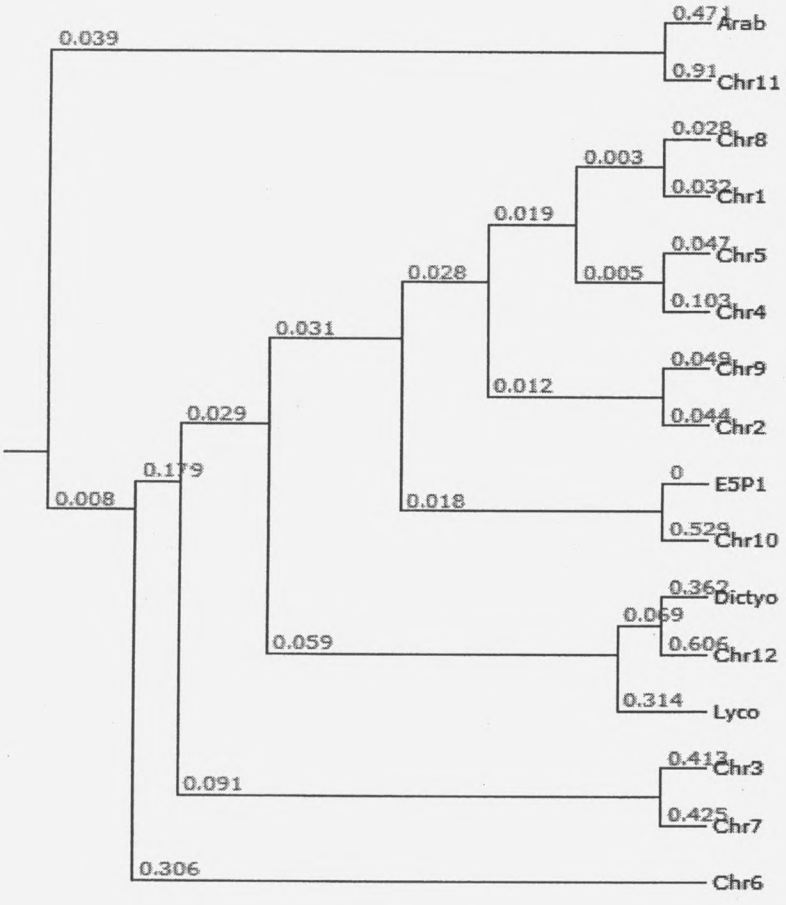

Figure 5: Phylogenetic relationship between all the centromeres used in this study

This figure shows the relationship among the wild rice, the domesticated rice, the *Arabidopsis*, the *Lycopersicon*, and the *Dictyostelium* centromeres.

Discussion

What defines a centromere? A centromere is the characteristic landmark and appears as a constriction on chromosomes during metaphase. The centromere plays an important role in proper segregation of sister chromatids to opposite poles during cell division. Errors in the centromere function lead to aberrant division and chromosomal instability, both of which are often observed in cancerous cell.

Why do we study the centromere? Considering the essential roles that centromeres play in mitosis, they are truly at the center of this process. Previous experiments have shown that the centromere is far more than simply a "primary constriction" in a chromosome. Later experiments proved that centromeres are dynamic assemblies of chromatin of which there are specialized functional regions existing within each centromere. Given the large amount of current research interest in centromere biology, it seems likely that the centromere has a few more surprises in store for us.

Centromere function is the same in all species. What is common in all these centomeres? This question has yet to be answered. In aiming to produce a comprehensive inventory of centromere components in plants and slime mold, I have selected *Germinae* (rice), *Solanaceae* (tomato), *Brassicaceae* (*Arabidopsis*) and *Dictyostelium discoidenum* (slime mold) species. I

compared all these sequences against the wild rice chromosome 8 centromere which came from Dr.Chunguang Du's lab.

The rice is the simplest monocotyledonous genome so far analyzed. The *Arabidopsis* is so similar to most other plants in many aspects. It has a smaller genome compared to other species in this family. Among the *Solanaceae* family, the tomato has the smallest diploid genome. The *Dictyostelium discoidenum* has a haploid genome.

The wild rice centromere fragments are aligned with the domesticated rice chromosome 8 centromere because the domesticated rice chromosome 8 centromere is the only one which has been sequenced fully so far. By doing these alignments, it was revealed that these two wild rice centromere fragments came from different region of the rice centromere and also observed some extent of overlap between these two fragments.

Preliminary annotations were done by RepeatMasker, GENSCAN, and FGENESH software programs. In addition, the Blast program was used to find the location of currently available centromere genes from a phylogenetically diverse species with complete or nearly complete genome sequences. Retroelements are mobile genetic elements. They can exist as either DNA or RNA. They contain a reverse a transcriptase gene. These retroelements are considered as primarily selfish or parasitic organisms which play an important role in evolution. Even though having these retroelements has not done any harm to the host, mutations caused by these elements lead to various diseases. From the preliminary annotations, it was revealed that the rice chromosome centromeress 2 and 9 centromeres have 0 repeats. And the rice chromosomes 1 and 5 centromeres, the wild rice E5P1 centromere fragment, have very few repeats. The percentage of repeats in *Arabidopsis, Lycopersicon,* and *Dictyostelium,* range from 0.5 to 10. There are no genes common in all the centromeres. However, most of the rice

centromeres have chloroplast precursors. Chloroplast proteins are synthesized on the cytosolic ribosomes as the precursor proteins. It was proven that these proteins are located near to the centromeric domain (Ma *et al*, 2007). All the centromeres used in this study aligned with the wild rice E5P1 centromere fragment. However, there were no alignments among the wild rice P2 centromere fragment and the domesticated rice chromosomes 2, 5, 9, 11 centromeres and also among the wild rice P2 centromere fragment and the *Arabidopsis* and the *Dictyostelium* centromeres.

Eventhough, all the centromeres have not had the same genes and repeats, all the centromeres showed alignment with the wild rice E5P1 centromere fragment. This stimulated me to further annotate the centromere sequences. In this process I came to know that each centromere has species specific characteristics. The rice centromere is characterized by having CentO, RCS2, and RCS1 satellite repeats in the core region which is interrupted by a centromere specific retrotransposon. Centromere flanking regions are characterized by having a centromere specific retrotransposon (Kurata *et al*, 2002; Dong *et al*, 1998; Cheng *et al*, 2002). The rice centromeres I used in this study have CentO and RCS2 satellite repeats and centromere specific retrotransposons. The *Arabidopsis* centromere is characterized by having a 180 BP tandem repeat in the core region, and *Ty3/gypsy*-type retrotransposons and middle repetitive sequences in the flanking region (Copenhaver et al 1999). The *Arabidopsis* sequence I used for this research did not have 180 BP tandem repeats. However, the *Arabidopsis* sequence had Ty3/*gypsy*-type retrotransposons. The *Lycopersicon* centromere is characterized by the presence of a TGR1 (162 BP) tandem repeat in the core region and retroelements in the flanking region. Having only retroelements in the *Lycopersicon* centromere, used in this study, indicates that this region is might be from the flanking region of centromere.

Of further interest, is the location of the retroelements, centromere specific tandem repeats and genes in each chromosome centromeric region. Previous studies showed that the core regions of the rice centromere have CentO satellite repeats and RCS2 repeats (Cheng *et al*, 2002; Kurata *et al*, 2002; Dong *et al*, 1998). Centromere specific retrotransposon (CRR) elements interrupt CentO elements irregularly (Cheng *et al*, 2002). One of my research objectives is to determine if the partially sequenced chromosome centromeres are from the core region or from the flanking region. Interestingly, by finding the location of repeats, genes and tandem repeats in the chromosomes, it was revealed that chromosomes 1, 2, 5, 7,9, and 12, centromeres, all have CentO satellite repeats and RCS2 repeats. This indicates that these sequences are from the core region of the centromere. Rice chromosomes 2 and 9 centromeres did not have retroelements because they are completely from the core region. There were fewer number of retroelements found in Rice chromosomes 1 and 5 centromeres because most of these centromere sequences are also from the core region. Chromosomes 6 and 11 centromeres have *LTR/Gypsy* type retroelements. This indicates that this region might be from the flanking region of the centromere. Chromosomes 3 and 4 centromeres have more centromere specific retrotransposon (CRR).

In order to detect similarities and differences in these centromeres, I used comparative tools. From the zPicture output, it was revealed that no evolutionary conserved regions (ECR) between centromere fragment P2 and domesticated rice chromosome 2, 5, 9, 11 centromeres, *Arabidopsis* centromere and *Dictyostelium* centromere. ECRs are more in between P2 fragment and the domesticated rice 3, 6, 8, 10, 12 chromosome centromeres and Lycopersicon centromere than the domesticated rice chromosomes 1, 4, 5 centromeres (Fig 3b). The wild rice E5P1

centromere fragment showed more ECR between all the centromeres except with *Dictyostelium* centromere (Fig 3a). There are no conserved genes, exons, and UTRs in these sequences.

Multipip maker results revealed that there is a conservation of CpG islands in all the sequences that were compared. Ikuo *et al* proved that genes are associated with CpG islands. Genes that are associated with CpG islands are mostly located in the promoter region. All these findings indicate that even though there are different genes in all the centromere sequences that were compared, these genes are located at the same location. Even though all the sequences do not have the same gene at the same location, all the sequences definitely have either one or the other gene at the same place. This was confirmed by observation of conservation of an intergene region within the same species. From this, it is clear that gene regions are mutated slowly compared to the other region.

Using the MAFFT software, a phylogenetic tree for the chromosomes was built. An interesting feature of phylogenetic relationships is that it allows one to obtain a phylogenetic relationship between species based upon similarities and differences in their physical and/or genetic characteristics. From these phylogenetic trees it was revealed that the *Lycopersicon*, the *Dictyostelium* and the domesticated rice are separated from the wild rice in less amount of time compared to the *Arabidopsis*.

Finally, my work shows that each centromere has species specific tandem repeats. Types of repeats are the same within the species. Both the domesticated and the wild rice centromeres share more structural similarities than the wild rice centromere and the Arabidopsis, The Dictyostelium , the Lycopersicon centromeres. Conservation of CpG islands and intergene region indicates that gene rich areas are evolving more slowly than gene rare areas. The phylogentic studies revealed that the *Arabidopsis* is distantly related to the wild rice compared to

the *Lycopersicon*, the *Dictyostelium* and the domesticated rice. Building on the structural similarities found in this study, additional sequencing of the wild rice genome, the domesticated rice genome may provide additional insight into their genome analysis.

References

Ashikawa, I. (2001). Gene associated CpG islands in plants as revealed by analyses of genomic sequences. *The plant Journal*, 26(9), 617-625.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acides Res*, 27(2), 573-80.

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol,* 268, 78-94.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, R.C., & Gu, M. (2002). Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. *The plant cell*, 14, 1691-1704.

Copenhaver, G.P. (2004). Who's driving the centromere?. *Journal of Biology*, 3, 17.

Copenhaver, G.P., Nickel, K., Kuromori, T., Benito, M., Kaul, S., & Lin, X. (1999). Genetic Definition and Sequence Analysis of *Arabidopsis* Centromeres. *Science,* 286, 2468-2474.

Dawe, R. K., Reed, l.M., Yu, H., Muszynski, M. G., & Hiatt, E. N. (1997). A Maize Homolog of Mammalian CENPC is a Constitutive Component of the Inner Kinetochore. *Plant cell*,11,1227-1238.

Dong, F., Miller, J. T., Jackson, S. A., Wang, G. L., & Ronald, P. C. (1998). Rice (Oryza sativa) centromeric regions consist of complex DNA. Pro Natl Acad Sci, 85, 8135-8150.

Dubchak, I., Brudno, M., Loots, G. G., Pachter, L.,..... Mayor, C. (2000). Active Conservation of Non coding Sequences Revealed by 3-way Species Comparisons. *Genome Research*, 10, 1304.

Fang, Y., & Spector, D. L. (2005). Centromere position dyanamics in living arabidopsis plant. *Mol Biol Cell*, 16(12), 5710–5718.

Glöckner, G., & Heidel, A. J. (2009). Centromere sequance and dynamics in the *Dictyostelium dicoideum*. Nucleic Acids Res, 37(6), 1809–1816.

Hall, S. E., Kettler, G., & Preuss, D. (2002). Centromere Satellites from *Arabidopsis* populations: Maintenance of Conserved and Variable Domain. *Genome research*, 13, 195-205.

Henikoff, S., Ahmad, k., & Malik., H. S. (2001). The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science Mag*, 293.

Jiang, J., Birchler, J. A., Parrott, W. A., & Dawe, R. K. (2003). A molecular view of plant centromeres. *Trends Plant Sci*, 8(12), 570-5.

Karniski, L. P. (2001). Mutations in the diastrophic dysplasia sulfhate transporter (DTDST) gene: correlation between sulfate transport activity and chondrodysplasia phenotype. *Human Molecular Genetics*, 10(14), 1485-1490.

Katoh, K., Misawa, k., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059-3066.

Kingsbury, J., & Koshland, D. (1993). Centromere function on mini chromosomes isolated from budding yeast. *Mol Biol Cell*, 4(8), 859–870.

Kitagawa, K., & Hieter, P. (2001). Evolutionary conservation between budding yeast and human kinetochores. *Nat. Rev. Mol. Cell biology*, 2, 678-687.

Kumekawa, N., Hosouchi, T., Tsuruoka, H., & Kotani, H. (2001). Size and sequence organization centromeric region of *Arabidopsis* chromosome 4. *DNA research*, 8, 285-290.

Kurata, N., Nonomura, K.,& Harushima, Y. (2002). Rice genome organization: the centromere and genome interactions. *Ann Bot (Lond)*, 90(4), 427-35.

Lamb, J. C., & Birchler, J. A. (2003). The role of DNA sequence in centromere formation. *Genome biology*, 4, 214.

Ma, J. (2007). Plant centromere organization: a dynamic structure with conserved function. *Trends in genetics*, 23(3).

Ma, j., Wing, R. A., Bennetzen, J. L., &Jackson, S. A. (2007a). Evolutionary History and Positional Shift of a Rice Centromere. *Genetics*, 177, 1217-1220.

Ma, J., wing, R. A., Bennetzan, J. L., & Jackson, S. A. (2007b). Plant centromere organization: a dynamic structure with conserved functions. *Trends Jenet*, 23(3), 134-9.

Meluh, P. B., & Koshland, D. (1997). Budding yeast centromere composition and assembly as revealed by in vivo cross-linking. *Genes Dev*, 11(24), 3401-3412.

Moore, P. (2004). Making sense of centromeres. *J Biol*, 3(4), 16.

Ovcharenko, I., Boffelli, D., & Loots, G. (2004). eShadow _A Tool for Comparing Closely Related Sequences. *Genome research*, 14(6), 1191-1198.

Ovcharenko, I., Loots, G. G., Hardison, R. C., Miller, M., & Stubbs, L. (2004). zPicture: Dynamic Alignment and Visualization Tool for Analyzing Conservation Profiles. *Genome Research*, 14(3), 472-477.

Saffery, R., Earle, E., Irvine, D. V., Kalitsis, p., & Choo, K. H. A. (1999). Conservation of centromere proteins in vertebrates. *Chromosome Research*, 7, 261-265.

Schwartz, S., Zhang, z., Frazer, K. A., Smit, A., & Riemer. C. (2000). PipMaker—A Web Server for Aligning Two Genomic DNA Sequences. *Genome research*, 10(4), 577.

Sharma, A., & Presting, G. G. (2001). Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity. *Mol Genet Genomics,* 279, 133–147.

Smith, M. F. (2006). Comparative genomics: The economies of evolution. *The Heridity*, 96, 109.

Sorrells, M.E., Rota, M. L., & Bermudez-Kandianis, C. E. (2003). Comparative DNA Sequence Analysis of Wheat and Rice Genomes. *Genome Res,* 13, 1818-1827.

Starr, D. A., Williams, B. C., Li, Z. B., Etemad-Moghadam, B., Dawe, K.R., & Goldberg, M. L. (1997). Conservation of the Centromere/Kinetochore Protein ZW10. *J Cell Biol,* 138, 1289-1301.

Talbert, P. B., Bryson, T. D., & Henikoff,S. (2004). Adaptive evolution of centromere proteins in plants and animals. *J Biol*, 3(4).

Villasante, A., Abad, J. P., & Mendez-Lago, M. (2007). Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. *Proc. Natl. Acad. Sci. USA,* **104,** 10542–10547.

Wu, J., Yamagata, H., Hayashi-Tsugane, M., Hijishita, S., Fujisawa, M., & Shibata, M. (2004). Composition and structure of the centromere of rice chromosome 8. *The plant cell*, 16, 967-976.

Yang, T. J., Lee, S., Chang, S. B., Yu, Y., Jong, H. D., & Wing, R. A . (2005). In depth sequence analysis of tomato 12 chromosome centromeric region: identification of a large CAA block and characterization of pericentromere retrotranposons. *Chromosoma*, 114, 103–117.

Yan.H., & Jiang, J. (2007). Rice as a model for centromere and heterochromatin research. *Chromosome Research*, 15, 77-84.

Yan, H., Talbert, P. B., Lee, H. R., Jett, J., Henikoff, S., Chen, F., & Jiang, J. (2008). Intergenic Locations of Rice Centromeric Chromatin. *Plos Biology*, 6(11).

Zhang, Y., Huang, Y., Zhang, L., Li, Y., Lu, T., Lu, Y., Feng, Q., Zhao, Q., Cheng, Z., Xue, Y., Wing, R. A., & Han, B. (2004). Structural features of the rice chromosome 4 centromere. *Nucleac Acids Res*, 32(6), 2023-2030.

http://en.wikipedia.org/wiki/Comparative_genomics

http://mips.helmholtz-muenchen.de/plant/tomato/

http://www.nsf.gov/pubs/2002/bio0202/model.htm