

WASHINGTON UNIVERSITY

Division of Biology and Biomedical Sciences

Biochemistry Program

Dissertation Committee:  
Sean R. Eddy, Chairperson  
Douglas E. Berg  
Warren R. Gish  
Stephen L. Johnson  
Allan Larson  
John E. Majors  
Timothy Schedl

Computational Identification and  
Characterization of Repeats  
in Sequenced Eukaryotic Genomes

by

Zhirong Bao

A dissertation presented to the  
Graduate School of Arts and Sciences  
of Washington University in  
partial fulfillment of the  
requirements for the degree  
of Doctor of Philosophy

December 2002  
Saint Louis, Missouri

Copyright © by

Zhirong Bao

2002

# Acknowledgments

I would like to first thank my thesis advisor, Dr. Sean R. Eddy, for taking me into the exciting new field of computational genome analysis. I have benefited tremendously from the fine balance of the academic independence and guidance that Sean nurtures for his group, especially the way he encourages his students to stand up to him in scientific discussions.

I would also like to thank members of my thesis committee, for their help in my research, their effort to keep me rooted in real biology and their tolerance of my impatience.

Part of the dissertation was done in collaboration with Dr. Susan Wessler's group at University of Georgia, Athens and Dr. Susan McCouch's group at Cornell University. I would thank Dr. Wessler, who initiated the collaboration when RECON was not much more than an idea, for her confidence in me. Much of my knowledge about transposable elements was learned from the Wessler group, especially from Dr. Ning Jiang.

I could not have done my research without the help from members of Eddy group. In particular, I would like to thank Dr. Elena Rivas, Dr. Todd Lowe, Robin Dowell, Robert Klein and Thomas Jones for their help with my computer and math questions and their enthusiastic discussion on so many fascinating scientific problems.

Finally, I would like to dedicate this dissertation to my parents, for their love and for nurturing my curiosity from the very beginning.

# Contents

<b>Acknowledgments</b>	<b>ii</b>
<b>Abstract</b>	<b>vii</b>
<b>1 Introduction: Repetitive Sequences And Their Impacts on The Genome</b>	<b>1</b>
1.1 Repetition is Not Necessarily Boring . . . . .	2
1.2 The Dynamic Nature of Repeats . . . . .	3
1.2.1 Tandem repeats . . . . .	3
1.2.2 Transposable elements . . . . .	4
1.2.3 Segmental duplications . . . . .	6
1.3 The Evolution of Repeats and the Evolution of the Genome . . . . .	7
1.3.1 The functionalist view . . . . .	7
1.3.2 Towards a more comprehensive view . . . . .	8
1.4 Challenges and Opportunities in the Genomic Era . . . . .	13
1.4.1 The temporal aspect: quantitating the evolutionary dynamics of repeats . .	13
1.4.2 The spatial aspect: genome organization . . . . .	15
1.5 Scope of Dissertation . . . . .	16
<b>2 Automated <i>de novo</i> Identification of Repeat Sequence Families in Sequenced Genomes</b>	<b>18</b>
2.1 Abstract . . . . .	19
2.2 Introduction . . . . .	19
2.3 Results . . . . .	21
2.3.1 The Existing Single Linkage Clustering Algorithms . . . . .	21
2.3.2 The RECON Algorithm . . . . .	29
2.3.3 Assessment . . . . .	30
2.4 Discussion . . . . .	35
2.5 Methods . . . . .	37
2.5.1 Components of RECON . . . . .	37
2.5.2 Implementation details . . . . .	40
2.5.3 Human Genome Analysis . . . . .	42
2.6 Acknowledgments . . . . .	42

<b>3</b>	<b>Rlib: a Database of Automatically Constructed Repeat Sequence Libraries for Sequenced Genomes</b>	<b>43</b>
3.1	Abstract . . . . .	44
3.2	Introduction . . . . .	44
3.3	Results . . . . .	45
3.3.1	Repeat Library Construction: Strategy and Practical Issues . . . . .	45
3.3.2	Pilot Experiments . . . . .	54
3.3.3	Use of Rlib: understanding repeat evolution . . . . .	55
3.4	Discussion . . . . .	60
3.4.1	More to Come . . . . .	60
3.4.2	Usage of the Rlib libraries . . . . .	63
3.5	Methods . . . . .	63
3.5.1	Input Sequences . . . . .	63
3.5.2	Identification of Repeat Families . . . . .	64
3.5.3	Inference of Consensus Sequences . . . . .	64
3.5.4	Genome Survey and Repeat Classification . . . . .	65
3.6	Acknowledgment . . . . .	65
<b>4</b>	<b>Repeats and Their Insertion Polymorphism in the Rice Genome</b>	<b>66</b>
4.1	Abstract . . . . .	67
4.2	Introduction . . . . .	67
4.3	Results . . . . .	68
4.3.1	Identification and classification of repeat families . . . . .	68
4.3.2	Physical distribution of repeats in the genome . . . . .	70
4.3.3	Insertion polymorphism between the two sequenced cultivars . . . . .	71
4.3.4	Transposon Display reveals extensive insertion polymorphism among cultivars and wild species . . . . .	75
4.4	Discussion . . . . .	78
4.5	Methods . . . . .	80
<b>5</b>	<b>Concluding Remarks</b>	<b>82</b>
5.1	Remarks on RECON and Rlib . . . . .	83
5.2	Remarks on the Genomics of Transposons . . . . .	83
5.2.1	Mechanisms for transposition . . . . .	84
5.2.2	Evolutionary timing of transposition . . . . .	86
	<b>Appendix A: List of Publications</b>	<b>88</b>
	<b>Appendix B: An Assessing Method for <i>de novo</i> Repeat Identification</b>	<b>89</b>

# List of Tables

2.1	Definition of the Alu Family . . . . .	32
2.2	The Larger Human Repeat Families Defined by RECON . . . . .	34
3.1	Summary of the Pilot Experiments for Rlib . . . . .	54
3.2	Similarity of Repeats Across Species . . . . .	58
3.3	The Initial Targets for Rlib . . . . .	61

# List of Figures

1.1	The original Britten plot . . . . .	10
1.2	History of transposition in Human and mouse . . . . .	12
2.1	Flowchart of the <i>de novo</i> strategy for repeat family identification . . . . .	22
2.2	Complications in the inference of syntopy . . . . .	25
2.3	Diagnosis of composite elements . . . . .	26
2.4	Complications due to sequence similarity between related families. . . . .	27
2.5	Diagnosis of false primary edges due to partial elements . . . . .	29
3.1	Flowchart of the strategy for repeat library construction . . . . .	46
3.2	Efficiency of the incremental approach . . . . .	49
3.3	Comparison of the Rlib consensuses and their corresponding canonical repeat sequences in mouse . . . . .	50
3.4	Coverage of the mouse genome by the Rlib library (I) . . . . .	52
3.5	Coverage of the mouse genome by the Rlib library (II) . . . . .	53
3.6	Comparison of the Rlib library and the built-in library of RepeatMasker . . . . .	56
3.7	Summary of the repeat families reported in the Rlib libraries . . . . .	57
4.1	Summary of the identified repeat families in rice . . . . .	69
4.2	Classification of the identified repeat families in rice . . . . .	70
4.3	Physical distribution of repeats in the rice genome . . . . .	72
4.4	Repeats aggregate in blocks in rice . . . . .	73
4.5	Comparison of copy numbers of the defined repeat families in two sequenced rice subspecies . . . . .	74
4.6	The <i>Dasheng</i> element in rice . . . . .	76
4.7	The <i>mPing</i> element in rice . . . . .	77

# Abstract

Repetitive sequences or repeats are often called “junk DNA”, for they do not seem to provide any sequence specific function in the genome in general. These sequences are ubiquitous and abundant in all species examined to date. It is generally believed that repeats have profound impact on genome evolution and genome organization. The recent availability of whole genome sequences has opened a new window for understanding this impact.

This dissertation aims to facilitate the study of repeats by whole genome sequence analysis. The main focus is to provide a two-fold solution to a basic yet critical problem: how to properly identify repeats from genomic sequences. First, I have developed a software package, RECON, which implements a new algorithm for *de novo* identification of repeat families from genomic sequences. Second, I have started a database, Rlib, which provides repeat sequence libraries for annotating sequenced genomes of higher eukaryotes. Based on the various tests performed and the feedback from users, RECON and Rlib, which are both freely available, should be useful for the initial characterization of repeats in newly sequenced genomes.

In addition, the dissertation also describes the identification of repeats and characterization of their insertion polymorphism between two sequenced subspecies of the rice *Oryza sativa*. In collaboration with the labs of Susan Wessler and Susan McCouch, we aim to characterize the microevolution of transposable elements in rice by systematically characterizing insertion polymorphism between various rice cultivars and wild species, using both computational and experimental techniques.

## **Chapter 1**

### **Introduction:**

# **Repetitive Sequences And Their Impacts on The Genome**

Repetitive sequences were first detected in the genomes of higher eukaryotes in the mid 60's, using the technique of DNA reassociation ( $C_0t$  curve studies) (Britten & Kohne, 1968). The finding that “[h]undreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms” was completely unexpected and immediately raised many questions: What are these repetitive sequences? What do they do in the genome? Why are there so many copies of them? The search for answers to these questions has led to some of the most profound discoveries in modern biology and has greatly changed our view on how the genome functions. In this genomic era, repetitive sequences continue to fascinate biologists and pose new questions concerning their roles in genome evolution and genome organization.

## 1.1 Repetition is Not Necessarily Boring

Although repeats were first found in higher eukaryotes, virtually all genomes contain repetitive sequences. These sequences primarily fall into three classes — local repeats (tandem repeats and simple sequence repeats), dispersed repeats (mostly transposable elements and retrotransposed cellular genes) and segmental duplications (duplicated genomic fragments). Generally, a genome contains many families of repeats. Copy number of a family may vary from several to several million in a genome.

In higher eukaryotes, repeats usually make up a significant portion of the genome. Among the three types of repeats, transposable elements, or transposons, are the most prevalent. For example, ~50% of the human genome is made of transposons. The most numerous family, *Alu*, has about one million copies and consists ~10% of the genome (Lander et al., 2001). In higher plant genomes, the fraction of transposons can be as high as 80% (Feschotte et al., 2002). In contrast, only a very

small fraction of the genome (e.g., < 5% in human) codes for genes (Lander et al., 2001).

Despite their overwhelming presence, repeats are usually dismissed, for they do not always easily fit into our common notions of how genes and genome function. Two features distinguish repeats from “typical genes”. First, the repeat content of a genome is dynamic — it changes very frequently. Second, apart from multi-copy genes, repeats in general do not appear to provide any sequence specific function to the cell. Their role in the genome can only be understood in the context of evolution. As discussed in the following sections, it is these features of repeats that made us rethink our ideas about what genomes are.

## **1.2 The Dynamic Nature of Repeats**

Repeats can rapidly change their copy number and/or positions in the genome, even from generation to generation or from cell division to cell division within a single individual. At the same time, they also promote sequence addition, deletion and translocation in the genome, which can have dramatic phenotypic consequences including cancer in human (Kazazian & Goodier, 2002; Emanuel & Shaikh, 2001). This constant change of content and structure of the genome is known as genome fluidity and plasticity. The dynamic/fluidal features of the three types of repeats are detailed below.

### **1.2.1 Tandem repeats**

When repeated DNA segments are positioned adjacent to each other in a direct orientation, it is called a tandem repeat (Lovett, 2002). The repeated DNA segment is often referred to as a monomer. The length of the monomer varies from several to several thousand base pairs. When the monomer is one or two base pairs, the tandem repeat is usually called a simple sequence repeat. A tandem

repeat may contain just two monomers, or thousands.

The number of monomers in a tandem repeat changes rapidly by unequal crossover between alleles or by intra-allele recombination (Charlesworth et al., 1994). Hence, different alleles of the same tandem repeat locus may have different lengths. Such length variation is so extensive in human populations that tandem repeats are widely used for genotyping (Vergnaud & Denoeud, 2000). In many species, the rRNA genes and the histone genes are arranged in the form of tandem repeats, so are sequences at their centromeres and telomeres. In the yeast *Saccharomyces cerevisiae*, the length of the rRNA tandem repeat is a marker for aging, for it shortens as the cell gets older (Sinclair et al., 1998). It is also found recently that the shortening of telomeres is a marker of age (Takahashi et al., 2000). Furthermore, many human diseases (e.g., Huntington disease, fragile X syndrome and myotonic muscular dystrophy) are due to the expansion of tri-nucleotide monomers in protein coding genes (Cummings & Zoghbi, 2000).

### **1.2.2 Transposable elements**

Transposable elements were first discovered by Barbara McClintock by genetic analysis in maize in late 1940's (Keller, 1993; McClintock, 1987; Fedoroff, 2002), before we knew the general existence of repeats. As indicated by their name, transposable elements are discrete DNA fragments with built-in mechanisms to change positions in the genome (Berg & Howe, 1989; Capy, 1998; Craig et al., 2002). Based on their mechanism of transposition, transposable elements fall in two classes. Class I elements are also known as retrotransposons. To transpose, retrotransposons are first transcribed into RNA. The RNA is then reverse transcribed into DNA, which is inserted back into chromosomes. Known retrotransposons include the LINES/SINEs (Long/Short Interspersed Nuclear Elements), Group II introns (homing introns, (Dickson et al., 2001)) and the LTR (Long

Terminal Repeat) elements, which are also called endogenous viruses due to their similarity to retroviruses. Class II elements are known as DNA transposons, which use DNA instead of RNA as the intermediate of transposition. Many of the known DNA transposons transpose by dissociating themselves from the chromosome and re-inserting at a different place. In both classes, there are autonomous elements which encode the proteins catalyzing their own transposition and non-autonomous elements which “hitch-hike” (relying on the proteins encoded by the autonomous elements). In chapter 4, I will describe an example of how non-autonomous elements arise from autonomous ones.

Transposable elements are arguably the most dynamic among the three types of repeats. The purple patches/stripes in the kernels of Indian corn, which are caused by a DNA transposon jumping in and out of the P gene (thus abolishing and restoring the function of the P gene), demonstrate how frequent transposable elements can change their positions in the genome (Becker et al., 2002). Because of this property, transposable elements are widely used as genetic tools, either for mutant screening or integrating transgenes into the genome (des Etages et al., 2002).

By transposition, i.e., by creating new loci, transposable elements can increase their copy number in the genome. This is the strategy that ensures the survival of transposable elements. In addition, transposable elements can also invade genomes of other species by horizontal transfer (Bushman, 2002). For example, the P element (a DNA transposon), which caused the so-called hybrid dysgenesis between the lab strain and a wild strain of the fruitfly *Drosophila melanogaster*, is believed to have invaded the wild strain from a sister species *D. willistoni* after the lab strain was collected (and isolated from the wild population) in 1930's (Daniels et al., 1990).

Besides their own transposition, transposons also promote other structural changes in the genome. For example, copies of the same transposons can mediate ecotopic recombinations which lead to

large scale sequence inversion, deletion or translocation (Caceres et al., 1999; Kazazian & Goodier, 2002). In addition, transposition may generate double-stranded breaks in chromosomes which lead to various mutations as a result of imperfect repair of the breaks, including segmental duplication of the sequences around the break (McClintock, 1978). Occasionally, cellular genes can be transposed by the protein machinery of retrotransposons, creating extra copies of the corresponding genes (Li, 1997; Kazazian & Goodier, 2002).

### **1.2.3 Segmental duplications**

Until recently, segmental duplications have been considered as rare and random events caused by errors in the repair of double-stranded breaks of DNA. However, new data suggest that it may not be so (Bailey et al., 2002). For example, about 5% of the human genome is made of recent segmental duplications (< 35 million years) (Samonte & Eichler, 2002). The newly created copies are enriched in the pericentromeric and subtelomeric regions, though the original copies are from different places in the genome (Lander et al., 2001). Many of these segments are several to several hundred Kb long. A segment called f7501 best demonstrates the dynamic nature of segmental duplication in human (Mefford & Trask, 2002). The segment has been detected in 15 subtelomeric regions of 12 chromosomes in the human genome, but is single-copy in non-human primates. Thus, it must have duplicated multiple times since human and chimpanzee diverged. Furthermore, a given allele of f7501 at a given locus might be more similar in sequence to some allele at another locus than other alleles at the same locus, suggesting frequent sequence conversion between the loci.

## **1.3 The Evolution of Repeats and the Evolution of the Genome**

Repeats are generally functionless to the cell, so why are there so many repeats in so many genomes? The question sparked a heated debate in the early 80's (Becker et al., 2002). One school believed that repeats are selfish parasites (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). The other believed that repeats are selected because the genome instability they promote can facilitate evolution (Nevers & Saedler, 1977; McClintock, 1984). To accommodate the merits of each side, Sydney Brenner revived the "junk DNA" hypothesis (Ono, 1972) in his recent lectures, with emphasis falling on the word junk: unlike garbage, junk may turn out to be useful some day. Given the large amount of repeats in genomes, the impact of this "sometimes useful" junk on genome evolution is actually quite extensive.

### **1.3.1 The functionalist view**

One possible way to look at the genome is to separate it in two parts: the "functional" genes and the junkish repeats. Consequently, the evolution of the genome is centered around the genes, and the impact of repeats on genome evolution can be understood from their impact on the genes. According to this view, repeats contribute to genome evolution in the following two ways.

First, repeats help create new genes with new functions. A constant theme in evolution is the invention of new functionalities for adaptation to the ever changing environment. Duplicated genes generated by segmental duplication play a critical role in such inventions because the extra copies can be freely modified due to the lack of natural selection (Ohno, 1970). Gene families, such as G-protein coupled receptors or kinases, presumably arose by this mechanism. Thus, segmental duplication might have paved the way for organisms to become more and more complex over evolu-

tion. Similarly, retrotransposed genes can also participate in this process by gene fusion and domain shuffling (Doolittle, 1995). In addition, transposons themselves can be co-opted by the genome for new functions. For example, over a dozen of transposon insertions have been assimilated into human genes, either as regulatory sequences (Britten, 1996) or coding sequences (Lander et al., 2001). In *D. melanogaster*, the telomere is composed of two retrotransposons (*TART* and *HeT-A*) that specifically target chromosome ends (Casacuberta & Pardue, 2002). The most striking example is probably the discovery that the RAG1 and RAG2 proteins, which catalyze the VD(J) recombination in the immunoglobulin genes, were derived from the transposase protein of a DNA transposon (Agrawal et al., 1998). Given the large amount of transposable elements in higher eukaryotes, one should expect to see more such incidental yet important examples.

Second, repeats can promote genetic diversity in natural populations. This can be achieved by adding, deleting and translocating genes in individual genomes as mentioned in the above section. It can also be done by polymorphic transposon insertions that abolish or modify the function of certain genes in certain individuals in a population (Kidwell & Lisch, 1997).

### **1.3.2 Towards a more comprehensive view**

The above functionalist view of the genome is, in certain ways, simplistic. One indismissible fact is that repeats make up a significant part of the genome, and sometimes an overwhelming majority of the genome. Therefore, the propagation of repeats is largely responsible for shaping genomes into their current forms. In fact, the size of a genome is not determined by the genetic complexity of the organism, but the amount of repeats in the genome. This phenomenon, known as the Constant-value paradox <sup>1</sup> or simply the C-value paradox (Macgregor, 2002), can be striking. For example,

---

<sup>1</sup>More commonly known today as the Complexity-value paradox.

the genome of the amoeba *Polychaos dubia* is over 200 times bigger than that of human (Li, 1997). We will encounter another example in chapter 3: the genome of the mustard *Brassica oleracea* is five times bigger than that of its close relative *Arabidopsis thaliana*.

Thus, the evolution of repeats deserves its own place as we approach the question of genome evolution: it should be viewed as another constant theme in evolution that is intertwined with the theme of gene evolution (McDonald, 1995). To start to understand the full impact of repeats on genome evolution, one needs to characterize the dynamics of these sequences over evolutionary time in various genomes, i.e., how fast repeats are generated and how fast they decay. The impacts of repeats on gene evolution as discussed above could then be approached in such a context.

The evolutionary dynamics of transposable elements has drawn much attention. The key to grasp the dynamics of transposons is their interaction with the genome <sup>2</sup> (Labrador & Corces, 1997): in order to minimize the deleterious mutations caused by transposons, genomes have evolved mechanisms to suppress transposition; but to ensure their own survival, transposons have adopted ways to escape from being suppressed (such as horizontal transfer). The arm-wrestling between transposable elements and the genome is best illustrated in the case of the I element (a LINE) in *D. melanogaster* (Jensen et al., 1999): when the I element is first introduced into a genome, it undergoes an initial burst of transposition, but is quickly put down. Such episodes must have happened constantly over evolution (with different transposons at different times), creating waves of transposition bursts (Fig 1.1).

Important progress has been made in this area in the past few years. Molecularly, people have identified two general mechanisms for transposon suppression, RNA interference (RNAi) (Tabara et al., 1999) and epigenetic control (such as methylation) (Singer et al., 2001), as well as the poten-

---

<sup>2</sup>This discussion is focused on the eukaryotic genomes. Bacterial transposons seem to rely more on self-regulation to control their transposition (Berg & Howe, 1989; Craig et al., 2002).

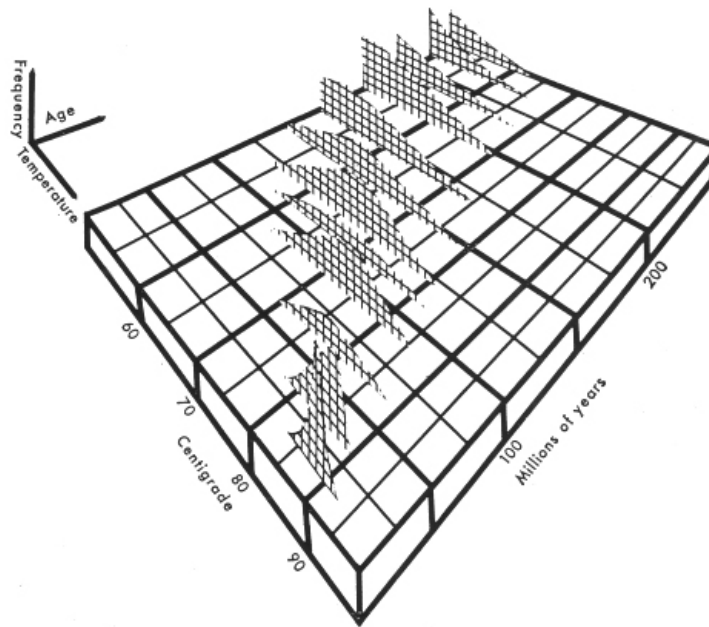


Figure 1.1: Reprinted with permission from (Britten & Kohne, 1968) (Figure 13), Copyright 1968, American Association for the Advancement of Science. This figure is a “schematic diagram intended to suggest the history of the families of repeated sequences”. The x axis (annealing temperature in DNA reassociation) can be translated as pairwise similarity. The z axis (Frequency) can be translated as copy number. Each family was assumed to “have originated in a sudden event (saltatory replication)”, the time of which is the y axis of age. Although the figure was produced for repeats in general, it is particularly appropriate for transposons. The saltatory replication can be understood as the burst-suppression of transposons.

tial link between the two (Volpe et al., 2002). Evolutionarily, people have been able to reconstruct the waves of transposition in the human and the mouse genomes (Fig 1.2), and should be able to do so for more and more genomes as they are sequenced. Deviations of these real waves from the hypothetical version of the “Britten plot” in Fig 1.1, which assumes burst-suppression on all occasions, will lead us to further understand how transposable elements interact with and propagate in the genome (see next section for more discussion).

In this regard, it is worth re-examining an early thought from Barbara McClintock — the “smart genome” hypothesis. Like many others, McClintock believed that transposons are tools for evolution. However, she believed that transposons are under the intentional control of the genome. That is, when genomes face stresses that they are unprepared to meet (i.e., no proper genetic programs exist), they would activate transposons in order to initiate massive reorganization (McClintock, 1984). The idea was originated from the observation that certain conditions (such as UV radiation (McClintock, 1984), culturing of plant cells (Hirochika et al., 1996) or interspecies hybridization (O’Neill et al., 1998)) can drastically elevate transposition and generate massive mutations.

It is becoming possible to test McClintock’s hypothesis. Given the molecular mechanism of suppression, one can test whether the observed de-suppression is a controlled process that involves a definite genetic program. In the meantime, studies of “Britten plots” could show whether global activation of transposons actually occurred in evolutionary history. Should this hypothesis be right, we would need to take a whole new look at evolution. According to this hypothesis, genomes are more active participants in evolution, rather than the passive receiver of mutations and natural selection as implied by the Darwinian theory. In the meantime, this is not to suggest that genome evolution is Lamarckian or directed; rather, that the organism may be able to control its rate of random variation and genome reorganization.

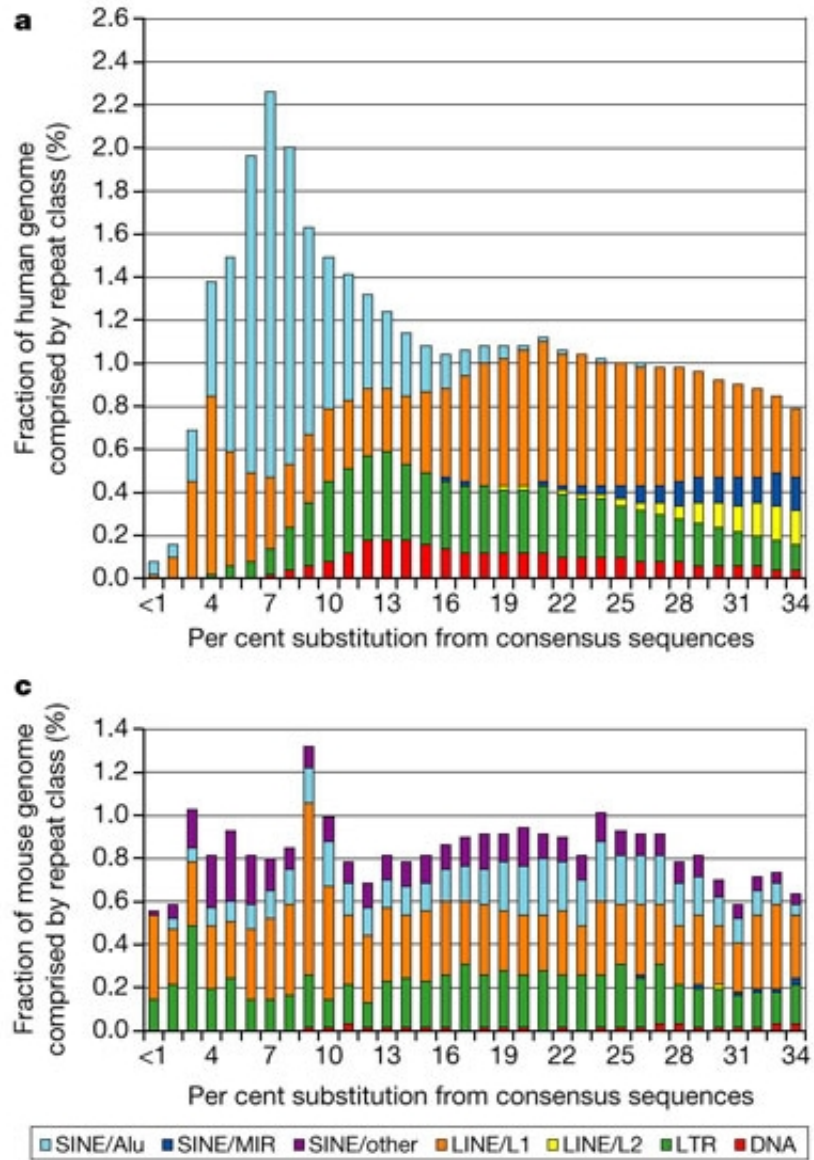


Figure 1.2: The history of transposition in the human (a) and mouse (c) genomes. Reprinted by permission from *Nature* (Figure 18 in (Lander et al., 2001)) copyright 2001, Macmillan Publishers Ltd. As shown, different transposons have different kinetics of transposition over evolutionary time (for detailed interpretations, see the original paper). Notice that the real history of transposition is quite different from the hypothetical one in Fig 1.1.

## 1.4 Challenges and Opportunities in the Genomic Era

In the past decade, genome sequencing projects have greatly facilitated the study of molecular biology. Similarly, the availability of large scale genomic sequences have greatly facilitated the study of repeats as well. After all, nature has done a thorough experiment of genome evolution, it is now up to us to read out the results and see how repeats have behaved. Generally, analysis of repeats in sequenced genomes can be summarized in two categories, concerning either the temporal or the spatial aspect of genome evolution.

### 1.4.1 The temporal aspect: quantitating the evolutionary dynamics of repeats

Much of our notion about the dynamic nature of repeats is qualitative. However, to fully understand genome evolution, we need more quantitative assessments of the frequency and extent of various events. Complete or nearly complete genome sequences are particularly helpful for these quantitative analyses.

One active topic of such quantitation focuses on the generation of duplicated genes. The most popular approach is to examine the age distribution of duplicated genes in a genome as inferred by pairwise distances between the duplicated copies. Thorough examinations of the available genomes suggest that generation of segmental duplications is frequent and constant, with the estimated average rate of 1% of the genome per million years (Lynch & Conery, 2000). Furthermore, the overall age distribution suggested that *A. thaliana* had undergone an ancient whole genome duplication, while the popular hypothesis of two whole genome duplications in vertebrate evolution was probably not true (Friedman & Hughes, 2001). Similarly, the recent hypothesis for a whole genome duplication in yeast (Wolfe & Shields, 1997) may not be true either.

A second topic concerns the dynamics of transposons and transposition. This topic has been relatively less advanced, partly due to the lack of convenient computational tools for systematic identification of transposons from genomic sequences (Holmes, 2002). Nonetheless, where some good tools are ready, such as in the human genome, people have started to reconstruct the history of transposition. As shown in Fig 1.2, “Britten plots” can be very informative for this matter. Yet, for more quantitative studies, one would like to take one step further and calibrate the kinetics of transposition over evolutionary time for each major family in a genome (Promislow et al., 1999). In principle, the age of a given copy of transposon (i.e., when this particular locus was created by transposition) could be inferred from sequence divergence between copies in a family. However, the proper methods for doing so are still an open question that requires careful integration of sequence comparison and population genetics (see Chapter 5 for more discussion).

A detailed picture of the history of transposons in a genome could tell us many things. For example, one could set out to search for correlations of timing of major evolutionary events: are there statistically significant peaks and valleys of transposition? Do these peaks or valleys coincide with other events in the genome, such as gene duplication and chromosomal restructuring? Furthermore, profiling the history of many transposon families would let us define the normal evolutionary pattern of transposition, which in turn would let us ask questions such as why certain families seem to have a longer lifespan than others. Sometimes, lack of history can also be informative — the lack of more divergent copies in a family would imply horizontal transfers.

A related topic concerns the dynamics of sequence deletion, which removes repeats from the genome. Recently, it was found that organisms which have smaller genomes compared to their close relatives tend to have higher rates of sequence deletion (Petrov et al., 2000). It is yet known how fast sequences turn over in a genome over evolution.

## 1.4.2 The spatial aspect: genome organization

At a first glance, genomes look like a mess — organisms have arbitrary numbers of chromosomes, genes can be ordered arbitrarily along the chromosomes (except for operons or some gene clusters), and on top of that is the mayhem of repetitive junk. However, a closer look at genomic sequences suggests that genome organization may not be completely random. Some of the non-random features are probably due to natural selection, such as the depletion of transposons in the Hox gene cluster in human (Lander et al., 2001). Other non-random events are created by biases in various molecular mechanisms. For example, certain LTR elements in yeast (*Ty1–4*) insert near tRNA genes (Kim et al., 1998); the LINE NeSL-1 in *C. elegans* inserts in front of the Spliced leader-1 genes (Malik & Eickbush, 2000); and in higher plants, a particular type of DNA transposons called MITEs (Miniature Inverted-repeat Transposable Elements) preferentially insert in genic regions (Jiang, 2002; Zhang & Hong, 2000).

Non-random distribution of various sequences in the genome has led to the proposal of the concept of “chromosomal domains” (Surzycki & Belknap, 2000). One example, as mentioned before, is that duplicated segments in human are enriched in the pericentromeric and subtelomeric regions (Lander et al., 2001). Another example is that repeats are enriched in the arms of the autosomes in *C. elegans* (The *C. elegans* Sequencing Consortium, 1998). We still do not have much clue about what creates chromosomal domains. In the case of *C. elegans*, the density of repeats is positively correlated with the frequency of recombination (Barnes et al., 1995), but no satisfying explanations have been offered. To make the picture more complicated, not all families in *C. elegans* are distributed in the same manner as the overall trend (Surzycki & Belknap, 2000).

Should there be any unknown principles governing genome organization, the search for non-

random distributions in sequenced genomes might be an excellent starting point to uncover them. In this regard, comparative studies on genomes at various evolutionary distances, as more and more genomes are being sequenced, would be particularly helpful: one could set out to search for the non-randomness that is preserved after massive translocations and genome restructuring.

## 1.5 Scope of Dissertation

My dissertation is set in the context of studying repeats by whole genome sequence analysis. Overall, the analysis of repeats in sequenced genomes is still in its infancy. Hence, my dissertation focuses on some of the groundwork that is needed for more advanced studies, i.e., developing relevant computational tools and exploring possible ways to take advantage of the large amount of genomic sequences.

Following this introduction, Chapter 2 of the dissertation focuses on a basic yet critical question: how to properly identify repeats from genomic sequences. Computational tools such as RepeatMasker (Smit & Green, 2002) have proven to be successful in finding copies of known repeats. However, to find new repeat families and to analyze the increasing number of less characterized genomes where little is known about their repeat content, one needs a different type of algorithm, a *de novo* algorithm that does not require any prior knowledge of the repeats to be found. Chapter 2 describes such an algorithm which only relies on the copy number of sequences to detect repeats (Bao & Eddy, 2002). The algorithm has been implemented in a software package dubbed RECON, which is freely available on the Web at <http://www.genetics.wustl.edu/eddy/recon>. Since its first release in January 2002, the package has been downloaded from nearly 100 different IP addresses outside Washington University.

*De novo* repeat identification in general requires tremendous computing power. For example, analyzing a 400Mb genome like the rice genome takes several days on a 128-CPU computer cluster. Thus, to better serve the interest of the research community, we have decided to distribute not only the RECON software, but also its results for various genomes. We are doing so by setting up a database named Rlib, which provides libraries of repeat sequences for sequenced higher eukaryotic genomes. One can then easily analyze his favorite sequence/genome by feeding the proper Rlib library into RepeatMasker. Chapter 3 of the dissertation addresses issues in building Rlib libraries and also describes the early results. Rlib is freely available on the Web at <http://Rlib.wustl.edu>. The goal of Rlib is to catch up and keep up with the ever expanding sequencing effort. Some of the libraries are now being used by various sequencing consortia for their initial genome analysis.

Chapter 4 studies repeats in the genome of rice, *Oryza sativa*. Rice is a very attractive system for studying the evolutionary dynamics of transposons. We formed a collaboration with Dr. Susan Wessler (University of Georgia, Athens) and Dr. Susan McCouch (Cornell University) to systematically identify transposons and characterize their insertion polymorphism in different rice cultivars and wild species. Chapter 4 summarizes the computational analysis for this project, with emphasis on estimating the overall level of insertion polymorphism between the two sequenced subspecies. It also illustrates how computational and experimental approaches will be combined to characterize the physical distribution of transposons in other rice cultivars and wild species. In the process of analyzing the identified repeats, we have found the first MITE that actively transposes in the lab (*mPing*) (Jiang et al., 2003).

Chapter 5 is the concluding remarks. Besides further comments on the previous chapters, I will discuss some open questions and challenges regarding the emerging field of transposon genomics.

## **Chapter 2**

# **Automated *de novo* Identification of Repeat Sequence Families in Sequenced Genomes <sup>1</sup>**

---

<sup>1</sup>This chapter was co-written with Sean Eddy, and appears in Bao & Eddy, *Genome Research* **12**:1269 - 1276,2002.

## 2.1 Abstract

Repetitive sequences make up a major part of eukaryotic genomes. We have developed an approach for the *de novo* identification and classification of repeat sequence families, based on extensions to the usual approach of single linkage clustering of local pairwise alignments between genomic sequences. Our extensions use multiple alignment information to define the boundaries of individual copies of the repeats and to distinguish homologous but distinct repeat element families. When tested on the human genome, our approach was able to properly identify and group known transposable elements. The program, RECON, should be useful for first-pass automatic classification of repeats in newly sequenced genomes.

## 2.2 Introduction

A significant fraction of almost any genome sequence is repetitive. Repetitive sequences fall primarily into three classes – local repeats (tandem repeats and simple sequence repeats), families of dispersed repeats (mostly transposable elements and retrotransposed cellular genes) and segmental duplications (duplicated genomic fragments). The role of repeated, transposed, and duplicated sequence in evolution is an interesting and controversial topic (Doolittle & Sapienza, 1980; Orgel & Crick, 1980; McClintock, 1984), but repetitive sequences are so numerous that simply annotating them well is an important problem in itself. This is particularly the case for repeat sequence families, which often carry their own genes (transposases, reverse transcriptases, and the like), and can confuse large-scale gene annotation.

Computational tools have been developed for systematic genome annotation of repeat families. Perhaps the best known is the program RepeatMasker (Smit & Green, 2002), which uses

pre-compiled representative sequence libraries to find homologous copies of known repeat families. RepeatMasker is indispensable in genomes where repeat families have already been analyzed. However, it does not pass the “platypus test” (Marshall, 2001): repeat families are largely species-specific, so if one were to analyze a new genome (like the platypus), a new repeat library would first need to be manually compiled. With sequencing efforts moving towards large-scale comparative genome sequencing of a wide variety of organisms, it is desirable to also have a *de novo* method that automates the process of compiling RepeatMasker libraries.

Several *de novo* approaches have been attempted, with limited success. They generally start with a self-comparison with a sequence similarity detection method to identify repeated sequence, then use a clustering method to group related sequences into families (Agarwal & States, 1994; Parsons, 1995; Kurtz et al., 2000). Detecting repetition by sequence alignment methods is relatively easy. Automatically defining biologically reasonable families is more difficult. Local sequence alignments do not usually correspond to the biological boundaries of the repeats, due to degraded or partially deleted copies, related but distinct repeats, and segmental duplications covering more than one repeat. Difficulty in defining element boundaries then causes a variety of subsequent problems in clustering related elements into families.

Similar problems arise in automated detection of conserved protein domains. Curated databases such as Pfam (Bateman et al., 2002) play a role equivalent to RepeatMasker by providing pre-compiled libraries of known domains. Automated clustering approaches are used to help detect new domains (Sonnhammer & Kahn, 1994; Gracy & Argos, 1998). These automated algorithms combine pairwise alignments with a variety of extra information to try to define biologically meaningful domain boundaries: most importantly, they look at multiple sequence alignments, not just pairwise alignments, in order to find significantly conserved boundaries.

Here, we describe an automated approach for *de novo* repeat identification. Our approach uses multiple alignment information to infer element boundaries, and also to infer biologically reasonable clustering of sequence families.

## 2.3 Results

Given a set of genomic sequences,  $\{S_n\}$ , our goal is to identify the repeat families therein (denoted by  $\{F_\alpha\}$ ), so that each family corresponds to a particular type of repeat, containing all and only copies of that repeat in  $\{S_n\}$ . Each individual repeat is a subsequence  $S_n(s_k, e_k)$ , where  $s_k$  and  $e_k$  are start and end positions in sequence  $S_n$ . Therefore, the output is  $\{F_\alpha = \{S_n(s_k, e_k)\}\}$ .

We define the following terms – *element*, *image* of an element, and *syntopy*. An individual copy of a repeat,  $S_n(s_k, e_k)$ , is called an *element*. A subsequence involved in an alignment is called an *image* (Fig 2.1). An element is the biological entity we are trying to infer. Images are observations from a pairwise comparison of the genome sequences  $\{S_n\}$ . One element forms many images, due to its repetitive nature. We call two images of the same element *syntopic images* (*syntopy* is a neologism from *syn* - 'same', *-topy* 'site'). Because observed alignments may extend well beyond the bounds of an element, and may even include unrelated elements (for example, because of segmental duplication or coincidental juxtaposition of abundant repeats), syntopy cannot be inferred just by image overlap – and this is the problem we must address.

### 2.3.1 The Existing Single Linkage Clustering Algorithms

The existing *de novo* repeat identification algorithms can be summarized in our terms as single linkage clustering algorithms, as follows:

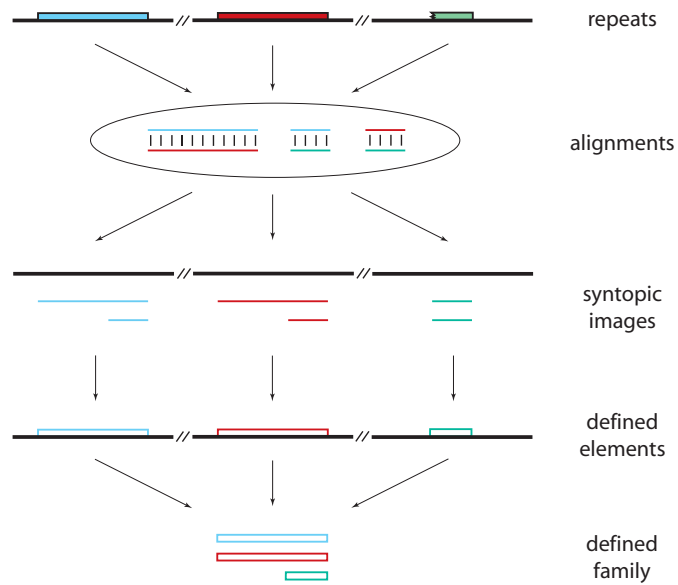


Figure 2.1: Flowchart of the *de novo* strategy. Input genomic sequences (black lines on top) contain a family of repeats with three copies (i.e., elements); two full-length (blue and red boxes) and one partially deleted (green box). These elements, unknown at this point, will yield three alignments in an all-*vs*-all pairwise comparison of the genomic sequences. The aligned fragments (i.e., images), colored as their corresponding elements for clarity, are sorted to their corresponding genomic region, and those coming from the same element (i.e., syntopic images) can be grouped together according to their overlaps. Based on the syntopic sets, elements can be defined. These defined elements are then clustered into one family as they are all similar to each other.

1. Obtain pairwise local alignments between sequences in  $\{S_n\}$ .
2. Define elements  $\{S_n(s_k, e_k)\}$  from the obtained alignments, or, images:
  - (a) Construct graph  $G(V, E)$ , where  $V$  represents all the images and  $E$  represents the syntopy between images. Two images are considered syntopic if they overlap, regardless of strand, beyond some threshold.
  - (b) Find all connected components in  $G$  (Skiena, 1997).
  - (c) For each connected component, define an element  $S_n(s_k, e_k)$  as the shortest fragment that covers all images in the component.
3. Group defined elements into families on the basis of their sequence similarity:
  - (a) Construct graph  $H(V', E')$ , where  $V'$  represents all the elements, and  $E'$  represents similarity (two elements are connected by an edge if they form alignments in step 1).
  - (b) Find all connected components of  $H$ .
  - (c) For each connected component, define a family as the set of all elements in the component.

***Problem 1: Inference of syntopy***

The main problems with this approach arise from the use of overlap to infer syntopy. If all repeat elements were full-length, well-conserved, and well-separated by unique sequence in the genome, all syntopic images would be equivalent to their corresponding element, and single linkage clustering would work fine. However, two major phenomena distort this ideal picture. One is drift (both deletion and substitution mutation), which causes partial images (Fig 2.2B). The other is segmental

duplication and juxtaposition of common repeats which produce images containing more than one element (Fig 2.2C).

Various strategies have been suggested for inferring syntopy from image overlap. Two typical measurements, termed *single coverage method* and *double coverage method*, require the overlap to be longer than a certain fraction of *either* or *both* of the images, respectively. When overlapping images are of different length, the two methods make different inference of syntopy which leads to different definitions of elements (Fig 2.2A). The single coverage method is suitable for the scenario in Fig 2.2B, while the double coverage is suitable for that in Fig 2.2C.

However, either strategy leads to errors. When the double coverage method is applied to partial images (Fig 2.2B), it yields many spurious, overlapping elements for one true biological copy. When the single coverage method is applied to multielement images (Fig 2.2C), it yields a composite element corresponding to the whole segmental duplication, which will lump families together later in family definition. Simply tuning the thresholds of these methods will not solve the problem; the two biological scenarios require opposite measurements of overlap in order to correctly infer syntopy (Agarwal & States, 1994). Furthermore, since these algorithms use only pairwise relationships between images, they are not able to distinguish the two biological scenarios and choose the proper criterion. The example in Fig 2.2 therefore suggests that no algorithm of this type can work.

However, one also sees in Fig 2.2 that there is useful information in the pattern of the multiple alignment of the images. In both cases, most image endpoints agree on the boundaries of an independent repeat. The key distinction lies in the endpoints of the shorter images. In Fig 2.2B, these endpoints are quasi-randomly dispersed throughout the multiple alignment, whereas in Fig 2.2C, the endpoints pile up. Biologically, this distinction will hold true so long as the independent replication of repeats is more frequent than the generation of composite elements (say, by segmental

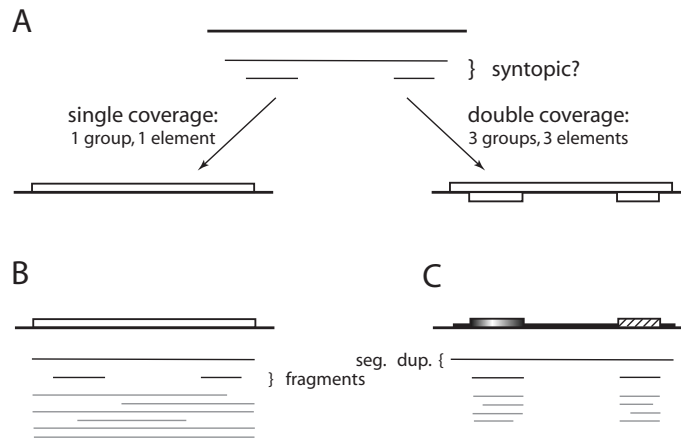


Figure 2.2: Different biological scenarios require different methods of syntopy inference. A. For three images (thin black lines) in a genomic region (top bold black line), the single and double coverage methods lead to different definitions of elements. B. A full-length element and its images (black and grey lines below). The top long image is formed with another full-length member in its family, while the shorter images are formed with the fragmented members. C. A segmental duplication covering two kinds of elements. The top long image is formed with the other copy of this segmental duplication, while the shorter images are formed with other members in the two families, respectively.

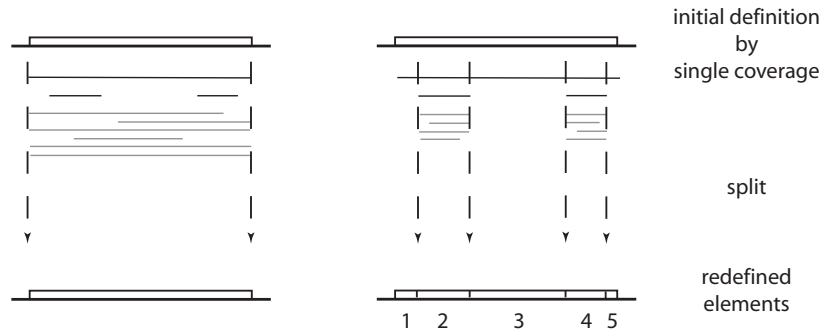


Figure 2.3: The RECON algorithm uses the aggregation of endpoints in the multiple alignment of images to distinguish between different biological scenarios.

duplications), and deletion is a random process, which are usually (but not always) the case.

Our approach to the problem is based on the above observation (Fig 2.3). After an initial definition using the single coverage method, elements are split according to significant aggregations of image endpoints. As shown in Fig 2.3, a composite element will be split into several pieces (right panel, five pieces in this case), while a full-length element will be preserved (left panel). Details are specified in the **element re-evaluation and update procedure** (see Methods).

Certain images complicate the above splitting process, such as those formed between related but distinct elements (Fig 2.4A and B), which may lead to an incorrect splitting of an element. Unlike those in Fig 2.2, these misleading image endpoints do not occur at the termini of either of the two elements involved (Fig 2.4C, open circles). We use this difference to filter the misleading endpoints prior to the element re-evaluation and update procedure (see **image end selection rule** in Methods).

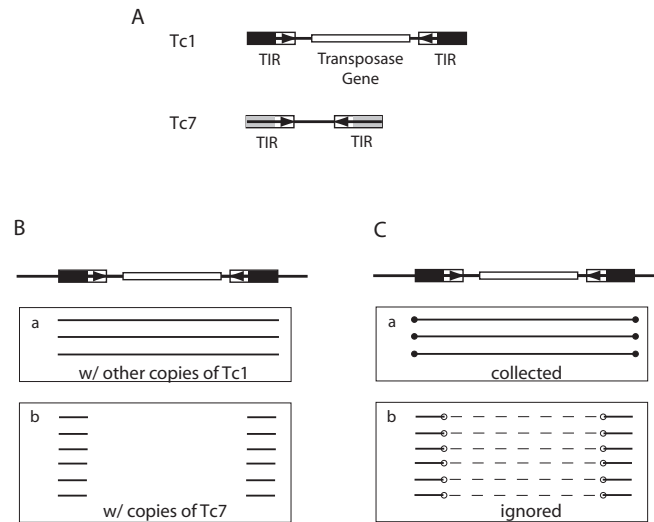


Figure 2.4: Complications due to sequence similarity between related families. A. The schematic structure of Tc1 and Tc7, two related DNA transposons which are similar at the end of their Terminal Inverted Repeats (black and grey blocks), but not in the rest of the sequences (Plasterk & von Luenen, 1997). B. A Tc1 element and its images. C. Images in B are filtered, and only those ends marked with closed circles will be collected to determine whether the element should be split. Open circles in box b mark the misleading ends. Dashed lines link the pairs of images formed with the same copy of Tc7 and represent the unalignable sequences between a Tc1 and a Tc7. Although not shown in the figure, the two TIRs of Tc1 also form alignments in the opposite strands, and images from these alignments are also filtered.

### ***Problem 2: Inter-family similarity***

Many repeat families are evolutionarily related (for example, the autonomous *C. elegans* Tc1 DNA transposons and the smaller nonautonomous Tc7 elements, Fig 2.4). Although the reality is that repeats, like Pfam's protein domain families or biological species, are a hierarchical evolutionary continuum that defies classification, it is still desirable to impose a simplistic classification that pretends that repeat families are distinct, for the purpose of practical genome annotation. Since related families may form significant sequence alignments, we will have to impose arbitrary criteria to avoid lumping related but "distinct" families together.

We consider two elements to be distinct if the length of the non-conserved regions adds up to more than certain ratio of both of the two sequences (Fig 2.4C, dashed lines). The **family relationship determination procedure** (see Methods) implements this definition. When constructing the graph for clustering (step 3 in the above algorithm), elements belonging to the same family are linked with *primary* edges, and those belonging to different families but still forming significant alignments are linked with *secondary* edges. Families (connected components) are defined by primary edges.

Incorrect primary edges can arise in the presence of certain partially deleted elements (Fig 2.5A). As shown in Fig 2.5B, primary and secondary edges are properly constructed between full-length copies of Tc1 and Tc7 by the family relationship determination procedure. However, edges between the partial copy of Tc7 and the Tc1s are rendered primary, as there are no non-conserved regions in this Tc7 compared to Tc1s. These false primary edges will lump the two families. Such a situation can be recognized by finding triangles involving two primary edges and a secondary edge (e.g. Tc1-2=>Tc7-1=>Tc7-partial). Once an element yielding incorrect primary edges is recognized, all its

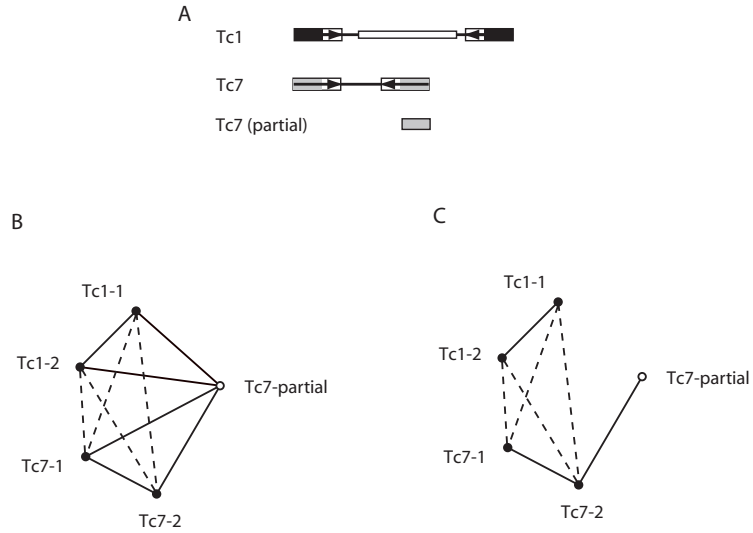


Figure 2.5: False primary edges due to partial elements. A. The schematic structure of full-length Tc1 and Tc7 (see also Fig 2.4) and a partially deleted Tc7, which preserves only the region similar to Tc1. B. Graph constructed for Tc1s and Tc7s. Closed nodes represent full-length elements. Solid and dashed lines represent primary and secondary edges, respectively. C. Certain primary edges are removed from the partial Tc7 in order to eliminate the false ones.

primary edges are removed except for the one linking to its most closely related element (Fig 2.5C).

More rules are specified in the **family graph construction procedure with edge re-evaluation** (see Methods).

### 2.3.2 The RECON Algorithm

Our algorithm is summarized as follows:

1. Obtain pairwise local alignments between the input sequences.

2. Define elements from the obtained alignments:
  - (a) Elements are first defined using the *single coverage method*, as described in step 2 of the existing algorithm;
  - (b) Each element defined is re-evaluated following the *image end selection rule* (Fig 2.4) and the *element re-evaluate and update procedure* (Fig 2.3);
  - (c) If an element defined is considered composite and is split, elements forming alignments with the composite element will be re-evaluated. The process continues till all definitions of elements stabilize.
  
3. Group elements defined into families on the basis of their sequence similarity:
  - (a) Elements and their family relationship are determined and converted to a graph  $H(V', E')$  according to the *family relationship determination procedure* and the *graph construction procedure with edge re-evaluation* (Fig 2.5);
  - (b) Find all connected components of  $H$  according to the primary edges constructed. For each connected component, define a family as a set of all elements in the component.

The algorithm has been implemented as RECON, a set of C programs and Perl scripts. The RECON package, including a demo and more materials, is available at <http://www.genetics.wustl.edu/eddy/recon/>.

### **2.3.3 Assessment**

In order to assess the performance of RECON, we used it to analyze a random sample of 3 Mb, or about 0.1%, of the human genome (Lander et al., 2001), and compared the results to RepeatMasker annotation as a “gold standard”. For purpose of comparison, we also implemented and tested the

basic single linkage clustering algorithm using both the single or double coverage element definition methods. All three *de novo* methods use the same set of 453,896 pairwise alignments generated by WU-BLASTN (Gish, 2002) (see Methods).

It took RECON 4 CPU hours and a maximum of 300 MB RAM to analyze this set of alignments on a single Intel Xeon 1.7GHz processor. A RECON analysis of a set of alignments from a three-fold larger sample (9 Mb) took 39 CPU hours and 750 MB RAM. We cannot give a useful asymptotic analysis of memory/cpu usage in terms of genome or sample size, because RECON's computational complexity is strongly dependent on repeat density and composition. For example, an analysis of the alignments from the same 3 Mb sample with known repeats masked out by RepeatMasker took less than 1 minute and 900 KB of RAM. This suggests that for a large, repeat-rich genome, it will be possible (and necessary) to carry out an iterative RECON analysis; e.g., first find the most abundant families in a small sample of the genome, then analyze progressively larger samples after masking element families that have already been confidently identified.

As to the quality of the results, we first looked specifically at the definition of Alu, which is the most numerous repeat element and therefore the most prone to many sorts of clustering artifacts (Table 2.1). We identified each *de novo* constructed family that contained one or more sequences that overlapped Alu elements defined by RepeatMasker. For the largest family defined by each method, we also counted how many of the defined elements contained non-Alu repeat sequences as defined by RepeatMasker. A "correct" result would be that a *de novo* method would identify a single family of 1,260 Alu elements covering 318,927 bases of the genome sample, exactly matching the RepeatMasker annotation.

The single coverage method defined 1,389 elements which overlapped the Alus defined by RepeatMasker. The number is larger than 1,260 because some Alu copies are broken into several

Table 2.1: Definition of the Alu Family

Method	# of elements	Total length, bp	Genomic coverage, bp	# of families	Largest Family	
					non-Alus	Alus
RepeatMasker	1,260	318,938	318,927	1	0	1,260
Single	1,389	357,830	331,593	1	576	1,378
Double	56,925	7,908,428	330,830	19	6	54,615
RECON	1,468	285,747	285,000	2	2	1,423

fragments by the method. The 1,389 elements covered too much of the genome (331,593 bp), because some of the “elements” are actually segmental duplications which happen to contain Alus. This method overclusters. In the largest family defined, it mixed 576 of non-Alu sequences (most of which are L1 elements, the second most abundant human repeat family) with the 1,389 Alu elements. The double coverage method underclusters images, defining many “elements” that completely overlap each other, leading to a huge number of “elements” (56,925) clustered into too many families (19). RECON minimizes both problems, leading to 2 Alu-containing families (one of which dominates) with 1,468 elements covering 285,000 bp, with minimal contamination from other repeat families. Some Alu elements are still inappropriately broken into two or more fragments (leading to significantly more than 1,260 elements). The somewhat lower genomic coverage of RECON compared to RepeatMasker results from the higher sensitivity of RepeatMasker’s similarity search algorithm and threshold (CROSSMATCH with an aggressive threshold, as opposed to RECON’s use of WU-BLAST with a conservative threshold).

In order to evaluate how reliable RECON annotation is overall, we systematically compared every RECON family containing  $\geq 10$  elements to RepeatMasker annotation (Table 2.2). Each RE-

CON family was labeled according to which RepeatMasker annotation made up the majority of its elements. Any element that was annotated as a different family or not annotated at all was considered as false positive elements (cluster fp1 and cluster fp2 columns in the Table, respectively). These results suggest that RECON's families are almost completely "pure", with very little contamination from unrelated repeat families. The families are slightly underclustered; for example, one large family with the majority of the L1 elements is found (f7), along with several smaller families of partial L1 elements (f8, f13, f22, f57, f146) which are not clustered with f7. f179, a "new" family, is a family of retroposed protein-coding genes, which are a class of repeats not annotated by RepeatMasker.

An important usage of a *de novo* method is to generate repeat libraries for the incremental analysis of a genome. In order to evaluate how useful RECON families would be for genome annotation of elements in a subsequent sample of human sequence, we compared the consensus sequence of each RECON family to their most similar sequences used in RepeatMasker (Table 2.2; see Methods). Bases in RECON's consensus that are not in RepeatMasker's sequence are counted as false positives (consensus fp column) – measuring to what extent RECON defines too large of a consensus element. Bases in RepeatMasker's sequence that are not in RECON's consensus are counted as false negatives (consensus fn column) – measuring to what extent RECON only recovers part of the consensus element. For four out of the six known transposable elements found, the canonical sequence is reconstructed essentially intact (f1/Alu, f7/L1, f46/MaLR and f28/MER41). For Tigger1 and MER1, however, only part of the canonical sequence is recovered in families f17 and f156. Manual inspection suggests that it is due to the truly fragmented nature of the copies in our sample, rather than erroneous splittings by RECON.

The canonical Alu sequence is dimeric, containing a left (L) and a right (R) monomer (Jurka &

Table 2.2: The Larger Human Repeat Families Defined by RECON

RECON family	RepeatMasker family	copy <sup>a</sup> number	cluster <sup>b</sup>		consensus <sup>c</sup>	
			fp1	fp2	fp	fn
<b>f1</b>	<b>Alu</b>	<b>1425</b>	<b>1</b>	<b>1</b>	<b>1/424</b>	<b>16/311</b>
f230	Alu	10	0	0	3/77	111/185
<b>f7</b>	<b>L1</b>	<b>292</b>	<b>2</b>	<b>1</b>	<b>0/6139</b>	<b>15/6152</b>
f8	L1	28	0	0	0/906	5391/6305
f13	L1	22	0	0	1/518	5668/6184
f22	L1	17	0	0	3/1481	4655/6146
f57	L1	14	0	0	1/690	5429/6146
f146	L1	13	0	0	2/273	6031/6305
f10	MaLR(LTR)	63	0	0	0/365	1/364
<b>f46</b>	<b>MaLR(LTR+internal)</b>	<b>44</b>	<b>0</b>	<b>0</b>	<b>3/2116</b>	<b>0/1935</b>
f12	MaLR(LTR)	17	0	0	3/211	218/426
<b>f28</b>	<b>MER41</b>	<b>18</b>	<b>0</b>	<b>0</b>	<b>2/559</b>	<b>1/554</b>
<b>f17</b>	<b>Tigger1</b>	<b>14</b>	<b>0</b>	<b>0</b>	<b>2/1021</b>	<b>1405/2418</b>
f179	New	13	0	13	n/a	n/a
<b>f156</b>	<b>MER1</b>	<b>10</b>	<b>0</b>	<b>0</b>	<b>3/199</b>	<b>99/297</b>

<sup>a</sup>number of defined elements in RECON family

<sup>b</sup>fp1: number of elements in RECON family corresponding to a different RepeatMasker family. fp2: number of elements in RECON family not annotated by RepeatMasker.

<sup>c</sup>fp: false positive positions vs length of the consensus. fn: false negative positions vs length of the RepeatMasker sequence. The consensuses of the L1-corresponding families match different L1 sequences in RepeatMasker. So do the MaLR-corresponding families.

Zuckerkindl, 1991). Interestingly, the consensus sequence identified by RECON family f1 contains exactly one and a half Alu elements, in the configuration LLR. The longest six elements in f1 are all in this configuration. Such trimeric Alus have been noted before (Perl et al., 2000), and RECON's annotation suggests that they have been actively transposed in the human genome.

## 2.4 Discussion

The problem of automated repeat sequence family classification is inherently messy and ill-defined, and does not appear to be amenable to a clean algorithmic attack. The heuristic approach we have taken in RECON appears to be satisfactory for many practical purposes. Our use of multiple sequence alignment information, specifically the clustering of observed alignment endpoints, is a significant improvement over single linkage clustering based on pairwise sequence relationships alone. Several aspects of the RECON algorithm are probabilistic in nature. For example, the split ratio  $n/m$  in the element re-evaluation and update procedure is correlated with the probability of two repeats being adjacent by chance. We could take a more formal approach evaluating the significance of  $n/m$ . However, a simple cutoff value appears to be sufficient.

The evaluation of RECON's performance suggests several issues which could use improvement. It slightly underclusters elements, failing to appropriately link some small fragmentary families to a large full-length family. This might be addressed by a post-processing step that merges RECON families when the consensus of one family covers the consensus of the other.

RECON is sometimes unable to recover a highly fragmented family in one piece. To overcome this, we could employ a statistical test to identify RECON families whose copies tend to be physically adjacent to each other. The more diverged families, such as the ancient human L2 family, was

not recovered in our test, due to the chosen sensitivity settings of WU-BLAST.

RECON can also fail when its simple assumptions about alignment end clustering are violated. For example, when a particular form of partial copy is generated preferentially (e.g., solo LTRs for retrovirus-like elements (Kim et al., 1998), formed by high-frequency deletion between the directly repeated LTRs), it can lead to an erroneous splitting of the full-length copies. Also, if a particular combination of repeat elements can itself be duplicated at high frequency (e.g., composite bacterial IS elements (Berg & Howe, 1989)), it may not be recognized as composite.

Different repeat composition may require tuning of parameters. For example, if elements are largely fragmented, one may lower the requirements of overlap between images, at the risk of producing more composite elements. When solo LTRs are predominant, one may raise the ratio cutoff for element splitting, at the risk of failing to break truly composite elements. One can only hope to optimize among conflicting situations in a genome-scale analysis.

We envision using RECON as a tool for initial analysis of a sequenced genome. Much like automated PRODOM protein domain family identification aids curated Pfam multiple alignment construction, the families identified by RECON can be the basis of a higher quality level of analysis, such as using RECON families to build a RepeatMasker library, or using RECON multiple alignments to build a library of profile hidden Markov models.

## 2.5 Methods

### 2.5.1 Components of RECON

#### *Image End Selection Rule*

This rule filters misleading images (Fig 2.4) by considering the length and arrangement of the aligned and unaligned sequences between two elements as follows:

1. For each pair of defined elements that form alignments, find all maximal groups of alignments in which all alignments are part of one (but not necessarily the optimal) global alignment of the two given elements. This is done by finding maximal cliques (Skiena, 1997) in a graph where the vertices represent the alignments and two vertices are linked if the two corresponding alignments can be seen as part of one global alignment of the two given elements.
2. For each group found above: order the alignments according to their coordinates; eliminate the group if the sequences outside the out-most alignment or between any two adjacent alignments in the group are longer than a given length cutoff in *both* elements; if not eliminated, assign a score to the group as the sum of scores of all alignments in the group. The length cutoff is chosen so that sequences shorter than the cutoff can be considered as generated by the random extension of true alignments by the pairwise alignment tool.
3. If more than one group remains, take the one with the highest score and discard the others. Ends of the images in the remaining group (if any) are collected for further analysis.

### ***Element Re-evaluation and Update Procedure***

This procedure updates the definition of a given element (Fig 2.3) by evaluating the aggregation of image endpoints collected according to the rule above.

1. Choose a length cutoff so that sequences shorter than the cutoff are considered as generated by the random extension of true alignments by the pairwise alignment tool.
2. Slide a window of the chosen length cutoff along the given element. Within each window, cluster the collected image ends as follows: seed a cluster with the leftmost end not yet clustered; if an end is within certain distance to any member in the cluster, it is assigned to the cluster; when no more ends can be assigned to the cluster, start a new cluster if necessary, till all ends in the window are clustered.
3. For each cluster found above, let  $n$  denote the number of ends in the cluster,  $c$  denote the mean position of these  $n$  ends, and  $m$  denotes the number of images of the given element spanning position  $c$ . If  $n/m$  is greater than a given threshold,  $c$  is considered a significant aggregation point.
4. If no significant aggregation point is accepted, the original definition of the given element is maintained.
5. Otherwise, update the given element as follows: split the element and its alignments at the aggregation points; discard the original definition of the given element; discard the split products (new elements and alignments) that are shorter than the chosen length cutoff at the beginning; assign alignments to proper new elements.
6. If more than one new element remains, the original element is considered composite.

### ***Family Relationship Determination Procedure***

This procedure determines for a given pair of defined elements that form alignments, whether the two belong to the same family, or to two related but distinct families (Fig 2.4). The procedure, which considers the relative length of the aligned sequences compared to the length of the elements, is as follows:

1. For a given pair of defined elements that form alignments, find all maximal groups of alignments in which all alignments are part of one (but not necessarily the optimal) global alignment of the two given elements. See step 1 in the image end selection rule for detail.
2. The total length of each group found above is calculated as the sum of the length of all alignments in the group. The longest total length among the groups is treated as the alignable length between the two elements.
3. If the alignable length is longer than a certain fraction of the length of *either* element, the two elements are considered to belong to the same family. Otherwise, not.

### ***Family Graph Construction Procedure with Edge Re-evaluation***

1. Each element defined is represented by a vertex.
2. Edges are constructed as follows: if two elements are considered to belong to the same family by the family relationship determination procedure, a *primary* edge is constructed between the two corresponding vertices; if two elements form significant alignments but do not belong to the same family, a *secondary* edge is constructed between the two; if two elements do not form significant alignments, no edge is constructed between the two.

3. For each vertex  $v$ , its primary edges are re-evaluated as follows (Fig 2.5): Let  $N(v)$  denote the set of vertices directly connected to  $v$  via primary edges. If any pair in  $N(v)$  are connected by a secondary edge, then  $\forall v' \in N(v)$ , the primary edge between  $v$  and  $v'$  is removed unless  $v'$  is the most closely related element to  $v$  in  $N(v)$  (based on alignment score and/or percent identity) or  $v$  is the most closely related element to  $v'$  in  $N(v')$ . In the latter case, the primary edges of  $v'$  will be updated as just described.
4. Remove all secondary edges.

### 2.5.2 Implementation details

RECON starts from a datafile containing pairwise alignments, which allows a user to choose a tool other than WU-BLAST to do the initial all-vs-all comparison of the genome to itself.

A major issue is memory usage. To avoid holding all alignments from a genome-scale analysis in RAM at once, RECON manipulates files on disk (including a separate file for each currently defined element). It is therefore extremely I/O intensive.

RECON is not useful for processing short-period tandem repeats; these are split down to shorter forms or even monomers, a process which can take many iterations to converge. To improve time efficiency, we filter these by ignoring the initially defined elements which have more than 1,000 images and where the number of partner elements is less 1/5 of the number of images. Furthermore, since we discard short elements generated during splitting (element re-evaluation and update procedure), the whole family can suddenly disappear when it falls below the minimum element length cutoff.

Besides the threshold and parameter choices in the initial pairwise comparison, RECON has four tunable parameters:

- The cutoff for fractional overlap between images which is used in the initial inference of syntopy by the single coverage method. (Default = 0.5.)
- The minimum length of an element, e.g. the maximal length that we expect the pairwise alignment tool to spuriously extend by chance from a true element boundary, used in the image end selection rule and the element re-evaluation and update procedure. (Default = 30 nt.)
- The ratio cutoff for splitting an element at a given position, used in the element re-evaluation and update procedure. (Default = 2.)
- The minimal fraction of alignable sequences between two elements before they are considered to belong the same family. (Default = 0.9.)

To guide parameter optimization, we have developed an objective function that measures errors in the recovery of known families in a genome (Z. Bao and S. Eddy, unpublished). For detailed description of the function see Appendix B.

The default parameters were hand optimized based on the recovery of four experimentally verified DNA transposons (Tc1, Tc2, Tc3, and Tc5 (Plasterk & von Luenen, 1997)) from the *Caenorhabditis elegans* genome sequence (The *C. elegans* Sequencing Consortium, 1998). The human genome is dominated by retro-transposons (Alu, L1 and MaLR) and old, fragmented DNA transposons (Lander et al., 2001), and these families yield different patterns in multiple alignments than the young DNA transposons in the *C. elegans* training set, so the test on human data was reasonably independent of our training of these few parameters.

### **2.5.3 Human Genome Analysis**

3 MB of sequence was randomly sampled as 20Kb chunks from the 796 contigs in the Dec 12, 2000 release of the human genome (Lander et al., 2001)

(<http://genome-test.cse.ucsc.edu/goldenPath/12dec2000/bigZips>). All-*vs*-all comparison of the sampled sequences was done using WU-BLASTN 2.0 (Gish, 2002)

(<http://blast.wustl.edu>) with options `M=5 N=-11 Q=22 R=11 -kap E=0.00001 wordmask=dust wordmask=seg maskextra=20 -hspmax 5000`.

Known repeats were identified using the 7 July 2001 version of RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>), with default options.

Consensus sequences of RECON families were made by aligning the ten longest members of the family with DIALIGN2 (Morgenstern, 1999), with default options, then selecting a simple majority rule consensus residue for each column.

## **2.6 Acknowledgments**

We thank Dr. Elena Rivas for discussions and advice. The sequence sampling tool was provided by Mr. Robert Klein. We gratefully acknowledge financial support from the National Science Foundation (grant no. DBI-0077709) and the Howard Hughes Medical Institute.

## **Chapter 3**

# **Rlib: a Database of Automatically Constructed Repeat Sequence Libraries for Sequenced Genomes**

### 3.1 Abstract

Proper identification of repeats is an essential step for both genome annotation and studies of repeats. However, it is usually bottlenecked by the lack of a comprehensive repeat library for the subject genome. Here, I introduce the Rlib database which provides automatically constructed repeat sequence libraries for sequenced eukaryotic genomes. Tests here suggest that the RECON-based computational approach is capable to reconstruct reasonable consensus sequences for known transposable elements from both assembled sequences and shotgun reads, and can handle large and repeat-rich genomes like the mouse's. The chapter further describes the pilot experiments on two pairs of closely related genomes, the Nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* and the Cruciferae *Arabidopsis thaliana* and *Brassica oleracea*.

### 3.2 Introduction

Currently, the major tool for repeat identification in genome analysis is the RepeatMasker program (Smit & Green, 2002). RepeatMasker has two functional components: the sequence comparison tool and a set of libraries of known repeat sequences. The libraries, organized according to genomes/species, are used to search for repeats in their corresponding genomes. Sequences in these libraries are derived from Repbase (Jurka, 2000), a manually curated database for repeats. While the careful curation of Repbase gives the results of RepeatMasker relatively high quality, it also slows down the construction of libraries for all sequenced genomes. For example, about 18% of the *C. elegans* genome is repetitive (Sulston & Brenner, 1974), but the RepeatMasker/Repbase library for *C. elegans* covers only 9% of the genome, even though the sequencing of the genome was finished three years ago. For many genomes that are sequenced or being sequenced, there are

no corresponding libraries at all.

A similar situation also existed for protein domain identification. The manually curated domain database Pfam (Bateman et al., 2002), combined with the sequence comparison tool HMMER (Eddy, 2002), provides highly trusted annotations. However, it does not provide a complete coverage of all domains in the ever expanding collection of protein sequences. Fortunately, there exist automatically-built domain databases, such as ProDom (Corpet et al., 2000). Although the automation lowers the quality, these databases provide timely annotation of new domains revealed by new protein sequences. In addition, they can also be used to guide the expansion of the curated databases.

In order to meet the demand for timely annotation of repeats, I have begun to construct a database, dubbed Rlib, to provide automatically-built repeat libraries for the sequenced higher eukaryotic genomes (see <http://Rlib.wustl.edu>). Much of the data processing for Rlib is still in progress. So instead of giving a description of the final database, this chapter will focus on the initial works that demonstrate the feasibility and usefulness of Rlib. Specifically, I will introduce and validate a RECON-based approach for constructing repeat sequence libraries from genomic sequences. I will then describe four pilot experiments on two pairs of closely related species, the Nematodes *C. elegans* and *C. briggsae* and the Cruciferae *A. thaliana* and *B. oleracea*, and compare the evolution of repeats in these genomes. Finally, I will list the current target genomes for Rlib.

## **3.3 Results**

### **3.3.1 Repeat Library Construction: Strategy and Practical Issues**

Fig 3.1 illustrates the overall strategy, starting from the input genomic sequences to the final product of a repeat sequence library. The library is a FASTA format file containing all the consensus

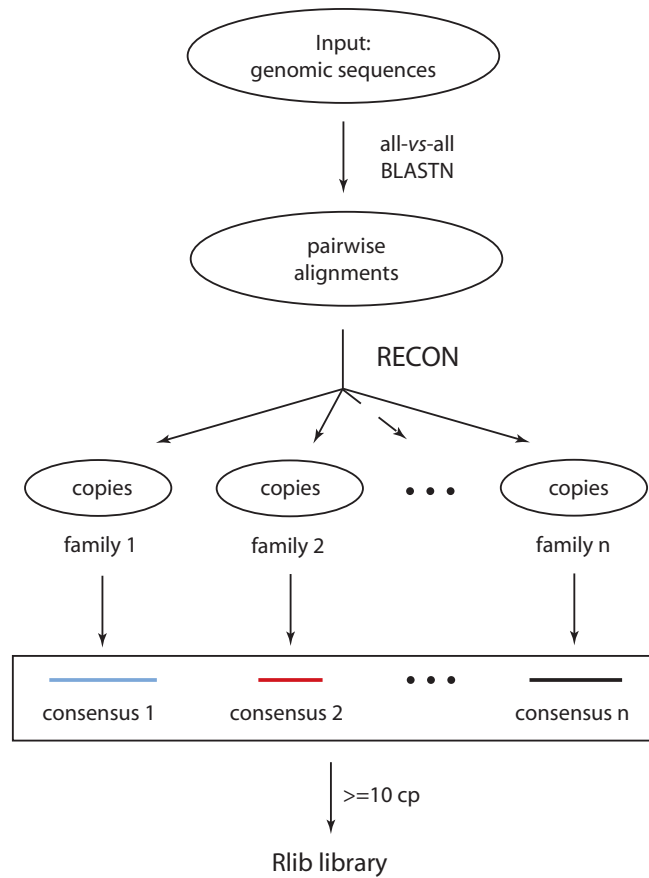


Figure 3.1: Flowchart of the strategy for repeat library construction. See Methods for details.

sequences built for the repeat families defined by RECON with ten or more copies. Such libraries, one per species, are provided at the Rlib database and can be used directly by RepeatMasker to annotate the corresponding genomes.

The same strategy and computational pipeline were used in the previous chapter to build consensus sequences for the RECON-defined repeat families in the 3Mb sample of the human genome. The close resemblance of those consensus sequences to their corresponding canonical sequences suggests that the strategy is promising. Yet, several questions remain to be answered. First, RECON was de-

signed with assembled sequences in mind. However, for timely construction of repeat libraries, one needs to deal with shotgun reads, as more and more genomes are being sequenced by the low-pass, whole genome shotgun strategy. Is this RECON-based approach able to analyze raw reads? Second, RECON is very memory intensive: the mere 3Mb sample of the human genome took 300MB of RAM (see Chapter 1). How should one approach the whole genome? Third, the approach obviously needs more extensive tests and assessments.

To answer these questions, I analyzed a 1 Gb sample of the mouse genome in shotgun reads, using the RECON-based strategy as illustrated in Fig 3.1. Detailed analysis of the performance of the approach follows. For further assessments, see also the pilot experiments.

### *Handling shotgun reads*

For repeats that are shorter than the typical length of shotgun reads (600 - 700 nt), raw reads should work as fine as assembled sequences. For longer repeats, each read is just a fragment of the whole repeat. Two problems stand in the way of identifying these repeats: whether RECON could properly cluster the fragments in a family and how to reconstruct full-length consensus from the fragments in a RECON-defined family.

There are two major components of RECON that could be affected by using raw reads. One is the detection and splitting of composite elements. As long as the ends of the reads are random relative to the repeats, this part of RECON should not be affected. The other component is to distinguish related but distinct families when clustering elements into families, and in particular to determine which of the related families a partial element belongs to. The test on the mouse reads (see below) suggests that the greedy algorithm used by RECON tends to undercluster in presence of many partial elements. That is, one biological family will be reported as several defined families,

each containing a subset of the members of the true family.

In theory, one can infer a full-length consensus by making a multiple alignment of the reads in a family. However, the available multiple alignment tools do not perform well on highly fragmented sequences (data not shown). Therefore, I have developed a mini-assembly algorithm to assemble the reads into longer sequences before the multiple alignment. The algorithm does not attempt to identify reads from the same repeat locus, which is known to be difficult. Instead, it assembles reads according to the relative position of a read to the full-length consensus and to other reads. For details, see Methods.

### ***Handling large genomes***

One possible way to analyze a large and repeat-rich genome like the mouse's is to take an incremental approach. That is, first find the most abundant families in a small sample of the genome, then analyze progressively larger samples after masking elements in the families that have already been confidently identified.

When tested on the 1Gb sample of mouse reads (Fig 3.2), this incremental approach greatly reduces the memory and time usage for the larger samples (compare the analyses of the 30Mb sample with and without the first round). Overall, one can sufficiently sample the  $\sim$  3Gb mouse genome with affordable memory usage and reasonable time by the incremental approach.

The analysis was stopped after the fourth round (sampling about one third of the genome) for the following reasons. First, the sampling should be sufficient to cover most of the moderately repetitive families (20 to 50 copies per genome). Second, the genome coverage yielded by each round has begun to plateau. Third, since shotgun reads were used, redundant reads would become more prevalent as the sampling gets deeper into the genome. The redundant reads will be detected

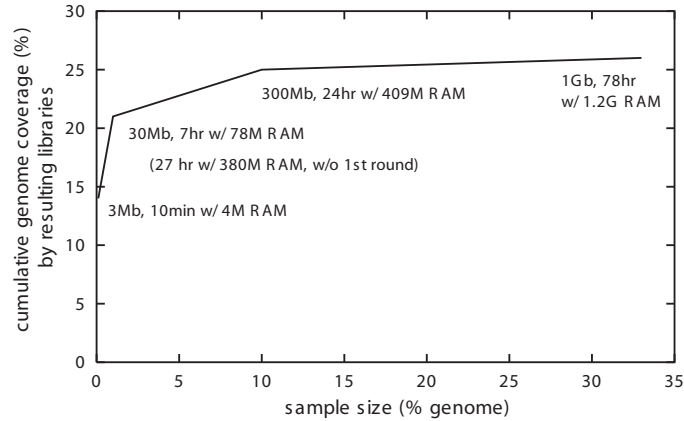


Figure 3.2: Efficiency of the incremental analysis on a 1 Gb random sample of mouse shotgun reads. Time and memory usage is based on a Linux machine with one Intel Xeon CPU (1.7 MHz) and 1 GB RAM.

as “repeats” and cause unnecessary complications for RECON.

### ***Assessments***

*Quality of the recovered consensus* The recovered consensus sequences tend to be fragmented compared to their canonical sequences. For example, the 6.6kb L1 element is recovered in five consensus (Fig 3.3A). As mentioned above, this is because of the underclustering caused by fragments.

Despite this, recovered consensus can be much longer than the typical length of reads (Fig 3.3B). In three cases, the recovered consensus sequence is longer than the corresponding sequence used by RepeatMasker (circled in Fig 3.3B). In one case, the sequence is a tandem repeat and our consensus contains more copies of monomers. In another case, the RepeatMasker sequence is a partial copy of L1. In the third case, the RepeatMasker sequence is a solo LTR while our consensus

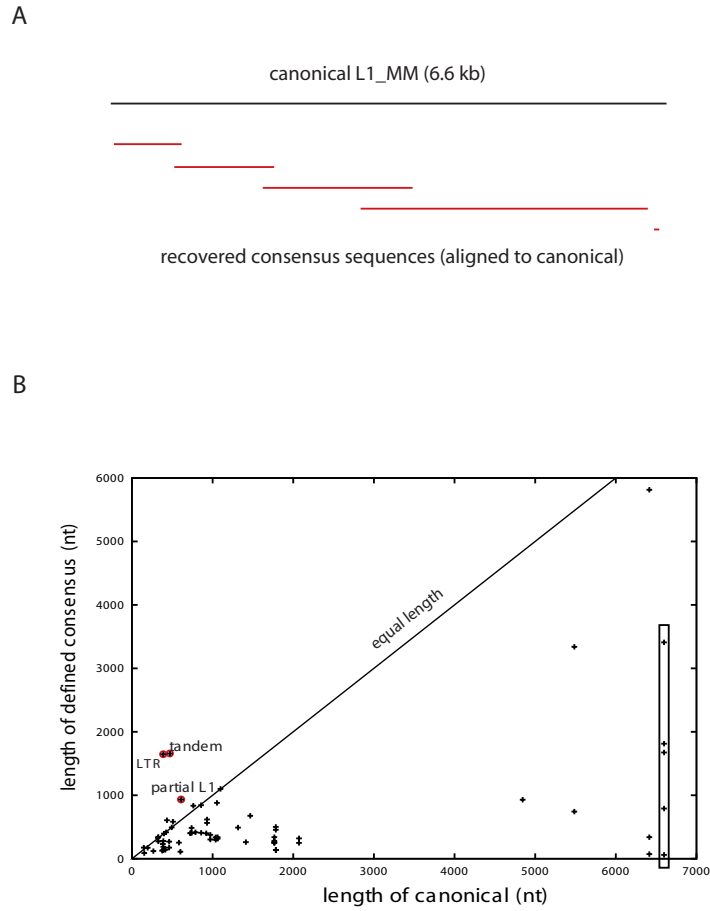


Figure 3.3: A. The recovery of the mouse L1 element. B. Comparison of the Rlib consensus and their corresponding canonical sequences in the built-in library of RepeatMasker. Circled points are discussed in text. Points in box correspond to consensus in panel A.

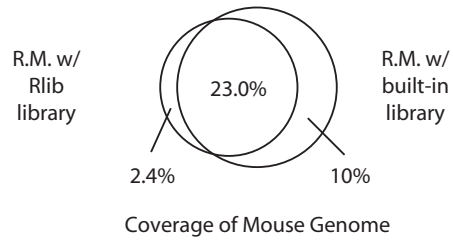
contains both the LTR and the internal sequence.

*Sensitivity: what have we missed?* Compared to the built-in library of RepeatMasker, the final library constructed above covered over two thirds of the known repeats in mouse, plus some new repeats (Fig 3.4A). About a third of the known repeats are not recovered in the above analysis. This is mostly because of the moderate sensitivity used in the initial BLASTN comparison of the reads: most of the repeats that are less than 75% identical to their most similar copies in the sample are not detected (Fig 3.4B). This only affects the identification of those families that have few or none young copies. When a family has both old and young copies, the old ones missed by the initial BLASTN comparison can be recovered by searching the genome with the consensus sequence built on the young ones, using RepeatMasker. (As mentioned in Chapter 2, RepeatMasker uses a different sequence comparison tool called CROSSMATCH with more aggressive sensitivity.) As shown in Fig 3.5, about 5.2% of genome can be recovered this way.

Some other families are missed because of the errors in RECON. As shown in Fig 3.5, a small fraction (1.4%) of the BLASTN-detected repeats (with  $\geq 10$  copies) are not recognized by the final library. Manual inspections suggest that in these cases, it is because sequences of several biological families were lumped in one defined “family”. Consensus of these mixed families are typically nonsense which can not mask any sequence.

*Conclusion* The test on the 1 Gb sample of the mouse genome suggests that the RECON-based approach can analyze shotgun reads and large genomes can be analyzed via the incremental approach. The yielded repeat library can mask the genome thoroughly (for purposes of other annotations). However, for the study of repeats, it still needs more curation. The main issue in such curation is the fragmentation of consensus sequences (Fig 3.3A). However, once these fragmented sequences are identified, it is relatively straight forward to merge them into full-length sequences.

A



B

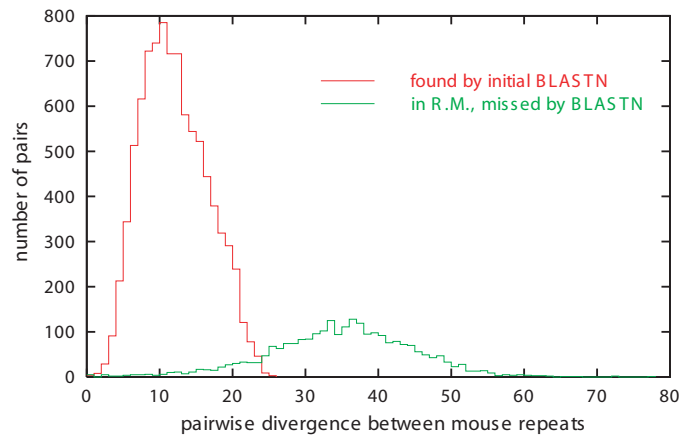


Figure 3.4: A. Comparison of genome coverage by the Rlib library and the built-in library for mouse in RepeatMasker. The Rlib library is used with RepeatMasker (with -nolow option). Built-in libraries are invoked by the “-mus” option plus the -nolow option. B. Distribution of pairwise sequence divergence of the mouse repeats.

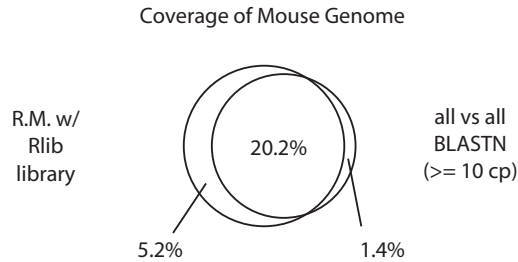


Figure 3.5: Comparison of repeats recognized by the Rlib library and by the initial all-*vs*-all BLASTN.

(For automatic identification of fragmented consensuses, see Discussion.)

Compared to RepeatMasker, our BLASTN-based detection of repeats is less sensitive. One could of course use CROSSMATCH and the sensitive settings of CROSSMATCH in RepeatMasker for the initial all-*vs*-all comparison. However, we have chosen not to do so — the time it would take makes whole genome comparisons impractical with current computers.

Generally, analyzing all available sequences at once and using assembled sequences would likely to yield better results. So one should only switch to incremental analyses and raw reads when the alternative is not available. In the test above, I only compared the final results to a “trusted answer”. For better controlled assessments, one may want to compare the result of an incremental analysis to that of a one-shot analysis of the same sequences, or to compare the result based on raw reads to that based on assembled sequences of the same genome. However, to certain extent, such benchmarks are not vital, so long as the final results are reasonable (like the test on mouse here).

Table 3.1: Summary of the Pilot Experiments for Rlib

Genome	Genome	Sample	Rlib Library		% genome unreported		
	Size (Mb)	Size (Mb)	# fam	% genome	$\geq 10$ cp	$< 10$ cp	tandem
<i>C. elegans</i>	99	99	554	16	0.4	5	0.2
<i>C. briggsae</i>	$\sim 110$	107	723	31	0.6	7	$< 0.1$
<i>A. thaliana</i>	125	116	973	16	2	13	0
<i>B. oleracea</i>	$\sim 650$	114	1198	41	1.4	13	0

### 3.3.2 Pilot Experiments

The four pilot experiments are summarized in Table 3.1. For *B. oleracea*, shotgun reads were used as the input, which corresponds to about one sixth of the genome. For the other three, assembled sequences were used, which represent essentially the whole genomes. The *C. briggsae* and *B. oleracea* genomes were analyzed in three rounds (1% of genome, 10% of genome then all available sequences). Based on lessons in the mouse genome, consensus with less than 10 hits in the genome/sample are eliminated: since consensus are built only for families with  $\geq 10$  copies, the consensus should have at least 10 hits in the genome/sample. Therefore, those with less than 10 hits are considered low quality consensus, either due to the conflation of biological families or errors in consensus inference. Table 3.1 is based on the filtered libraries.

The unreported repetitive fraction of each genome was estimated by all-vs-all BLASTN comparison after masking the genome with the respective library using RepeatMasker. In each genome, there was a small fraction of repeats with  $\geq 10$  copies that are not represented in the reported library, due to the limitation of RECON's ability to correctly cluster loci into families (see also the mouse

test and Fig 3.5).

To trade for efficiency, RECON ignores long tandem arrays and does not include those in its output (see Chapter 2). However, as shown in Table 3.1, only a small fraction of the tandem repeats were not represented in the reported library. This is because consensus sequences of most of the tandem repeat families can be recovered from the shorter loci, which are not ignored by RECON.

Compared to the built-in libraries of RepeatMasker, the libraries built here contain many new families in *C. elegans* and *A. thaliana* and have doubled the genome coverage (Fig 3.6). More families are missed in *A. thaliana* because families in this genome tend to be less numerous, as indicated by Table 3.1 (almost twice as many families as *C. elegans* for comparable repeat content, and significantly higher fraction with <10 copies) and Fig 3.7.

Fig 3.7 summarizes the length and copy number of each family reported in each genome. Due to the limited sequence available for *B. oleracea*, families with 10 to 60 copies in the genome are not covered in the library. Notice that our definition of repeats is solely based on copy number, so some of the reported repeats are multi-copy cellular genes. For example, the histone gene cluster is reported in *C. elegans*. Also, the largest families in *B. oleracea* are fragments of the chloroplast genome, or the mitochondrial genome, or the rRNA gene cluster.

### **3.3.3 Use of Rlib: understanding repeat evolution**

While genome annotation is an important issue, the ultimate goal for studying repeats is to understand the evolution of repeats and its impact on genome evolution and genome organization. Even at a very crude stage, the Rlib libraries could shed lights on the dynamic nature of repeat evolution, as demonstrated below by the comparative studies of the worm and mustard genomes.

The four Rlib libraries constructed in this chapter confirm the early observation that closely

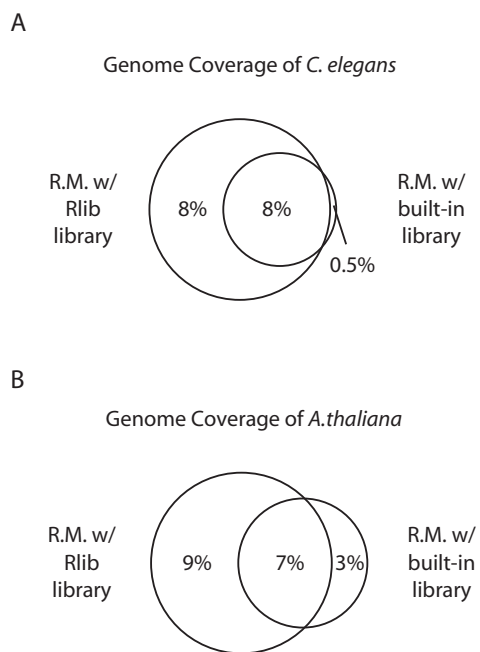


Figure 3.6: Comparison of the Rlib library and the built-in library of RepeatMasker. Rlib libraries are used to search the corresponding genome with RepeatMasker (with `-nolow` option). Built-in libraries are invoked by the `-el` option for *C. elegans* and `-ar` for *A. thaliana*, plus the `-nolow` option.

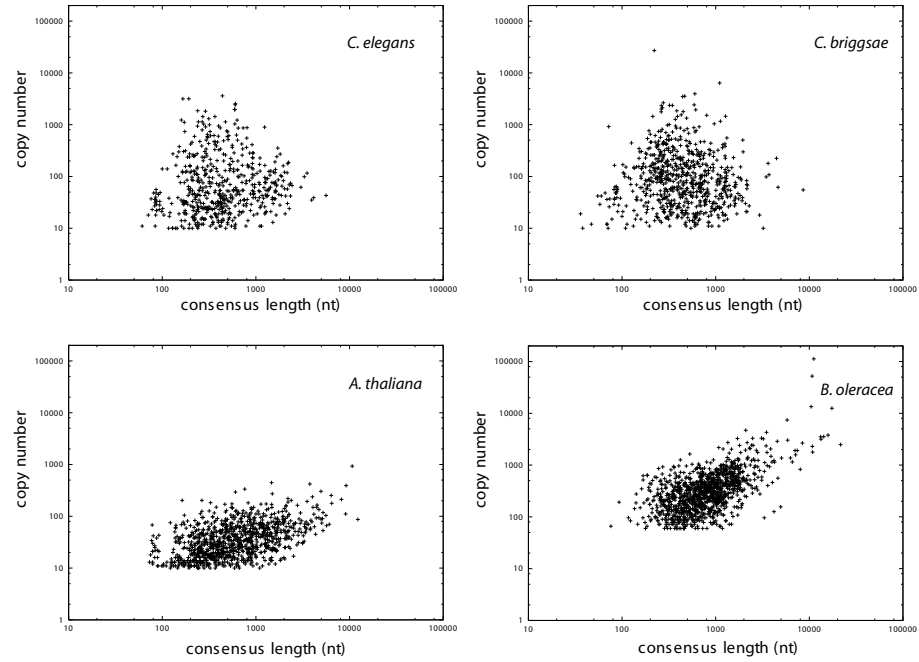


Figure 3.7: Summary of the repeat families reported in the Rlib libraries. Each point represents a defined family. For *B. oleracea*, copy number is estimated as number of reads, then projected to the whole genome, in which case the copy number cutoff of 10 in the sample is equivalent to about 60 in the whole genome.

Table 3.2: Similarity of Repeats Across Species

Target Genome	% Masked By Library of			
	<i>C. elegans</i>	<i>C. briggsae</i>	<i>A. thaliana</i>	<i>B. oleracea</i>
<i>C. elegans</i>	<b>16</b>	<b>3.5</b>	0.3	0.3
<i>C. briggsae</i>	<b>8</b>	<b>31</b>	0.2	0.2
<i>A. thaliana</i>	0.2	0.1	<b>16</b>	<b>4.2</b>
<i>B. oleracea</i>	0.1	0.1	<b>13.5</b>	<b>41</b>

related species contain similar repetitive sequences (Britten & Kohne, 1968). For example, 130 families in *A. thaliana* and 385 in *B. oleracea* are similar to known LTR elements (40% and 74% of the classifiable families, respectively. see Methods for classification). In contrast, only one family in *C. elegans* and none in *C. briggsae* are recognized as LTR elements. (The known LTR elements in *C. elegans* (Ganko et al., 2001) are not reported due to low copy number.) Instead, the predominant type are MITE-like (Miniature Inverted-repeat Transposable Element) repeats, which are 200 to 500nt long and flanked by short inverted repeats. According to manual inspections, seven out of the ten most numerous families in *C. elegans* and nine out of ten in *C. briggsae* belong to this class.

Overall, the libraries can mask about one fourth of the repetitive fraction in the closely related genome compared to the native library (Table 3.2), while cross recognition between the worm and mustard genomes is minimal. This does not mean that one fourth of the repeats are common to each closely related pair. Rather, it is due to the similarity between homologous families. Generally, the similarity between homologs in two genomes is restricted to parts of the consensuses, which are presumably functionally conserved regions. It is yet clear if the significant extant of homologs between the closely related species is due to the “founder effect” in the ancestral genome, or it is

because related genomes tend to be permissive to similar types of repeats in terms of horizontal transfer.

Despite the similarities described above, there are considerable differences between the closely related genomes. Although the two nematode genomes are similar in size, *C. briggsae* contains twice as many recognizable repeats as *C. elegans* (Table 3.1). The recognized repeats are presumably the younger additions to the genome. Therefore, if one assumes comparable mutation rates in the two organisms, this result suggests that *C. briggsae* acquired twice as many repetitive sequences in the recent past as *C. elegans*. If one could further assume that the size of the two genomes remained more or less constant in the recent past, then repetitive sequences must have been deleted more aggressively in *C. briggsae* (in order to balance the faster acquisition). That is, repetitive sequences recycle faster in *C. briggsae* than in *C. elegans*.

The difference between *A. thaliana* and *B. oleracea* is more significant. Even with the current under estimate, the total length of repeats in *B. oleracea* is 15 times more than that in *A. thaliana* (Table 3.1). Since both genomes are believed to be diploid, as much as 5/6 of the *B. oleracea* genome, or five times worth of the *A. thaliana* genome could be repeats. Interestingly, not all types of repeats in *A. thaliana* are over-amplified in *B. oleracea*. For example, 29 families in *A. thaliana* are similar to known Helitrons, a newly identified class of DNA transposon (Kapitonov & Jurka, 2001), adding up to 1935kb in the 116Mb genome. However, only 7 families in *B. oleracea* are similar to known Helitrons, adding up to only 337kb in the 114Mb sample.

Drastically different abundance between homologous families is also seen in the nematodes. For example, the Cb000010 family in *C. briggsae* is a nonautonomous DNA transposon in the Tc1 transposon superfamily and has over 3,000 copies. The known Tc1-type transposons in *C. elegans* on the other hand usually have 10 to 30 copies (up to 300 copies in certain strains). Thus, the

amplification of homologous families in related genomes must be independent events shaped by various stochastic factors.

## 3.4 Discussion

### 3.4.1 More to Come

The pilot experiments show that even for the widely studied genomes like *C. elegans* and *A. thaliana*, repeat analysis is still less than adequate. Furthermore, as shown in Table 3.2, using repeat libraries of closely related species are not sufficient for the newly sequenced genomes. Therefore, we plan to cover all sequenced higher eukaryotic genomes in the Rlib database, and expand as new genomes are released. Table 3.3 lists our current targets. Besides the four genomes discussed above, three other genomes have also been analyzed (species names in bold in the table): the zebrafish *D. rerio* (in collaboration with Dr. Stephen Johnson), the sea squirt *C. intestinalis* and the rice *O. sativa*. The analysis of the rice genome will be discussed in the next chapter. For the human and the mouse genome where curated libraries are already substantial, we will focus on finding repeat families not yet identified.

Following the incremental approach, RECON is no longer the bottleneck in building Rlib libraries. Instead, the most time consuming step becomes the subsequent masking of genomes/samples using RepeatMasker. Roughly speaking, the whole process of analyzing 1Gb of sequences takes about two weeks on the Eddy lab Linux cluster, which has 128 Intel Pentium III 1GHz CPUs.

A vital issue in high-throughput genome analysis is quality control. So far, the quality assessments in this and the previous chapters have relied heavily on the comparison to the curated libraries from the same genome. However, such assessments would not be applicable to future analyses —

Table 3.3: The Initial Targets for Rlib

Species	Common Name	Genome Size (Mb)	Available Sequences (Mb)	Sequence Status
<i>Homo sapiens</i>	human	3,000	3,000	assembled
<i>Mus musculus</i>	mouse	2,700	2,500	assembled
<i>Rattus norvegicus</i>	rat	3,000	16,000	reads
<b><i>Danio rerio</i></b>	zebrafish	1,700	5,000	reads
<i>Tetraodon nigroviridis</i>	pufferfish	400	319	assembled
<i>Fugu rubripes</i>	pufferfish	375	350	assembled
<i>Ciona savignyi</i>	sea squirt	180	390	assembled
<b><i>Ciona intestinalis</i></b>	sea squirt	175	1,000	assembled
<i>Drosophila melanogaster</i>	fruitfly	150	137	assembled
<i>Anopheles gambiae</i>	mosquito	260	278	assembled
<b><i>Caenorhabditis elegans</i></b>	worm	99	99	assembled
<b><i>Caenorhabditis briggsae</i></b>		100	108	assembled
<b><i>Arabidopsis thaliana</i></b>		125	116	assembled
<b><i>Brassica oleracea</i></b>		650	114	reads
<i>Medicago truncatula</i>	annual alfalfa	500	14	assembled
<b><i>Oryza sativa</i></b>	rice	430	366	assembled

the very reason to construct a repeat library is because there is not (a comprehensive and reliable) one available. Three alternative methods are worth considering as they do not rely on the existence of a “trusted answer”.

The first is to measure the fraction of the genome which is detected by the initial all-*vs*-all BLASTN as repeats with  $\geq 10$  copies but is not covered by the Rlib library (see Fig 3.5 and the sixth column in Table 3.1). Since this fraction is composed of the families that RECON fails to resolve properly, it can be used to measure the overall quality of RECON’s definition of families.

The second method aims to evaluate the reliability of individual families. As mentioned in the pilot experiments, if a consensus sequence has less than 10 hits in the genome/sample, it should be considered low quality and removed from the library.

The third method measures the other type of common errors, fragmentation of consensus sequences, which exists for analyses of both assembled sequences (see Chapter 2) and raw reads (see Fig 3.3A). Copies of the fragmented families, such as those (red lines) in Fig 3.3A, would tend to be next to each other in the genome, more so than expected by chance. Therefore, one can devise a statistical test to identify these families. Such identification is not only a quality measurement, but also valuable information for manual curations of the library.

Another relevant post-processing is to identify families of multi-copy genes by comparing the sequences in an Rlib library to the known genes in Genbank. Although it is not a quality measurement, it would be useful for users of the library.

All these methods will be implemented as part of our future plan of Rlib.

### 3.4.2 Usage of the Rlib libraries

The Rlib libraries have already been put to use by the genome research community. The *C. elegans* library has been incorporated in WormBase (version 77 and beyond, D Larson and K Jekosch, personal communication). The *C. briggsae*, zebrafish and *Ciona intestinalis* libraries have been adopted by their sequencing consortia for the initial genome analysis. In addition, a RepeatMasker server using the Rlib library has been set up for zebrafish at the Sanger Center for the public to analyze their own favorite sequences (see [http://www.sanger.ac.uk/Projects/D\\_rerio/fishmask.shtml](http://www.sanger.ac.uk/Projects/D_rerio/fishmask.shtml)).

Obviously, producing the automatically-built repeat libraries is only the beginning for the studies of repeats. The immediate next step is to develop manually curated repeat databases with better sequence models and classification. To support such effort, Rlib will be freely available to the public.

## 3.5 Methods

### 3.5.1 Input Sequences

Genomic sequences were downloaded as follows.

mouse: [ftp.ncbi.nlm.nih.gov/TraceDB/mus\\_musculus/ClipReads](ftp.ncbi.nlm.nih.gov/TraceDB/mus_musculus/ClipReads), 06/27/2001;

*C. elegans*: [ftp.sanger.ac.uk/pub/C.elegans\\_sequences/CHROMOSOMES/001602](ftp.sanger.ac.uk/pub/C.elegans_sequences/CHROMOSOMES/001602), 11/07/2000;

*C. briggsae*: from the sequencing consortium, 02/05/2002;

*A. thaliana*: [ftp.tigr.org/pub/data/a\\_thaliana/ath1/SEQUENCES](ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES), 11/06/2001;

*B. oleracea*: [ftp.cshl.org/pub/sequences/brassica\\_shotgun/cshl](ftp.cshl.org/pub/sequences/brassica_shotgun/cshl), 11/06/2001.

Assembled sequences were split into 20Kb fragments as input for the initial all-vs-all BLASTN.

For the incremental approach, sequence sampling was done by randomly choosing reads or the 20Kb

fragments.

### **3.5.2 Identification of Repeat Families**

For the initial all-*vs*-all comparison of the input sequences, I use WU-BLASTN 2.0 (Gish, 2002) with options `-kap E=0.00001 wordmask=dust wordmask=seg maskextra=20 -hspmax 5000`. For mouse, a built-in scoring matrix was used (M=5 N=-11 Q=22 R=11). For the others, a custom AT-rich matrix was used (available at Rlib website).

In an incremental analysis, I first mask a new sample with consensus built in the previous rounds, using RepeatMasker (version 6.8) with the `-nolow` option.

All parameters for RECON were default(see Chapter 2).

### **3.5.3 Inference of Consensus Sequences**

When using assembled sequences, I infer the consensus by aligning the ten longest members of the family with DIALIGN2 (Morgenstern, 1999), with default options, then selecting a simple majority rule consensus residue for each column in the multi-alignment.

When using shotgun reads, I first assemble the reads into “templates”, then infer the consensus from the ten longest templates as described above. Templates are assembled from reads in a greedy approach — starting from the longest read in the family, find the two reads, one for each end of the chosen longest read, that overlap with the chosen read and extend the furthest from it; keep extending till no more reads can be added. Thus, the mini-assembly is based on the relative positions of the reads to the full length consensus, regardless of which genomic loci they are from.

Finally, I discard consensus sequences with less than 10 hits in the final sample of the genome, using RepeatMasker (see below), for these are likely low quality consensus produced by erroneous

clustering and/or consensus inference.

### **3.5.4 Genome Survey and Repeat Classification**

Copy number and genome coverage of the identified families are based on the annotation of the genome/sample by RepeatMasker (version 6.8) using the corresponding library with the -nolow option.

Classification of the identified repeats is based on their similarity to known repeats in the RepeatMasker libraries, using RepeatMasker as the alignment tool.

## **3.6 Acknowledgment**

I thank the mouse, *C. briggsae* and *B. oleracea* sequencing consortia for releasing their sequences prior to publication.

## **Chapter 4**

### **Repeats and Their Insertion**

### **Polymorphism in the Rice Genome**

## 4.1 Abstract

The rice *Oryza sativa* appears to be a unique system to study the microevolution of transposable elements. In this chapter, I evaluate the potential of rice for studying such questions by assessing the level of transposon insertion polymorphism among rice cultivars and related wild species. A total of 1256 repeat families were identified from the genomic sequences of the subspecies *japonica*. These families cover  $\sim 37\%$  of genome and give rise to an estimate of  $\sim 550,000$  insertions in the genome, most of which are transposons. About six to ten percent of the insertions are polymorphic between *japonica* and another sequenced subspecies, *indica*. Insertion polymorphism in the unsequenced cultivars and wild species was assessed using an experimental technique called Transposon Display (TD). TD of two identified transposable elements, *Dasheng* and *mPing*, suggests that insertion polymorphism is extensive at all levels: between species, between subspecies and between cultivars within subspecies. Based on these results, we have formed a collaboration to systematically characterize transposable elements and their insertion polymorphism in various rices.

## 4.2 Introduction

The rice *Oryza sativa* is a unique model to address questions of microevolution because of the wide and well-documented collection of cultivars and related wild species (Matsuo et al., 1997). In addition, rice is currently the only higher organism for which the genomes of two subspecies are being sequenced. Recently, a draft sequence covering 92% of genome has been produced for the subspecies *indica* (cultivar 93-11) by whole-genome shotgun sequencing (Yu et al., 2002). In the meantime, the International Rice Genome Sequencing Project (IRGSP) has generated finished

sequences for about half of the genome for the subspecies *japonica* (cv Nipponbare) (Leach et al., 2002). In addition, two draft sequences were also produced for Nipponbare by private companies (Butler & Pockley, 2000; Goff et al., 2002), albeit with limited access.

At roughly 430Mb, rice has the smallest genome among the cereal crops. Yet, at least 40% of the genome is repetitive sequences (Yu et al., 2002). The known repeat families, which are largely transposable elements, only account for about 13% to 16% of the genome. Evidences suggest that repeats might be highly polymorphic between the two subspecies: many kilobase-sized indels were found in the overlapping sequences between *indica* and *japonica*. These indels coincide with clusters of highly repetitive 20mer-words (Yu et al., 2002). In addition, the genome sizes of the two subspecies are different. Based on whole genome shotgun sequencing, the *indica* genome is estimated to be 466 Mb (Yu et al., 2002), while the *japonica* is only 422 Mb (Goff et al., 2002).

In this chapter, I will evaluate the potential of rice as a model system to study the evolution of repeats and their impact on the evolution of the genome. I will conduct a systematic identification of the repeat families in the rice genome and assess their insertion polymorphism among the rice cultivars, especially between the two sequenced cultivars, using both computational and experimental approaches.

## **4.3 Results**

### **4.3.1 Identification and classification of repeat families**

The IRGSP sequences were used to identify the repeat families and also for the rest of the analysis concerning the *japonica* genome. Using the computational strategy described in the previous chapter, a total of 1256 repeat families were identified from the 187Mb of the IRGSP sequences

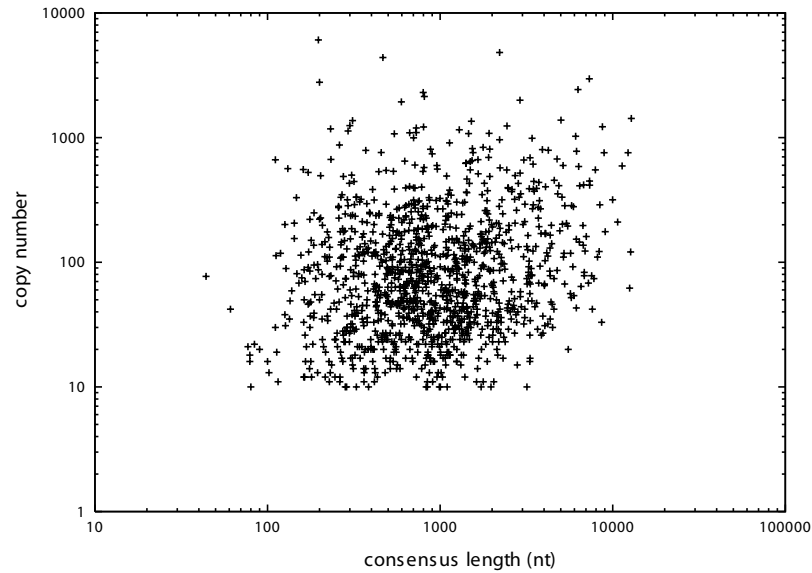


Figure 4.1: Summary of the 1256 identified repeat families. Each point represents a family. The horizontal axis is the length of the consensus sequence of a family. The vertical axis is the number of hits of a family in the 187 Mb *japonica* sequences, using RepeatMasker with the consensus sequence.

available in December 2001, with a minimal cutoff of ten or more copies in the sequences. These families cover ~37% of the sequences, including the 13% covered by the 130 previously known families<sup>1</sup>. In total, the 1256 families give rise to about 250K insertions in the 187Mb sequences, leading to an estimate of about 550K insertions in whole genome. The consensus length and copy number of these families are summarized in Fig. 4.1.

The identified families were classified based on their sequence similarity to the known families

<sup>1</sup>The *indica* genome paper reported that 16% (Yu et al., 2002) of assembled sequences were derived from “known” transposable elements. However, the repeat library used is not published. The *japonica* genome paper (Goff et al., 2002) did not provide summary statistics on genome coverage by repeats. Dr. Ning Jiang at University of Georgia, Athens conducted a thorough literature search and collected 130 rice repeat families (mostly transposable elements). These families cover 13% of 187Mb *japonica* sequences released by IRGSP.

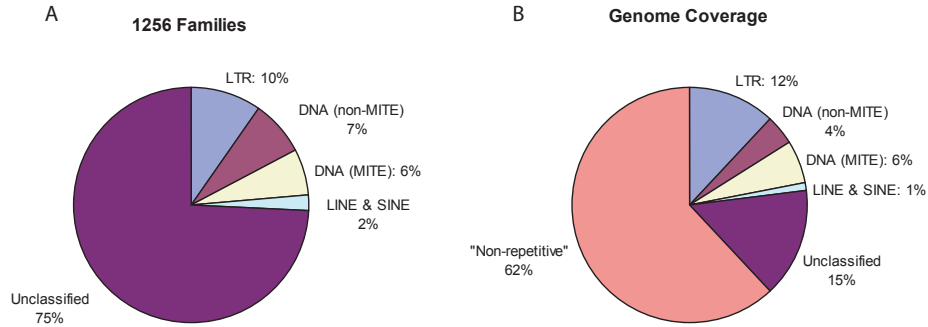


Figure 4.2: Classification of the identified families. Left panel shows the percentage of the 1256 families in each class. Genome coverage by each class is estimated in the 187 Mb *japonica* sequence. MITEs are list separately for historical reasons.

(see Methods). As in many other eukaryotic genomes, transposable elements are a major component. About 25% of the identified families can be recognized as transposable elements (Fig. 4.2A), and these families account for more than 60% of the repetitive fraction of the genome (Fig. 4.2B). The unclassified families are less numerous. Some of these families could as well be transposable elements. For example, MITEs are not easily recognizable on the sole basis of similarity to related MITEs, as the conserved regions are short. For more sensitive classification, one may need to rely on structural features of various transposable elements, such as terminal inverted repeats.

### 4.3.2 Physical distribution of repeats in the genome

To characterize the physical distribution of the repeats, we measure the length distribution of the inter-repeat regions (Fig. 4.3A). If repeats insert randomly into the genome, then the measured distribution can be approximated by a geometric distribution (red line in Fig. 4.3B). The green points in Fig. 4.3B represent the distribution from a computer simulation, assuming the same genome size

and same number of insertions (see Methods). As seen in the figure, the variance (reflected by the deviation of green points from the red line) is small, because the number of insertions is large.

The observed distribution in the 187 Mb *japonica* sequences (blue points in Fig. 4.3B), however, deviates significantly from the random distribution. In particular, there are more inter-repeat regions that are longer than 2.8 Kb, compared to the expectation by chance (red line). It has been reported that repeats seem to avoid genic regions in rice (Yu et al., 2002). Therefore, the large inter-repeat regions could be genes. Unfortunately, such correlation cannot be tested at the moment, as the complete gene prediction in the *japonica* sequences is yet available.

In the meantime, there are also more inter-repeat regions that are shorter than 50 bp compared to the random distribution. This suggests that many repeat insertions in rice aggregate into more or less contiguous blocks. In practice, we define a block as a maximal set of repeats where the inter-repeat regions between each adjacent pair of repeats are less than 50 bp. Under such a definition, the 250K or so insertions fall into about 120K blocks, half of which contain more than one insertion. These blocks can be as long as 70 Kb, containing several dozens of insertions (Fig. 4.4).

### **4.3.3 Insertion polymorphism between the two sequenced cultivars**

The level of insertion polymorphism between the two sequenced cultivars was estimated in two ways. First, highly polymorphic families were identified by comparing the copy numbers of each family in the two genomes (Fig. 4.5). About 10% of all the identified families have more than two-fold difference in copy number between the two genomes (green lines). Specifically, 40 families have twice or more copies in *japonica* than in *indica*. For these families, there are 2,100 more copies in *japonica*. At the other end, 85 families have twice or more copies in *indica*, containing 29,000 more copies.

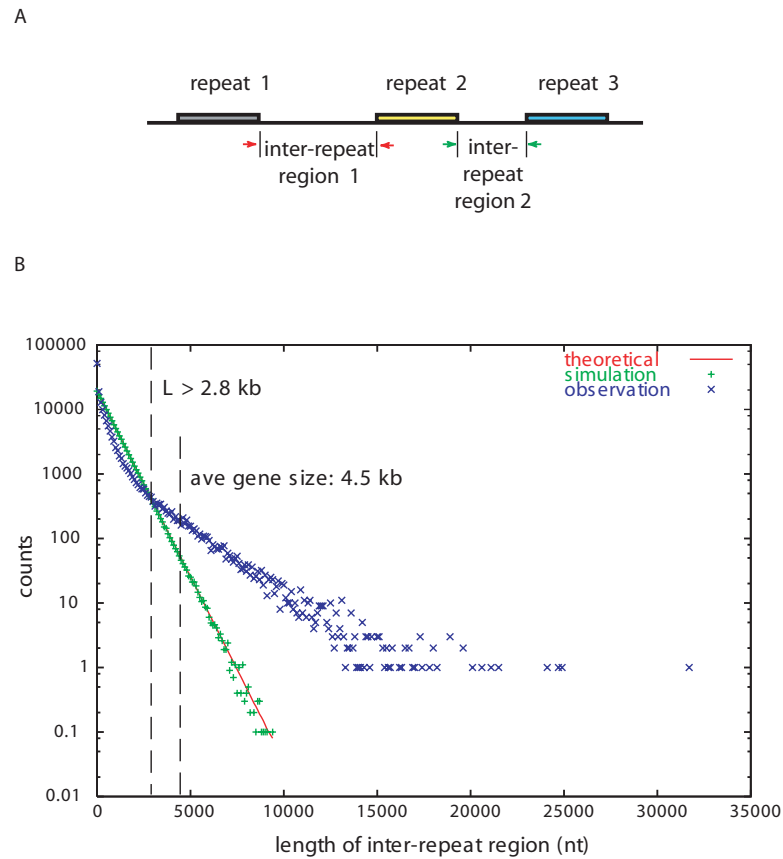


Figure 4.3: Physical distribution of repeats in the rice genome. A. Physical distribution is measured by the distances between adjacent insertions, i.e., the length distribution of the inter-repeat regions. B. The physical distribution of repeats in rice is not random. Inter-repeat regions are binned every 50 nt (horizontal axis). The vertical axis is the number of regions in each bin. Blue points represent the distribution in the 187 Mb *japonica* sequences. Green points represent the random simulation. Red line represents the theoretical approximation of random insertion, which is a geometric distribution.

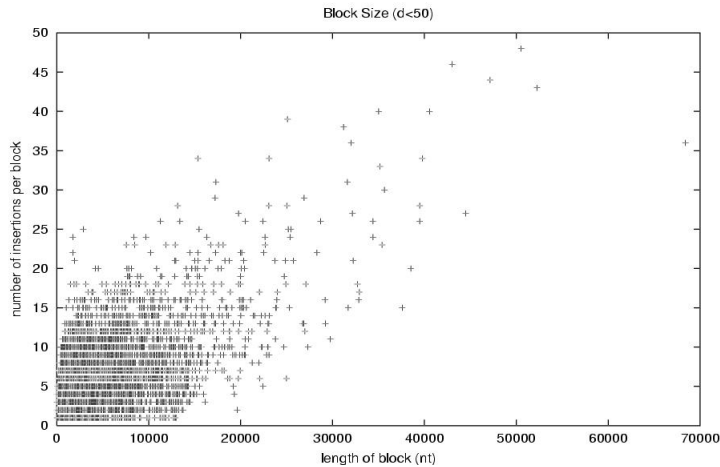


Figure 4.4: Blocks of repeats in the 187 Mb *japonica* sequences. Each point represents a defined block.

One could see in Fig. 4.5 that families generally have more copies in *indica* (the average ratio of copy number in *indica* vs copy number in *japonica* is 1.24). This is consistent with the observations that *indica* has a bigger genome. However, one also has to consider the biases in the *japonica* sequences — *japonica* is sequenced on a BAC-by-BAC basis starting with repeat-poor regions, while *indica* is sequenced by the whole-genome shotgun strategy, which is a more random sampling of the genome.

Second, I estimated the overall rate of polymorphism by identifying the polymorphic insertions. The basic strategy is to take the flanking sequences of a repeat locus in one of the genomes, map the position of the flanking sequences in the other genome and see if the same repeat is there (see Methods for details). In total,  $\sim 6\%$  of the repeat insertion loci in *japonica* are not in *indica*, and  $\sim 10\%$  of the repeat insertion loci in *indica*, are not in *japonica*. This gives the estimate of 12,000 *japonica*-specific insertions and 16,000 *indica*-specific insertions in the whole genome.

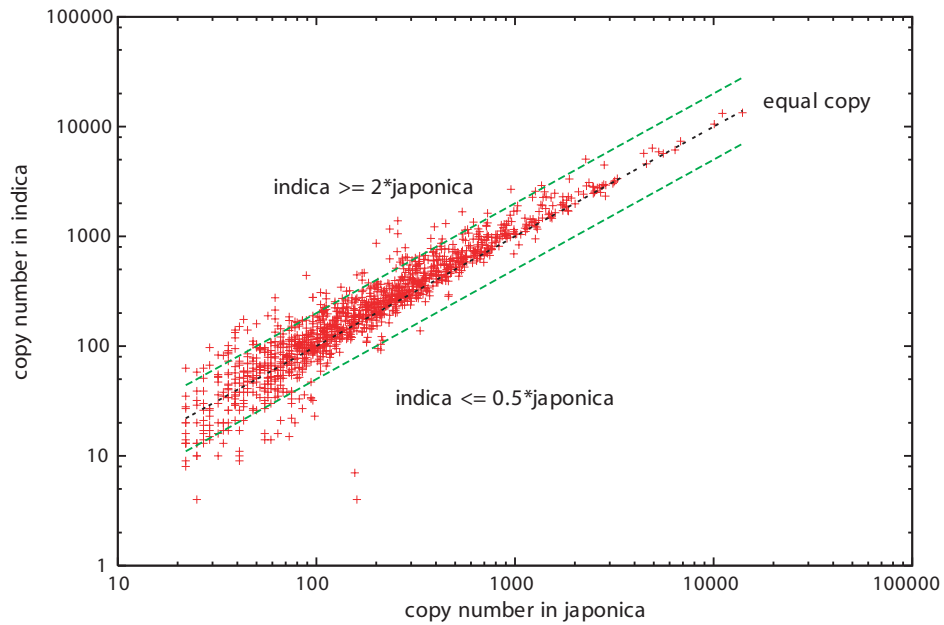


Figure 4.5: Comparison of copy numbers of the defined repeat families. Each point represents a family. Copy numbers were estimated based on the number of hits in the available sequences, projected to a 430 Mb genome. If a family is completely monomorphic, it will fall onto the black equal-copy line or close to it considering sampling errors as both genomes are partial. Green lines mark two-fold difference between the two genomes.

#### **4.3.4 Transposon Display reveals extensive insertion polymorphism among cultivars and wild species**

Insertion polymorphism between the two sequenced cultivars can be detected by computational sequence analysis. To study insertion polymorphism in other rice cultivars and species, one can take advantage of an experimental technique called Transposon Display (TD). TD is modified from the technique of Amplified Restriction Fragment Length Polymorphism (AFLP). Just as in AFLP, genomic DNA is first digested by a chosen restriction enzyme, and the DNA fragments produced are amplified by PCR. However, for TD, only one of the PCR primers is based on a short sequence ligated to the restriction fragments. The other primer is based on the internal sequence of a given transposon (or any repeat) (Waugh et al., 1997). If the insertion of the given repeat is random relative to the distribution of restriction sites, then each insertion locus will give an amplified fragment with a unique length, which can be distinguished as different bands on a gel.

Demonstrated here is the TD characterization of two transposable elements in rice, namely *Dasheng* and *mPing*. The *Dasheng* element (Fig. 4.6A), which is a non-autonomous LTR element, was identified from the BAC end sequences of *japonica*, using a semi-automatic prototype of the RECON package (Jiang et al., 2002). The *mPing* element (Fig. 4.7A) is family No. 1031 in the 1256 families identified, and is the first MITE to show transposition activity in the lab (Jiang et al., 2003).

Both elements show extensive polymorphism at all levels: between species, between subspecies and between cultivars within subspecies (Fig. 4.6B and Fig. 4.7B, courtesy of Ning Jiang). (The differences between the cultivars are unlikely to be due to site mutations in the restriction sites because the average sequence difference between *indica* and *japonica* is  $\sim 1\%$ , which should only affect

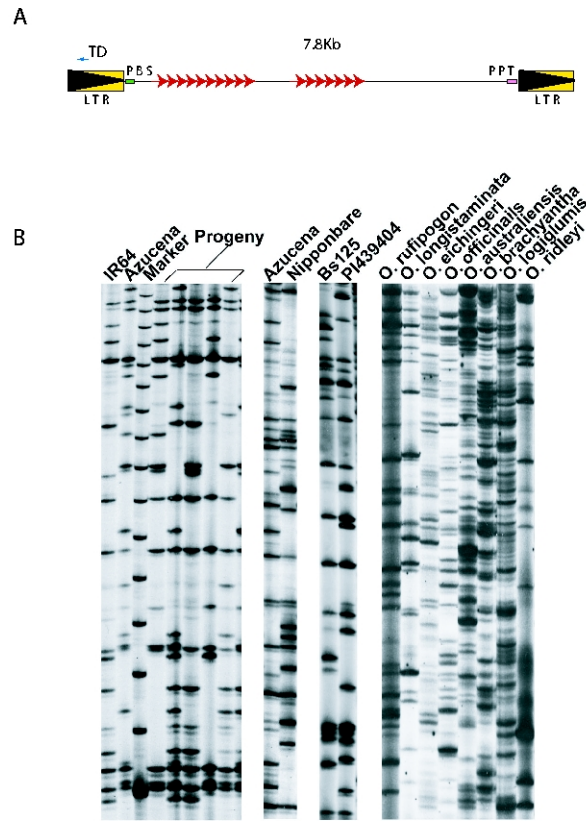
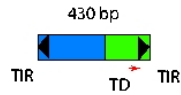


Figure 4.6: A. Schematic structure of *Dasheng*. LTR, PBS (primer binding sequence) and PPT (poly-purine tract) are structural features of LTR elements. In between the LTRs are two arrays of tandem repeats (red arrows), instead of the typical protein coding genes of autonomous LTR elements. Blue arrow on the left LTR marks the position and direction of the TD primer. B. TD gel of *Dasheng*. First panel shows an *indica* cultivar (IR64) and a *japonica* cultivar (Azucena) and their hybrids (used as mapping populations). Second and third panel compare two *japonica* and two *indica* cultivars, respectively. Fourth panel shows various wild species. *O. rufipogon* is considered the direct ancestor of *O. sativa*.

A



B

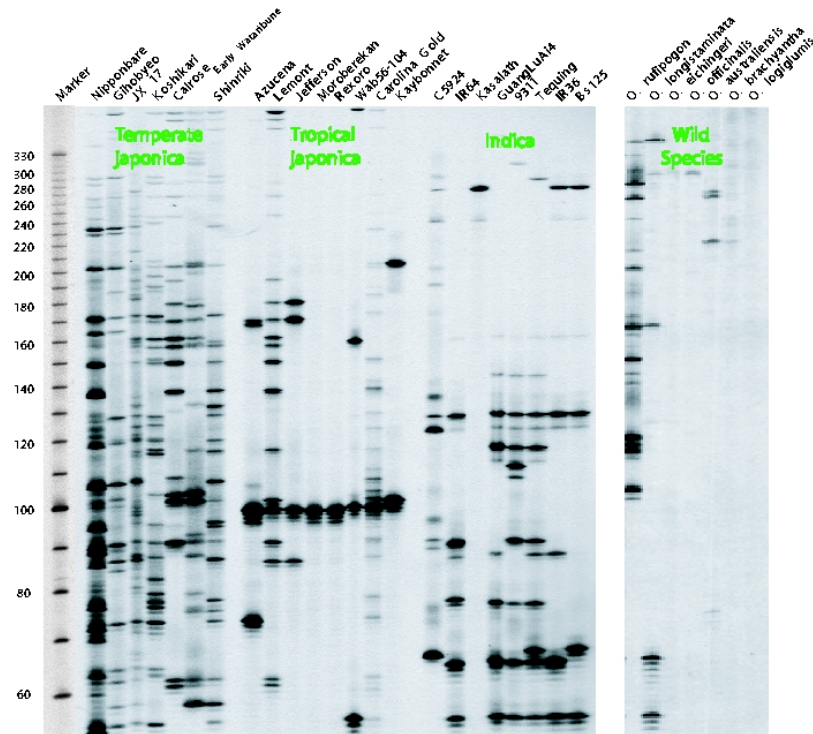


Figure 4.7: A. Schematic structure of *mPing*. Black arrows represent its 15 nt terminal inverted repeats (TIRs). *mPing* is a deletion derivative of an autonomous DNA element, *Ping*. Blue and green regions are inherited from the 5' and 3' end of *Ping*, respectively. Red arrow under the right TIR marks the position and direction of the TD primer. B. TD of *mPing*.

~ 4% of the bands). In particular, *mPing* shows a dramatic difference in copy number between the temperate and the tropical *japonica* cultivars, though the two groups are believed to be derived from a single domesticated ancestor five to seven thousand years ago (Matsuo et al., 1997). In contrast, *Dasheng* does not show such a difference (second panel in Fig. 4.6B). Furthermore, *Dasheng* exists extensively in wild species, while *mPing* does not.

*Dasheng* and *mPing* are two very different elements. *Dasheng* is an LTR element, long (7.8 Kb), has one of the highest copy numbers (~1000) among the LTR elements in rice and is biased towards pericentromeric regions. *mPing* is a DNA element, short (430 bp), has very low copy numbers (several to several dozen) compared to many other MITEs in rice and targets genic regions. The fact that both elements are highly polymorphic suggests that there would be many other repeats with comparable level of insertion polymorphism in rice cultivars and wild species.

## 4.4 Discussion

Presented in this chapter are some preliminary, yet promising results on the identification of repeats and the characterization of their polymorphism in the rice genome. The consensus sequences built for the 1256 identified repeat families are a useful tool for analyzing the rice genome. In addition, the estimates of insertion polymorphism suggest that rice is a valuable system to study the evolution of repeats and their impacts on population diversity. Based on these results, a consortium has been formed to study repeats in the rice genome. The identified repeats will be used to annotate the rice genomes for the Gramene database, a multi-species genomic database for rice and related crops (Ware et al., 2002). During the annotation, our collaborators will also identify all the polymorphic insertion loci in the two sequenced cultivars by aligning the two genomes. Meanwhile, in order to

facilitate gene discovery and breeding, we will create useful genetic markers from the polymorphic insertions. Two types of insertions would be of particular interest. One is those closely associated with genes, such as the likes of *mPing*. The other is those in regions with less known markers, such as the *Dasheng* element for the pericentromeric regions.

The diversity of repeats (largely transposable elements) found in eukaryotic genomes suggests that the genome is like a micro-ecosystem: elements interact with each other in a given genomic environment. For example, *mPing*, a deletion derivative of an autonomous DNA element named *Ping*, can be cross-mobilized by certain related but distinct elements when *Ping* is not available (Jiang et al., 2003). Since *mPing* has no further mutations from *Ping* except for the deletion, whatever element that mobilized *mPing* must have the capacity of moving the autonomous *Ping* as well. That is to say, these autonomous elements work as a collaborative community, taking care of not only one's own copies, but also the close relatives. The interactions among transposable elements, which may go well beyond the simple collaboration of *Ping* and its relatives, could have played an important role in shaping a genome to its current form. However, it has not been possible to fully characterize these interactions due to the limited number of active transposable elements known in model genomes. *mPing* was chosen from our identified families to test for activity because it has many identical copies in the genome, indicating very recent activity (Jiang et al., 2003). The same criterion can be used to screen for other active transposable elements. At the same time, the highly polymorphic families identified in this chapter are also plausible candidates for activity. A comprehensive list of active elements could help initiate the study of "genome ecology". Meanwhile, some of these active elements may become useful genetic tools for insertion mutagenesis.

A potential impact of transposable elements on evolution is to help create diversity in populations. The level of insertion polymorphism seen here is much higher than our naive expectation.

However, is rice a special case? Could our effort of domestication and continuous breeding have elevated or reduced polymorphism? In other words, what is the evolutionary norm of population homogeneity? As genome projects extend to more and more species, it will become more and more practical to address this question by following the research model presented here, i.e., by *de novo* repeat identification and subsequent TD characterization. Perhaps the more interesting question here is to what extent the insertions in QTLs help to create phenotypical diversity. In this regard, it is intriguing to note that an apparent gamma-ray induced insertion in an intron of the rice Hd1 gene was reported to be responsible for a quantitative change in flowering time. The insertion turned out to be an *mPing* element.

## 4.5 Methods

The *japonica* genomic sequences were download from the IRGSP website at <http://rgp.dna.affrc.go.jp> on Dec 24, 2001. The *indica* genomic sequences were downloaded from <http://btm.genomics.org/cn/rice> on Feb 25, 2002.

Repeat library was built on the *japonica* sequences (see Chapter 3 for details on methods), and is available from the Rlib website at <http://Rlib.wustl.edu>. Classification of the repeats identified is based on their sequence similarity to the known transposable elements in rice collected by Ning Jiang (N Jiang and S Wessler, unpublished), using RepeatMasker with the -nolow option. A repeat is assigned to the same class as its top hit (highest alignment score according to RepeatMasker).

The repeat library constructed here is used to annotate the *japonica* and the *indica* genomes, using RepeatMasker with the -nolow option. Further analyses of the physical distribution and insertion polymorphism are based on this annotation.

In estimating the length distribution of inter-repeat regions, regions at the very ends of a contig (outside the first and last repeat loci in a contig) are ignored ( $\sim 3000$  out of a total of  $\sim 250,000$ ). Four artificial runs of  $N$ 's, two  $N_{50}$ s and two  $N_{100}$ s, were detected in the *japonica* sequences, but not removed from the estimate. The theoretical length distribution, assuming random insertion, is a geometrical distribution (mathematical proof omitted). For computer simulation, I assumed same numbers of non-repetitive positions and repeat loci as annotated in the *japonica* sequences, and repeats insert randomly in between non-repetitive positions (a new insertion can be adjacent to a previously inserted repeat, but not in the middle of it).

To estimate the level of insertion polymorphism, 1% of the total insertions were randomly chosen from each genome. The flanking 50 bp at each side of an insertion were used to search the other genome. If either of the flanking sequences has one and only one hit with  $> 97\%$  identity, the hit is considered the homeologous position in the other genome. If the same repeat exists at this homeologous position, the site is considered monomorphic. Otherwise, polymorphic.

For technical details of Transposon Display, see (Jiang et al., 2002) (for *Dasheng*) and (Jiang et al., 2003) (for *mPing*).

## **Chapter 5**

# **Concluding Remarks**

## **5.1 Remarks on RECON and Rlib**

RECON and Rlib are paired tools for repeat identification, somewhat similar to the paired tools of HMMER (Eddy, 2002) and Pfam (Bateman et al., 2002) for protein domain identification. Judging from the positive reaction from the research community to RECON and Rlib, we plan to continue developing these tools. As mentioned in Chapter 3, our current plan for Rlib is to catch up and keep up with the ever growing sequencing effort. To ensure continuity in the development of Rlib, I have implemented a computational pipeline for the Eddy lab computer cluster, which requires minimal human intervention in constructing a repeat library from raw genomic sequences.

In the meantime, I have begun to receive feedback from users, which will allow me to improve RECON and produce better repeat libraries for Rlib. In addition, we are open to potential collaborations. The availability of RECON and Rlib will facilitate the production of manually curated, high quality repeat libraries.

## **5.2 Remarks on the Genomics of Transposons**

With tools like RECON and Rlib in hand, one could start addressing questions concerning repeats and their impact on genomes. Personally, I am most interested in transposons. Transposons are the most abundant and most volatile of repeats. To some extent, they are almost mysterious. From molecular studies, we know that transposons have the capacity of doing many things in the genome, yet we know little about what they have actually done. It is difficult to predict what we will learn about transposons from genome analysis by the end of the day, but the following questions seem to be promising and feasible.

### 5.2.1 Mechanisms for transposition

Like many basic concepts in biology such as genes, there is not a clear definition for transposable elements. Instead, they are recognized on a we-know-when-we-see basis. We know three major types of transposons (LTR elements, LINEs/SINEs and DNA transposons), each having a distinct transposition mechanism. Although these three types account for most of the dispersed repeats in all examined genomes to date, they are probably not all the transposons and transposition mechanisms there are. Recently, Kapitonov and Jurka reported a new type of transposon by analyzing the genomes of *C. elegans* and *A. thaliana*. This new type, dubbed Helitron, might transpose via a “rolling circle” mechanism (Kapitonov & Jurka, 2001). Homologues of Helitrons were also identified in the Rlib library of *B. oleracea*. While examining the repeats in *C. briggsae*, I have noticed possibly another new type of transposon. Based on sequence similarity, part of the 8Kb element encodes its own DNA-directed DNA polymerase, and is probably related to single-stranded DNA viruses. Given the large number of repeat families in Rlib that can not be easily classified, I am comfortable predicting that we will find ample new types of transposons with new mechanisms for transposition.

One aspect of transposition mechanism is target site specificity, i.e., how transposons choose where to insert in the genome. Most of the known transposable elements recognize certain DNA sequence motifs as the insertion sites, such as the TA dinucleotide for the *Tc1* type DNA transposons (Plasterk & von Luenen, 1997). However, some elements require more than the DNA motif. For example, *Ty3* in yeast recognizes a component of the RNA PolIII complex, which directs *Ty3* to the vicinity of PolIII-transcribed RNA genes (e.g., tRNAs) (Kirchner et al., 1995). Many other transposons show spatial association to certain genes or certain regions in the chromosome, though

the underlying mechanisms are still unknown (for a comprehensive list, see (Craig, 1997; Malik & Eickbush, 2000)). Since this type of target site selection will inevitably lead to non-random distribution of transposons in the genome, potentially all such cases could be identified through genome analysis (see also Chapter 1).

Needless to say, genome analysis can only suggest the existence of new mechanisms of transposition and target site selection. Fully understanding these mechanisms requires detailed molecular studies, which in turn require elements that can actively transpose in one's experimental system. Traditionally, active transposons are identified fortuitously when the insertion of a particular element causes phenotypic changes. Today, active elements can be sought via sequence analysis. As shown in the case of the *mPing* element in rice (see Chapter 4 and (Jiang et al., 2003)), families with multiple nearly identical copies are very likely to be active. In addition, "dead" families can be resurrected by carefully constructing consensus sequences for these families, which corrects the mutations in individual copies (Ivics et al., 1997). The tricky problem is how to re-activate a family, which is probably under suppression by the host genome. One possible solution is to introduce the element into a closely related organism, which may provide a similar genomic environment minus the suppression. If the strategy works well, it also provides a convenient way to construct transposon-based genetic tools. For example, some of the potentially active elements in *C. briggsae* (Z Bao and S Eddy, unpublished data) can be tested for activity in *C. elegans* where people have had trouble with both native elements and elements introduced from distantly related organisms (Bessereau et al., 2001).

## 5.2.2 Evolutionary timing of transposition

A central task in genome analyses of transposons is “molecular archeology”, that is, to reconstruct the history of transposition in a given genome (see also Chapter 1). In order to do so, we need to be able to determine when the individual transposition events occurred (i.e., the age of a given copy of transposon, see also Chapter 1). Common practice is to use the sequence divergence of a given copy from its most similar copy in the genome to approximate its age. However, it is obviously inaccurate: such an approximation assumes that the most similar copy was either the parent of the given copy or identical to the parent when the given copy was created, thus the sequence divergence reflects the age of the given copy. This is not necessarily true in a population context — the parent of the given copy, as well as those identical to the parent at the time, could have subsequently been lost from the population due to random drift. Therefore, the coalescence time between a given copy and its most similar copy in the genome is not necessarily the time when the given copy was generated.

It is unlikely that we could ever overcome this lack of information. However, it is possible for us to estimate the statistical bounds of the errors and understand the intrinsic limit of accuracy of the method. A reasonable statistical model should take into account three stochastic processes: transposition, site mutation in individual copies and the drift of the frequency of occurrence of individual copies in the population. Integrating three stochastic processes is not a trivial task, but should be within the reach of theoretical biologists. Some ground works have been laid in the early studies of population genetics of transposons (Brookfield, 1986; Ohta, 1986; Brookfield & Badge, 1997; Charlesworth et al., 1994), but need updating in light of our new knowledge about these elements.

Two types of studies could help us in constructing the statistical models. The first is to char-

acterize the evolution of LTR elements. Due to their particular mechanism of replication, the two LTRs in a given copy of LTR element are identical when the copy is first generated. Thus, sequence divergence between the two LTRs reflects the age of the copy (Jordan & McDonald, 1999). This independent way of timing could give us the empirical estimate of the intrinsic errors of the general method, which in turn allows us to evaluate our statistical models. The second one is to characterize the frequency spectrum of individual copies in the population (see the TD technique in Chapter 4 and (Waugh et al., 1997; Wright et al., 2001)), which allows us to better estimate the parameters in our models.

Besides the questions above, I have also discussed several other questions concerning transposons in the previous chapters, including testing McClintock's "smart genome" hypothesis (Chapter 1), characterizing vertical evolution versus horizontal transfer of transposons (Chapter 1), turnover of transposon sequences (influx by transposition and efflux by deletion) (Chapters 1 and 3) and the "genome ecology" of transposons (Chapter 4). In all, the availability of large scale genomic sequences has opened a new window for the studies of transposable elements. To capture the liveliness of this emerging field of transposon genomics, my labmate, Thomas A. Jones, has proposed a new term — mobilomics. Although it was first created to mock the frenzy of creating "-omes" and "-omics", mobilomics seems to be a highly appropriate term that indeed has a substantial body of research underneath.

# Appendix A

## List of Publications

1. Jiang, N., **Bao, Z.**, Temnykh, S., Cheng, Z., Jiang, J., Wing, R.A., McCouch, S.R. & Wessler S.R. (2002). *Dasheng*: A recently amplified nonautonomous Long Terminal Repeat element that is a major component of pericentromeric regions in rice. *Genetics*, 161, 1293-305.
2. **Bao, Z.** & Eddy S.R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res*, 12, 1269-76.
3. Jiang, N.\*, **Bao, Z.\***, Zhang, X., Eddy, S.R., McCouch, S.R., Hirochika, H. & Wessler S.R. (2003). An active DNA transposon family in rice. *Nature*, 421, 163-7.

\* co-first authors

## Appendix B

# An Assessing Method for *de novo* Repeat Identification

We assess the result of a *de novo* repeat family identification by comparing it to a trusted result. Comparing two results of repeat family identification is similar to comparing two results of gene prediction, with elements corresponding to exons and families corresponding to genes. However, the former can be more complicated, as one nucleotide position may be assigned simultaneously to multiple elements and families. Here, we present our method of comparing two results of repeat family identification, which was used to train the RECON package.

### *Notations and Definitions*

Let  $\{F_\alpha = \{E_i^\alpha\}\}$  denote the true repeat families and the copies/elements of each family in the input sequences, and  $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$  denote the families and their copies identified by a given method. Also, let  $\{E_i\}$  and  $\{\mathcal{E}_j\}$  denote the set of true and identified elements (regardless of which family they belong to).

A true element  $E_i$  and an identified element  $\mathcal{E}_j$  are considered to correspond to each other if

the overlap between the two is longer than 50% of either of the elements. A true family  $F_\alpha$  and an identified family  $\mathcal{F}_\beta$  are considered to correspond to each other if any  $E_i^\alpha$  corresponds to any  $\mathcal{E}_j^\beta$ . Let  $m(E_i, \mathcal{E}_j) = 1$  if  $E_i$  and  $\mathcal{E}_j$  correspond and  $m(E_i, \mathcal{E}_j) = 0$  if  $E_i$  and  $\mathcal{E}_j$  do not correspond.

## ***The Strategy***

Our goal is to quantitate the difference between  $\{F_\alpha = \{E_i^\alpha\}\}$  and  $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$ . The comparison would be straightforward if a nucleotide position can be assigned to at most one  $\mathcal{E}_j$  and there is one-to-one/zero correspondence between  $\{E_i\}$  and  $\{\mathcal{E}_j\}$  and between  $\{F_\alpha = \{E_i^\alpha\}\}$  and  $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$ . Unfortunately, it is usually not the case in practice.

We capture the following three types of differences/errors that could occur in  $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$  when compared to  $\{F_\alpha = \{E_i^\alpha\}\}$ :

1. lack of sensitivity, including: (a) failing to identify an element, i.e., for an  $E \in \{E_i\}$ , there is no  $\mathcal{E} \in \{\mathcal{E}_j\}$  corresponding to it; (b) failing to identify part of an element, i.e., certain nucleotide positions in an  $E \in \{E_i\}$  are not in its corresponding  $\mathcal{E} \in \{\mathcal{E}_j\}$ ; and (c) breaking a true family into several identified families, i.e.,  $\mathcal{E}_j$ s that correspond to  $\{E_i^\alpha\}$  are assigned to different  $\mathcal{F}_\beta$ s;
2. redundancy, i.e., simultaneously assigning a position in  $\{F_\alpha = \{E_i^\alpha\}\}$  to multiple  $\mathcal{F}_\beta$ 's;
3. lack of specificity, including: (a) identifying an  $\mathcal{E} \in \{\mathcal{E}_j\}$  that does not correspond to any  $E \in \{E_i\}$ ; (b) some positions in an  $\mathcal{E} \in \{\mathcal{E}_j\}$  not in its corresponding  $E \in \{E_i\}$ ; and (c) lumping families, i.e.,  $\mathcal{E}_j$ s in an  $\mathcal{F}_\beta$  corresponding to true elements that are in different  $F_\alpha$ s.

This is of course not a complete list of possible differences/errors, but rather those that could

potentially affect the quality of the sequence models (consensus sequences, etc.) built on top of each  $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$ . To quantify, we count the amount of nucleotide positions involved in each of the three types of errors.

### ***The Formula***

For a set of nucleotide positions  $E$ , we define  $|E|$  as the number of nucleotide positions in  $E$ . For a set of elements  $F = \{E_i\}$ , we define  $|F| = |\bigcup_{E_i \in F} E_i|$ , which is the number of nucleotide positions in  $F$  with redundant positions being counted only once. Furthermore, we define

$$u(F_\alpha, \mathcal{F}_\beta) = u(\mathcal{F}_\beta, F_\alpha) = \bigcup_{m(E_i^\alpha, \mathcal{E}_j^\beta)=1} E_i^\alpha \cap \mathcal{E}_j^\beta$$

which is the set of nucleotide positions in  $F_\alpha$  which are properly identified in  $\mathcal{F}_\beta$ . Also, let  $\mathcal{B}(F_\alpha)$  denote  $F_\alpha$ 's best match in  $\{\mathcal{F}_\beta\}$ , so that

$$\mathcal{B}(F_\alpha) = \arg \max_{\{\mathcal{F}_\beta\}} \left( \left| u(F_\alpha, \mathcal{F}_\beta) \right| \right),$$

or  $\mathcal{B}(F_\alpha) = \phi$  if  $F_\alpha$  does not match any  $\mathcal{F}_\beta$ . Similarly, let  $B(\mathcal{F}_\beta)$  denote  $\mathcal{F}_\beta$ 's best match in  $\{F_\alpha\}$ , so that

$$B(\mathcal{F}_\beta) = \arg \max_{\{F_\alpha\}} \left( \left| u(\mathcal{F}_\beta, F_\alpha) \right| \right),$$

or  $B(\mathcal{F}_\beta) = \phi$  if  $\mathcal{F}_\beta$  does not match any  $F_\alpha$ .

For a given true family  $F_\alpha$ , the error for the lack of sensitivity (denoted by  $Err_{1,F_\alpha}$ ) is calculated as

$$\begin{aligned} Err_{1,F_\alpha} &= \left| F_\alpha \right| - \left| \bigcup_{\{\mathcal{F}_\beta\}} u(F_\alpha, \mathcal{F}_\beta) \right| \\ &\quad + \left| \bigcup_{\{\mathcal{F}_\beta\}} u(F_\alpha, \mathcal{F}_\beta) \right| - \left| u(F_\alpha, \mathcal{B}(F_\alpha)) \right|. \end{aligned}$$

The first two terms measure error 1(a) and 1(b). The last two terms measure error 1(c) by counting the number of identified nucleotide positions in  $F_\alpha$  which are not found in its best match  $\mathcal{B}(F_\alpha)$ .

After canceling the second and the third terms,

$$Err_{1,F_\alpha} = \left| F_\alpha \right| - \left| u\left(F_\alpha, \mathcal{B}(F_\alpha)\right) \right|$$

which is equivalent to the number of nucleotide positions in  $F_\alpha$  which are not properly identified in  $\mathcal{B}(F_\alpha)$ .

The error of redundancy concerning  $F_\alpha$  ( $Err_{2,F_\alpha}$ ) is calculated as

$$Err_{2,F_\alpha} = \sum_{\{\mathcal{F}_\beta\}} \left| u\left(F_\alpha, \mathcal{F}_\beta\right) \right| - \left| \bigcup_{\{\mathcal{F}_\beta\}} u\left(F_\alpha, \mathcal{F}_\beta\right) \right|.$$

For a given identified family  $\mathcal{F}_\beta$ , the error for the lack of specificity ( $Err_{3,\mathcal{F}_\beta}$ ) is calculated as

$$Err_{3,\mathcal{F}_\beta} = \left| \mathcal{F}_\beta \right| - \left| u\left(\mathcal{F}_\beta, \mathcal{B}(\mathcal{F}_\beta)\right) \right|.$$

The total error ( $Err$ ) is defined as

$$Err = \sum_{\{F_\alpha\}} Err_{1,F_\alpha} + \sum_{\{F_\alpha\}} Err_{2,F_\alpha} + \sum_{\{\mathcal{F}_\beta\}} Err_{3,\mathcal{F}_\beta} \quad (5.1)$$

### ***For Optimization***

Typically, we know the representative sequence of certain families in the subject genome. We can carefully locate copies of these families in the input sequences, and optimize the whole analysis by reducing the errors concerning these families/copies. Accordingly, in equation 1, the first two terms will sum over these known families, and the last term will sum over identified families which correspond to the known families, using the corresponding known family as the best match for the identified families concerned.

# Bibliography

- Agarwal, P. & States, D. J. (1994). The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. In *Proc Int Conf Intell Syst Mol Biol* (pp. 1–9).
- Agrawal, A., Eastman, Q., & Schatz, D. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, 394, 744–51.
- Bailey, J., Gu, Z., Clark, R., Reinert, K., Samonte, R., Schwartz, S., Adams, M., Li, P., & Eichler, E. (2002). Recent segmental duplications in the human genome. *Science*, 297, 1003–7.
- Bao, Z. & Eddy, S. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, 12, 1269–76.
- Barnes, T., Kohara, Y., Coulson, A., & Hekimi, S. (1995). Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics*, 141, 159–79.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S., Howe, K., Marshall, M., & Sonnhammer, E. (2002). The Pfam protein families database. *Nucleic Acids Res*, 30, 276–80.
- Becker, H., Saedler, H., & Lonnig, W. (2002). Transposable elements in plants. In *Encyclopedia of Genetics* (pp. 2020–33). London, UK: Academic Press.
- Berg, D. E. & Howe, M. M., Eds. (1989). *Mobile DNA*. Washington, D.C.: American Society for Microbiology.

- Bessereau, J., Wright, A., Williams, D., Schuske, K., Davis, M., & Jorgensen, E. (2001). Mobilization of a *Drosophila* transposon in the *Caenorhabditis elegans* germ line. *Nature*, 413, 70–4.
- Britten, R. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A*, 93, 9374–7.
- Britten, R. & Kohne, D. (1968). Repeated sequences in DNA. hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science*, 161, 529–40.
- Brookfield, J. (1986). A model for DNA sequence evolution within transposable element families. *Genetics*, 112, 393–407.
- Brookfield, J. & Badge, R. (1997). Population genetics models of transposable elements. *Genetica*, 100, 281–94.
- Bushman, F. (2002). *Lateral DNA transfer: mechanisms and consequences*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
- Butler, D. & Pockley, P. (2000). Monsanto makes rice genome public. *Nature*, 404, 534.
- Caceres, M., Ranz, J., Barbadilla, A., Long, M., & Ruiz, A. (1999). Generation of a widespread *Drosophila* inversion by a transposable element. *Science*, 285, 415–8.
- Capy, P., Ed. (1998). *Dynamics and evolution of transposable elements*. New York, New York: Chapman & Hall.
- Casacuberta, E. & Pardue, M. (2002). Coevolution of the telomeric retrotransposons across *Drosophila* species. *Genetics*, 161, 1113–24.
- Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371, 215–20.
- Corpet, F., Servant, F., Gouzy, J., & Kahn, D. (2000). Prodom and prodom-cg: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*, 28, 267–9.

- Craig, N. (1997). Target site selection in transposition. *Annu Rev Biochem*, 6, 437–74.
- Craig, N., Craigie, R., Gellert, M., & Lambowitz, A., Eds. (2002). *Mobile DNA II*. Washington, D.C.: American Society for Microbiology.
- Cummings, C. & Zoghbi, H. (2000). Trinucleotide repeats: mechanisms and pathophysiology. *Annu Rev Genomics Hum Genet*, 1, 281–328.
- Daniels, S., Peterson, K., Strausbaugh, L., Kidwell, M., & Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*, 124, 339–55.
- des Etages, S., Kumar, A., & Snyder, M. (2002). Transposons as tools. In *Encyclopedia of Genetics* (pp. 2034–40). London, UK: Academic Press.
- Dickson, L., Huang, H., Liu, L., Matsuura, M., Lambowitz, A., & Perlman, P. (2001). Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proc Natl Acad Sci U S A*, 98, 13207–12.
- Doolittle, R. (1995). The multiplicity of domains in proteins. *Annu Rev Biochem*, 6, 287–314.
- Doolittle, W. F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284, 601–603.
- Eddy, S. (2002). HMMER. unpublished. Website <http://hmmer.wustl.edu>.
- Emanuel, B. & Shaikh, T. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat Rev Genet*, 2, 791–800.
- Fedoroff, N. (2002). Barbara McClintock. In *Encyclopedia of Genetics* (pp. 1161–2). London, UK: Academic Press.
- Feschotte, C., Jiang, N., & Wessler, S. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*, 3, 329–41.

- Friedman, R. & Hughes, A. (2001). Pattern and timing of gene duplication in animal genomes. *Genome Res*, 11, 1842–7.
- Ganko, E., Fielman, K., & McDonald, J. (2001). Evolutionary history of Cer elements and their impact on the *C. elegans* genome. *Genome Res*, 12, 2066–74.
- Gish, W. (2002). WU-BLAST. unpublished. Website <http://blast.wustl.edu>.
- Goff, S., Ricke, D., Lan, T., & et al (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, 296, 92–100.
- Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. II. delineation of domain boundaries from sequence similarities. *Bioinformatics*, 14, 174–187.
- Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H., & Kanda, M. (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci U S A*, 93, 7783–8.
- Holmes, I. (2002). Transcendent elements: whole-genome transposon screens and open evolutionary questions. *Genome Res*, 12, 1152–5.
- Ivics, Z., Hackett, P., Plasterk, R., & Izsvak, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, 91, 501–10.
- Jensen, S., Gassama, M., & Heidmann, T. (1999). Taming of transposable elements by homology-dependent gene silencing. *Nat Genet*, 21, 209–12.
- Jiang, N. (2002). *PhD thesis*. University of Georgia, Athens.
- Jiang, N., Bao, Z., Temnykh, S., Cheng, Z., Jiang, J., Wing, R., McCouch, S., & Wessler, S. (2002). Dasheng, a recently amplified nonautonomous Long Terminal Repeat element that is a major component of pericentromeric regions in rice. *Genetics*, 161, 1293–305.
- Jiang, N., Bao, Z., Zhang, X., Eddy, S., McCouch, S., Hirochika, H., & Wessler, S. (2003). An active DNA transposon family in rice. *Nature*, 421, 163–7.

- Jordan, I. & McDonald, J. (1999). Tempo and mode of Ty element evolution in *Saccharomyces cerevisiae*. *Genetics*, 151, 1341–51.
- Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, 16, 418–20.
- Jurka, J. & Zuckerkandl, E. (1991). Free left arms as precursor molecules in the evolution of Alu sequences. *J Mol Evol*, 33, 49–56.
- Kapitonov, V. & Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A*, 98, 8714–9.
- Kazazian, H. & Goodier, J. (2002). Line drive. retrotransposition and genome instability. *Cell*, 110, 277–80.
- Keller, E. (1993). *A Feeling for the Organism : The Life and Work of Barbara McClintock*. New York, New York: W.H. Freeman & Co.
- Kidwell, M. & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A*, 94, 7704–11.
- Kim, J., Vanguri, S., Boeke, J., Gabriel, A., & Voytas, D. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res*, 8, 464–78.
- Kirchner, J., Connolly, C., & Sandmeyer, S. (1995). Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science*, 267, 1488–91.
- Kurtz, S., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2000). Computation and visualization of degenerate repeats in complete genomes. In *Proc Int Conf Intell Syst Mol Biol* (pp. 228–238).
- Labrador, M. & Corces, V. (1997). Transposable element-host interactions: regulation of insertion and excision. *Annu Rev Genet*, 3, 381–404.
- Lander, E. S., Linton, L. M., Birren, B., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

- Leach, J., McCouch, S., Slezak, T., Sasaki, T., & Wessler, S. (2002). Why finishing the rice genome matters. *Science*, 296, 45.
- Li, W. (1997). *Molecular evolution*. Sunderland, Massachusetts: Sinauer Associates.
- Lovett, S. (2002). Tandem repeats. In *Encyclopedia of Genetics* (pp. 1932–3). London, UK: Academic Press.
- Lynch, M. & Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290, 1151–5.
- Macgregor, H. (2002). C-value paradox. In *Encyclopedia of Genetics* (pp. 249–50). London, UK: Academic Press.
- Malik, H. & Eickbush, T. (2000). Nesl-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics*, 154, 193–203.
- Marshall, E. (2001). Genome teams adjust to shotgun marriage. *Science*, 292, 1982–1983.
- Matsuo, T., Futsuhara, Y., Kikuchi, F., & Hamaguchi, H., Eds. (1997). *Science of the rice plant*. Tokyo: Ministry of Agriculture, Forestry and Fisheries.
- McClintock, B. (1978). Mechanisms that rapidly reorganize the genome. In *The discovery and characterization of transposable elements : the collected papers of Barbara McClintock (1987)* (pp. 593–615). New York, New York: Garland Publishing Inc.
- McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226, 792–801.
- McClintock, B. (1987). *The discovery and characterization of transposable elements : the collected papers of Barbara McClintock*. New York, New York: Garland Publishing Inc.
- McDonald, J. (1995). Transposable elements: possible catalysts of organismic evolution. *Trends Ecol Evol*, 10, 123–6.
- Mefford, H. & Trask, B. (2002). The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet*, 3, 91–102.

- Morgenstern, B. (1999). DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15, 211–218.
- Nevers, P. & Saedler, H. (1977). Transposable genetic elements as agents of gene instability and chromosomal rearrangements. *Nature*, 268, 109–115.
- Ohno, S. (1970). *Evolution by gene duplication*. Berlin: Springer-Verlag.
- Ohta, T. (1986). Population genetics of an expanding family of mobile genetic elements. *Genetics*, 113, 145–59.
- O’Neill, R., O’Neill, M., & Graves, J. (1998). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature*, 393, 68–72.
- Ono, S. (1972). So much “junk” DNA in our genome. In *Evolution of Genetic Systems* (pp. 366–70). New York, US: Gordon and Breach.
- Orgel, L. E. & Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284, 604–607.
- Parsons, J. D. (1995). Miropeats: Graphical DNA sequence comparisons. *Comput Appl Biosci*, 11, 615–619.
- Perl, A., Colombo, E., Samoilova, E., Butler, M. C., & Banki, K. (2000). Human transaldolase-associated repetitive elements are transcribed by RNA polymerase III. *J Biol Chem*, 275, 7261–7272.
- Petrov, D., Sangster, T., Johnston, J., Hartl, D., & Shaw, K. (2000). Evidence for DNA loss as a determinant of genome size. *Science*, 287, 1060–2.
- Plasterk, R. H. A. & von Luenen, H. G. A. M. (1997). Transposons. In *C. elegans II* (pp. 97–116). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
- Promislow, D., Jordan, I., & McDonald, J. (1999). Genomic demography: a life-history analysis of transposable element evolution. *Proc R Soc Lond B Biol Sci*, 266, 1555–60.
- Samonte, R. & Eichler, E. (2002). Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*, 3, 65–72.

- Sinclair, D., Mills, K., & Guarente, L. (1998). Aging in *Saccharomyces cerevisiae*. *Annu Rev Microbiol*, 5, 533–60.
- Singer, T., Yordan, C., & Martienssen, R. (2001). Robertson's Mutator transposons in *A. thaliana* are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev*, 15, 591–602.
- Skiena, S. S. (1997). *The Algorithm Design Manual*. New York: Telos/Springer-Verlag.
- Smit, A. F. A. & Green, P. (2002). RepeatMasker. unpublished. Website <http://ftp.genome.washington.edu/RM/RepeatMasker.html>.
- Sonnhammer, E. L. & Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci*, 3, 482–492.
- Sulston, J. & Brenner, S. (1974). The DNA of *Caenorhabditis elegans*. *Genetics*, 77, 95–104.
- Surzycki, S. & Belknap, W. (2000). Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc Natl Acad Sci U S A*, 97, 245–9.
- Tabara, H., Sarkissian, M., Kelly, W., Fleenor, J., Grishok, A., Timmons, L., & Fire, A. (1999). The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell*, 99, 123–32.
- Takahashi, Y., Kuro-O, M., & Ishikawa, F. (2000). Aging mechanisms. *Proc Natl Acad Sci U S A*, 97, 12407–8.
- The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282, 2012–2018.
- Vergnaud, G. & Denoeud, F. (2000). Minisatellites: mutability and genome architecture. *Genome Res*, 10, 899–907.
- Volpe, T., Kidner, C., Hall, I., Teng, G., Grewal, S., & Martienssen, R. (2002). Regulation of heterochromatic silencing and histone H3 Lysine-9 methylation by RNAi. *Science*, epub ahead of print.

- Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Cartinhour, S., McCouch, S., & Stein, L. (2002). Gramene: a resource for comparative grass genomics. *Nucleic Acids Res*, 30, 103–5.
- Waugh, R., McLean, K., Flavell, A., Pearce, S., Kumar, A., Thomas, B., & Powell, W. (1997). Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (s-sap). *Mol Gen Genet*, 253, 687–94.
- Wolfe, K. & Shields, D. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387, 708–13.
- Wright, S., Le, Q., Schoen, D., & Bureau, T. (2001). Population dynamics of an Ac-like transposable element in self- and cross-pollinating Arabidopsis. *Genetics*, 158, 1279–88.
- Yu, J., Hu, S., Wang, J., Wong, G., & et al (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, 296, 79–92.
- Zhang, X. & Hong, G. (2000). Preferential location of MITEs in rice genome. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)*, 32, 223–228.