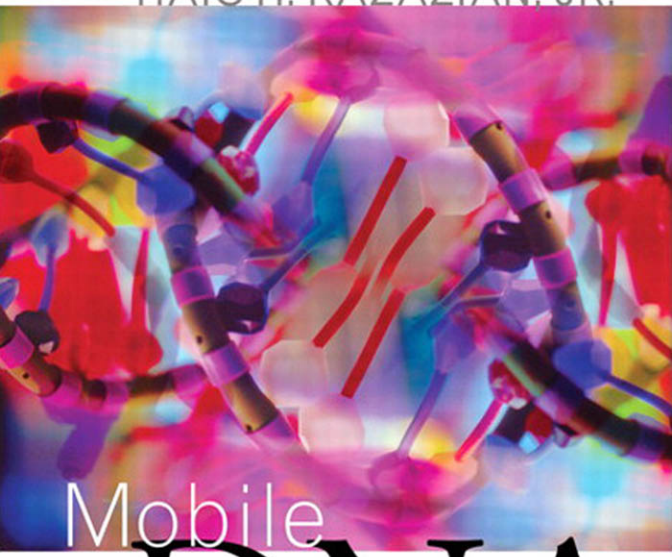
HAIG H. KAZAZIAN, JR.



Mobile DNA

Finding Treasure in Junk

Mobile DNA

This page intentionally left blank

Mobile DNA

Finding Treasure in Junk

Haig H. Kazazian, Jr.

Vice President, Publisher: Tim Moore
Associate Publisher and Director of Marketing: Amy Neidlinger
Acquisitions Editor: Kirk Jensen
Editorial Assistant: Pamela Boland
Operations Manager: Gina Kanouse
Senior Marketing Manager: Julie Phifer
Publicity Manager: Laura Czaja
Assistant Marketing Manager: Megan Colvin
Cover Designer: Gary Adair
Managing Editor: Kristy Hart
Senior Project Editor: Lori Lyons
Copy Editor: Language Logistics, LLC
Proofreader: Sheri Cain
Indexer: Angela Martin
Compositor: Nonie Ratcliff
Manufacturing Buyer: Dan Uhrig

Publishing as FT Press
Upper Saddle River, New Jersey 07458

FT Press offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales. For more information, please contact U.S. Corporate and Government Sales, 1-800-382-3419, corpsales@pearsontechgroup.com. For sales outside the U.S., please contact International Sales at international@pearson.com.

Printed in the United States of America

First Printing April 2011

ISBN-10: 0-13-707062-4
ISBN-13: 978-0-13-707062-6

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.

Library of Congress Cataloging-in-Publication Data

Kazazian, Haig H. (Haig Hagop), Jr., 1937- author.
Mobile DNA : finding treasure in junk / Haig Kazazian.
p. ; cm.
Includes bibliographical references.
ISBN-13: 978-0-13-707062-6 (hardback : alk. paper)
ISBN-10: 0-13-707062-4 (hardback : alk. paper)
1. Insertion elements, DNA. 2. Transposons. 3. Recombinant, DNA. 4. Genomics. I. Title.
[DNLM: 1. DNA Transposable Elements. 2. DNA, Recombinant—physiology. 3. Genomics—methods. 4. Sequence Analysis, DNA. QU 470]
QH462.I48K39 2011
572.8'69—dc22

2010050899

To Lilli, my loving wife of so many years,
who put up with my idiosyncrasies
and encouraged me to write this book.

This page intentionally left blank

Contents

Acknowledgments

I thank my early mentors, Lafayette Noda and Lucille Smith, at Dartmouth; Barton Childs at Johns Hopkins; and Harvey Itano at the NIH. I thank colleagues who helped me in entering the mobile DNA field, especially Maxine Singer, Allan Scott, and Jef Boeke. I thank all the lab members, both those mentioned and those unmentioned, for their tireless efforts to contribute to the mobile DNA field. I thank other colleagues, John Moran, Nancy Craig, Vivian Cheung, and Eric Ostertag, for helpful comments and critiques of the book. I also thank Adam Ewing, Dustin Hancks, and Lauren Zeitel for their advice and comments.

About the Author

Haig Kazazian was born and raised in Toledo, Ohio. After attending public schools there, he received his A.B. degree from Dartmouth College in 1959. He then attended Dartmouth Medical School, a two-year school at the time, and finished his M.D. degree at Johns Hopkins University School of Medicine. At Hopkins, he met his wife of nearly 50 years and married during his internship in Pediatrics at the University of Minnesota Hospital. After two years training in Minneapolis, he returned to Johns Hopkins for a two-year fellowship in genetics with Barton Childs, M.D. He then trained for two years in molecular biology in the lab of Harvey Itano, M.D., at the NIH. After a third year of Pediatric training at Johns Hopkins, he joined the faculty there in 1969. He rose through the ranks to become a full professor in 1977, and at that time, he headed the Pediatric Genetics Unit. In 1988, he became Director of the Center for Medical Genetics at Johns Hopkins.

After 25 years on the Hopkins faculty, he was recruited to the University of Pennsylvania School of Medicine as Chair of the Department of Genetics in 1994. At Penn, he recruited 10 young faculty to the department. In 2006, he stepped down as department chair, but remained as the Seymour Gray Professor of Molecular Medicine in Genetics until 2010. In July 2010, he returned to Johns Hopkins as a Professor in the Institute of Genetic Medicine.

Dr. Kazazian is still heavily involved in molecular genetic research, concentrating for the past 20 years on mammalian and human transposable elements, or "jumping genes." Prior to 1988, he characterized much of the variation in the cluster of genes involved in production of the beta chain of human hemoglobin. With Stuart Orkin at Harvard, his work led to the nearly complete characterization of the mutations causing the β-thalassemias, common anemias in regions of the world endemic for malaria.

Dr. Kazazian is a member of a number of national organizations, including the Institute of Medicine of the National Academy of Sciences and the American Academy of Arts and Sciences. He has received a number of honors for his research, most notably the 2008 William Allan Award, the top honor of the American Society of Human Genetics.

Preface
Thoughts on doing science

Before diving into the subject of mobile DNA and my adventures in the field, I'd like to provide a few personal tips from my experience on working with some success in science for the past 45 years. Doing science is often very difficult and extremely hard work, requiring long hours. In my view, the first thing necessary is very high personal motivation. My original and long-time mentor, Barton Childs, an esteemed Professor of Pediatrics and the "father of pediatric genetics," always used to say, "You've got to burn to do research!" You can't go at it with a half-hearted enthusiasm or self-doubt.

If you do have high personal motivation, you then need to get excellent training, both in didactic class work and in the nuts and bolts of how to do research. You need to find a subject or area that really interests you, no matter what the field. Then find the investigator who is doing excellent research in your field of interest, hopefully someone at the forefront, but also consider that the person's lab is not too large so that he or she will have sufficient time to spend with you. You want someone who will discuss your research with you on a daily basis, perhaps so much so that you feel that he or she is pestering you all the time for new data. That kind of attention means that that individual has great interest in your work. You need that kind of person for both your predoc and your postdoc training. Your trainers also need to be available for discussion of all kinds of problems, both those that you face in the lab and those that are related to other aspects of your life.

Next, you need a dependable mentor. Your mentor could be either your predoc or your postdoc trainer, or it could be a member of your thesis committee or another senior investigator from down the hall. However, you'd like a mentor that you can carry over from your

training days into your first 5–10 years as a faculty member. That mentor can help you with all kinds of problems and questions, giving advice for how to approach various professional and daily life situations. Having an interested, accessible, and experienced mentor is crucial to success in science. Behind every good scientist is an outstanding mentor. I certainly had one in Barton Childs, even if I didn't follow his advice at every turn. He was probably my major mentor for at least 20 years.

I've talked about mentorship from the aspect of the trainee, but what about the importance of being a good mentor? From your first academic job to becoming a long-term lab director, you have the responsibility for mentoring predoc and postdoc trainees. I have usually found it rewarding to give trainees considerable independence, letting them pick their own problems from a wide variety of problems available in the lab. This works well when the trainee is very bright and picks one or more problems that are of real interest to the lab director. When the problem is of little interest to the lab director, there is a good chance that the work will flounder. However, if the problem is important to the mentor, the mentor will add ideas and enthusiasm to the work. I have dealt with both situations over my career, as the reader will discover in this book. A third situation occurs when the student or postdoc needs to finish a period of training and hasn't had much success up to that point. The trick at that time is to find a project that is important for the field (so that the student will take pride in his or her accomplishment), has a clear-cut endpoint, and uses techniques already available in the lab. The design of this project usually requires considerable input from the lab director.

Then there is the question of how one should approach other scientists. Should one be open in discussing new data even with colleagues in the same field, or should one be secretive to avoid being "scooped?" My view has always been that it is better to be open but prudent. It is good to discuss your unpublished work at meetings. If your work is important, your colleagues will respect you for talking

about new data and not rehashing work that has been published and that they've heard previously. Moreover, it is very, very rare that another investigator can start a new experimental tack or line and actually beat you to publication. After all, you've probably been working on that same question for a year or two, so you've got a major head start. It's a rare investigator indeed who would sail off in a new direction hoping to beat you to publication of data that you've just presented. As a general rule, openness in presenting and discussing new data is the best approach.

A corollary to openness is to discuss your science with a wide range of other scientists, including those within your immediate field, e.g., mobile DNA, those in the broader field, e.g. human genetics, and those in other fields of biomedical science, e.g., immunology or developmental biology. You never know from where the next good experimental idea will come. The reader will find throughout this book that members of my lab and I personally have gotten ideas from a wide variety of sources who are mentioned at some future time. This plethora of good ideas has come from discussing the work with a large number of other scientists and sources and being as open as possible to new ideas.

I once knew a well-trained, smart young researcher who had a great deal of trouble gaining traction in his field. I always thought that his problem was that he stayed in his lab and did not seek discussion of his science with colleagues. At the other extreme was and still is the Medical Research Council (MRC) laboratory at Cambridge, England, whose investigators have had enormous success over many decades. The Cambridge MRC labs have housed a number of Nobel Laureates, including Francis Crick, Fred Sanger, Sydney Brenner, Aaron Klug, and others. After spending a few months at the MRC early in my career, I felt that a major factor in the success of that lab was the English tradition at that time of a common coffee break in the morning and a common tea break in the afternoon. At 10:30 AM, every investigator, from the trainees to the most senior people, would

gather in the cafeteria for morning coffee and, importantly, discuss science for 30 minutes over coffee. A similar gathering would occur at 3:30 PM over afternoon tea. The number of great new ideas passed from one investigator to another, from past and future Nobel Prize winners to beginning postdocs, and vice versa, was astonishing. Open discussion of science is wonderful for the development of new ideas.

Now I'd like to make a general comment on picking problems in your field on which to work. I've always believed that the problem should be important but potentially solvable with hard effort. All researchers are gamblers. A colleague used to tell me to pick problems with 5 to 1 or 10 to 1 odds of success. Those problems were about right in terms of difficulty. Odds of 50 to 1 or 100 to 1 were too long, and success on those problems was too unlikely. Odds of 2 to 1 meant that the problem was too easy and relatively unimportant, so called "low-hanging fruit." I've also felt that it is best to pick problems that are logical next steps in the project but are important to the field and have those reasonable odds of success, which would be 5–10:1.

My last point is to keep one's mind alert for possible collaboration. Collaborations with other scientists should be welcomed as a way to broaden one's scientific outlook and scope. If two investigators have differing expertise that can be applied to solve a particular problem, this is an ideal situation for collaboration. I once heard it said that collaboration finds its own level, meaning that in order to work best, collaborators should be on the same level of experience and respect in the field. In this way, I've had good collaborations as a postdoc with another postdoc and as a senior scientist with other senior scientists. Many of these collaborations are discussed throughout the book.

This page intentionally left blank

1

Introduction to Mobile DNA

Charles Darwin would be surprised. Indeed, even present day scientists are surprised by the existence of mobile DNA. Consider the skepticism within the scientific community that greeted Barbara McClintock, already a highly-respected scientist, when she announced that she had found what appeared to be mobile DNA in maize plants (McClintock, 1950). DNA was the genetic material, so it must be static, stable, and immobile. The mutation rate had been determined to be ~10^{-8} per nucleotide, or building block, of DNA per generation—very low indeed. How and why would some DNA move from place to place in a genome? Scientists are still grappling with these questions. Two hundred years removed from Darwin's birth, and we're still wondering how mobile DNA with all its detrimental effects on organisms could have reached such high proportions in the genomes of mammals and plants. Yet mobile DNA is found in all forms of living things, including plants, animals, bacteria, and archea. The genome seems to cherish its ability to make rapid changes by rearranging some of its parts as opposed to the slow change afforded by the nucleotide mutation rate.

One theme of this book is that biological scientists have come to expect the unexpected. The study of living things is full of surprises. One of them is the prevalence of mobile DNA in genomes. Another is that most genes are broken up by sections of DNA called introns that need to be removed at the RNA stage in order for the genes to function. A third is that the protein-coding regions of genes make up a very small fraction of mammalian genomes. A fourth surprise is the importance of reverse transcriptase, the enzyme that synthesizes DNA from an RNA template. These are just a few examples of old surprises, or unexpected findings, that have now become hard facts in

all biology textbooks. Many more will be highlighted in the research adventures outlined in this book. These "unexpected observations" provide excitement and anticipation for even the most experienced researchers. What finding will be the next to shatter our present view of the biological world? One can be sure that the future will bring many more surprises to delight the graduate student just beginning his or her studies.

Prior to 1970, scientists thought that the genome was composed mostly of genes lined up like balls on a string with some repetitive DNA in between the balls. Then in the late 1970s, introns were found to break up the regions of genes that encode proteins (Berget et al., 1977; Chow et al., 1977). Protein-coding regions were disrupted by intervening sequences (introns) that required removal from pre-messenger RNA before the intact protein could be synthesized. Soon, we knew that introns were much larger than protein-coding regions, then called exons. The DNA between the genes along with most of the intronic sequences of genes was thought to be functionless, and was called "junk DNA" (Orgel and Crick, 1980). However, now we know that introns make up about 30% of human and mammalian genomes, and exons encode only between 1 and 2% of the human genome (Lander et al., 2001). What a comedown for protein-coding regions! Thus, over 98% of human DNA had been dismissed as "junk."

Transposable elements were then found in human DNA, and this active mobile DNA along with the remnants of many transposition events over hundreds of millions of years is now known to account for at least 50% of human genomic DNA. This transposable element DNA, both those relatively few sequences that are presently mobile, and the many remnants of old events are now demonstrating function. However, this function is evident only in the many ways mobile DNA can modify the genome over evolutionary time. It can be co-opted for useful purposes but has not yet been definitively shown to have a useful function in the individual organism. Moreover, DNA encoding small RNAs of different types and functions has been discovered amidst the "junk." Enhancer sequences at great distances from the genes upon which they act are being found continually. Segmental duplications of hundreds to many thousands of nucleotide pairs of DNA are strewn around the genome and are further grist in

the mill of genome plasticity and malleability. The bottom line is that "junk" DNA is gradually being eroded away as function is found for a greater and greater fraction of genomic DNA. In this book, I concentrate on the "junk" DNA that is mobile or has been over the millennia. This is the DNA that those of us in the mobile DNA field have come to treasure.

In the next several chapters, I provide details on important topics in the mobile DNA field as well as discuss a number of top scientists who have been pioneers in many areas involving mobile DNA. I then discuss the state of the human mobile DNA field prior to my involvement in it, what led to my fascination with mobile DNA, and why I jumped at the chance to work on it when the opportunity presented itself. Later, I discuss many of the people who worked in my lab up to the present time, their most important work, and the relationship of that work to what is known about L1 biology today. This is followed by important findings of other labs working on mammalian mobile DNA, ending with some thoughts about the future of the field. Yes, DNA as genetic material would have surprised Charles Darwin, but mobile DNA would have really made his head spin!

This page intentionally left blank

2

Varieties of mobile DNA

Mobile DNA comes in many different flavors, the most common of which are DNA transposons and two main varieties of retrotransposons (Figure 2.1) (Goodier and Kazazian, 2008). Another curious observation is that even though mobile DNA is present in every living organism, its proportion of the genome of an organism may vary from a few percent in yeast and pufferfish to a huge 60–70% in maize and barley (Table 2.1) (Kazazian, 2004). In fact, there is a striking direct correlation between genome size and the percentage of any genome that is mobile DNA. Thus, the pufferfish genome is quite small, while the human and all other sequenced genomes of mammals (Table 2.1) contain ~50% mobile DNA and are comparatively large. On top of that, the type of mobile DNA that predominates in the genome of an organism can vary greatly from one organism to the next, from over 90% DNA transposons in *C. elegans*, a round worm, to 100% of one type of retrotransposon in *S. cerevesiae*, a budding yeast, to 75% of a second type of retrotransposon in *H. sapiens*, human beings (see Table 2.1). It is still a mystery why one type of mobile DNA is tolerated in the genome of one organism, while another type is favored in the genome of another. In this chapter, I summarize the various types of mobile elements, and what is known about the mechanisms behind their mobility. I will also present some information concerning their effect on genomes, and hypotheses as to why they have acquired such a major role in so many genomes.

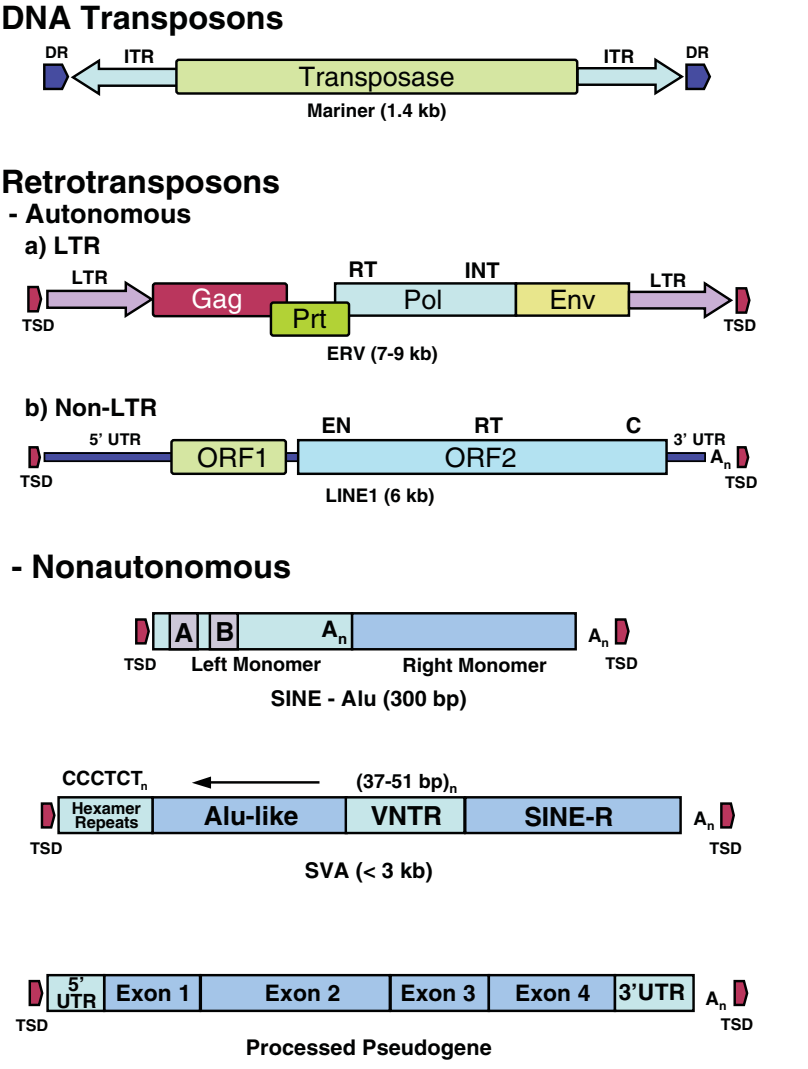


Figure 2.1 Classes of mobile DNA. DNA transposons move by a "cut and paste" mechanism with no duplication of sequence. They are removed from a genomic donor site and inserted into a new target site by their transposase. Retrotransposons move by a "copy and paste" process involving an RNA intermediate and reverse transcription of the RNA into DNA. Retrotransposons are called autonomous when they provide key activities necessary for their mobility. Long terminal repeat (LTR) retrotransposons have direct repeats at their two ends and encode proteins. Non-LTR retrotransposons also encode proteins, but end in a poly A tail. Nonautonomous retrotransposons do not encode any protein, and all nonautonomous retrotransposons require the reverse transcriptase and endonuclease supplied by LINE1 or L1 elements.

Table 2.1 Transposable elements make up a large proportion of some genomes

Species	Common Name	Genome (Mb)	Number of Protein-Coding Genes	% of Genome Occupied by Transposable Elements
Saccharomyces cerevisiae	Baker's yeast	12	5,773	3
Dictyostelium discoideum	Slime mold	34	9,000	10
Caenorhabditis elegans	Nematode worm	100	18,400	6
Arabidopsis thaliana	Thale cress	125	25,498	14
Drosophila melanogaster	Fruit fly	180	13,600	15
Anopheles gambiae	Malaria mosquito	278	13,000	16
Takifugu rubripes	Pufferfish	400	38,000	2
Oryza sativa	Rice	400	37,544	35
Mus musculus	Mouse	2,500	30,000	40
Zea mays	Maize (corn)	3,200	50,000	60
Homo sapiens	Humans	3,000	25,000	44

DNA transposons are usually composed of short inverted repeat sequences at their front and rear ends. (The so-called 5' and 3' ends named by convention for the unattached free sites on the deoxyribose moieties at the ends of a string of nucleotides that compose a DNA molecule.) Between the inverted repeats is a sequence encoding a transposase protein that recognizes the inverted repeats and cuts the transposon out of its genomic site. For most transposases, the transposase then holds the ends of the transposon together while it finds another site in the DNA to cut and into which to insert the transposon. Thus, the process is a "cut and paste" one, and the insertions usually, but not always, occur in the DNA close to the original site of the transposon. This phenomenon is called "local hopping." For Sleeping

Beauty (a "resurrected" fish DNA transposon used in making insertional mutations in mice), "local hopping," including insertions into the DNA of the same chromosome on which the transposon is located, accounts for 80% of transposition events (Carlson et al., 2003). The biochemistry of DNA transposon excision and insertion has attracted considerable attention for a number of years, and the detailed enzymology and structural biology of a number of transposases is well known (Schweidenback and Baker, 2008; Yanagihara and Mizuuchi, 2003). DNA transposons predominate in bacteria and some animals (see Table 2.2); however, at present, there are almost no known active DNA transposons in mammals, such as mice or primates. (See exception in bats, discussed in Chapter 3.) Why DNA transposons have generally lost their ability to hop over many millions of years of mammalian evolution is another mystery.

Table 2.2 Proportion of different types of transposable elements in different organisms

Species	Common Name	LTR Retrotransposons	Non-LTR Retrotransposons	DNA Transposons
Saccharomyces cerevisiae	Baker's yeast	100	0	0
Dictyostelium discoideum	Slime mold	45.8	38.5	15.6
Caenorhabditis elegans	Nematode worm	1.7	6.9	91.4
Arabidopsis thaliana	Mustard weed	46	5	48.9
Drosophila melanogaster	Fruit fly	69.2	22.7	8.1
Homo sapiens	Humans	18.6	75.2	6.3
Oryza sativa	Rice	56.2	3.7	40.1
Zea mays	Maize (corn)	95	1.7	3.3

The second main variety of mobile DNA is retrotransposons. The old view that is still commonly emphasized in didactic lectures is that information contained in DNA is transferred to RNA that is then

decoded into protein. This view is so ingrained that it is called the "Central Dogma of Biology." Retrotransposons make a major modification in the Central Dogma of information transfer, essentially turning it on its head. They make the enzyme reverse transcriptase, an RNA-dependent DNA polymerase that catalyzes the synthesis of DNA from RNA. This is backwards from the canonical view of gene action. The discovery of reverse transcriptase was a surprising finding and led to David Baltimore and Howard Temin receiving the Nobel Prize in 1975 (Baltimore, 1995; Temin, 1976).

For a short review, DNA is composed of four different nucleotides in a long string of different combinations. Those four nucleotides are deoxyadenosine monophosphate (A), deoxycytidine monophosphate (C), deoxyguanosine monophosphate (G), thymine monophosphate (T), or A, C, G, and T, for short. In RNA, the nucleotide combinations are similar, but the sugar ribose replaces deoxyribose, and uridine monophosphate (U) replaces thymine monophosphate (T). DNA is in the form of a double helix with the two anti-parallel strands having complementary sequences, A pairing with T, and G pairing with C (see Figure 2.2). The two strands have specific and opposite polarity. Each strand of the double helix is a long string of As, Ts, Gs, and Cs connected by phosphate bridges between the 5' carbon of one deoxyribose and the 3' carbon of the next deoxyribose. One strand has a free phosphate attached to the 5' carbon of a deoxyribose moiety (the 5' end) while at its other end is a free hydroxyl at the 3' carbon of the final deoxyribose (the 3' end). The second strand of the helix runs in the opposite direction such that the nucleotide with the free phosphate at its 5' end (the first nucleotide) pairs with the last nucleotide (or 3' end) of the first strand. The last nucleotide of the second strand (the 3' end) pairs with the first nucleotide of the first strand (the 5' end).

RNA contrasts in other ways from DNA beyond the altered sugar and uridine (U) replacing thymine (T). Much of RNA is in the form of a single strand, not a double helix. Parts of many RNA molecules are double-stranded because a portion of the single strand of RNA can pair with another region of that same single RNA strand, but their structure is very different from the double helix of DNA. Most

types of RNA can traverse the nuclear membrane into the cell's cytoplasm. Almost all cellular DNA is stuck in the nucleus, packaged with proteins in chromosomes. The only DNA in the animal cell cytoplasm is mitochondrial DNA, small ~15,000–17,000 nucleotide pair circles found in many copies in each mitochondrion. A major class of RNA, messenger RNA or mRNA, can be decoded or translated in the cytoplasm into protein with many other smaller RNAs (many proteins also help in this process). DNA cannot be directly decoded into protein. Thus, the canonical pathway of the "Central Dogma" has been the decoding (transcription) of DNA into RNA (mostly messenger RNA), this messenger RNA, or mRNA, making its way into the cytoplasm where it is decoded or translated into protein by the protein synthesizing machinery.

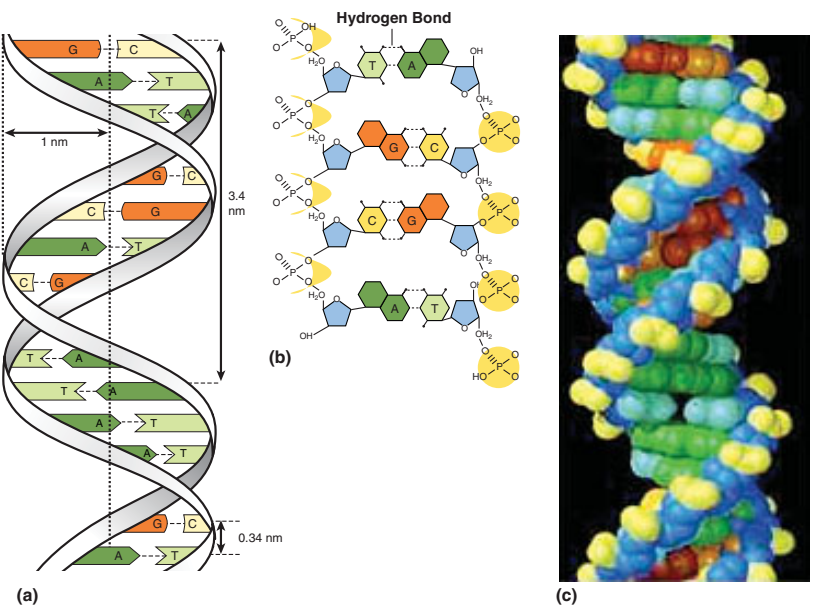


Figure 2.2 Schematic structure of DNA (a), (b), and (c) present different views of the DNA double helix. (Used with permission from Pearson Education.)

However, all the evidence on the origin of life on Earth has been pointing to the first nucleic acid being RNA, not DNA. That would indicate that very early on, probably billions of years ago, an enzyme was needed to synthesize DNA from an RNA template. Thus, a reverse transcriptase must have been one of the earliest of enzymes

as life forms appeared on the planet. A primitive and the oldest reverse transcriptase known was perhaps a descendant of that earliest reverse transcriptase. This reverse transcriptase is the enzyme encoded by the Mauriceville plasmid in the mitochondria of *Neurospora crassa*, the slime mold (Kuiper and Lambowitz, 1988). All known reverse transcriptase enzymes except the one encoded by the Mauriceville plasmid use a short nucleic acid primer to get their enzymatic activity started. In contrast, the Mauriceville reverse transcriptase has the ability to synthesize full-length DNA copies of RNA without a primer of any kind.

Retrotransposons are in striking contrast to DNA transposons. They have taken over very large portions of the genomes of most plants and animals. In plants, the so-called long terminal repeat (LTR)-retrotransposons predominate, while in mammals the majority of the retrotransposons are non-LTR or poly A elements. The LTR-retrotransposons are "copy and paste" elements that have many characteristics similar to retroviruses. They are called LTR-retrotransposons because they have direct repeat sequences of 300 to 1000 nucleotides at their two ends. (These direct repeats contain the same sequence in the same order, say ABCD. In contrast, inverted repeats at the ends of DNA transposons are ABCD at one end and DCBA at the other.) These LTRs contain promoters that stimulate expression (transcription) of the RNA of the element. Using a protein encoded by the element, they make a "coat" for their cytoplasmic viral-like particles (see Figure 2.3). Within the particle, the element RNA is reverse transcribed into DNA in a complicated multi-step process. The double-stranded DNA is then carried back into the nucleus where it is integrated into the host DNA using an integrase enzyme also encoded by the retrotransposon (Garfinkel et al., 1985).

Retroviruses, which likely evolved from LTR-retrotransposons, go through a similar life cycle with reverse transcription occurring in cytoplasmic viral particles. Retroviruses and retrotransposons encode similar enzymatic activities, with one major difference being that retroviruses encode a functional envelope (env) gene that helps them travel from one cell to another, while LTR-retrotransposons do not encode a functional env gene (If they do have an env gene, it is defective and non-functional.) Thus, LTR-retrotransposons cannot traverse cell membranes and are stuck within their original cell.

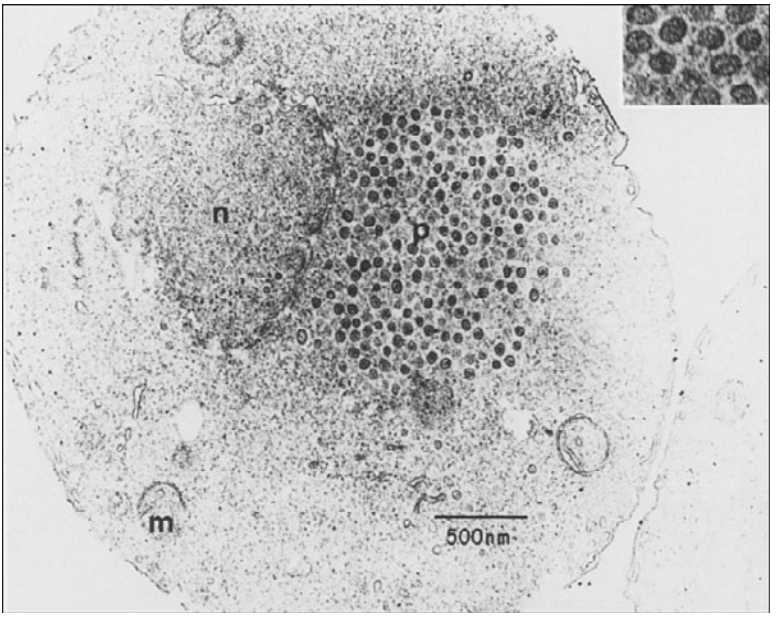


Figure 2.3 Retrotranspositionally active yeast cells contain many cytoplasmic virus-like particles (VLPs). Electron microscopy demonstrates the particle (p), nucleus (n), and mitochondria (m). Reverse transcription of Ty1 retrotransposons is carried out in the cytoplasmic VLPs. (Used with permission from Garfinkel et al., © 2008 Elsevier.)

The non-LTR retrotransposons of many organisms, including mammals, are quite different from LTR-retrotransposons and retroviruses in structure and replication cycle. Non-LTR retrotransposons contain an internal promoter at their beginning or 5' end that is important for starting expression or transcription of the element RNA (Mizrokhi et al., 1988; Swergold, 1990). Active non-LTR retrotransposons usually, but not always, encode two proteins necessary for their retrotransposition. These elements also contain sequence at their rear or 3' end that does not encode protein, and they end in a poly A tail (a region containing many A nucleotides in a row). Usually upon insertion into the genome, this poly A tail is long (50–100 A residues in length). However, in subsequent generations of the organism, the poly A tail length is gradually shortened during DNA synthesis so that the average poly A tail of non-LTR retrotransposons, e.g., LINE-1 elements, in the human genome is 10–20 As. All retrotransposons, both

LTR and non-LTR types, are surrounded by short duplications of the genomic sequence at their insertion sites. These are called target site duplications, and they can either be of fixed or variable length, depending upon the type of element. Retrotransposons are transcribed into an RNA intermediate that is translated into protein in the cytoplasm. The element RNA and its proteins (along with other components) form a ribonucleoprotein particle (RNP) that makes its way back into the nucleus (Martin 1991; Kulpa and Moran 2006).

In contrast to LTR-retrotransposons, reverse transcription of non-LTR retrotransposons occurs in the nucleus on the DNA itself (Luan et al., 1993) and is preceded by a nicking of the so-called bottom strand of the DNA by an endonuclease encoded by the element (Feng et al., 1996). Reverse transcription and integration are then coupled together in one process. The 3' OH at the endonuclease cut site serves as a primer for reverse transcription, while the element RNA serves as a template (Luan et al., 1993). Further steps in the process of integration are not clear, but work involving insect non-LTR retrotransposons is providing clues.

Some retrotransposons are site-specific, which means that they insert into the genome only at very specific sites. For example, the R1 and R2 non-LTR retrotransposons insert only at specific sequences within the ribosomal RNA genes of insects (Burke et al., 1987; Xiong and Eickbush, 1988a). In contrast are a variety of non-LTR retrotransposons of the LINE-1 or L1 type that insert at very many different sites that are merely characterized by being AT-rich, of the type 5'-TTTT/AA-3', where / signifies the cut site (Jurka, 1997).

The LINE-1 or L1 types of retrotransposons have been called autonomous because these elements supply key enzymes for their retrotransposition, namely reverse transcriptase and endonuclease. However, the term autonomous in their description is something of a misnomer because L1 type elements must also require various host factors in order to be mobilized. In contrast to L1 elements are the various SINEs, such as Alus and SVAs in humans, and B1 and B2 in mice. These elements are called non-autonomous because they do not encode any proteins. They rely on the endonuclease and reverse transcriptase of the autonomous L1 elements to assist in their retrotransposition.

One would think that in order for mobile DNA to make up a large fraction of so many diverse genomes that it must either be or have been under positive selection. It must have provided some function to individuals of a species that enhanced their reproduction, or, in other words, increased their genetic fitness. Yet we still don't know what that function might be. Hypotheses abound! Perhaps their reverse transcriptase provides a function during embryonic development. Perhaps they provide promoter activity for transcription of sequences between genes. Perhaps the small RNAs that many of them produce are important for the survival of the host. However, we still don't know for sure whether transposable elements of any kind provide any function to any individual organism. That is to say that we don't know how on an individual basis mobile elements have improved genetic fitness, the ability of individuals to reproduce.

In any case, we do know that mobile DNAs, especially non-LTR retrotransposons, have been major drivers of genome evolution by providing diversity and plasticity to the genome (Kazazian, 2004). They have had effects on genomes for up to 500 million years through a wide variety of mechanisms. These myriad mechanisms will be discussed in a later chapter. Evidence for these effects on genome evolution has come from a number of sources.

1. Bioinformatic analyses of the large number of genomes of various species that have been sequenced to date (over 190), including many bacteria, yeast, *D. melanogaster*, *C. elegans*, *Arabidopsis thaliana*, and mammals, such as the platypus, mouse, rat, dog, cow, opossum, chicken, chimpanzee, and the human being have been invaluable. See Figure 2.4 for mobile DNA content of various mammalian genomes. These genome sequences have yielded a treasure trove of valuable information. This information ranges from when particular types of elements appeared in various organisms, to how frequently they have been mobilized in one species relative to another from the time that the two species diverged, to how much genome sequence and genes have been gained, lost, or altered from one species to another by transposable elements.

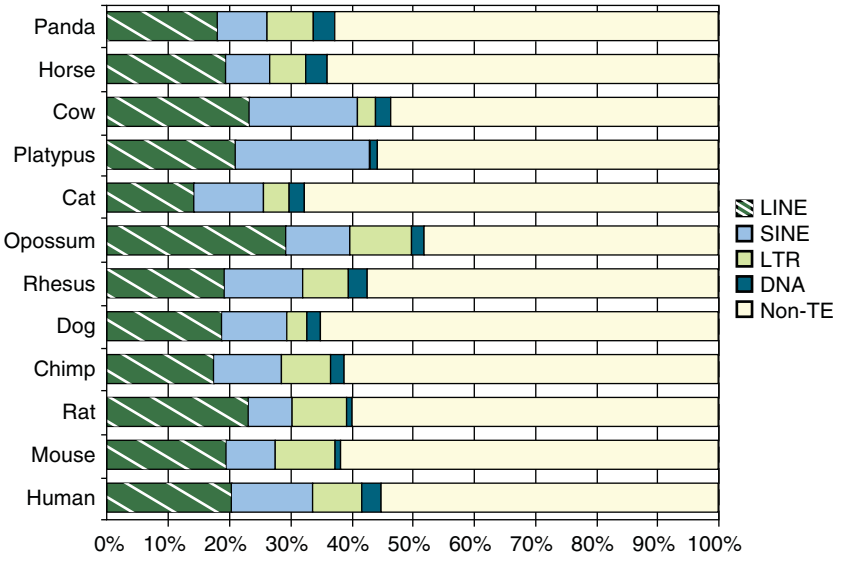


Figure 2.4 The percentage of mobile DNA remnants among sequenced animal genomes. LINEs are represented in hatched dark gray (dark green in e-book version), SINEs in medium gray (blue), LTR-retrotransposons in light gray (green), DNA transposon sequences in black, and the remainder of non-transposable element (TEs) sequence in white.

2. Cell culture assays have elucidated which elements are active. They have told us from analysis of the reference human genome that there are about 80–100 active LINE-1 elements in the average human diploid genome. The diploid genome contains two copies of non-sex, autosomal chromosomes. In addition, analysis of mutations, both natural and man-made in the L1 sequence has told us which sequences are critical for retrotransposition in cultured cells. Cell culture assays have provided evidence that retrotransposon insertion can lead to significant deletion of genomic DNA at the insertion site. These assays have also shown that active L1 elements can supply the required proteins for retrotransposition of non-autonomous human and mouse retrotransposons (see later chapter).
3. Likewise, analysis of transgenic mice and rats carrying L1 elements known to be active in cell culture has given insights into the process and mechanism of retrotransposon integration.

Reassuringly, the insertions characterized from transgenic animals have all the hallmarks of insertions that have been present in genomes for millions of years. Analysis of germ cells and transgenic embryos has given further clues as to when most retrotransposition events occur during germ cell and early embryonic development.

4. Natural insertions in mice, humans, and other animals, such as dogs, that produce a disease phenotype have demonstrated which types of elements are active and whether their activity is mediated in *cis* or in *trans*. (In a *cis* event, the specific retrotransposon encodes its own proteins important for its mobility, while in a *trans* event the element is mobilized using proteins originating from another retrotransposon.) These insertions tell us that retrotransposition events can cause isolated cases of disease in a variety of animals. They have also allowed us to estimate the frequency of retrotransposition in natural mammalian populations.
5. Biochemical analysis of the proteins and RNA involved in mobility of retrotransposons has become very important. We have a good idea about the replication cycle of non-LTR retrotransposons like L1s, but a number of questions are still unanswered (Figure 2.5). Some of these key questions are: "How do the proteins encoded by the retrotransposon work?" "What is their structure?" "Where are they located within the cell, and with what other factors are they associated?" The same questions have been asked for the RNA and the ribonucleoprotein (RNP) complex of RNA and protein that has been found in the cytoplasm as a likely intermediate in non-LTR retrotransposition. The biochemistry of retrotransposition has been difficult to study mainly because one of the proteins (ORF2p of L1) is in very low quantity, even after transfection of cells by an active element. However, firm data and answers to key questions are gradually being obtained and will be discussed in Chapters 18 and 28.

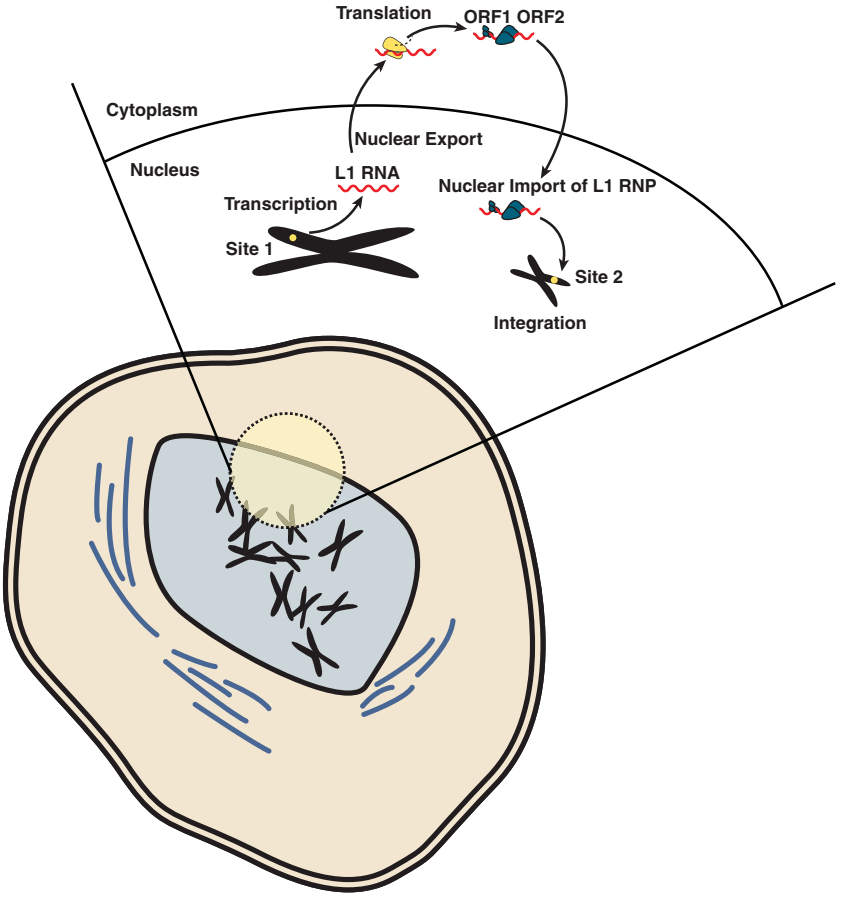


Figure 2.5 Life cycle of a human L1. A full-length L1 is transcribed from Site 1 (inset), and its RNA is transported to the cytoplasm by nuclear export. Bicistronic L1 RNA is translated into two proteins, forming an L1 ribonucleoprotein (RNP) particle. The ORF1p and ORF2p proteins tend to associate with their encoding RNA (so-called *cis* preference). At least some of the L1 RNP re-enters the nucleus where reverse transcription and integration of the newly formed L1 DNA occurs as a single step at a new chromosomal site (Site 2).

6. In the final chapter of this book, I discuss new methods featuring microarray technology and high-throughput sequencing of DNA that are opening new vistas in determining further the role of retrotransposition in human populations and in disease etiology. These new methods hold great promise for delivering exciting new information in the very near future.

Because of the usefulness of these various general methods, we can now present certain key facts about human retrotransposon biology compared to that of other mammals. Humans have active non-LTR-retrotransposons, called L1s, and these active retrotransposons make an endonuclease and a reverse transcriptase that drives the retrotransposition of themselves and of other elements, called Alu and SVA. They also drive the retrotransposition of cellular mRNAs to make processed pseudogenes. Processed pseudogenes are copies of cellular mRNAs that have been reverse transcribed and inserted into the genome at new locations. These sequences lack introns, and they land at their new sites without any promoters. Most processed pseudogenes lose function. Others hijack a promoter near their landing sites and retain their gene function. L1s are present in over 500,000 copies in the human genome, and there are over 1 million copies of Alu present. There are about 3,000 copies of SVA and about 8,000 processed pseudogenes. In human beings, only L1s, Alus, SVAs, and processed psuedogenes are capable of being duplicated and retrotransposed to a new genomic site. Although humans carry remnants of DNA transposons and LTR-retrotransposons in their genome, these latter elements are presently inactive.

In contrast, the mouse has between 1500 and 3000 active L1s, or 15–30 times the number in the human genome. The mouse also has active LTR-retrotransposons, which are mostly retroviral-like and are defective and inactive on their own. However, mouse genomes also contain a few active copies of these retrotransposons, and these few active copies can mobilize the many defective ones in *trans*.

Retrotransposon insertions account for ~10% of all mutations in the mouse but only ~0.1% in human beings. At the extreme, mobile element insertions, mostly insertions of retrotransposons, cause about 80% of the spontaneous mutations in *D. melanogaster*. Genome analyses indicate that there was a substantial burst of retrotransposition of both L1 and Alu elements in the primate lineage about 40 million year ago with a slowing of activity since that time (Khan et al., 2006). At the present time, retrotransposition in primates and humans appears to be ebbing further.

3

DNA transposons

In the late 1940s and early 1950s, Barbara McClintock studied the mosaic color patterns of maize (corn) kernels and the unstable inheritance of this mosaicism. She found "controlling elements," Ac (activator) and Ds (dissociator), that could be mobilized from one chromosomal position to another, leading to changes in kernel color when a suppressed gene containing a Ds element was reactivated and the Ds element moved to another genomic site. McClintock believed that these movable DNAs could regulate gene action and their mobility would in turn be regulated by environmental conditions, such as stress. In 1982, Nina Fedoroff, Sue Wessler, and M. Shure characterized the Ac and Ds elements, sequencing them and showing that Ac is an autonomous element, but Ds requires Ac for transposition (Fedoroff et al., 1983). This was an auspicious beginning for DNA transposons.

Transposons are DNA sequences that encode functions that promote their movement to new locations in the genome. This movement of DNA could potentially occur into genes, thereby disrupting gene expression and compromising viability. There are so many different varieties of transposons that use slightly different mechanisms of removal from their present home (donor site) and insertion into their new one (target site) that I feel compelled to limit the discussion to a few of the more interesting and more "famous" DNA transposons that have gained fame due to their present and potential utility. These include insect Hermes, a hAT element in insertional mutagenesis, insect piggyBac in insertional mutagenesis, fish Tol2 in insertional mutagenesis. Drosophila P-element in gene identification, a fish consensus element, Sleeping Beauty, in insertional mutagenesis and gene

therapy, and bacterial Tn7 in functional gene analysis and DNA sequencing.

DNA transposons are classified by their transposition mechanisms and by the transposases that mediate their movement. While the details of the chemistry behind the transposition reaction vary among the families of DNA transposons, the critical steps are the exposure of 3' OH groups at the transposon ends at the donor site and a strand transfer reaction to integrate the element at the target site (Craig, 1995, 1997). Integration occurs not by cleavage, but via nucleophilic attack on the target site by an exposed 3' OH group.

The Hermes transposon of the housefly is part of the eukaryotic hAT superfamily that includes **h**obo of *Drosophila*, McClintock's maize **A**ctivator, and **T**am3 of snapdragon. Because the sequence of hAT superfamily transposases differs from that of the other elements, it seemed likely that these elements use a distinct mechanism for their mobility. The insect hAT element Hermes, like other transposons, excises itself from DNA via double-strand breaks between the donor site DNA and the transposon ends, and the newly exposed transposon ends join to the target DNA and are inserted. Interestingly, Nancy Craig's lab has shown that the double-strand ends of the donor Hermes form hairpin intermediates (Zhou et al., 2004). These intermediates are similar to those seen during V(D)J recombination, the process that underlies the combinatorial formation of antigen receptor genes. In addition, significant similarities exist in the catalytic amino acids of Hermes transposase, the V(D)J recombinase RAG1/2, and retroviral integrase superfamily transposases. These similarities appear to link the movement of transposable elements and V(D)J recombination. It had previously been shown that RAG1/2 had sequence similarities with known ancient DNA transposons, suggesting that the evolutionary progenitor of RAG1/2, the key V(D)J recombinase, was a DNA transposon. (See the roles David Schatz, Marty Gellert, and Marjorie Oettinger take in Chapter 5, "Exceptional Scientists Working on Mobile DNA in Lower Organisms.")

Two other DNA transposons are presently vying for the title of most valuable for insertional mutagenesis. The *piggyBac* transposon is an insect transposon that is being used increasingly for genome manipulation in a variety of systems including mammalian cells and rodents.

PiggyBac transposase is a member of the DDE superfamily of recombinases, an unanticipated result because of the lack of sequence similarity between *piggyBac* and DDE family of recombinases. DDE superfamily members have aspartic acid residues (D) and glutamic acid residues (E) in key positions that are important for integrase activity. *PiggyBac* is touted as an element that is not subject to "local hopping" (it may be so active that it "skips"), that is highly active in germ cells, and that makes clean excisions and entries, meaning that it rarely, if ever, makes deletions or additions of genomic DNA (Ding et al., 2005; Wu et al., 2006). This last attribute may allow for its use in reversible genetic engineering. Thus, *piggyBac* appears promising as an agent for making inherited insertional mutations in mice and other rodents and is now more popular for this purpose than the hyperactive Sleeping Beauty, mentioned later in this section.

Another transposon, Tol2, is derived from medaka fish. Like *piggyBac*, Tol2 has the ability to make germ line and somatic insertions in mice and is being tested to determine whether it is superior to *piggyBac* (Keng et al., 2009). It is truly amazing that these DNA transposons from insects and fish have the ability to mobilize in mammals, such as mice. They are far from limited in activity to the species in which they originated.

Another famous DNA transposon is the P-element of *Drosophila melanogaster*. In 1973, Margaret Kidwell and colleagues described the condition called hybrid dysgenesis in which particular male flies (P for paternal) crossed to particular female flies (M for maternal) had offspring that died or were sterile with high mutation rates and increased chromosome rearrangement and recombination (Kidwell et al., 1973). Kidwell went on to characterize many of the genetic and environmental factors involved in hybrid dysgenesis. Later, Gerry Rubin and colleagues demonstrated that P-bearing flies carried copies of a DNA transposon, the P element, which M females lacked (Bingham et al., 1982; Rubin et al., 1982). In addition, P males carry a somatic protein inhibitor of transposition that is lacking in M females. At the same time, Allan Spradling and Gerry Rubin developed this element into a useful tool for identifying genes in the fruit fly (Spradling and Rubin, 1982). The native element has inverted repeats at its ends with DNA encoding a transposase between the inverted repeats. Spradling and Rubin replaced the transposase gene with a

marker gene for eye color between the inverted repeats and then mated flies carrying the transposase gene on a chromosome with transgenic fruit flies carrying the marker and inverted repeats on one of their chromosomes (Figure 3.1). In this way, they could mobilize the inverted-repeat transgene to a new site, but this site was nearly always close in the DNA to its original site. This is an example of "local hopping" that is explained by a direct DNA-to-DNA transposition event. Another important point is that most of the time when the element moved, the ins and outs were not precise. Usually, a few hundred to thousands of DNA nucleotides were deleted at the donor and target sites.

Spradling and Rubin then made a large library of fly strains each carrying the movable, inverted-repeat bearing transgene at a different site in the Drosophila chromosomes. Other investigators then expanded the library of flies marked with these defective P-elements that could be mobilized by mating with transposase-bearing flies. For the ten years preceding the sequencing of the Drosophila genome and to some extent since then, fly investigators would map a trait of interest to a position on the fruit fly chromosome map and then go to the atlas of P-element bearing fruit flies and order the strains that had P-element insertions mapping close to the trait of interest. When the element jumped and produced the trait, the investigator could easily find the gene into which it jumped or the gene that was deleted by the jump. Thus, the P-element has been of critical importance to fruit fly genetics and gene characterization, particularly before the Drosophila genome sequence was completed in 2001.

Another important DNA transposon is Sleeping Beauty. Salmon DNA contained an interesting transposon that had sequence similarity to the transposases of other DNA transposons, but it was inactive. Perry Hackett and others sequenced a number of these salmon transposons. Then Hackett developed a consensus sequence for this salmon transposon, similar to Alan Scott's consensus sequence for the human L1 element that is discussed in Chapter 7. Hackett then bet that the consensus sequence would encode an active DNA transposon. He went ahead and painstakingly made mutations in the element that was already closest to the consensus sequence, restoring

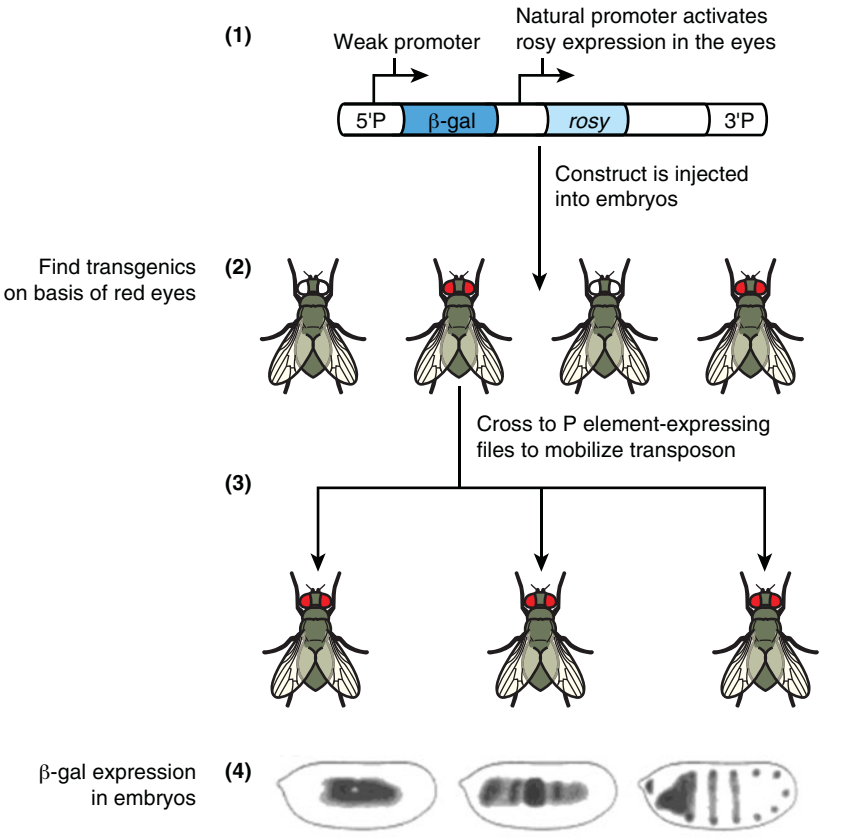


Figure 3.1 Transposition of a P element in Drosophila. A transgene comprising P element 5' and 3' ends and β-galactosidase (β-gal) and rosy genes is transfected into Drosophila eggs (1). Transgenic flies are identified on the basis of red eyes (dark in figure) due to expression of the rosy gene (2). Transgenic red-eyed flies are then mated to P element carrying flies expressing transposase (3). Offspring have mobilized their P elements, and the new transposition events are detected by different patterns of β-gal staining in the embryo (4).

various activities, such as DNA binding and transposase activity, one at a time. After he made roughly 20 different mutations in the element and swapped various segments, it was very close to consensus, and, *voila*, the element was a very active DNA transposon (Figure 3.2). This "reawakened" element he called Sleeping Beauty (Ivics et al., 1997).

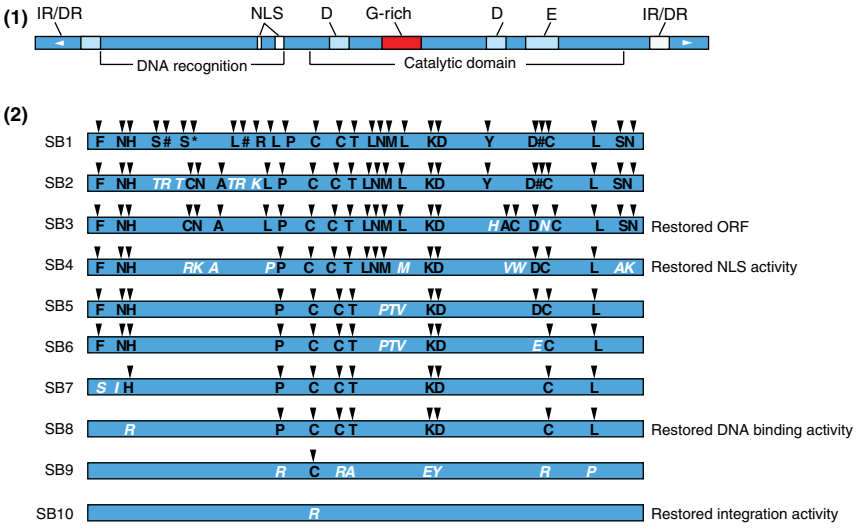


Figure 3.2 Restoring function to an inactive salmon transposon by multiple mutation steps. The original inactive Sleeping Beauty is shown (1). In ten steps, Sleeping Beauty was "reawakened" by sequence changes (downward arrows) that restored the transposase open reading frame (ORF), the nuclear localization signal (NLS), DNA binding, and finally integration activity (2).

Now hyperactive mutants of Sleeping Beauty have been made in the lab. The transposase has become many times more active than that of the original consensus Sleeping Beauty. The element has shown considerable utility in insertional mutagenesis in the mouse in a system similar to the fruit fly P-element system, in which a transposon-bearing mouse is bred with a transposase-bearing animal (Carlson et al., 2003) This system has worked well for both germ line and somatic mutagenesis, but about 80% of the transposition events are into DNA of the same chromosome that harbors the transposon. That is, they demonstrate "local hopping." Sleeping Beauty has also shown promise as a vector for gene therapy in delivering genes to the liver in mice (Kay et al., 2000). Whether this potential becomes realized clinically is unclear at the moment.

Other important transposons are the Tn series that are prevalent in bacteria. Of these, perhaps Tn7 is the most interesting and the most useful. Tn7s natural host is *Escherichia coli*, and it is a somewhat atypical transposon with inverted repeat ends (typical) but

encoding a handful of proteins in addition to a transposase (atypical). Many bacterial DNA transposons do encode another protein, usually an antibiotic-resistance gene, in addition to transposase, but rarely do they encode four or five proteins as does Tn7. Also atypical is the fact that the great majority of the time Tn7 inserts into a specific site in the *E. coli* chromosome, called attTn7, at high frequency (Bainton et al., 1993). Most transposons insert at low frequency with little in the way of target site-specificity. The machinery used for Tn7 insertion includes not only a substantial number of Tn7 proteins, but also a number of host-encoded proteins and *cis*-acting sequences at both the ends of the transposon and at the target site. To insert into the attTn7 site, the transposon uses sequence-specific binding by a protein encoded by the transposon, TnsD, that selects the target. The target DNA sequence, attTn7, is located in the region of the glucosamine synthetase gene (glmS) that encodes the C-terminus of the protein. Interestingly, Tn7 can also transpose into the analogous site in the evolutionarily conserved human gene (gfpt-1) in human cells in culture (Kuduvalli et al., 2005). Whether or not this ability of Tn7 along with its ability to hold considerable DNA could someday be used to deliver genes clinically in a site-specific manner is unclear.

On the other hand, Tn7 also has the ability to insert into essentially random sites in DNA at much lower frequencies than at its attTn7 site. Random insertion has led to production of an insertional library of the yeast (*S. cerevisiae*) genome useful for gene knockouts in haploid yeast (Kumar et al., 2004). The library can then be screened for genes that have an effect on sensitivity to various agents or other environmental conditions. This type of library is quite useful for genome-wide functional analysis of genes. Random insertion by Tn7 has also been used for genomic sequencing in which the genome is flooded with Tn7 insertions and then sheared or cut into small pieces by a restriction endonuclease (Kumar et al., 2004). Sequencing is then carried out from the ends of the transposon into genomic DNA. Obviously, this method has now been overtaken by next-generation sequencing in which millions of sequences of 100 or more nucleotides are generated in a single lane, but prior to 2008, transposons were quite useful in DNA sequencing.

An important general point about DNA transposons is that until quite recently, we thought that there were none that were active in

mammals. Humans and mice do not appear to harbor any active DNA transposons, although their genomes contain many relics and many families of inactive transposons. It was believed that DNA transposons have not been active in the mammalian lineage for at least 40 million years. That is, until recently when eight different families of DNA transposons that were recently active, including hAT family members and *piggyBac*-like elements, were found in bats. At least, one of these eight families appears to be expanding at the present time (Ray et al., 2008). A key question is what makes the bat different and able to harbor these elements in its genome when other mammalian species seem to have driven them to extinction such a long time ago.

Another important point about transposons is that there are many examples, particularly among insects, of their capacity for horizontal transmission. We used to think that all transmission of genomic DNA sequence must occur by vertical transmission, i.e., the heritable transmission of sequence from generation to generation through sexual transmission. Now we know that DNA sequences can be transferred horizontally, meaning from one species to another, particularly in insects, but also in vertebrate animals. In the case of DNA transposons, that means transmission from the germ line or somatic cells of one animal to the germ line of another. Recently, a role for host-parasite interactions has been found for horizontal transfer of DNA transposons. A bug that feeds on the blood of various animals and is a vector of Chagas' disease in humans has in its genome four distinct transposon families that have also invaded the genomes of a diverse set of four-legged animals. The bug transposons have striking sequence similarity to those of the opossum and squirrel monkey, two preferred mammalian hosts in S. America, strongly suggesting that horizontal transfer between vertebrates has been facilitated by the invertebrate parasite (Gilbert et al., 2010). Such horizontal transfer events have also been documented for the P element that is present in many *D. melanogaster* strains but absent in nearly all other Drosophila species. Evidence suggests that the P element entered the melanogaster line from *D. willistoni*, which contains a P element of nearly identical sequence to the melanogaster P element (Daniels et al., 1990). Another example of horizontal transmission is that of

mariner, a DNA transposon that has likely moved horizontally among many arthropod species.

On the other hand, the evidence for horizontal transmission of retrotransposons is essentially non-existent. Presumably, the requirement of an RNA intermediate in retrotransposition makes horizontal transmission difficult to achieve.

This page intentionally left blank

4

Mobile DNA of model organisms

Model organisms have contributed greatly to our knowledge of mobile DNA. Among those organisms are various yeast species, Drosophila, worms, Arabidopsis (the mustard weed), bacteria, mouse, and rat. Mobile DNA in the mouse is discussed in a later chapter.

Mobile elements have been studied extensively in *S. cerevisiae* (the budding yeast), *Schizosaccharomyces pombi* (*S. pombe*, the fission yeast), and *Candida albicans* (the pathogenic fungus). These yeast species are extremely far apart on the evolutionary tree, meaning that they diverged from one another hundreds of millions of years ago. *S. cerevisiae* have a relatively small number of Ty elements that are LTR-retrotransposons. In an analysis of the yeast genome in the late 1990s, Kim et al. found 217 Ty1, 34 Ty2, 41 Ty3, 32 Ty4, and 7 Ty5 elements. Importantly, the vast majority of these elements were solo LTRs or LTR fragments. (Solo LTRs are derived from misalignment and unequal crossing over between two LTR-retrotransposons on separate chromosomal homologues or mispairing and crossing over between the two LTRs of a single LTR-retrotransposon.) Thus, of the 217 Ty1s present in the yeast genome, only about 30 are full-length (Kim et al., 1998).

Retrotransposition frequencies of Ty elements are quite low. Ty1 has been estimated to undergo one retrotransposition event in every 10^5 cell divisions (Garfinkel et al., 2005). This low rate of retrotransposition is in stark contrast to the estimated rate of human L1 retrotransposition (~1 in 140 meioses) and human Alu retrotransposition (~1 in 50 meioses), which is discussed in Chapter 28. The distribution of many of these Ty-type elements is non-random in the genome.

Eighty to ninety percent of Ty1s, Ty2s, Ty3s, and Ty4s are located within 750 nucleotides of a tRNA gene or some other RNA polymerase III-transcribed gene. Ty3 especially inserts in a very restricted region close to the transcription initiation site of RNA polymerase III-transcribed genes and requires host factors along with its encoded integrase for insertion (Kirchner et al., 1995).

Meanwhile, another Ty element, Ty5, specifically inserts in heterochromatin (compacted, relatively inactive, gene-poor regions of chromosomes) near telomeres, or the ends of chromosomes, and the yeast mating type locus. Insertion at these specific sites of heterochromatin is controlled by phosphorylation of the targeting domain of the Ty5 integrase protein. This phosphorylation allows interaction with the host protein, Sir4, and specific integration of Ty5 into heterochromatin. However, when phosphorylation of the targeting domain of Ty5 is inhibited by stress conditions, such as starvation of amino acids, specific targeting of Ty5 is greatly reduced and instead the retrotransposon inserts throughout the genome (Dai et al., 2007). This suggests that mobile elements can alter genome structure as an adaptive response to environmental challenge, a satisfying proof-of-principle for McClintock's earlier postulate.

In *S. pombe*, the major mobile element is a single family of LTR-retrotransposons, Tf. The most studied and characterized of these Tf elements is Tf1. In contrast to the targeting of Ty1-Ty4 of *S. cerevisiae* to RNA polymerase III-transcribed genes, Tf1 finds its safe haven by targeting the promoters of RNA polymerase II-transcribed genes using host transcription activators (Leem et al., 2008). The sequence window used by Tf1 is 100–400 nucleotides upstream of the open reading frame encoding a protein. Notice that all of the LTR retrotransposons of yeast, including Ty5, which inserts into heterochromatin and Tf1, have devised systems in cooperation with their host to insert non-randomly away from genes. Host factors essentially dictate where these retrotransposons will insert, thus protecting the host genome from the damage that would result from insertions of mobile DNA into genes. This diversion of mobile elements in yeast to tolerable locations in the genome is an important aspect of host-mobile DNA symbiosis or cooperation.

Candida albicans, an asexual yeast species very far removed in evolution from *S. cerevisiae*, has some 34 different retrotransposon families, all containing a very small number of members. Most of these are fragments of LTR-retrotransposons and solo LTRs, but a small number of a non-LTR retrotransposon family, called *zorro*, have also been found. One of these was essentially full-length and found to be active in cell culture (Goodwin et al., 2007). This important element is discussed further in the chapter on host factors (Chapter 26).

While *S. cerevisiae* has a very small number of transposable element families, *Drosophila melanogaster* has a very large number of such families of all three types, DNA transposons, LTR-retrotransposons, and non-LTR retrotransposons. Interestingly, none of these families contains a very large number of elements, but each variety of elements is abundant. Of total mobile elements, LTR-retrotransposons make up 69%, non-LTR retrotransposons are 22%, and DNA transposons compose 8%. A recent analysis of the fruit fly genome sequence estimates that 22% is transposable elements,, much greater than previously thought. Roughly 80% of all spontaneous mutations in Drosophila are due to mobile DNA insertions, a far cry from the 10% estimated for the mouse and 0.1% for human beings. The DNA transposon, P element, was discussed in the previous chapter. Other DNA transposons are mariner (closely related to Sleeping Beauty), hobo (related to Hermes), transibs, and many others. These elements appear to be relatively young, ~10 Myr old. The most prominent non-LTR retrotransposon is I factor. Experimental retrotransposition of I factor has been achieved in the whole fly after transfection of fly eggs (Jensen et al., 1999). Knockout of piRNA associated proteins, such as piwi, is associated with a marked increase in expression and retrotransposition of I factor in the ovary (Brennecke et al., 2008; Chambeyron et al., 2008). There are a number of other non-LTR retrotransposons, such as F, G, Doc, and so on, but these have been studied much less than I factor. Other non-LTR elements, HeT-A and TART, are found in long tandem arrays at the telomeres of Drosophila chromosomes. Each HeT-A and TART copy is added by target-primed reverse transcription (TPRT) so that the 5' or front end of a retrotransposon forms the end of the chromosome. The Pardue lab has shown that different mechanisms are used by *D. melanogaster* and *D. virilis* to keep their chromosome ends

intact (George et al., 2010). LTR-retrotransposons in the fruit fly are numerous and have been studied for many years. They also include a large number of families, such as copia (related to Ty1), Tom, 17.6, and so on. For unknown reasons, recent insertions of LTR-retrotransposons appear to have concentrated in and around active genes in the fly.

The numbers of endogenous retroviruses, cousins of LTR-retrotransposons, vary from species to species. The Drosophila element, gypsy, was at one time considered to be an LTR-retrotransposon, but then in various Drosophila species, copies of gypsy were found with active *env* genes and the capacity to infect cells (Song et al., 1994). Gypsy is now considered to be an active endogenous retrovirus in Drosophila. In sharp contrast to the small number of endogenous retroviruses in the fruit fly, the mouse has a large number of endogenous retroviruses. Indeed, many of the numerous mouse endogenous retroviruses are still active. On the other hand, although human beings also have a very large number of endogenous retrovirus sequences comprising some 8% of the human genome, none of these elements are presently active. One or more of their *gag*, *pol*, and *env* genes are defective. A few human endogenous retroviruses, so-called HERV-K (Human Endogenous RetroVirus-K; K stands for lysine as the tRNA at the primer binding site) have *gag*, *pol*, and *env* open reading frames but are unable to retrotranspose *in vivo* or in cell culture. ERVs are discussed in more detail in Chapter 24.

In contrast to the fruit fly and yeast, nearly all the mobile DNA (>91%) in *C. elegans*, a round worm, is DNA transposons, making up about 6% of the worm genome. The study of transposons in *C. elegans* began with the identification of Tc1, the founding member of the Tc1/mariner superfamily. Transposon research led to much-needed genomic tools for *C. elegans* research, including the means to inactivate and clone genes of known function. *Mos1*, a mariner-like element of Drosophila, has also been used to generate single-copy transgenic insertions and engineer alterations in the worm genome by homologous recombination (Robert et al., 2009). Mutation analysis has provided evidence that a large number of host genes are crucial for protection against transposition in the worm (Pothof et al., 2003). Study of how transposition is regulated in *C. elegans* has demonstrated a link between transposition, genome surveillance, and RNA

interference (RNAi). In fact, RNAi was discovered in *C. elegans*, and this discovery led to the 2006 Nobel Prize for Andrew Fire and Craig Mello (Fire, 2007; Mello, 2007).

In another model organism *Arabidopsis thaliana*, the mustard weed, transposable elements account for roughly 14% of the genome. These elements are divided nearly 50-50 between DNA transposons and LTR-retrotransposons with very few non-LTR retrotransposons. *Arabadopsis thaliana* has been used to follow the evolutionary biology of transposable elements and the effects of demethylation on mobility. Recent lab experiments have shown that knockout of the DDM1 gene (decrease in DNA methylation1) leads to a burst of transposition of a class of DNA transposons. Moreover, a similar recent burst of transposition was observed in natural populations, with most insertions occurring into non-genic centromeric repeat regions of the chromosomes (Tsukahara et al., 2009).

In this chapter, I've mentioned a number of different mobile elements in model organisms. It is interesting that certain mobile elements are related to one another in DNA sequence and their mechanism of mobility. We group these very similar transposable elements into "superfamilies." Despite their enormous diversity and abundance, all currently known eukaryotic DNA transposons belong to only 15 superfamilies (Bao et al., 2009). For example, the bacterial DNA transposons Tn5 and Tn10 are related. Mu and Tn7 are related in their core machinery even though Mu is a bacteriophage and Tn7 is a "cut and paste" transposon. The P element has its own "superfamily," while other "superfamilies" of DNA transposons are Tc1/mariner, piggyBac, and hAT (named for **h**obo, **A**c, and **T**am3). The hAT superfamily also includes Tol2 and Hermes. Among LTR-retrotransposons, there are also superfamilies. These include Ty1/copia and Ty3/gypsy. Of course, this classification is muddied by the finding that gypsy is really an endogenous retrovirus.

This page intentionally left blank

5

Exceptional scientists working on mobile DNA in lower organisms

For the mobile DNA subjects that I do not discuss in great detail, I'd like to provide some names of key players whose papers the reader may readily find in PubMed. These scientists have made crucial contributions to the mobile DNA field. In DNA transposition, in addition to Barbara McClintock, Nina Fedoroff, Sue Wessler, Nancy Craig, and Gerry Rubin mentioned in Chapter 3, "DNA Transposons," see the work of Nancy Kleckner, Bill Reznikoff, Koichi Mizuuchi, Nancy Craig, Tania Baker, Mick Chandler, and Nigel Grindley. In related retroviral work, Pat Brown, Harold Varmus, and J. Michael Bishop have made important contributions. In Group I and Group II introns, the stars are Alan Lambowitz, Marlene Belfort, and Phil Perlman. In LTR retrotransposons, look at the contributions of Jef Boeke, David Garfinkel, Joan Curcio, Dan Voytas, Suzanne Sandmeyer, Henry Levin, and Jeff Benetzen. Mary Lou Pardue should be noted for her work on the non-LTR retrotransposons that make up the telomeres of Drosophila chromosomes. In V(D)J recombination and transposition, look for Marty Gellert, Marjorie Oettinger, and David Schatz. In transposition silencing and the role of RNAi, look to the work of Ron Plasterk. In plant transposon work, I especially admire the work of Sue Wessler and Rob Martienssen.

As mentioned in Chapter 3, DNA transposons have been heavily studied in bacteria where they are major drivers of genome remodeling. They also play an important role in horizontal gene transfer (the transfer of DNA from one organism to another). In bacteria, they can

take up and transmit different genes involved in accessory cell functions, such as resistance to antibiotics, catabolism of unusual compounds, and pathogenicity or virulence. Their passing of genes conferring resistance to various antibiotics from one bacterium to another has been a major source of consternation to physicians trying to eradicate bacterial infections. DNA transposons are also used as tools to identify specific gene regulatory regions by insertion.

I consider the giants in the DNA transposition field those individuals who discovered and characterized these transposons in plants and flies: McClintock, Fedoroff, Kidwell, Rubin, and Spradling. Others did pioneering work on the molecular mechanism of transposition in prokaryotes. The ways in which transposons are removed from donor DNA and enter target DNA with the aid of transposase enzyme have been worked out in amazing detail for different transposons. In fact, individuals who pioneered in the molecular mechanism of transposition, such as Nancy Kleckner, and their trainees have gone on to work on the molecular mechanism of meiotic recombination and other forms of DNA recombination and transfer of one DNA strand to another DNA molecule. The pioneers in prokaryotic DNA transposition are Nancy Kleckner (just mentioned), working on bacterial transposon Tn10, Bill Reznikoff working on bacterial transposon Tn5, Koichi Mizuuchi, working on bacteriophage Mu, and Nancy Craig, working on transposon Tn7.

Kleckner showed in a series of elegant studies that the transposition process for Tn10 does not involve DNA replication (Sakai and Kleckner, 1997). 3'OH termini are created at both transposon ends by hydrolytic nicking, and then these termini engage in direct nucleophilic attack upon the two strands of the target DNA in a symmetrical pair of transesterification reactions. Excision of the Tn10 transposon involves first-strand nicking, hairpin formation, and hairpin resolution (Kennedy et al., 1998; Mizuuchi, 1997). The nontransferred strands (the DNA strands that are not being transferred from the donor to the target site) are also cleaved prior to strand transfer, so the double-stranded transposon segment is completely removed from the donor DNA. Transposition results in simple insertion of the excised transposon at the new site.

Mizuuchi pioneered the development of *in vitro* transposition systems (Mizuuchi, 1983). Using his system, he found a different

mechanism for transposition of bacteriophage Mu from the Tn10 and Tn5 mechanism. In this mechanism, the non-transferred strands are not cut prior to strand transfer. Then processing of the branched strand transfer intermediate by host enzymes gives one of two outcomes: 1) a complicated replicative cointegrate that is seen during lytic growth of Mu or 2) simple insertion, as occurs during lysogenic insertion of Mu into its host genome. Other major aspects of Mu transposition have been worked out by this group, including donor DNA cleavage and strand transfer that occurs by a Mu transposase situated on the other end of the transposon (Han and Mizuuchi, 2010).

Reznikoff has characterized many aspects of Tn5 transposition in bacteria, including the synaptic complex of transposase, transposon, and DNA. He has characterized the effect of various transposase active site mutants. In 2000 with Rayment, he published the first three-dimensional crystal structure of a transposase complexed with its transposon and DNA (Davies et al., 2000; Han and Mizuuchi, 2010).

Craig found even more complications with Tn7, another bacterial transposon, including a specialized insertion site in the bacterial chromosome under some conditions and more random sites under other environmental conditions. Also the transposon Tn7 can synthesize a number of different proteins to aid its movement, one of which, TnsD, is crucial for insertion into the specialized site, attTn7, mentioned in Chapter 3 (Bainton et al., 1993; Mitra et al., 2010).

Other important players in the DNA transposon field who have provided important information about the biochemistry of transposition are Mick Chandler, Tania Baker (a trainee of Mizuuchi), George Chaconas, Rasika Harshey, and Nigel Grindley. Mick Chandler in Toulouse, France, has worked for many years on insertion sequence (IS) transposition in bacteria. Insertion sequences are small DNA transposons found in many bacteria. They are the simplest of autonomous DNA transposons in that they encode only a transposase and often a regulatory protein between inverted repeat ends. A variety of structurally and mechanistically different transposase enzymes have evolved to carry out transposition by several different pathways (Beauregard et al., 2008; Curcio and Derbyshire, 2003; Turlan and Chandler, 2000). These transposases all contain an endonuclease activity, allowing them to cleave target DNA and insert the transposon into the new site. In various systems (Curcio and Derbyshire, 2003),

different nucleophiles are used by the transposase to attack a phosphorus atom of a backbone phosphodiester bond and cleave DNA. These nucleophiles include water activated by metal ions, a hydroxyl group at one end of a DNA strand, or the hydroxyl group of a serine or tyrosine located within the transposase itself.

Although DNA transposons generally move by a "cut and paste" mechanism involving removal of a double-strand DNA copy from one site and insertion into another, there are some bacterial insertion sequences (ISs) that use a different mechanism. Indeed, in these special ISs, their transposase is also unusual, and it recognizes only one strand of the transposon, cleaves it from donor DNA, and transfers it to target DNA. The second strand does not transpose. The process of transposition of these ISs has now been worked out in detail. Chandler has found that the transposition of these elements in single-stranded form is linked to host DNA replication (Ton-Hoang et al., 2010).

Tania Baker has been dissecting the role of Mu produced proteins in Mu transposition (Schweidenback and Baker, 2008). Specifically, she has considered how these proteins affect the synaptonemal complex of transposon, DNA target site, and transposase. Grindley works on site-specific recombination in DNA, researching resolvases that come in two types, tyrosine and serine resolvases. These enzymes hold four strands of DNA together and facilitate recombination in bacteria by different mechanisms (Grindley et al., 2006). Recently, another group headed by Phoebe Rice has provided an x-ray crystallographic view of the synaptic complex of target DNA and resolvase (Mouw et al., 2008). Chaconas and Harshey, working independently, have also made numerous important contributions over many years to the biochemical dissection of Mu transposition.

In related work, Brown, working with Varmus and Bishop, set up an *in vitro* system of retroviral integration (Brown et al., 1987). Using this system, they showed that retroviral integration proceeds using the same chemistry as that of DNA transposons, i.e., attack of 3'OH ends at staggered target positions. This observation makes sense of the related active sites of DNA transposases and retroviral integrases.

Moving into another bacterial and lower eukaryotic retroelement we have Alan Lambowitz, Marlene Belfort, and Phil Perlman working in the field of group I and group II introns. These individuals have

played major roles in various aspects of the biology of these fascinating elements. Group I and II introns are interesting because they are special types of retrotransposons that move by different mechanisms. Group I introns contain an endonuclease activity that helps them mobilize from one RNA site to another RNA site. They are self-spicing, found in bacteria, lower eukaryotes, and higher plants, and interrupt ribosomal RNA, mRNA, or transfer RNA, depending on the host. While Group I introns move into RNA, the target of Group II introns is DNA. Group II introns insert into specific sites in a gene by a mechanism that suggests their close relationship to non-LTR or LINE-like retrotransposons. Group I introns are found in certain bacteria, while group II introns are prevalent in bacteria, such as *Lactobacillus lactis*, and in mitochondria and chloroplasts of lower organisms. They are thought to be the precursors of spliceosomal (self-spicing) introns. They have mobility into specific genes at high frequency nearing 100%, called "retrohoming," and into sequences that resemble their retrohoming sites at much lower frequencies of 10^{-4}, termed *retrotransposition*. In Group II mobility, the RNA transcript of the Group II intron inserts into the DNA site by reverse splicing into top strand DNA with the RNA acting as a ribozyme (an RNA enzyme). Then an endonuclease encoded by the Group II intron nicks the bottom DNA strand, and the element encoded reverse transcriptase carries out target primed reverse transcription (TPRT) in a very similar reaction to the TPRT of L1 retrotransposons. The RNA of the Group II intron is then removed, and the second DNA strand is synthesized [see a recent review by (Lambowitz and Zimmerly, 2010) for details].

In the area of LTR retrotransposons, the major work has been done in yeast. Seminal contributions have been made in studies of the Ty elements of *S. cerevisae*, budding yeast. In the Ty1 field, the big guns are Jef Boeke (who is also important in the human L1 field), David Garfinkel, and Joan Curcio who trained with Garfinkel. In Ty3, it's Suzanne Sandmeyer, and, in Ty5 work, it's Dan Voytas. Much of the major work in Ty1 has already been mentioned, including reverse transcription of element RNA in a viral-like particle (VLP), a large number of particles in the cytoplasm of a Ty1 overproducing yeast cell (Garfinkel et al., 1985; Eichinger and Boeke 1988), a particle coat protein and an integrase encoded by Ty1. Ty1 insertions are located

in the yeast genome mainly in the 5' flanks of polIII transcribed genes, like tRNAs (Devine and Boeke, 1996). Sandmeyer has shown that Ty3 inserts at precise positions close to polIII genes (Kirchner et al., 1995), while Voytas has worked on the mechanism of Ty5 insertion into heterochromatic regions of the yeast genome (Dai et al., 2007). Garfinkel and Curcio have made their mark by characterizing a number of yeast proteins, or host factors, that are important for stimulating or attenuating Ty1 retrotransposition (Beauregard et al., 2008; Checkley et al., 2010).

In another yeast, the fission yeast *S. pombe*, Henry Levin has made great progress in work on the biology of Tf1, another LTR retrotransposon. Levin has shown that Tf1, in contrast to Ty1 and Ty3 of *S. cerevesiae*, prefers to insert at the 5' side of polII transcribed genes (Leem et al., 2008).

LTR retrotransposons make up a large fraction of the genome of many plants. Insertions of one retrotransposon into the sequence of another retrotransposon are commonplace in various plant genomes. Jeff Benetzen has been a major contributor to the analysis of retrotransposons and their effect on gene function and the evolution of a number of plant genomes, including maize and rice.

In the area of site-specific recombination, Marty Gellert, Marjorie Oettinger, and David Schatz have reproduced V(D)J recombination in the test tube, and shown that it has many features of DNA transposition reactions. The genes encoding immunoglobulins and T-cell receptors are assembled from the multiple variable (V), joining (J), and occasionally diversity (D) genes present in germline loci. V(D)J recombination is the major source of immune system diversity in vertebrates. The recombinase that initiates V(D)J, recombination-activating genes 1 (RAG1) and 2 (RAG2), belongs to a large gene family that includes transposases and retroviral integrases. Oettinger in a collaboration with Gellert, showed that purified RAG1/2 are sufficient to cleave the DNA adjacent to the gene segments to be recombined (McBlane et al., 1995), and Oettinger went on to demonstrate the key sequences in RAG1/2 important for this process. After cleavage, the segments are then joined together by DNA repair enzymes (Grundy et al., 2007; Jones and Gellert, 2004; Chatterji et al., 2004).

There are many similarities between V(D)J recombination and transposition. Gellert's lab has shown that RAG1/2 can carry out transpositional strand transfer *in vitro* (Hiom et al., 1998), while Schatz's group has demonstrated RAG1/2 binding to sites of recombination called "recombination centers" *in vivo* (Chatterji et al., 2004; Ji et al., 2010).

Ron Plasterk has been a major player in the study of DNA transposons of *C. elegans*, the round worm. Plasterk characterized host proteins in the worm that suppressed transposition of Tc1, a mariner-like DNA transposon. He found a number of silencers, and among them were components of the RNAi machinery (Pothof et al., 2003). His group showed how RNAi produced by read-through transcription of Tc1 and foldback of the RNA could have a direct silencing affect on Tc1 transposition *in vivo* (Sijen and Plasterk, 2003).

As for plant transposons, I should mention a couple of players, Sue Wessler and Rob Martienssen. Working in rice genomes, Wessler has found miniature inverted repeat transposable elements (MITEs) that are non-autonomous and dependent on mariner-like elements (MLEs) for transposase. In the rice genome, there are tens of thousands of MITEs and only tens of MLEs (Jiang et al., 2004). MITEs have less affinity for the MLE transposase than MLEs but also lack MLE sequences that inhibit transposition. Thus, MITEs are able to use the MLE transposase for mobility (Yang et al., 2009).

Martienssen has elucidated a role for small RNAs produced by transposable elements in *Arabidopsis thaliana*. These small RNAs, which differ in size between developing pollen and sperm, appear to limit or silence transposition in developing germ cells (Martienssen, 2010). Perhaps there is a similar role for small RNAs in limiting retrotransposition in developing germ cells in mammals.

These are a few of the areas outside of mammalian mobile DNA and non-LTR retrotransposons that are important for our understanding of the diversity of the field. In this chapter, I've highlighted the seminal work of the main contributors to this very broad field, but I realize that other worthy investigators may have been overlooked. To those individuals, I offer a sincere apology.

This page intentionally left blank

6

Role of bioinformatics in genome analysis

As DNA sequencing has improved, the number of species that have had at least one genome sequenced has expanded enormously. The genomes of an ever-growing number of individual human beings have been sequenced, including James Watson, Craig Venter, 2 Africans, and a Korean. In addition, the genomic DNA of 1000 other human beings is being sequenced, though most are presently scheduled for low coverage sequencing. (Coverage of 2–4 genomes of sequence (2–4x) is considered low coverage, while 40 genomes of sequence (40x) is considered high coverage.) The total number of human genomes sequenced as of September 2010 is >500. A large number of mammals and other vertebrates have had their genomes sequenced. In most cases, these sequences do not yet extend beyond a single individual of a species. This means that little is known about DNA polymorphisms in those species. All of these DNA sequences have gone into large computer databases, and these databases provide a treasure trove of information for analysis. Indeed, the computer analysis of genome sequence, particularly the comparison of genome sequences among a variety of different species, has become an extremely important and productive undertaking. The biological sciences now need many more individuals trained not only in wet bench research, but also in bioinformatics. Indeed, in my view and the view of many biologists, the two most important broad scientific areas in biology today are genome science and bioinformatics. These two areas join "hand in glove" in DNA sequence analysis.

What analyses are being carried out with regard to repetitive DNA? Analysis of human genome sequence alone has identified L1s

of different subfamilies based on their sequence, some of which are very young and still active, while others are of varying ages based on their divergence from the youngest sequence. Taking a look at Figure 6.1, this analysis has found some 17 different L1 subfamilies in the human genome with evidence that at any time over the past 40 million years (Myr) only one subfamily has been active (Boissinot and Furano, 2005). In the past 25 Myr, five different subfamilies of L1s have arisen and died away in primates. In other words, 25 Myr ago, a single active L1 subfamily expanded and then died out after being replaced by a new L1 subfamily, which itself died out in a few million years and was replaced by another L1 subfamily, and so on.

Comparison of the chimpanzee genome with the human genome has shown that in the chimp genome, two L1 subfamilies have continued to expand over the past 5 Myr. L1s specific to either human or chimp genomes are similar in number. However, the chimp genome has roughly twice as many polymorphic L1s as the human genome, suggesting that the effective ancestral population size of chimps was greater than that of humans (Lee et al., 2007).

Another type of genome analysis by bioinformatics involves the study of the sequences of the two long terminal repeats (LTRs) of particular LTR-retrotransposons. Because of the mechanism of reverse transcription of LTR-retrotransposons, the left and right LTRs of a single element have the identical sequence at the moment of insertion. Thus, analysis of the sequence differences between the two LTRs of a particular insertion provides data on the length of time that has expired since that element inserted into the genome. This timing of LTR insertions has been particularly valuable in the study of the evolutionary biology of plants (SanMiguel et al., 1998).

In plant genomes in which 60–70% of the genome is transposable element DNA, most of the transposable elements are LTR-retrotransposons; many of these elements have inserted into other older transposable elements (SanMiguel et al., 1996). Thus, one finds many locations in the rice or barley genomes in which transposable element sequence has been disrupted by a second retrotransposition event whose sequence has in turn been disrupted by a third insertion. This phenomenon of multiple retrotransposon insertions at the same location is another way to determine the order and timing of insertions.

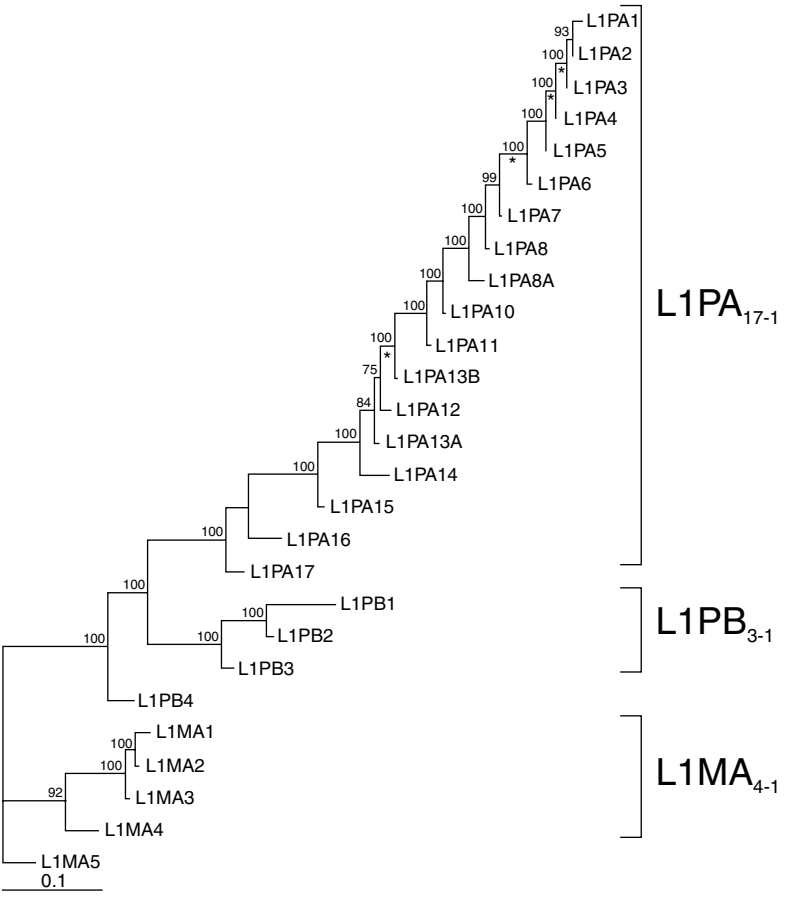


Figure 6.1 Analysis of L1 families in human genomic DNA by Khan et al., 2005. L1 PA17-1 have evolved over the past 40 million years one L1 family at a time. The entire L1 phylogeny shown spans 70–74 million years. The present human L1 family is L1PA1 or Ta and is 2–3 million years old.

Comparison of genome sequences has also led to changes in the taxonomic classification of certain organisms. For example, analysis of SINE and LINE insertions has been very useful for inferring phylogenetic relationships. In one study, use of these insertions led to the conclusion that the closest living relative of the whale is the hippopotamus (Nikaido et al., 1999). Whales turn out to be deeply nested phylogenetically within Artiodactyla, or even-toed ungulates, and whales and hippopotamuses form one phylogenetic group. Besides hippopotamus, the even-toed ungulate order contains pigs, camels, deer, giraffes, antelopes, sheep, goats, and cattle. Thus, it was

completely unexpected that whales are also members of this order. In addition to this reclassification, other examples of taxonomic classification and reclassification have come from comparative analysis of SINE and LINE sequences in different genomes.

Because we know the ancestral state of any particular transposable element, that is, absence of the element from the site, SINEs and LINEs are useful genetic markers for analyzing primate phylogeny and human population genetic relationships. Batzer's group has used 350 Alu insertions to infer the phylogenetic relationships of the macaque genus (Li et al., 2009). They suggest that this genus contains four different species groups, that the Asian group diverged first, and that the relationships of the three other macaque species to each other can be determined. This work demonstrates that transposable elements can be used not only to place animals, such as the whale, within an order, but also to resolve evolutionary relationships among taxa of closely related species within a genus.

Transposable elements can be used to study species diversification not only among living species, but also among extinct species. One such extinct species is the wooly mammoth. Analysis of the transposable elements in this species has revealed a novel pattern of transposable element diversification. The mammoth genome contained just over 30% LINEs, but a whopping 12% of the genome, or 40% of those LINEs, was BovB LINEs, an element found in widely diverged species, such as cattle, snakes, and marsupials (Zhao et al., 2009). Because these other species have much less BovB LINE in their genomes, BovB may have entered the mammoth genome by horizontal transfer, suggesting that the mammoth and other vertebrates may have acquired BovB from another organism rather than by inheritance.

Computational analysis of genomes has led to finding unusual mechanisms to make new genes. One example is SETMAR, a new chimeric gene in primates resulting from fusion of a SET histone methyltransferase gene to the transposase gene of a DNA transposon, Hsmar1. The fusion occurred between 40 and 60 million years ago (Cordaux et al., 2006). Another example is a fusion gene caused by unusual splicing of an RNA followed by retrotransposition. This chimeric gene formed ~17 million years ago in the primate lineage (Babushok et al., 2007), and is discussed in Chapter 20.

Another use of the analysis of retrotransposon insertions is in determining human origins and migrations in human history. Many studies using various marker loci have found that at most human loci, ancestral alleles (the oldest ones) reach their highest frequency among Africans, suggesting that they are of African origin. Conventional wisdom holds that this reflects a recent African origin of modern humans. Allele frequencies of retrotransposon (Alu) polymorphisms challenge that view by showing that the empirical pattern of elevated allele frequencies within Africa is not as pervasive as once thought. Although there is an African bias in a set of protein-coding loci, this bias is much smaller in Alu insertion polymorphisms and even smaller in non-coding loci. Thus, the strong bias for an African origin of protein-coding loci that was originally observed might reflect some other factor that varies among data sets. This factor may be the mutation rate per locus, given the African bias is most pronounced in loci where the mutation rate is high. However, Adam Ewing in our lab has recently found from the 1000 Genomes Project data that L1 polymorphisms specific to African populations are far more prevalent than those specific to European and Asian populations (Ewing and Kazazian, 2011), as discussed in Chapter 28. Thus, in contrast to the abovementioned Alu data that suggest some ambiguity, the L1 results strongly corroborate previous mitochondrial and nuclear gene data, indicating an African origin of modern humans.

Mobile DNA elements also represent an excellent group of molecular markers for identity testing and forensic applications. For any particular position in genomic DNA, as previously mentioned, essentially only one insertion has occurred in human history. Therefore, a key characteristic of Alus and L1s is that the ancestral state of the insertion site is known to be the absence of the transposable element. In addition, the transposable elements of any one species can be distinguished by their sequence from the transposable elements of any other species, that is, mobile DNA is lineage-specific. Analysis of SINEs and LINEs in forensic samples can provide quantitative species-specific DNA detection, meaning that one can quantify the fraction of any DNA sample that is human in origin (Ray et al., 2007). Transposable element analysis can also unravel the source(s) of complex biological samples. Moreover, analysis of various human transposable elements can yield information on the geographic origin of any

human DNA, meaning that one can determine whether a human sample is from an Asian, Caucasian, or other ethnic group. Thus, mobile DNA can be very useful in forensics and probably could be used soon for identity testing because the number and location of active retrotransposon families (L1s, Alus, and SVAs) is different from one individual to another.

Now that I've provided some background information on mobile DNA, I'd like to recount my experience in the field, beginning with how I happened to get into it and the state of the field at that time. This discussion is also meant to inform the reader about the vagaries and chance occurrences that influence how one does science. Because my trainees and I have picked significant problems on which to work, the people whose work is discussed have experienced considerable frustration and failure before finding success. Their flexibility and perseverance to move from disappointment to new attempts fraught with uncertainty have characterized these young scientists. I admire all of them greatly! They've really come through in the clutch! Later, I provide more information about mammalian mobile DNA in particular, its role in driving genome evolution, the ways in which the host controls its activity, and thoughts concerning the future of the field.

7

The prologue

Rarely does one find Maxine Singer outside a public place without a lit cigarette. Once, a scientific colleague described Maxine walking up the winding road near the Carnegie Institution building, named after her, on the edge of the Johns Hopkins campus as "a diminutive white-haired lady surrounded by a cloud of smoke." Yet she can survive hours on end in a conference room or a meeting hall without a cigarette.

I first met Maxine when Alan Scott invited her to Hopkins to present a seminar on her work on repetitive DNA. It was the early 1980s, and Maxine had immersed herself in a new field. She had been universally recognized as a world-class nucleic acid biochemist, working on RNA enzymes. Recently, she had begun an effort to understand the structure and function of repetitive DNA (DNA present in many, many copies of very similar sequence) in human and primate genomes. She was a very quick mind, dedicated to Science with a capital S, but loved talking about her family almost as much as discussing nucleic acids. In her seminar, she pointed out that her lab had found that one kind of repetitive DNA was present in the human genome in roughly 100,000 copies. (We now realize that that number is a 5-fold underestimate.) She called these sequences LINEs for *l*ong *in*terspersed *e*lements. The repeated short interspersed sequences she called SINEs. The most prominent class of LINEs (the ones she was working on), she called LINE-1 or L1 for short. She said these elements were mostly fragments of one end of the element, but perhaps 5% of them were full length, and the full-length element was 6,000 base pairs long (6 kb) (Grimaldi et al., 1984).

Yet certain aspects of these long interspersed sequences had been known for a few years. Scientists are continually "climbing on the shoulders" of their predecessors. They use whatever previous information is available to advance knowledge of our universe and of humankind. The buildup of information passed on from many, many past researchers, many of whom we now consider colleagues, provides the stepping-stones for future discoveries.

In the 1960s, Roy Britten at Cal Tech had invented a DNA hybridization technique and had used it to determine the fraction of the human genome that was either highly repetitive, moderately repetitive, or single-copy DNA (see Figure 7.1). He had found that a large fraction of our genome was repetitive, and only a minority of the genome was single copy (Britten and Kohne, 1968). (Most genes were thought to be present in only one location in the genome and therefore to be single-copy DNA. Note that all genes except for those on the sex chromosomes are actually present in two copies, one on each chromosomal homologue of a chromosome pair.)

In 1980, Art Nienhuis, a colleague from the hemoglobin field, had visited Hopkins from NIH and told a seminar audience that near the gene encoding β-globin, one of the proteins that make up adult hemoglobin, was a roughly 6 kb sequence that was repeated a few thousand times in the human genome. Most interestingly, this repeated sequence appeared to have sequence similarities to the DNA sequence of retroviral reverse transcriptases (Adams et al., 1980). To repeat, reverse transcriptase is the enzyme that reverses a section of the canonical information route from DNA to RNA to protein by passing information back from RNA to DNA. Ten years previously, David Baltimore and Howard Temin had surprised the biological world by discovering reverse transcriptase activity in retroviruses, some of which like avian myeloblastosis virus (AMV) can cause cancer. For this discovery, Baltimore and Temin had shared the Nobel Prize in 1975 (Baltimore, 1995; Temin, 1976). Thus, the suggestion that perhaps there was a reverse transcriptase activity encoded not only in viruses, but also in the human genome itself was very exciting! Not only that, but the human reverse transcriptase activity might be present in thousands of places and copies in the genome. Wow! Potentially exciting stuff, but at the time I had no clue that it would ever have any connection to my research interests.

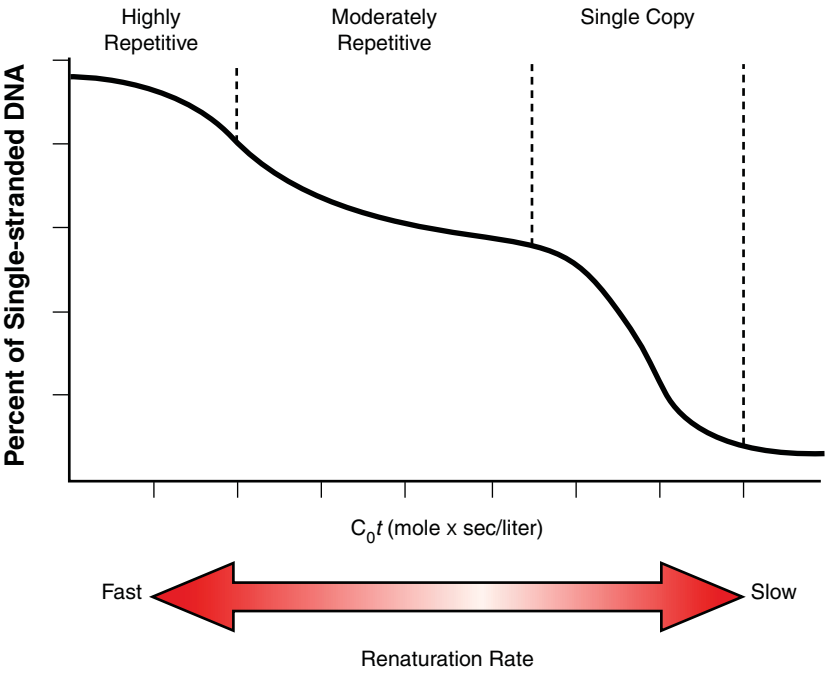


Figure 7.1 Renaturation of DNA shows that the majority of the genome is repetitive sequence. Britten and Kohne, 1968, denatured the DNA by heating and then cooled the preparation to renature. Highly repetitive sequences renature rapidly, while single copy sequences (about 40% of the genome) renature very slowly. The percent of the DNA that remains single stranded over time (not renatured) is plotted against the DNA concentration multiplied by time (C_0t). LINEs and SiNEs are present in the moderately repetitive sequences (roughly 30% of the total).

Meanwhile, at the University of North Carolina in Chapel Hill, Marshall Edgell and Clyde Hutchison had teamed up to sequence the whole region of a cluster of genes encoding β-globin-like proteins of mouse hemoglobins. This was a huge undertaking at the time. The region was over 80 kb in size, and interspersed between the hemoglobin genes the region was packed with repetitive DNA, most of which was L1. By 1981, Edgell and Hutchison had obtained much of the sequence that was published in its enitirety in 1989. In analyzing the L1s in the mouse genome, they found one that was over 7 kb long and appeared to contain two regions capable of encoding proteins (Shehee et al., 1989). They postulated that these LINE-1s of the mouse were transposable elements, and it appeared as though the mouse

genome had just as many L1s as the human genome. So the mouse became another important source of data on L1 elements. Edgell and Hutchison even hypothesized that these potential transposable elements have as one function to fill in gaps in DNA (Voliva et al., 1984). We now know that this hypothesis was correct but that this "band aid" effect happens in only a small fraction of mobile element "hops."

But I digress. Let's get back to Maxine Singer. My second encounter with Maxine was fleeting. I was scheduled to speak right after her at the Cold Spring Harbor meeting of 1986 on the *Biology of Homo Sapiens*. Although I was nervous at presenting next, I was able to concentrate on her presentation, and I did learn a lot from her. Jaceb Skowronski in the Singer lab had succeeded in isolating from an embryonic tumor cultured cell line a batch of L1 RNAs. He then made DNA copies of them and sequenced a number of these DNA copies. Making these DNA copies in the lab was a handy way to discover the sequences of the L1 RNAs, that is, L1s that were expressed into RNA. Some of the RNAs were nearly full length, and some could almost be decoded into two good-sized proteins. They had a few substitutions of nucleotides that prevented them from encoding proteins. Another interesting point was that 20% of the L1s that were transcribed into RNA had an unusual sequence at their 3' end. Skowronski and Singer called these special L1s, Ta, for the first class of L1s discovered that was transcribed into RNA (Skowronski et al., 1988). So Singer had shown that some of these repeated DNA sequences made their way into RNA, a first step in demonstrating their potential importance. They weren't just sitting in the genome doing nothing. Some L1s were being expressed. Perhaps some of these L1 RNAs, those that didn't contain protein-coding mutations, were getting translated, or decoded, into protein.

So what did we know about mammalian mobile DNA in 1986, and what didn't we know that we know now? We knew that there were a lot of L1 and Alu sequences present in mammalian genomes, and we thought that those two repeats might account for between 5 and 10% of the human genome. We now know that that number is at least 33% and probably 50%. We knew that some L1s were transcribed into RNA, and some might make protein that had the possibility of encoding a reverse transcriptase. We thought that some repeats might still be actively transposing, but no natural insertions

had been found in any animals, although natural insertions and cultured cell assays were available for the yeast retrotransposon, Ty1. So we didn't know much about mammalian mobile DNA, but we did know quite a bit about bacterial, yeast, and Drosophila mobile DNA.

In 2010, we've learned a lot more about mammalian mobile DNA. We now know that at least 70 different natural insertions of retrotransposable elements have caused isolated cases of disease in human beings (Chen et al. 2005; Goodier and Kazazian 2008). We know about a similar large number of disease or phenotype-altering insertions in mice of L1s and of the LTR-retrotransposons, IAP (Intracisternal A Particle), Etn (Early transposon), and MaLR (Mammalian apparent LTR-Retrotransposon). We have robust cell culture assays for determining retrotransposition of human and mouse L1s, human Alus, human processed pseudogenes, mouse IAPs and Etns, and the small mouse elements, such as B1s (the Alu-like element in mice) and B2s. We can use an *in vitro* cell culture assay to determine the relative retrotransposition capability of an L1 or alleles of a particular L1. We can make mutant L1s and test them in the cell-culture assay. We can also retrotranspose human L1 elements in transgenic mice and rats from many transgenic chromosomal locations or from one predetermined chromosomal site. We can retrotranspose a mouse L1 in transgenic mice. (See Goodier and Kazazian, 2008, for review of many of these points.) We can isolate L1 ribonucleoprotein particles (RNPs) from cells and show that they contain L1 encoded proteins, L1 RNA, and reverse transcriptase activity. We can also find these RNPs by immunohistochemistry within cells. However, we still don't know anything about non-L1 proteins and RNAs associated with L1 RNPs. Amazingly, we can analyze the sequence of a wide variety of animals for changes in retrotransposon content over evolutionary time (see Figure 2.4 in Chapter 2). We can follow the evolution of a particular element over millions of years using sequences present in the human genome. For example, we now know that over the past 40 million years there was only one active family of L1 retrotransposons at any one time (see Figure 6.1 in Chapter 6). As one family took over, its predecessor lost its activity. We've learned about ten or more different mechanisms by which L1s and other retrotransposons can alter mammalian genomes, potentially shuffle gene exons,

and affect gene expression. We've learned about ways by which the host works to control retrotransposon mobility. Most prominently, the fact that sequences are repeated means that very similar sequences are present at many genomic locations. This setup leads to homologous unequal crossing over or recombination, producing either deletion or duplication of the sequence between two copies of the repeat. These examples represent just a small sampling of the new information available since 1986 on mammalian mobile DNA. They will all be discussed later in Chapters 25, "Effects of Retrotransposons on Mammalian Genomes," and 26, "Host Factors Involved in L1 Retrotransposition." In 1986, the field was in its infancy but poised for a breakthrough.

Around the same time in 1986–1987, two groups, one led by Alan Scott at Johns Hopkins, and a second led by Yoshi Sakaki at the University of Tokyo, provided further clues on L1 biology. Scott had become a good friend of mine since coming to Hopkins in 1975 as a postdoc with Ned Boyer. He was a soft-spoken, unassuming man who thought hard about what he wanted to say before speaking. Alan had earlier helped me scientifically by training one of my postdocs, and I enjoyed and respected him greatly. In fact, I was so high on Scott's ability that I talked my Chair of Pediatrics, John Littlefield, into bringing him onto the Pediatric Genetics faculty at Hopkins after he finished his postdoc training with Boyer in the Department of Medicine. Our lab space was tight, so Scott went into a small "attic" lab on the top floor of the Children's Center at Hopkins. He had become interested in repetitive DNA in the human genome in the early 1980s, so he decided to sequence a large number of L1 elements from various chromosomal locations in the human genome to try to develop a consensus sequence of these elements.

A consensus sequence was made by finding the sequence of a number of L1 DNAs at each particular nucleotide in the 6 kb element and then deciding which of the four nucleotides was most frequent at that nucleotide position. For instance, let's say Scott had eight sequences that included position 4000 and that six sequences contained an A at position 4000, while the other two had a G at that site. The consensus for nucleotide 4000 would be A, and so on for the roughly 6,000 nucleotides in the element. Because many of the L1s in

the human genome contain only their 3' end, Scott, essentially working alone, sequenced nearly twenty 3' ends, but only a handful of 5' ends. However, when Scott analyzed his consensus sequence, he found something very interesting indeed. Although none of the individual elements that Alan had sequenced were capable of encoding one or more proteins, the consensus sequence could be decoded or translated into two proteins (Figure 7.2).

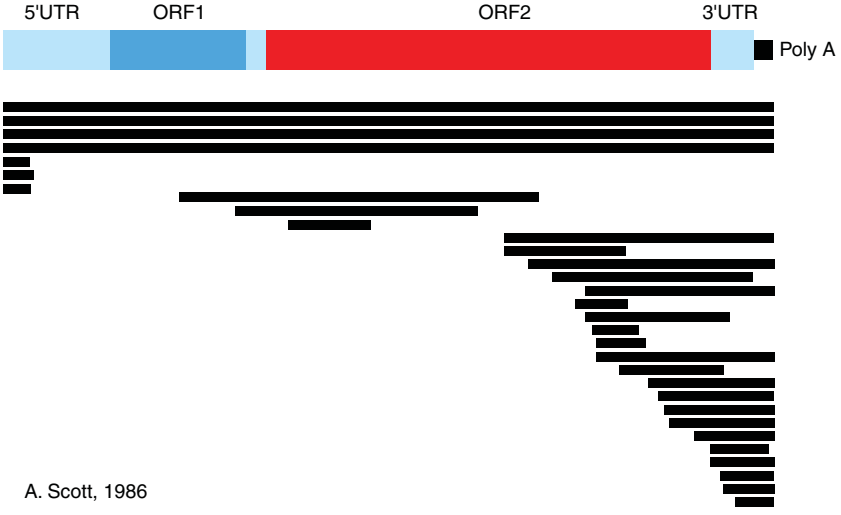


Figure 7.2 A consensus sequence for human L1 elements. Scott isolated and sequenced full-length, truncated, and L1 fragments from 31 locations in the human genome. He then developed a consensus sequence for the human L1 at every one of the 6,000+ nucleotides. His major finding was that the consensus sequence contained two intact open reading frames or ORFs. It also contained 5' and 3' untranslated regions or UTRs. See text for further details. (© 2008 with permission from Elsevier)

One of these proteins, called ORF1p for the protein decoded from the first open reading frame of the DNA, was predicted to be about 300 amino acids long, while the other protein, called ORF2p, was projected to contain around 1300 amino acids. The term ORF refers to an *o*pen *r*eading *f*rame, meaning that the protein-synthesizing machinery of the cell has the potential to synthesize a protein by decoding the information contained in the ORF. Scott published his consensus L1 sequence in 1987 (Scott et al., 1987).

A bit earlier in 1986, M. Hattori in Sakaki's group in Tokyo analyzed the second ORF region of L1s of various mammals. He found that there was a sequence with significant similarity to reverse transcriptase of retroviruses (Hattori et al., 1986). This sounded very much like the conclusion reached by Nienhuis earlier from his analysis of the L1 near a human hemoglobin gene, so it was reassuring that the same conclusion was now coming from multiple sources. Hattori's observation that a region of the LINE-1 sequence had sequence similarity to reverse transcriptase was important, but it was not based on any data indicating that human L1 had one or more ORFs. Scott's work showed that the consensus sequence of human L1 indeed contained two ORFs, suggesting that there were likely individual elements with intact ORFs, and some of these might be capable of making a reverse transcriptase enzyme.

Why was it important to realize that L1 sequences might have the potential to encode a reverse transcriptase activity? As mentioned in Chapter 3, "DNA Transposons," about 30 years previously, Barbara McClintock, working with maize at the Cold Spring Harbor labs on Long Island, had found what she called "controlling elements," "mutable loci" that caused mosaic coloration of maize kernels. She thought these sequences might turn out to be important in gene regulation (McClintock, 1950). Nobody believed her at the time, but gradually over the next 20 years, her work gained in favor. Finally in 1983, Nina Federoff and colleagues at the Carnegie Institution in Baltimore isolated and characterized Ac and Ds, the DNA transposons that McClintock had discovered some 30 years earlier.

To reiterate, a DNA transposon is a piece of DNA that has the ability to move from one place in the genome to another. McClintock's transposable elements were DNA transposons (discussed in Chapter 3), pieces of DNA that could be cut from one genomic site and pasted into another genomic site. These elements did not duplicate themselves upon mobility. They were lost from their original site. In distinct contrast was the second recognized class of transposable or mobile DNA. This class was called retrotransposable elements (see Chapter 4, "Mobile DNA of model organisms") because within the mechanism of its mobility was an RNA intermediate. The retrotransposable piece of DNA is first decoded into RNA. In 1985, Jef Boeke, David Garfinkel, C.A. Styles, and Gerry Fink showed that

a yeast retrotransposable element called Ty1 had this RNA intermediate step in its mobility (Boeke et al., 1985). After the RNA of a retrotransposable element is synthesized, it needs to be reverse transcribed back into DNA, and the DNA copy inserted into the genome at a new site. (Indeed, in working on the Ty1 paper Boeke coined the term "retrotransposon" in homage to the RNA intermediate and its reverse transcription.) The retrotransposable element is a "copy and paste" element. Thus, when moving, the number of retrotransposons increases from one copy to two, while a DNA transposon like a McClintock "controlling element" remains a single copy. By 1987, in the back of the mind of everyone in the repetitive DNA field was the thought that L1 was a retrotransposable element. Thus, it was very important that there be a reasonable chance that some L1s encode a reverse transcriptase that could help them to mobilize. This was the state of the mammalian transposable element field in 1987.

This page intentionally left blank

8

"Welcome to the wonderful world of LINEs"

Now for a change in the cast of characters and a description of the convoluted path that I took to enter the mobile DNA field. Don't worry, I don't plan to go all the way back, but I'll give you a brief outline of my path beginning with college.

I attended Dartmouth College as a premed major, and although I wasn't sure about becoming a physician, after three years in the college, I entered Dartmouth Medical School. Because I hadn't had any experience with sick people, I thought I needed to work in the hospital as an orderly, so I went to the Dean for advice. He noted that I had done well in organic chemistry and told me to speak to a biochemist, Lafayette Noda, about doing research. Noda was just setting up his lab and was eager for help, so I worked with him one summer on creatine kinase enzymology. I enjoyed this first taste of research work, so the next two summers, I worked with another biochemist, Lucille Smith, on electron transport in bacteria. Dartmouth Medical School was a two-year school at the time, so I transferred for the clinical years to Johns Hopkins in Baltimore. After medical school, I trained in pediatrics for two years at the University of Minnesota Hospitals. While at Minnesota, I decided that I wanted to be a medical geneticist, so I asked Barton Childs, with whom I had taken a seminar elective in genetics at Hopkins, for advice. He gave me a few options of other places, but he also suggested returning to Hopkins for a fellowship. Because I admired Childs greatly and he had recently published a landmark paper in PNAS confirming Mary Lyon's hypothesis of X chromosome inactivation in humans, I decided to return to Baltimore. For the next 20 months, I worked with Childs and Bill Young on

dosage compensation in Drosophila with some success. But this was the time of escalating U.S. involvement in Vietnam, and the doctor draft loomed large. I was fortunate to find a position in the U.S. Public Health Service with Harvey Itano at NIH. To obtain this position, I was greatly assisted by a reference from Lafayette Noda to Itano, a fellow Japanese-American who had also been interred during World War II. With Itano, I worked on regulation of human hemoglobin synthesis. In 1968, Bob Cooke, the Chair of Pediatrics at Hopkins, offered me a faculty position, which I took in 1969 after another year of pediatric training at Hopkins. From my NIH days through my first 20 years on the faculty at Johns Hopkins, my lab worked on globin synthesis in the thalassemia syndromes, severe anemias common in many parts of the world, but best studied at that time in Italians and Greeks.

Now I need to introduce another key player in my story. For that introduction, we go back to early 1980, when a young Greek national named Stylianos Antonarakis, wrote me a handwritten letter from Athens. He had finished medical school at the University of Athens and done clinical training and army service and was still only 26 years old. He seemed eager and bright (he stated that he had had the top score among applicants to medical school at the University of Athens), but I had no money to pay a foreign applicant for a postdoctoral position in the lab. (Foreign postdocs without permanent U.S. residency status are ineligible for NIH funding.) A month passed, then six weeks, and I had not answered his letter. I then received a call from a Greek-American, George Stamatoyannopoulos, who was a respected colleague of mine in the hemoglobin field. George urged me to find a way to hire Antonarakis as a postdoc, saying that Antonarakis was highly motivated and desperately wanted to train at Johns Hopkins. So I acceded to George's request and was never disappointed. Stelios, as he was called, was exceedingly hard working, eager to do research and stretch his mind, and as brilliant as advertised. On top of all that, he was willing to work without pay, and his productivity was phenomenal.

With some effort on his part and mine, we were able to find him some money. In those days, a postdoc made about $15,000 per year. Stelios went to the Greek Orthodox Church in Baltimore and got the church to donate $5,000 toward his salary. By letter, I solicited a wealthy Armenian-American industrialist from Detroit whose family had been good friends of my family for many years. On Christmas

Day 1980, as we were celebrating the holiday at my uncle's home in Detroit, the industrialist called to say he would give another $5,000 for Stelios. So Stelios did have some outside income, and the Greek Church continued its support.

Stelios and I decided that he would search for new DNA polymorphisms (changes in DNA sequence present in some human beings but not in others) in a cluster of genes, the β-globin gene cluster. To be a true polymorphism, the change or mutation would be common in the population, having a gene frequency >.01 or 1%. The globin genes encoded proteins that were part of hemoglobin at various stages of life, embryonic, fetal, and adult. In short order, Antonarakis was quite successful. He found a number of these polymorphisms, and before long we had discovered all kinds of interesting things about recombination in the human genome. Recombination didn't occur equally everywhere, but there were "cold spots" and "hotspots." There were DNA polymorphisms that associated with each other (haplotypes). These were cold spots. Then there were regions of the genome in which DNA polymorphisms were close together but not associated with each other. These were hotspots for recombination (Antonarakis et al., 1982).

Then, in collaboration with Stuart Orkin, we successfully used the haplotypes to predict the occurrence of undiscovered mutations in a common hemoglobin disease called β-thalassemia (Orkin et al., 1982). Whenever a β-globin gene cluster containing a β-thalassemia gene had a novel haplotype, there was an excellent chance that the thalassemia mutation in the β-globin gene was novel and previously unknown. (We went on to characterize β-thalassemia mutations using this technique first in Mediterranean peoples, Asian Indians, Chinese, and African-Americans. (See Orkin and Kazazian, 1984 for a review.) Later, we successfully used the technique in Egyptians, other populations from India, Kurdish Jews, Mestizo Mexicans, among others. Antonarakis was so productive that after three postdoctoral years he had published some 30 papers. In 1983, he joined the faculty at Hopkins in Pediatric Genetics. Antonarakis is now Chair of the Department of Genetic Medicine and Development at the University of Geneva in Switzerland.

Meanwhile, in October 1983 at the American Society of Human Genetics meeting in Norfolk, Virginia, I had engaged in a telling lunchtime conversation with Stuart Orkin. We both saw the writing

on the wall for characterization of β-thalassemia mutations. It would continue, but without many more surprises and major hurdles. The work would quickly become humdrum. We needed to find fresh big problems to bite into. Stu had likely found his already in hemopoietic transcription factors, beginning with GATA-1. I wanted an open, new, and interesting field to pursue, but I needed to be patient. Indeed, it would come but not until the summer of 1987 from an unexpected source—mobile DNA.

In 1984, Antonarakis and I decided that it was a good time to move on from the study of hemoglobin genes to a new project. The hemoglobin genes were unusual. They were tiny and simple, and many mutations or changes that occurred in the DNA of these genes were favorably selected in many parts of the world. Falciparum malaria is a very common human disease in many regions of the world, including throughout the Mediterranean basin (Spain, Italy, Greece, N. Africa), Sub-Saharan Africa, the Middle East, India, S.E. Asia, South China, and Indonesia. The parasite that causes falciparum malaria, *Plasmodium falciparum*, spends a portion of its life cycle in the human red blood cell. Often when the hemoglobin in the red cell is abnormal, the red cell becomes a relatively inhospitable environment for the malarial parasite. Many mutations of the hemoglobin genes, such as those that cause sickle cell anemia and the various thalassemias, alter the hemoglobin quality or quantity within the red cell and thus are protective against falciparum malaria. In those regions of the world that are endemic for malaria, individuals carrying these mutations, that is, having one normal and one mutant gene at the β-globin locus, reproduce better than those carrying normal hemoglobin genes. Because of increased reproduction in carriers of many mutant hemoglobin genes, these mutant genes increase in frequency in malarial regions. They are said to be under positive selection, and their gene frequencies are high, from .01 to .10, meaning that in some populations 10% or more of β-globin genes are abnormal. Positive selection makes the hemoglobin genes unusual and special.

We had been characterizing mutations in the small hemoglobin genes that were under positive selection. Now we wanted to characterize mutations in a contrasting gene, one that was very large, located on the X chromosome (all males with a single X chromosome would show the disease if a deleterious mutation were present), and

not under positive selection. This kind of gene should give us a full, unbiased view of the whole spectrum of human mutations. In 1984, the ideal candidate gene, factor VIII, was cloned and characterized. Factor VIII is the gene that is mutated in hemophilia A, by far the most common hemophilia, affecting 1 in every 5,000 males in all parts of the world.

In 1935, the noted population and statistical geneticist J.B.S. Haldane had postulated that since hemophilia A was a genetic lethal (affected males did not reproduce) and the incidence of the disease was not changing worldwide, that each mutant hemophilia gene would be lost from the population in roughly three generations and would be replaced by ongoing mutation to new hemophilia genes (Haldane, 1935). He also predicted that nearly every unrelated affected male would have a different mutation unless particular mutations tended to recur. (Indeed, it turned out that a single mutation did tend to recur. In 1993, we learned that an inversion triggered by mispairing of nearly identical ~10 kb sequences located both ~500 kb upstream of the factor VIII gene, and within an intron in the gene followed by crossing over recurs frequently in male meiosis. This inversion mutation accounts for almost 50% of severe hemophilia A cases. So Haldane did miss on his prediction.)

Based on Haldane's ideas and the availability of the factor VIII gene for analysis in hemophilia A patients, Antonarakis and I decided to characterize mutations in this gene. We set up a meeting with the factor VIII gene cloners at the Genetics Institute, a biotech company in Cambridge, Massachusetts. With the help of Antonarakis' Greek connection with Tom Maniatis, one of the company's founders, we got the factor VIII gene probes we needed. Antonarakis then began to collect blood samples on a large number of hemophilia A patients, mainly from his connections in Greece and from Carol Kasper, a prominent hemophilia doctor in Los Angeles.

Then another character entered the scene. His name was Hagop Youssoufian. I had known Youssoufian since the spring of 1982, but I did not know that his story and that of my family were interconnected. Youssoufian was a medical student at the University of Massachusetts who had taken a year out of medical school to do research, spending the summer in my lab at Johns Hopkins. I was impressed with his work

that summer. After he left, he kept me informed of his training activities. In July, 1983, he had begun residency training in Internal Medicine at the Cleveland Clinic. I knew he might be looking for a postdoctoral fellowship, perhaps as early as July, 1985. In the fall of 1984, I wrote him and asked him to join us at Hopkins for Genetics Fellowship Training. I told him that I had had outstanding trainees every 5 years at Hopkins, John Phillips in 1975 and Stylianos Antonarakis in 1980, and now it was his turn to star. He wrote back that he was planning to train with Art Nienhuis at the NIH beginning in July, 1985. I shot back that he should reconsider his decision and come to Hopkins instead where he could get genetics training and finish a third year of Internal Medicine training all in two years. After this exchange, Youssoufian was convinced and started genetics training and work in the lab in July 1985. From July 1985, until July 1986, he also did enough Internal Medicine training to get a year's credit. From July 1986, until July 1987, he did his Clinical Genetics training so that after those two years Youssoufian was able to complete his boards in both Internal Medicine and Clinical Genetics. On top of all that, his research flourished, and he published ten papers in those two years.

In 1985, I learned his family's story. The Youssoufians lived in Kayseri, Turkey, my father's hometown, at the time of World War I. My father's family also lived in Kayseri until April, 1915. On the night of April 24, 1915, my grandfather, an import-export merchant, was taken from his home, imprisoned, and not seen by his family again. The next day, my father, his mother, maternal grandmother, and three siblings, along with the rest of the Armenian community of Kayseri, were put on a forced march and later taken by train to a concentration camp in far eastern Turkey. Over the next two years, most inmates in the camp, including all members of my father's family except him, perished of typhus. He contracted the disease twice but luckily survived. He later escaped from the camp, and after a number of attempts to enter the United States, he finally arrived at Ellis Island in 1923. This was my family's involvement in the Armenian genocide. I'm sure that the Youssoufian family had a similar fate, though we never discussed it. After World War I, the surviving Youssoufians settled in Aleppo, Syria, where Hagop was born. At age 12, he and his immediate family had immigrated to the U.S. and settled in the Armenian community of Watertown, MA. In late 1985, I asked

my father, then age 85, whether he remembered the Youssoufian family from Kayseri. Indeed, he had walked past the Youssoufian home on his way to school. In addition, he distinctly remembered that one adult member of the family (an uncle of Hagop) was hung in the town square by the Turkish authorities for revolutionary activities. Because the Youssoufian's and the Kazazian's of Kayseri had a shared experience in the Armenian genocide of 1915, I felt a special connection to Hagop Youssoufian.

As you can tell from his productivity, Youssoufian did not disappoint. He was hard working and ambitious, a typical immigrant eager to achieve success in America. He was very bright and already knew most molecular biology techniques available at the time because he had taken the year off from medical school to do molecular biology research. He began to study mutations in the factor VIII gene in the blood samples collected from hemophilia A patients. In the end, mostly through the efforts of Antonarakis and because we received patient samples for gene diagnosis of hemophilia A by use of DNA polymorphisms in and near the FVIII gene, we had 240 patients to analyze. Antonarakis gave them each a number with the prefix JH for Johns Hopkins.

Youssoufian used a technique called Southern blotting, named for Ed Southern, its inventor. He cut whole genomic DNA into fragments of a size that can readily be separated by electrophoresis (1,000–20,000 nucleotide pairs in length), and after gel electrophoresis, he hybridized these fragments as single DNA strands to radioactive factor VIII gene probes. He then visualized all the fragments closely related in sequence to the factor VIII gene from among the millions of fragments in genomic DNA. When a fragment from the factor VIII gene was abnormal in size, it meant that the disease-causing mutation was likely in the gene region of that fragment. But while Southern blotting is excellent for finding DNA rearrangements, such as deletions, it is inefficient at finding single nucleotide substitutions, such as an A for a G. Moreover, most mutations are single nucleotide substitutions and not DNA rearrangements. However, Youssoufian did find about 10% of the mutations in factor VIII-deficient patients, including a few nucleotide substitutions and a number of deletions. Then in May 1987, we had the surprise results that pulled me into the mobile DNA field.

The DNA of two patients had factor VIII fragments of unusual sizes. When Youssoufian mapped them to the factor FVIII gene, it appeared that these patients did not have deletions, but instead had extra DNA in their factor VIII genes (see Figures 8.1 and 8.2). Perhaps they had insertions. The patients were JH-27 and JH-28, the 27th and 28th patients in our collection at Johns Hopkins. Interestingly, both insertions had occurred on X chromosomes in the last

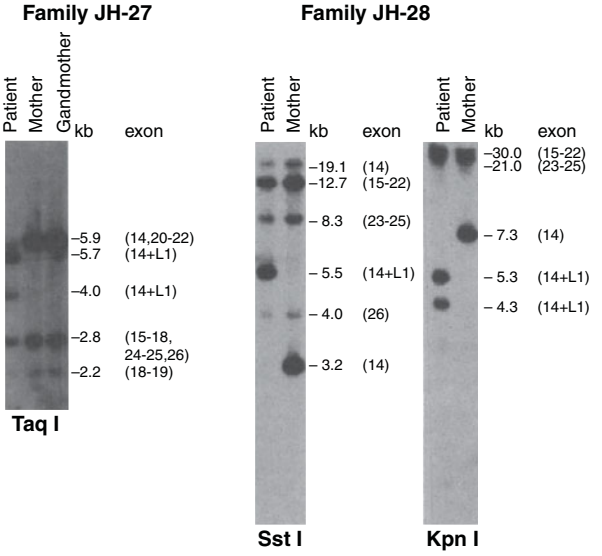


Figure 8.1 Restriction endonuclease digests and Southern blots of portions of the factor VIII gene from families JH-27 and JH-28. On the left is a TaqI digest of JH-27 patient, mother, and maternal grandmother hybridized with a 3' FVIII cDNA probe spanning exons 14–26. The cDNA probe is a copy of a portion of the FVIII mRNA. The patient is lacking a 5.9kb band but contains new 5.7 and 4.0kb bands. The mother and grandmother both have the normal 5.9kb band and lack the abnormal 5.7 and 4.0kb bands. The minor band at 5.9kb in the patient is derived from exons 20–22. Further analysis showed that an insertion of the 3' roughly 3.8kb of an L1 element accounted for the abnormal bands. Two bands are expected since the L1 fragment contains a TaqI site near its 5' end. On the right are SstI and KpnI digests of patient JH-28 and his mother hybridized with the same FVIII cDNA probe. In patient JH-28, the 3.2kb SstI band is replaced by a 5.5kb band, not seen in the mother. In the KpnI digest, the normal 7.3kb band is replaced by 5.3 and 4.3kb fragments in patient JH-28. Again, the abnormal bands are not seen in his mother. Further analysis showed that roughly 2.2kb of the 3' end of an L1 was inserted into exon14 of the FVIII gene in patient JH-28.

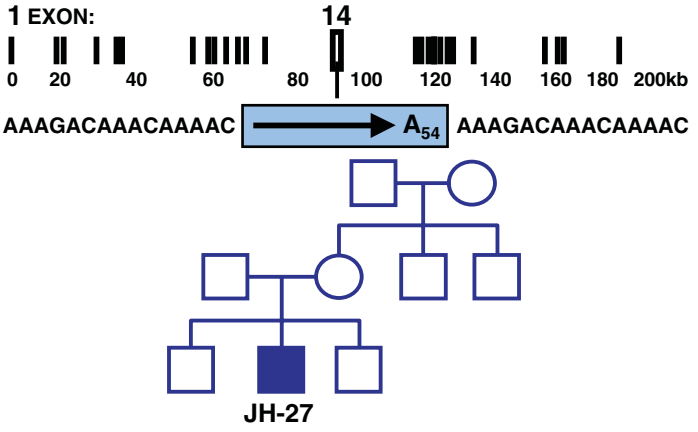


Figure 8.2 Schematic drawing of the L1 insertion into exon14 of the FVIII gene of patient JH-27. The pedigree shows that the mother does not carry the insertion, so it must have occurred either in one of her germ cells or early in embryonic development of JH-27. The insertion was discovered by Hagop Youssoufian.

generation, that is, neither mother had the insertion on either of her two X chromosomes (see the pedigree of JH-27 in Figure 8.2). So these insertions were brand new, either occurring in a developing egg of the mother or early in embryonic development of the children. One abnormal band from patient JH-27 looked promising for cloning a small fragment. However, Youssoufian had decided to leave the lab in July, 1987, for a second postdoc at MIT and Harvard. He only had one month to go. He pleaded with me to let him clone that abnormal DNA fragment, and I agreed. Amazingly, within a week, he had succeeded! One Monday morning in June 1987, I entered the lab to find Hagop exuberant. "I got the clone," he exclaimed. "I ran a little out on a gel (gel electrophoresis), got Alu (another repetitive DNA) and L1 probes from Alan (Scott), and look at this. It's an L1 element." Indeed, this *was* exciting! A piece of repetitive DNA suspected of being a transposable element had inserted into the coding region of a gene and likely caused hemophilia A in patient JH-27. Not only that, but it happened in the last generation. The insertion was less than one generation (25 years) old. I thought to myself, "This is what I'm going to work on from now on. I'm going to change the direction of the lab to work on mobile DNA."

Although this was the first discovered human mobile DNA insertion, I should not have been surprised. Mobile DNA insertions had been found in bacteria, maize, yeast, fruit flies, and many other organisms. We knew that there were a large number of potential mobile elements in the human and mouse genomes. Of course, we weren't sure of their identity and whether they were indeed mobile elements. We also didn't know if indeed they were mobile elements, whether they were all dead for mobility or whether some were still active in mammalian genomes. Now in 2010, we know that mobile element insertions still account for a small fraction of human mutations and about 10% of mouse mutations. In humans, there are ~20 known insertions of L1s, ~40 known insertions of Alus, and 8 known insertions of SVA elements, or a total of ~70 known disease-producing insertions.

Because Youssoufian was leaving for Cambridge Massachusetts, a new person had to pick up the L1 project. Luckily, an outstanding graduate student named Corinne Wong was eager to do so. Corinne was a Chinese-American who had entered our human genetics graduate program at Hopkins in 1982 after finishing undergraduate work at Wellesley. She was nearly finished characterizing an unusual chromosome abnormality for her Ph.D. project, but she had the ability to do two or three interesting projects at the same time. She quickly cloned the complete insertion from JH-27. Then she cloned the insertion from JH-28 that also turned out to be a portion of an L1 element. Then Wong set out to sequence the two insertions that were roughly 3,800 (3.8 kb) and 2,200 (2.2 kb) nucleotides in size. In those days, DNA sequencing was done by hand using a procedure devised by Fred Sanger at the Medical Research Council (MRC) labs in Cambridge, England. Painstakingly, Wong sequenced the two insertions over about six weeks, and the sequences turned out to be very interesting (Figure 8.3).

First, the 3.8 kb JH-27 insertion stretched from the tail or 3' end of L1 almost to the beginning of the 4,000 nucleotides of ORF2 that had been postulated to encode a reverse transcriptase. Importantly, as far as it went, the ORF was intact and could be translated into a large protein. It did not contain any missing or added nucleotides or nucleotide substitutions that would stop protein production. That was big news because all of the human L1 elements characterized by Alan Scott and others up to that point could not be translated into protein.

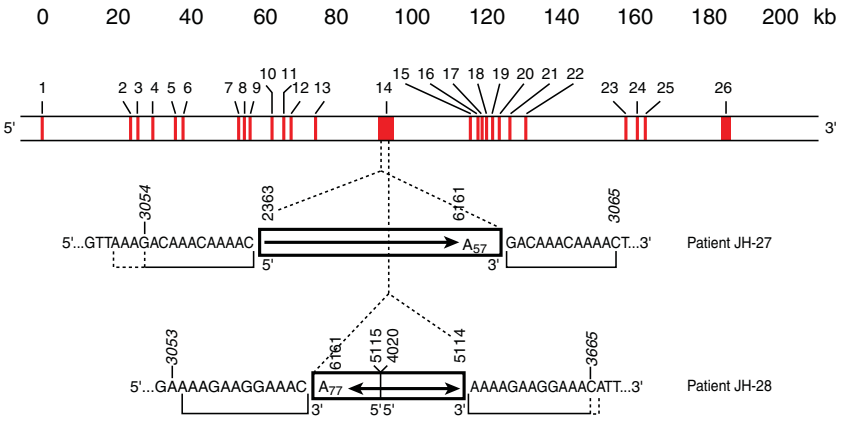


Figure 8.3 L1 insertions into exon14 of the FVIII gene in patients JH-27 and JH-28. In patient JH-27, the insertion is the 3' 3.8kb of L1, while in patient JH-28, the insertion includes 2.2kb at the 3' end of L1, but from roughly nucleotide 3800 to nucleotide 5000, the L1 is inverted with no gain or loss of nucleotides at the inversion site. Both insertions have long poly A tails of 57 and 77 nucleotides and target site duplications of 15 and 12 nucleotides, respectively.

The sequence of JH-28 was interesting because it contained an inversion. From the 3' end, the sequence proceeded from nucleotide 6000 to roughly nucleotide 5000, and then it started up again at roughly nucleotide 3800 and proceeded to nucleotide 5000. Thus, the front or 5' half of the L1 sequence was inverted. Last, while the sequences of the protein-coding regions were similar to Scott's consensus sequence, the last 200 nucleotides that Scott called the 3' untranslated region diverged significantly. This sequence divergence of a short region outside of the protein-coding region was very puzzling. At this point, I needed help!

Who else to call but Maxine Singer, the L1 guru? When I told Maxine about the sequences, she immediately came to the rescue. First, the inverted sequence in JH-28 had been observed often in human L1s, perhaps in 25% of the elements. Second, the 200 nucleotides at the 3' end of the two insertions were reminiscent of the Ta subset of expressed L1s that she had reported a year earlier at the 1986 Cold Spring Harbor meeting. Singer pointed out that her four Ta subset L1s all had a telltale trinucleotide, ACA, replacing GTG, about 90 nucleotides from their 3' end, and a telltale G replacing an A

about 10 nucleotides from that end. Indeed, Maxine had hit the nail on the head! Upon closer inspection, the JH-27 insertion was a Ta element with the telltale ACA and G nucleotides, and the JH-28 insertion was a Ta variant (called pre-Ta) with ACG and G at the key sites. Thus, not only did we have two new L1 insertions, but also the insertions came from a special family of L1s that was relatively rare in the genome, yet was commonly expressed into RNA (see Figure 8.4). This special Ta subfamily of L1 still has importance in the field. Tony Furano and Stephane Boissinot have shown that this subfamily of L1s makes up the only active L1 subfamily today. It has existed for about 2–3 million years, and it appears to be increasing in size at the present time.

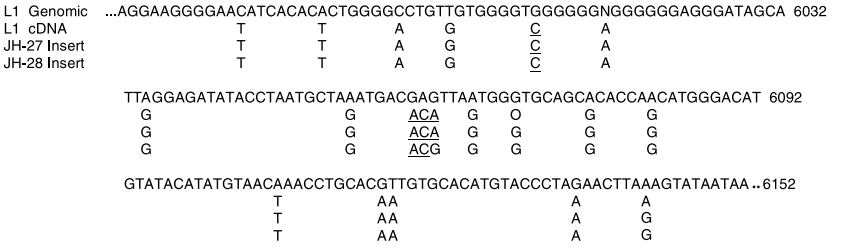


Figure 8.4 3' untranslated regions (UTRs) of the JH-27 and JH28 L1 insertions. The 206 nucleotides of the insertions differ at 20 nucleotide positions from the Scott consensus. However, they are very similar in sequence to the Ta cDNA sequences found by Skowronski and Singer (Skowronski et al., 1988). The ACA trinucleotide roughly 90 nucleotides from the 3' end and the G nucleotide 9 nucleotides from the 3' end were later used to mark human-specific L1 elements.

Maxine Singer then sent me her paper on the expressed LINE-1s with the greeting on the top of the title page, "Welcome to the wonderful world of LINEs." Coming from her, that was a very exciting welcome to the field!

I then quickly wrote a draft of our paper and circulated it to Antonarakis, Wong, Scott, and others. After some editing, we sent it off to the journal *Nature*. The peer reviewers liked the paper, and it was published in early 1988 (Kazazian et al., 1988).

Youssoufian had also found a third patient with an L1 insertion in his factor VIII gene, but this insertion was located in an intron, not in

sequence that coded for factor VIII protein. This fact made it questionable that it had caused the disease. In addition, it had been inherited from an X chromosome of the mother. If the deceased maternal grandfather also had the insertion but didn't have hemophilia, then the insertion did not cause the disease. How to find out whether the deceased grandfather had the insertion? The 90-year-old great-grandmother of the patient and mother of the deceased grandfather was still alive. I telephoned the great-grandmother and learned that she was willing to provide a blood sample for analysis. After attending a meeting in Manhattan, I drove about 100 miles up the New York State Thruway to the great-grandmother's house. She was very hospitable, and I obtained a small blood sample. We later found the insertion in her DNA, making it highly likely that the grandfather also had the insertion, so in this case, the L1 insertion was not the cause of hemophilia A in the patient (Woods-Samuels et al., 1989). However, because the L1 was present in some individuals but not others and did not cause hemophilia A, it was the first example of a non-disease producing, dimorphic L1 insertion.

In 1988, I was a rank neophyte when it came to mobile DNA and transposable elements. However, I really wanted to enter the field. It was very interesting and of fundamental importance to biology in general. I knew that if I were going to contribute to the field, I had to learn about it quickly. What better way to acquire the necessary knowledge than from Singer herself! I proposed quarterly lab meetings to Maxine, and she agreed. The Singer lab at NIH wasn't far from the Hopkins medical campus—only about 40 miles. So for the next six years, Singer and I had quarterly meetings rotating between Hopkins and NIH. When she became president of the Carnegie Institutions and I moved to the University of Pennsylvania, we met exclusively at the Carnegie Institution's Department of Embryology labs on the Homewood campus of Johns Hopkins University. These meetings were very productive. Singer's lab was interested in the biochemistry of L1s, and mine was concentrating on the genetic aspects of these elements, so there was little or no overlap of our interests.

But what was the next step in my L1 research? What direction should I take? Then a good friend, Larry Shapiro, a respected human geneticist at UCLA and presently Dean at Washington University School of Medicine in St. Louis, asked me to visit UCLA and present a seminar. Although Los Angeles is a long way from Baltimore, I

agreed because I thought I might learn from the excellent faculty that I knew there. After the seminar, while taking me to my next appointment, Shapiro mentioned that if we could isolate the precursor full-length (6kb) L1 of one of our insertions that would likely allow us to better study the biology of these mobile DNAs. I had been contemplating this very possibility, and now Shapiro had crystallized it for me. Could we possibly succeed in isolating a precursor from among the 100,000 or more L1s in the human genome? It would be challenging, but perhaps it was possible. Moreover, if we succeeded, it would be a major advance. I decided to give it a shot. I had just received a 7-year MERIT award for work on hemoglobin genes from the NIH. Why not redirect that grant to L1 work and attempt to isolate an active transposable element, the precursor of the JH-27 or JH-28 insertion? (Note that the NIH granting system allows considerable freedom to change a project's focus. However, in order to stay competitive in peer review, I'd have to demonstrate success in isolating an active L1 within 1–2 years.)

Then Corinne Wong received her Ph.D. degree and entered an accelerated medical school program at the University of Miami. Her departure for medical school disappointed me greatly because I thought she had significant scientific talent and an impressive ability to get experiments to work. I wanted her to stay in research and become a principal investigator. However, her Chinese-American parents were telling her that an MD is a real doctor and her Ph.D. only made her a second-class doctor. They wanted her to get an MD like her sister to become a first-class doctor. (In my opinion, the Ph.D. degree is worth at least as much as the MD. The Ph.D. degree teaches one how to do research and how to think critically. The MD degree teaches many things, including some critical thinking, but not how to do research.) So Corinne went to medical school, and Beth Dombroski entered the lab. Beth's timing couldn't have been better. I was ready to undertake the project to isolate an active, full-length L1, and Beth had the perfect temperament to take it on.

9

An experimental breakthrough

Beth Dombroski was an attractive, dark-haired young lady who had recently finished her Ph.D. in the Chemistry department at Johns Hopkins University with Tom Tullius. Her first paper had been a first-author paper in *Science* on a new method of DNA "foot-printing," discovering where on a DNA molecule proteins interacted. She came from Reading, Pennsylvania, and had done her undergraduate degree at Shippensburg, a small, highly regarded, liberal arts university in Eastern Pennsylvania. She was eager to work on the biology of L1.

However, we still needed a hook to get into the problem. How would we separate a small group of L1 elements that contained the precursor to either the JH-27 or JH-28 element from the remaining very large number of irrelevant L1s? In other words, we needed a way to find the proverbial needle in a haystack. I knew of one reasonable possibility. Over the previous four years, the lab had used a technique devised by Bruce Wallace at City of Hope in which one could carry out a Southern blot using a very short, labeled probe (Itakura et al., 1984). Previously, Southern blot probes were 500 nucleotides or longer, but Wallace had shown that one could find in genomic DNA any exact match for a short probe of 18–20 nucleotides in length under the right conditions. These probes were called oligonucleotides, or oligomers, because they contained a relatively small number of nucleotides (but enough so that they would hybridize specifically). Moreover, they could be synthesized for any desired nucleotide sequence in special core labs. From 1983 to 1987, we had perfected Wallace's technique of hybridizing an oligonucleotide probe to DNA fragments in a dried agarose gel. By this time, we knew the conditions that would allow hybridization of a 20-nucleotide

oligomer to only its exact match in genomic DNA. If the match were 19 nucleotides out of 20, no hybridization signal would result.

The next step was to determine whether within either the JH-27 or JH-28 insertion sequence there was a sequence that might be rare among human L1 sequences. Now Scott's consensus sequence became invaluable! The entire 6kb consensus sequence with all the component L1s that were used to make it was circulating in the lab on five or six pieces of paper scotch-taped together. (We didn't have computers at this time.) Because there was more sequence available for the JH-27 insertion (3.8kb) than for the JH-28 insertion, I first compared it to the consensus sequence. *Voila!* There was a 20-nucleotide stretch roughly in the center of the 3.8kb insertion sequence that had three nucleotide substitutions from the consensus sequence. After a conference with Dombroski and Scott, we decided that it was worth the gamble to have the core synthesize a 20-nucleotide oligomer (called a 20-mer for short) with the three changes from consensus sequence.

Dombroski then radioactively labeled one end of the 20-mer and carried out the experiment. She digested the DNA from the JH-27 patient, his mother, his father, and two other individuals and ran the five digested DNAs in separate lanes in a gel electrophoresis. She then hybridized the short, radioactively-labeled 20-mer probe to the dried gel in solution, washed off the unhybridized probe, and put it up against an X-ray film to visualize the radioactivity (Figure 9.1). Her result was amazing! Instead of a smear of thousands of fragments of DNA hybridizing to our short probe, we saw only a handful of DNA fragments, and most gratifyingly the patient had a unique fragment that was not seen in either of his parents or the two controls. This fragment represented the L1 insertion that was present only in the patient and not in his parents. JH-27 was about to become the toast of our lab because this experiment told us that there was a reasonable chance that we could isolate the precursor of his insertion.

Who was JH-27? In the ensuing years, I occasionally communicated with his parents and his doctor, Dr. Donna DiMichele at Cornell Medical Center. Luckily for us, his parents were very cooperative in providing both samples and information about their son. Unluckily, JH-27 was born with hemophilia A in 1980—absolutely the worst time to be born with hemophilia A. In the early 1980s, the

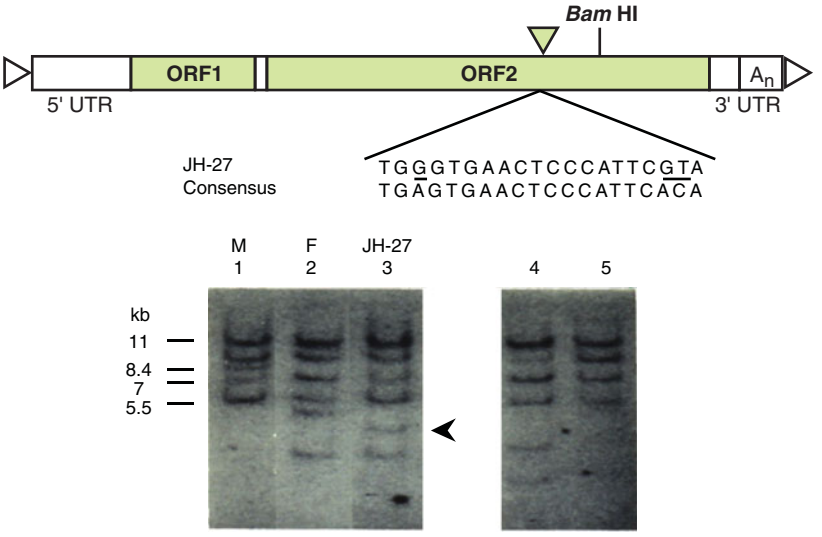


Figure 9.1 An oligonucleotide detects the L1 insertion in JH-27 on Southern blot. What is shown here is the human L1 with the marked 20 nucleotide region at roughly nucleotide 4000 in the 6000 nucleotide human L1. The JH-27 insertion sequence differs from consensus in this region by the three underlined nucleotides. After agarose gel electrophoresis of BamHI digested genomic DNA of members of the JH-27 family (lanes 1–3) and two controls (lanes 4 and 5), the radioactively labeled JH-27 oligonucleotide was hybridized to the gel in solution. The gel was washed and placed against X-ray film. Instead of a smear of many L1 fragments, only a handful of bands are seen in each individual. The JH-27 patient has a new band (arrow) not present in either parent. This new band represents his L1 insertion.

plasma supply used to treat hemophiliacs was contaminated with human immunodeficiency virus (HIV). Sadly, JH-27 contracted HIV-AIDS by age 3, yet he continued to do well until his late teens. His mother sent me a picture taken of his family at his Bar Mitzvah when he was age 13 (Figure 9.2). At the time he was a budding actor, and at age 14, he played a leading role in a feature movie that was critically well received. However, by age 19, AIDS was taking its toll. He became very depressed, estranged from his parents, and died in his early 20s, ending a very depressing story for a promising, but unlucky, young man.

Now I return to the effort to isolate the precursor of JH-27's insertion. From the DNA fragment sizes on Dombroski's gel, we

Figure 9.2 Patient JH-27 shown with his parents at his Bar Mitzvah.

could tell that there were no more than four full-length L1s hybridizing to the oligomer that we now named JH-27. We thought that the precursor of the JH-27 insertion should be among those potentially four full-length L1s. So the next step in finding the precursor was to clone the four full-length 6kb L1s to find the one that contained the exact sequence of the insertion over the 3.8kb of the insertion. We knew from Scott's work that the average L1 differed from other genomic L1s by about 5%. So an exact match over 3.8kb was going to be a tall order.

For biological reasons, it was an even taller order than we realized! We were asking that the precursor be an exact match to the insertion. In other words, we were expecting the proteins that were potentially synthesized from the RNA of the precursor would act on that same RNA to retrotranspose it to a new genomic site. This effect is called *cis*-preference in contrast to the situation in which proteins made from one L1 would come off the ribosomes and act to retrotranspose another L1. Under this latter scenario of *trans*-preference, the precursor could have been any L1 that had the ability to encode the necessary

proteins. In this latter case, the full-length element that was the actual precursor could have a sequence very different from the insertion sequence. In fact, for nearly every virus or other transposable element known, the retrotransposition mechanism involves *trans*-preference, not *cis*-preference. Thus, the information in the scientific literature at the time suggested that it was unlikely that we'd find a precursor to the insertion with exactly the same sequence as the insertion over its length. However, I was ignorant of this fact! If I had been better informed, I might not have done the experiment. On the other hand, this was the best means at the time to isolate a potentially active human transposable element, so I may have considered the reward worth the risk. After all, at the time, I was already in mid-career and had much less to lose than an early career investigator. My NIH support was pretty much guaranteed for seven years on the MERIT award given for hemoglobin work.

So we went ahead. Beth Dombroski obtained a bacteriophage lambda library from Clontech to attempt the cloning of the potentially four full-length L1s that had the sequence of the JH-27 oligomer. A phage library is one that contains essentially all of the human genome cut into fragments of 9–23kb with each fragment present in a different phage that is able to hold that much extra DNA. Thus, the library contains millions of different phage, each one with a different human DNA fragment. Beth spread the phage on agar plates that contained bacteria within which the phage could replicate, aiming for about 50,000 phage/plate. Under the appropriate conditions, each bacterium would contain only one type of phage. She then hybridized the human DNA-containing phage with the specific JH-27 oligomer. In order to find full-length L1s containing the JH-27 sequence, she also hybridized the phage with an oligomer corresponding to a common sequence at the 5' end of the L1. She then picked phage colonies that hybridized with both oligomers and purified them by further plating, growing, and hybridizing. In this first cloning experiment, Dombroski isolated two L1s that hybridized with both the JH-27 and the 5' end oligomer. Beth then proceeded to sequence these two L1s using non-automated DNA sequencing. The first, called L1.1, was ruled out rather quickly as the precursor because it had a critical change in sequence that prevented it from encoding an ORF1 protein. The second full-length L1, later called L1.2A, looked

much more promising. Progressing from the 5' end toward the 3' end, Beth's sequence was identical to the JH-27 insertion sequence. However, when Dombroski approached the 3' end of the sequence from nucleotides 5500 to 6020, she found two nucleotide substitutions from the JH-27 insertion sequence. So it was close, but no cigar!

My immediate first thought when I saw the two altered nucleotides on the sequencing gel was perhaps there are sequences of this L1 in the population that contain a few nucleotide changes. We call two sequences that sit at the same position in the genome but differ from each other, *alleles* at the *locus* of interest. Perhaps the Clontech library sequence of this particular L1 was an allele of the L1 that retrotransposed in JH-27. Of course, I was again playing the optimist, but we had analyzed the sequence of this L1 from a commercial library, not from the parents of the patient. All of my previous genetic training told me that we needed to study this specific L1 in the parents of JH-27. That was the next step—to look at the sequence of this specific L1 between nucleotides 5500 and 6020 in the parents' DNA. Luckily, the parents were willing to provide blood samples. Beth isolated DNA from their lymphocytes and did a PCR reaction using the specific JH-27 oligomer as one primer, the forward primer, and a primer from downstream of the L1 as a second or reverse primer. (Dombroski had obtained some DNA sequence downstream of L1.2A from her L1.2A genomic clone.) We obtained clean products from this PCR on both mother's and father's DNA, and upon sequencing the key region we found that both changes were absent. Both parents had the sequence of the JH-27 insertion at the key nucleotide positions! In fact, both parents appeared to be homozygous for the allele corresponding to the JH-27 insertion allele. Now we were convinced that this L1 from one of the parents was indeed the precursor of the insertion in the patient and was an active human transposable element. However, we still needed to clone it out from a bacteriophage library of one of the parents and sequence that clone.

Dombroski chose the mother's DNA, cloned the specific L1, and our newly developed DNA sequencing core facility carried out the DNA sequencing. This time, as predicted, the sequence was an exact match over the entire 3.8kb with the JH-27 insertion. In addition, both ORF1 and ORF2 were intact and capable of encoding proteins. Now I was sure we had our active precursor element (Figure 9.3).

Then we asked Marcia Budarf at UPenn to find the chromosomal location of this L1. Marcia told us that this L1, that we now called L1.2B, mapped to chromosome 22. The element sitting in chromosome 22 had been expressed into RNA, used its own proteins for its reverse transcription, and a 3.8kb portion of it had inserted into the factor VIII gene, disrupted it, and caused hemophilia A in JH-27. We wrote up the paper and sent it to *Nature*, but as so often happens for unknown reasons, it was returned without review. How disappointing!

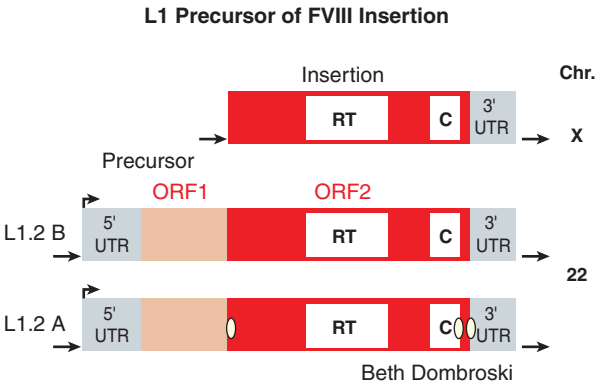


Figure 9.3 Isolation of the precursor full-length L1 of the insertion in JH-27's FVIII gene on his X chromosome. The JH-27 oligonucleotide was hybridized to bacteriophage libraries, and two potential precursor L1s were isolated that were alleles at a locus on chromosome 22. L1.2B has the exact sequence of the insertion over the insertion's 3.8kb. L1.2A has a nucleotide change (white oval) at the 5' end of ORF2 that is outside of the insertion sequence. L1.2A also has two nucleotide changes that alter amino acids at the 3' end of ORF2. We later learned that these two changes reduced the retrotransposition of L1.2A to ~10% that of L1.2B in the cell culture assay (Chapter 13). RT is the region of ORF2 that encodes the reverse transcriptase, and C is an evolutionarily conserved region that is critical for retrotransposition but whose function is still unknown.

This page intentionally left blank

10

Reverse transcriptase to the rescue

A few months after Dombroski had isolated L1.2A but before her isolation of L1.2B from the mother's DNA, Abram Gabriel had approached me with an experimental idea. I knew Gabriel from his days as a medical student at Hopkins, and I knew he was another of the bright, serious young minds attracted to Johns Hopkins to carry out biomedical research. Abram had done a research elective with Alan Scott and me on hemoglobin genes one summer while in medical school. Now he was working on a transposable element from a Trypanosome, a parasite. This transposable element, called CRE-1, had a similar structure to that of human L1, and it too was thought to be a retrotransposon, acting through an RNA intermediate. Abram had hooked up with Jef Boeke, the yeast retrotransposon expert from the Department of Molecular Biology and Genetics at Hopkins, to carry out a very interesting experiment.

At the time, I knew Boeke pretty well but would get to know him much better over the next several years. Jef joined the Hopkins faculty in 1987 after doing the groundbreaking work with Gerry Fink (mentioned in Chapter 7) demonstrating that Ty1 was a yeast retrotransposon with an RNA intermediate in its life cycle. Boeke was brilliant, hard driving and ambitious—a really great academic scientist who knew Ty1 and retrotransposon biology inside and out. After my lab had found the JH-27 and JH-28 insertions in patients with hemophilia A, I had given a seminar in Boeke's department about the work. I was asked in the question period whether the potential precursor L1 had a primer-binding sequence used to start the process of reverse transcription in many retrotransposons. I didn't know the answer because I hadn't looked for that specific sequence in L1.2. I didn't yet know about the complicated process of reverse transcription used by

retroviruses and LTR-retrotransposons and thought likely to be used by all retrotransposons. (Soon we were to learn differently.) After the seminar, Jef took me to his office and gave me a blackboard description of reverse transcription as carried out in the yeast LTR-retrotransposon, Ty1. It was very illuminating but complicated!

Meanwhile, Abram Gabriel had done an experiment in yeast, attempting to demonstrate that his Trypanosome retrotransposon, CRE-1, encoded a reverse transcriptase activity. Boeke and David Garfinkel at NCI-Frederick had thoroughly characterized the reverse transcriptase in TYB, the second protein encoded by the Ty1 retrotransposon. So Gabriel removed Ty1's reverse transcriptase domain and replaced it with the presumptive reverse transcriptase domain of CRE-1. He then transfected this new hybrid element into growing yeast and showed that it encoded a reverse transcriptase activity. He then made mutations in key sites in the CRE-1 reverse transcriptase domain, and as predicted, these mutants failed to make reverse transcriptase. From these experiments, Abram concluded that CRE-1 encoded a reverse transcriptase activity (Gabriel and Boeke, 1991).

Now the experiment with L1 was obvious: Do the same experiment using the ORF2 region of L1. That is, replace the Ty1 reverse transcriptase domain with ORF2 of L1.2A, grow the Ty1–L1 hybrid in yeast, and see if it too made reverse transcriptase. (Note that L1.2B had not yet been isolated when this experiment was proposed.) Steve Mathias, a graduate student with Alan Scott, was elected to carry out the experiment for his thesis work. He would be guided by Scott, Boeke, and Gabriel and obtain the L1 materials from my lab. Mathias did an admirable job, and the experiment worked! Mathias showed that L1 ORF2 did encode a reverse transcriptase activity, and mutations in a critical region of ORF2 knocked out the activity. He also did a number of biochemical characterizations of the reverse transcriptase activity of L1.2.

At this point, the case for L1.2B indeed being an active human transposable element had become compelling. L1.2B was one full-length L1 element picked from at least 7,000 full-length and over 100,000 total L1s in the human genome that had the identical sequence to a disease-causing human insertion over the 3.8kb of the insertion, and an allele with minor changes possessed reverse transcriptase activity. After Dombroski cloned the other two full-length

L1s that hybridized with the JH-27 oligomer a year later, we knew for certain that L1.2B was indeed the only full-length L1 in the genome of both parents of JH-27 that contained the exact sequence of the insertion.

The reverse transcriptase paper was quickly written and submitted to *Science* along with the paper on the isolation of L1.2B. This time, the two papers were quickly accepted and published in the last issue of 1991. It had been almost four years from the publication of the two L1 insertions in hemophilia A patients to the publication of the isolation and initial characterization of an active human transposable element. After a lot of hard work, the gamble had paid off! Although it is now 19 years later, those 1991 papers have lost none of their importance to the mobile DNA field (Dombroski et al., 1991; Mathias et al., 1991).

This page intentionally left blank

11

A quirk of L1 elements—a lousy 3' end is important for genome evolution

In 1991, another graduate student, Susan Holmes, joined the lab. Since Holmes had entered the human genetics graduate program at Hopkins because of a strong interest in neurogenetics and schizophrenia, I'm not sure how she got to my lab and the study of L1 biology. She was a rather quiet person who spoke softly but with confidence. She had a pale complexion, betraying her Celtic heritage. She had obtained her BA degree in Biology from Swarthmore, a fine liberal arts school on the Philadelphia Main Line. Susan's parents were members of the intelligentsia. Her father was an esteemed Professor of the History of Science at Yale, and her mother also had the air of an academic. Holmes wanted to continue their tradition. At the time, it was just prior to Dombroski's isolation of the JH-27 precursor, so I decided to send Holmes down to the Singer lab to become immersed in the biochemistry and molecular biology of the ORF1 protein. There she did very well, finding that the L1.2A ORF1 protein had the same electrophoretic mobility as the ORF1 protein of a teratocarcinoma cell line, suggesting further that L1.2A could be an active retrotransposon (Holmes et al., 1992). She published this work with Maxine Singer and Gary Swergold, a previous Singer trainee who had earlier found the internal promoter of L1 transcription at the 5' end of L1. The Holmes publication also contained the fact discovered by Singer that the ORF1 protein had a special protein-interacting motif within its sequence.

Holmes then initiated a difficult project to demonstrate retrotransposition in cell culture. She began by attempting to clone the L1.2A sequence, the allele of the potential JH-27 precursor, into a vector to assay for retrotransposition in cells. This project proved very challenging, and after many unsuccessful attempts she finally got the desired clone just before she defended her Ph.D. thesis.

However, serendipity finally brought Holmes a good thesis project that she could accomplish. At this time, our genetic diagnostic lab at Hopkins was part of my research lab. The gene encoding the protein that is defective in Duchenne muscular dystrophy, a disorder mapped to the X chromosome, had been cloned, first by Lou Kunkel and colleagues at Harvard and soon thereafter by Ron Worton and his colleagues at the Hospital for Sick Children in Toronto. It turned out that most of the mutations in the disease were deletions of a portion of an extremely large gene, which they called dystrophin (Kunkel, 1989; Worton and Thompson, 1988). We obtained the gene probes from Kunkel and carried out diagnostic testing for 1) female carriers of the condition, 2) affected males, and 3) affected male fetuses prenatally. One of the affected males studied had what appeared to be an abnormally large band for a dystrophin exon. Then Dombroski made PCR primers to amplify the exon and found that the product from the patient was larger than the product from a normal individual by roughly 2kb. Sequence of the PCR product demonstrated an insertion of about 1.4kb of the 3' end of an L1 plus an extra roughly 600-nucleotide single-copy sequence at the 3' end of the L1. The single-copy sequence ended with a poly A signal and poly A tail (Figure 11.1). Moreover, the entire non-dystrophin sequence, including the truncated L1 sequence and the 600-nucleotide extra sequence, was surrounded by typical target site duplications.

The target site duplications signified that all the non-dystrophin sequence was part of a single insertion event. Where did this extra single-copy sequence come from? Was it present next to the precursor L1 in the genome, or was it added at the RNA level by some novel mechanism not previously described? We were betting on the former, but the only way to find the answer was to clone the precursor of this insertion. But this time around, it looked to be much easier than the cloning of the JH-27 insertion because the extra 600-nucleotide sequence was single copy or unique. It could be used as a probe in

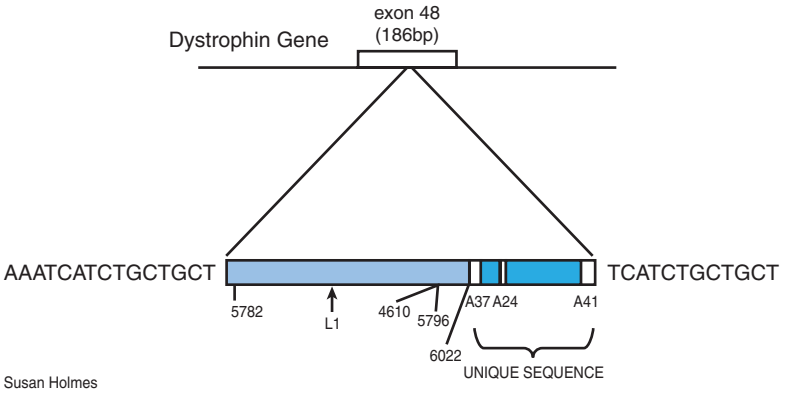


Figure 11.1 An unusual insertion into exon 48 of the dystrophin gene on the X chromosome of a male with Duchenne muscular dystrophy. The insertion contains the 3' end of L1 from nucleotide 5796 to the end at 6022 preceded by an inversion of L1 sequence from nucleotides 4610 to 5782. After a 37-nucleotide poly A tail at the 3' end of the L1, there are roughly 600 nucleotides of unique or single-copy sequence. The entire insertion is surrounded by a 15-nucleotide target site duplication of dystrophin gene sequence. This insertion was characterized by Susan Holmes.

phage cloning, and we thought it likely that the precursor L1 would be located just upstream to this extra sequence. But if there were no L1 sequence upstream of the single copy sequence, then we would be dealing with a new mechanism, splicing together of two different RNAs derived from different genomic locations. Susan used that unique extra sequence as a probe against a phage genomic library.

Within a short time, Holmes isolated a phage clone that contained the unique sequence and found that it did contain a full-length L1 just upstream of it. The hypothesis that the unique sequence was present in the genome downstream of the precursor L1 was correct. This precursor was located on chromosome 1, and, like L1.2, also had two intact ORFs, suggesting that it could make the two critical proteins, ORF1p and ORF2p. It was the second L1 element isolated that contained intact ORFs. L1.2 was the first. However, this likely precursor element that had the identical nucleotide sequence as the insertion lacked the first 21 nucleotides at the 5' end of the element. This observation indicated that the first 21 nucleotides were not critical for L1 transcription. [Gary Swergold had shown in 1990 that L1 contained an internal promoter for transcription and that much of the promoter activity was in the first 50–100 nucleotides of the element (Swergold,

1990).] We concluded that the insertion was derived from an RNA transcript that contained the full-length L1 minus the first 21 nucleotides along with the roughly 600 nucleotides from the 3' flanking region of the L1. We knew that for RNA transcripts synthesized by RNA polymerase II there was in the DNA a signal sequence, AATAAA, called the poly A signal that signaled cleavage of the transcript about 20 nucleotides downstream of the signal and addition of a poly A tail. In this case, the poly A signal sequence was variant, AATTAAA, in the DNA, suggesting that the RNA readthrough and failure of cleavage (called 3' transduction) might be due to this variant sequence and not a common occurrence in the genome.

On close inspection of the L1 sequence at its 3' end, it was clear that the L1 signal for cleavage of the RNA transcript and addition of the poly A tail was quite weak. The transcript was frequently uncleaved after the poly A signal and continued until the next poly A signal in flanking DNA. This second poly A signal could be up to 1000 to 2000 nucleotides downstream. However, it wasn't until 1999 when the cell culture assay was used to demonstrate this effect (Moran et al., 1999), called 3' transduction, and 2000 when the human and mouse genome sequences were analyzed by Eric Ostertag and John Goodier in our lab (Goodier et al., 2000) and Oksana Pickeral analyzed the human sequence in the Boeke lab (Pickeral et al., 2000) that it became clear that 3' transduction was a relatively common effect. When the transduced flanking sequence was single-copy sequence as was the case for this muscular dystrophy L1, one could readily trace its precursor, the L1 from which it originated. This was accomplished by phage cloning in 1993, but after 2001 with a little bit of luck it could be done by simple database inspection. (The need for luck stems from the fact that active L1s of the Ta subfamily like L1.2A, L1.2B, and the precursor of the dystrophin insertion, are often polymorphic as to presence or absence in any human genome. Thus if the precursor is absent from the genome that comprises the database, then, while it is present in some individuals, it won't be found by a database search.) Holmes wrote up her 3' transduction story, published it in *Nature Genetics*, and soon got her Ph.D. degree (Holmes et al., 1994). To repeat, although the biology has changed little since Holmes made her observation of 3' transduction, the availability of genome sequences has made demonstration of 3' transduction much easier.

12

A tour de force from Tom Eickbush

In 1993, Tom Eickbush and his graduate student, Dongmei Luan, at the University of Rochester, made a truly amazing discovery. Since Tom had completed his graduate work for his Ph.D. in the Johns Hopkins' Department of Biology 13 years earlier, he had been invited to present a seminar at the Homewood campus soon after his breakthrough paper was published in *Cell*. That is when we first met. Tom was a pleasant, extremely knowledgeable man. He had been working on *D. melanogaster* (fruit fly) and *Bombyx mori* (silk worm) biology since his postdoc with Fotis Kafatos at Harvard. Tom specialized in the non-LTR retrotransposons R1 and R2, each of which inserted at a different particular site within the ribosomal RNA genes of insects. In other words, R1 and R2 were site-specific non-LTR retrotransposons. In 1988, his lab had shown that the only ORF of the R2 element encoded an endonuclease that could cleave DNA at the specific R2 site in a ribosomal RNA gene (Xiong and Eickbush, 1988b). Now Luan with Eickbush had purified the protein made by the R2 ORF in *E. coli*. Then they carried out an experiment in a test tube that contained that purified protein, the 3' portion of the R2 RNA, a portion of the ribosomal RNA gene encompassing the R2 insertion site, and nucleotides used to synthesize DNA. Using PCR, they showed that the ribosomal RNA gene was nicked at the R2 site, new DNA was synthesized beginning at that site, and the new DNA was attached to the 3' end of the nick site (Figure 12.1). They also isolated the growing DNA around the nick, and showed that it had a branched structure. Mutations in the R2 protein at positions critical for reverse transcriptase activity eliminated the growing DNA at the nick site. All of these data were excellent evidence that reverse transcription was being carried out by the protein of R2 using the 3' OH at the DNA

nick site as a primer and the R2 RNA as a template (Luan et al., 1993).

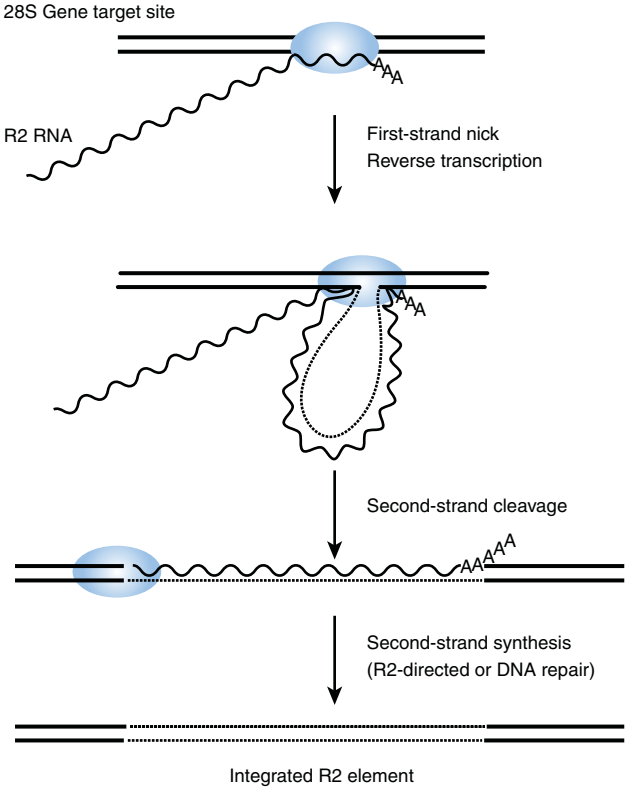


Figure 12.1 Model for R2Bm retrotransposition. The R2 protein associates near the 3' end of the R2 transcript. The R2 protein makes a nick on the bottom strand at the 28S gene target and uses the 3'OH at the nick as a primer for reverse transcription. After reverse transcription, top strand cleavage occurs, and second strand synthesis follows, presumably by the R2 protein. DNA is solid lines, R2 RNA is wavy lines, and cDNA is dotted lines. From Luan and Eickbush 1993 with permission.

The R2 reverse transcriptase carried out its activity on the genomic DNA itself. This finding was not only a big surprise, but it was also truly revolutionary! It signaled a completely different mechanism from the reverse transcription mechanism of LTR-retrotransposons and retroviruses. There was no complex, multi-step, cytoplasmic

process occurring in viral-like particles. For this non-LTR retrotransposon, the evidence indicated that it all happened right in the nucleus on the DNA! Not only that, but it was likely that Eickbush had found a general mechanism for reverse transcription of all non-LTR retrotransposons. He called the mechanism "target-primed reverse transcription, or TPRT" for short, and the term has stuck to the present. TPRT is the mechanism by which non-LTR retrotransposons, such as mammalian L1s are reverse-transcribed. In fact, many of the findings of the Eickbush lab from studies of R2 have been generally useful for our understanding of non-LTR retrotransposons (See the discussion of Het-A and TART in Chapter 4, "Mobile DNA of Model Organisms").

Luckily for my lab, one of his findings on reverse transcription of R2 did not hold for L1 elements. Eickbush found that the last 250 nucleotides at the 3' end of the R2 element were critical for reverse transcription (Luan and Eickbush, 1995). Without those key nucleotides, reverse transcription did not occur. Luckily for the L1 field, the 200 nucleotides upstream of the poly A tail in L1 are not critical for reverse transcription of L1. In 1993, the field of non-LTR retrotransposons was small, perhaps between 5 and 10 labs. However, Eickbush's *tour de force* and work of other labs over the next several years began to draw the attention of a number of scientists and increase the enthusiasm of those already in the field.

This page intentionally left blank

13

"I don't believe all those colonies represent retrotransposition events."

In 1993, Dombroski isolated the other two full-length L1s that hybridized with the JH-27 oligomer from the phage library of JH-27's mother. Both of these elements were Ta subfamily members and had intact ORFs. However, they each had many nucleotide differences (about 1 in every 200) from L1.2B, the JH-27 precursor L1. These latest potentially active L1s were located on different chromosomes and were also polymorphic as to presence in human genomes. Some individuals carried them, while others did not (Dombroski et al., 1993). All of this was fine, but the field still badly needed a cell culture assay for retrotransposition. Enter John Moran.

John Moran was a tall, gregarious Long Islander of Irish/Finnish heritage. He had attended Rochester Institute of Technology (RIT) in Upstate New York and then went for his Ph.D. degree to Ohio State. His mentor was Phil Perlman, an excellent molecular biologist and experimentalist whom I had met a few years earlier on a professional visit to Columbus. A couple years after Moran entered Perlman's lab, Phil had moved his lab to Southwestern Medical School in Dallas, Texas, and John finished his Ph.D. work there. John worked on L1-like mobile elements in yeast mitochondria, called Group II introns (mentioned in Chapter 5, "Exceptional Scientists Working on Mobile DNA in Lower Organisms"). Group II introns have many similarities to LINE-1 elements in their structure and biochemistry, so Group II introns may be ancestors of mammalian L1s. Moran knew about this connection and wanted to get into the study of human transposable elements. Jef Boeke had met Moran at a conference, and Jef confided in me that Moran wanted to work with me and that he had great

potential as a postdoc. Jef had been very impressed! Indeed in early 1993, Moran applied to my lab for a postdoc position, saying that he wanted to set up a cell culture assay for retrotransposition. I was quite impressed that John had prepared to the point that he knew the next big step in our project and that he was keen to be the one to take it.

At first, I thought Moran would join us in early 1994, but he kept putting off his start date because he wanted to come to a good finishing point in his work with Perlman. Finally, he signaled that he would move to Baltimore in April, 1994, but there was a complication. I was starting a new position as Chair of the Department of Genetics at UPenn in Philadelphia. Moran's wife had secured a great position with the Environmental Protection Agency (EPA) in Washington, D.C, thinking that John would be starting in Baltimore.

In the fall, she could go to another EPA position in Philadelphia, so we arranged that John would start for a few months in Jef Boeke's lab at Johns Hopkins. This was a very appropriate place, and Moran began his project by researching the various vectors that might be used to deliver L1 to cultured cells. He wanted a vector that would be long-lived within cells and would attain a copy number of 10–20 copies per cell in the nucleus as an episome, that is, a piece of DNA that would not insert into the chromosomes. A postdoc in the Boeke lab suggested pCEP4, and Moran decided to try it.

When Moran came to the lab at Penn in the fall of 1994, he made quite an impression on everyone. First, he was a big guy! Second, he had a real gift for gab. He could talk his way out of any situation. Third, this guy was really smart and very quick. He would come up with a dozen experimental ideas at a moment's notice. Fourth, he had a facility for finding the right word at the right time, even in his joking manner. But Moran was definitely the real deal! He not only had all those ideas, but he could deliver experimentally. His first decision after coming to Penn was to determine the sequences he would use as a retrotransposition indicator cassette. Previously, Thierry Heidmann from the Institut Gustave Roussy in a suburb of Paris had made a retrotransposition indicator cassette using a backward neomycin resistance (*neo*) gene that worked in demonstrating integration of a retrovirus into a mammalian genome (Heidmann et al., 1988). Joan Curcio, working with David Garfinkel at NCI-Frederick, had used a

backward *his-3* gene disrupted by an artificial intron to demonstrate retrotransposition of the yeast retrotransposon, Ty1, in *S. cerevesiae* (Curcio and Garfinkel, 1991). When the artificial intron was removed from the *his-3* gene upon retrotransposition of a marked Ty1 element, yeast cells would grow without addition of histidine to the medium. This system with the backward gene relative to the retrotransposon disrupted by a forward intron, a la Curcio and Garfinkel, was the one John decided to try first.

He obtained a neomycin resistance gene that contained a human γ-globin intron from Dixie Mager at the University of British Columbia. Dixie was an old friend of mine from her graduate student days working on hemoglobin genes with Oliver Smithies. She knew that the intron could be removed *in vivo* from the neomycin resistance (*neo*) gene, and that would convert cells that contained the construct from neomycin sensitive to neomycin resistant. In other words, once the intron was removed, cells containing the *neo* gene would grow in the presence of neomycin (Freeman et al., 1994). Moran decided to use this disrupted gene with a strong promoter to drive RNA transcription and a strong poly A signal to terminate transcription. So he added an SV40 promoter to the 5' end of the *neo* gene and a thymidine kinase (TK) poly A signal to the 3' end of the *neo* gene.

Now Moran needed to decide where to put his cassette. He certainly didn't want to disrupt either of the potential ORFs in the L1, and he didn't want to place the cassette far from the 3' end of the element. He wanted to detect as many retrotransposition events as possible, and most natural insertions only extended a short distance from the 3' end of the element. They began at the 3' end of L1 but stopped soon thereafter. Many natural insertions were less than 1kb in length. Moran needed to have *neo* expression in order to detect a retrotransposition event. Because the *neo* gene minus the intron but plus the SV40 promoter and TK poly A was about 1.6kb in size, even if the cassette was placed very close to the 3' end of the L1, the insertion would need to be at least 1.6kb in order to be detected. So Moran decided to engineer a restriction endonuclease site very close to the L1 3' end and put the cassette in the backward *neo* orientation into that site. He could then use a very rare restriction endonuclease site that was present in nearly every human L1 just upstream of the cassette to easily exchange essentially any human L1 into the vector.

That new L1 would still retain the retrotransposition cassette (see Figure 13.1).

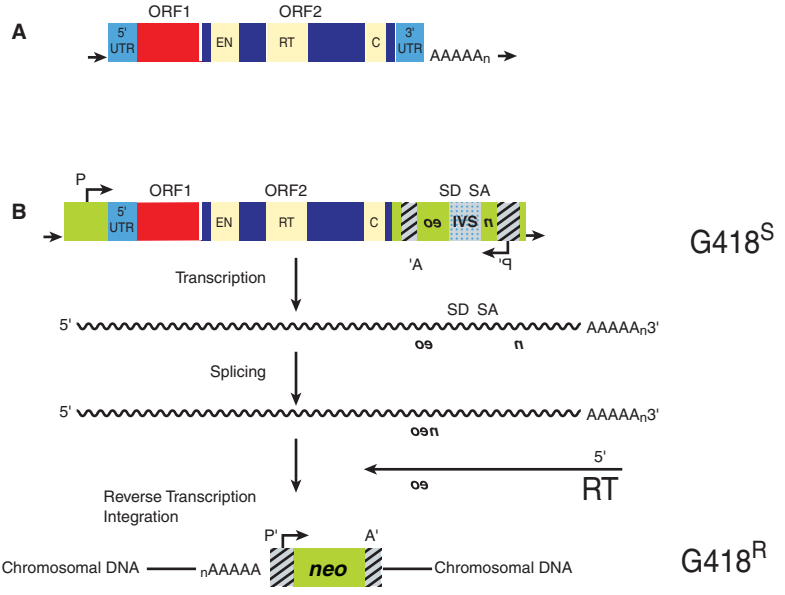


Figure 13.1 An L1 Retrotransposition Assay. (A) Organization of a 6.0 kb human L1 element. The approximate positions of the endonuclease (EN), reverse transcriptase (RT), conserved cysteine-rich (C) motif, and poly(A) tail (AAAAAn) are indicated. Arrows indicate the target site duplications flanking the element. (B) An overview of the L1.2mneoI retrotransposition assay. L1.2 was tagged with an indicator gene (mneoI) containing an antisense copy of the neo gene disrupted by intron 2 of the γ-globin gene in the sense orientation. The splice donor (SD) and spice acceptor (SA) sites of the intron are indicated. The neo gene is also flanked by a heterologous promoter (P') and a polyadenylation signal (A') denoted by the hatched rectangles. Transcripts originating from the promoter driving L1.2mneoI expression (P) can splice the intron but contain an antisense copy of the neo gene. G418-resistant (G418R) colonies should arise only when this transcript is reverse transcribed, integrated into chromosomal DNA, and expressed from its own promoter, P'. Although the sequence to the left of the chromosomally integrated neo gene is actually poly(T) on the strand depicted, for consistency it is shown as poly(A).

After all this preparation, by March, 1995, or six months after arriving in Philadelphia, he had a vector to test-a pCEP4 plasmid

containing a presumptive active human L1, L1.2A, that contained in its 3' end his retrotransposition indicator cassette. Here's how it would work: If transcription were initiated from a promoter driving the L1 (either the L1 promoter itself or a heterologous promoter that he added to the L1), the transcript would contain the *neo* gene plus the γ-globin intron. The intron would be in the forward orientation so it could be spliced out of the L1 transcript. The *neo* gene would be in the backward or reverse orientation so it would not be expressed. However, if the L1-*neo* RNA transcript were reverse transcribed and integrated into the cell's genome (a retrotransposition event), the *neo* gene would then be in the right orientation. If the inserted DNA were transcribed from the SV40 promoter and translated into protein, the cell containing that retrotransposition event would become resistant to the chemical G418, a neomycin analogue. It would be *neo* resistant. On the other hand, any transcript derived from the *neo* gene could not make the *neo* protein. This is because the *neo* gene would remain disrupted by the intron in the wrong orientation relative to it, and that intron could not be removed. In theory, this was a clever assay for retrotransposition. He used L1.2A, an allele of the likely precursor in JH-27, as a test L1. As a negative control, he used an L1.2A containing a mutation in the reverse transcriptase domain that Mathias had shown would eliminate reverse transcriptase activity.

Moran decided to try the assay first in HeLa cells, an excellent transformed human cell line. Because he was a novice in mammalian cell culture, he got help from Roger Kennett, a professor in the Department of Genetics and a cell culture expert. Moran carried out the cell transfections, selected for transfected HeLa cells that were resistant to hygromycin (a hygromycin resistance gene was present on the pCEP4 vector), then he added G418 to detect potential retrotransposition events. After two weeks, he looked at the plates and found 300 to 600 G418 resistant colonies per plate of 10^6 transfected cells. When he showed me the plates, I was amazed. "These can't all be cells with retrotransposition events. I don't believe it!" I exclaimed. "The controls look good. Almost no colonies in the cells carrying a deletion mutant," John replied. John and I worried that perhaps recombination had occurred between the reverse transcribed DNA copy of the L1 and the intron-containing L1 in the transfected plasmid. This could produce the positive result that was

observed. That type of recombination was common in yeast. However, when Moran checked out this possibility, he found that it was not the case. The G418-resistant cells were all the products of retrotransposition events (see Figure 13.2). To me, it was amazing that the experiment had worked so well on the first try—even one as carefully thought out as this one. Moreover, as Moran repeated the experiment, it worked better and better. And the negative controls, including an RT- (reverse transcriptase-) mutant, remained negative.

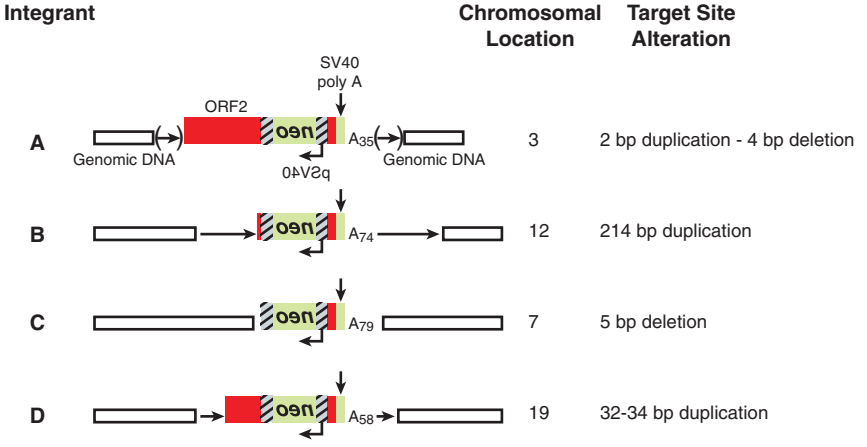


Figure 13.2 L1 insertions from a transfected marked L1 in HeLa cells (A–D). Each insertion was compared with its corresponding empty site, which was independently cloned from HeLa genomic DNA. Truncated portions of L1.2mneoI (the retrotransposition cassette) are shown, and the nucleotide position of the truncation in L1.2 is noted. Closed rectangles are L1.2 sequences, and hatched rectangles are the SV40 promoter and TK poly A signal at the two ends of the antisense neo gene. Stippled rectangles are transduced sequences between the 3' end of L1.2 (the transfected human L1) and the SV40 poly A site derived from the pCEP4 vector. Open rectangles represent genomic DNAs. Right arrows indicate target-site duplications. The length of the poly A tracts and the sizes of the target site duplications and deletions are indicated. The arrow flanking insertion A is marked parenthetically because the target site could be a 1–2-bp duplication, a blunt insertion, or an up to 4-bp deletion.

Then John made mutants in L1.2A in a conserved region of ORF1 and in a conserved region of ORF2. These mutations reduced the number of G418-resistant colonies by two orders of magnitude or 100-fold (see Figure 13.3). Then he made a mutation that merely

changed a common restriction endonuclease site, and, as expected and hoped for, it had no effect on retrotransposition because it did not affect a key activity of the element. At that point, it looked like Moran had an effective assay for retrotransposition in cell culture.

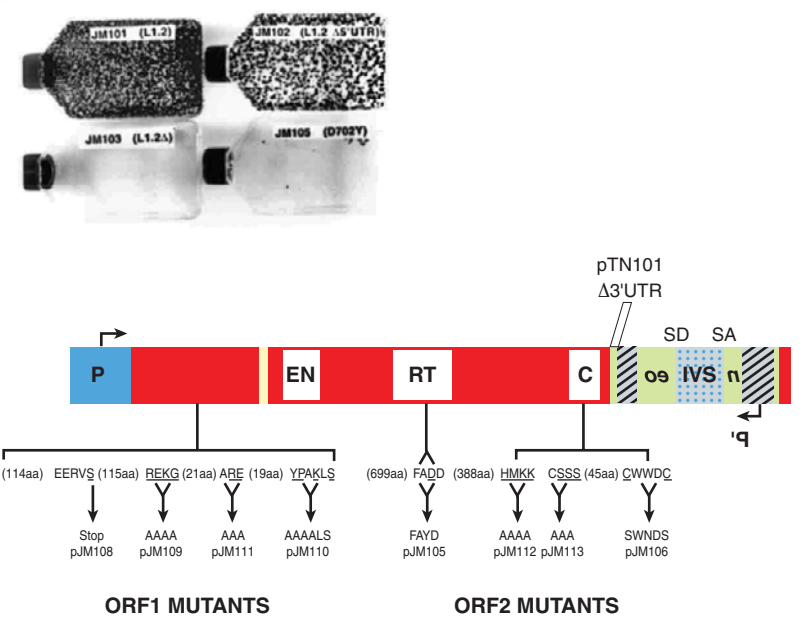


Figure 13.3 L1.2 mutants knock down retrotransposition activity in cell culture. At the top are assays of L1.2 driven by both its 5' UTR promoter and a heterologous CMV (cytomegalovirus) promoter, L1.2 driven only by its 5' UTR, a deletion of much of L1.2, and a reverse transcriptase mutant D702Y (aspartic acid to tyrosine at position 702). On the bottom are ORF1 and ORF2 mutants that knocked down retrotransposition by roughly 2 orders of magnitude.

This page intentionally left blank

14

L1 encodes an endonuclease

Now it was Boeke's turn to re-enter the story and add luster to it. Although I was now working in Philadelphia at Penn, my wife and I still resided in our house in North Baltimore. I would take the train to Philadelphia twice a week and spend two or three nights per week in an apartment close to the University. Because I was in Baltimore anyway on Monday mornings, on a few occasions I attended Boeke's regular weekly Monday lab meeting. In addition, on a few occasions, Jef would join me in a 2–3 mile jog on Sunday morning. On one of those occasions, I told him about John's success with the cell culture assay of human L1 retrotransposition. Jef was now branching out from studies on the yeast Ty1 retrotransposon to work in the human L1 field and had decided to put a graduate student who recently joined his lab on an L1 problem. I think on one occasion Moran suggested to Boeke that considering L1 would need an endonuclease to insert into DNA, perhaps it carried its own endonuclease in its DNA sequence.

One Sunday, Jef picked me up at home for another run, but this time was different for a couple reasons. First, he drove an old pickup truck that carried with it the odor of his dog that had often been transported in it. For me, that made for an unpleasant ride. Second and more importantly, Jef was really excited! He had done a computerized analysis called a pileup of sequences in the first 400 nucleotides (the 5' end) of ORF2 of L1.2A and the 5' ends of ORF2 (or the single ORF) of a number of non-LTR retrotransposons of various organisms. The region in all of them strongly resembled an apurinic-apyrimidinic (AP) endonuclease. Although I later learned that other investigators had done the analysis previously and reported similar results (Martin et al., 1995), Boeke's analysis was news to me at the time. He had found that the DNA sequence close to the 5' end of L1

ORF2 appeared to encode an apurinic-apyrimidinic endonuclease of a type previously crystallized from *E. coli* (Mol et al., 1995). Moreover, all the key amino acid residues in the *E. coli* protein appeared to be encoded by the L1 DNA. This was really big news! The obvious next experiment was to show that this endonuclease activity existed in ORF2 and then to make mutations in the crucial nucleotides that would change amino acids in key residues. If the endonuclease were important for retrotransposition, those mutations would eliminate retrotransposition of the element in cell culture.

This was the perfect research project for the new student, Qinghua Feng. She quickly did some lovely biochemical studies to show that indeed ORF2 of L1.2A did contain a specific endonuclease activity that would nick a single strand of double stranded DNA (Figure 14.1). The nick would leave a 5'-phosphate and a 3'-hydoxyl residue. Then John Moran did the acid test. He made mutations altering the encoded amino acids in key residues of the endonuclease domain. Moran then put the L1.2A elements each with a different endonuclease mutant into the cell culture assay and showed that all of these key mutants dramatically reduced retrotransposition activity to 1% or less of wild-type levels. Moran's cell culture assay and the information that he had obtained showing that both ORFs of L1 were critical for retrotransposition along with Qinghua Feng and Boeke's demonstration of endonuclease activity in the element were submitted as two papers to *Cell* and published back-to-back (Feng et al., 1996; Moran et al., 1996). They were both important. Moran's assay is still used either with a *neo* indicator cassette or modified with an enhanced green fluorescent protein (*EGFP*) cassette by many investigators in the field. From whichever cassette, the loss of the intron from the backward gene signifies the occurrence of a retrotransposition event. Boeke's demonstration of an endonuclease activity in L1 pointed out the similarity of this mammalian non-LTR retrotransposon to the insect R2 retrotransposon. They both encoded endonuclease (although they differ in their types) and reverse transcriptase activities, so it was likely that if R2 used a TPRT mechanism for coupled reverse transcription and integration, L1 did also. A few years later, Greg Cost in Boeke's lab presented *in vitro* evidence that L1 did indeed utilize the TPRT mechanism (Cost et al., 2002).

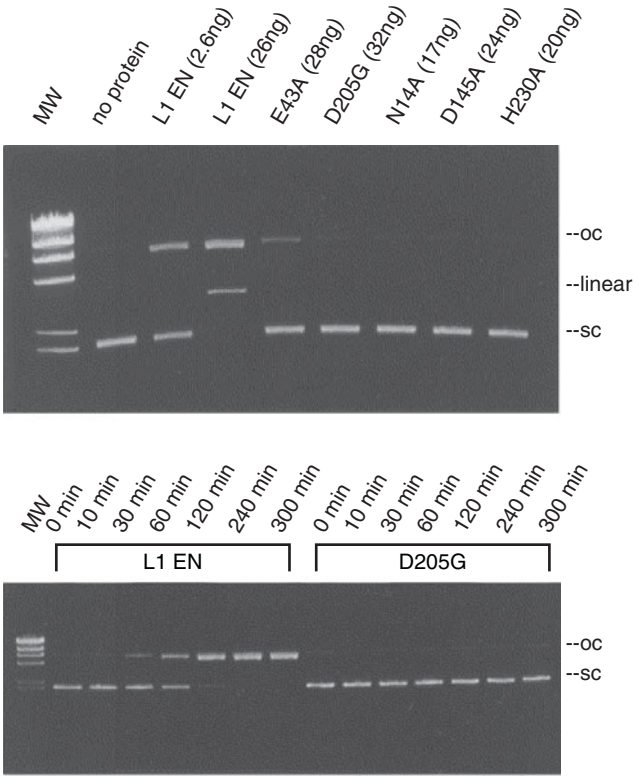


Figure 14.1 Nicking Activities of L1 ENp and Mutant Proteins. Nicking activities. Lane 1 (left to right), phage λ; HindIII digest MW marker; lane 2, substrate pBS DNA, no protein added; lane 3, with 2.6 ng of wild-type L1 ENp; lane 4, with 26 ng of wild-type L1 ENp; EN mutants-lane 5, E43A mutant; lane 6, D205G; lane 7, N14A; lane 8, D145A; lane 9, H230A. Sc, supercoiled plasmid; oc, open (nicked) circular plasmid; l, linear plasmid. (C) Time course, 50 fmol of L1 ENp (or D205G mutant) was used to digest 500 fmol of pBS. Note complete conversion of supercoiled DNA to closed circle with L1 EN and no conversion with the D205G mutant.

Both Moran and Boeke have had great success since those heady days in the mid-1990s. Both have published in the most prestigious journals. Moran has concentrated on L1 retrotransposons, and Boeke has worked on both yeast Ty1 and human L1 biology. Moran is a tenured full professor in Human Genetics at the University of Michigan Medical School. He has won a number of institutional awards and is now a Howard Hughes Medical Institute investigator and

holds an endowed chair. Moran and I communicate by phone frequently, and as his mentor, I am extremely proud of his accomplishments. Boeke is a full professor in Molecular Biology and Genetics at Johns Hopkins University School of Medicine and is highly ranked nationally in total NIH support. Jef is a good colleague at Johns Hopkins where since July 2010, our labs are only one floor apart. (After 16 years at Penn, I returned to the Institute of Genetic Medicine at Johns Hopkins in July 2010.)

15

The jocks

Donna Sassaman and Brook Brouha are inextricably linked together in my mind. They never met because Sassaman left the lab in 1996, and Brouha joined us in 2001. However, they both worked on the same problem, getting solutions appropriate to the time, and both were superb athletes. I've decided to put them together in this chapter because of those similarities, even though they were five years apart in their work.

First, let me tell you about Sassaman. Donna was an outstanding scholar-athlete at Drew University in New Jersey. She graduated magna cum laude in biology and was a star in both field hockey and lacrosse. I once saw her play third base in a softball game, and she was amazing. I thought she was the best player on the field, including the guys. Needless to say, she was later elected to the Drew University Athletics Hall of Fame. She was a dark-haired, enthusiastic young lady when I met her in the summer of 1992. Sassaman had completed her three lab rotations in the human genetics graduate program at Johns Hopkins but was not convinced that any of those three labs was right for her. She asked to do a fourth rotation in my lab. I thought she should start on a project that could lead to her degree. On occasion, as you see later, I would ask a student to choose his or her own project. However, because Sassaman was doing a fourth lab rotation, I took an active role in her project's design. Donna would clone full-length L1s of the Ta subset and determine whether they had reverse transcriptase activity as a first step to estimating the number of active L1s in the human genome. (At that time in 1992, it was pre-John Moran and the cell culture assay.) By 1992, Dombroski, with considerable help from Boeke, had set up an *in vivo* assay for reverse transcriptase activity in yeast (Dombroski et al., 1994). Hopefully, Donna

could clone a number of full-length, young L1s and use the assay to determine what fraction of them had reverse transcriptase activity. I knew the project was risky. The number of active human L1s could be very small, and Sassaman might happen to clone only those that were inactive. I tried hard to be positive, but deep down I was worried that the project might fail completely.

Luckily, Sassaman wasn't afraid of hard work. She hybridized oligomers complementary to the 5' end sequence and 3' untranslated region containing the ACA tri-nucleotide characteristic of Ta subset elements to a phage library of human genomic fragments. She succeeded in cloning 13 Ta subset L1s that were full-length. Gary Swergold, a colleague then at the FDA, had used another hybridization technique to estimate that there were about 200 full-length Ta elements in the diploid genome. Meanwhile, Dombroski had succeeded in cloning out the other two full-length L1s present in the library of the mother of JH-27 using the specific JH-27 oligomer. I recall the night before Donna was expecting results from her reverse transcriptase assay of the first six Ta elements she had isolated. I told her then that I was worried that she might find very few or no elements with activity. Her response was, "Now you tell me! We'll see tomorrow." Indeed, there were four positives among the six L1s tested! I knew then that Donna would get her Ph.D. from this project. It turned out that eight of thirteen L1s that she isolated had reverse transcriptase activity—a remarkable number.

Then in 1994, the lab moved to Philadelphia. Sassaman's mother lived in Cherry Hill, New Jersey, just across the Delaware River from Philly. Now Donna could live at home and commute to Penn, a short distance. Soon Moran had set up his cell culture assay for retrotransposition. Again, Donna worked hard to clone her 13 Ta elements into the pCEP4 plasmid vector to assay the L1s for retrotransposition. She also had Dombroski's two full-length L1s from JH-27's mother to test. We knew she was getting close to finishing her project, so she applied for medical school to start September 1996. Of course, I was again disappointed that Donna was not going to stay in science, but I knew she'd make a great physician. Time was getting short, so many nights Sassaman would work late, not go back home to Jersey, but sleep on the soft bench in the meeting area at the end of the lab hall. Finally, she got all her clones and carried out the retrotransposition assays.

Three of her 13 Ta elements were active in the assay, and both of the L1s from the JH-27 mother were very active (Figure 15.1). She had shown that many L1s (mostly Ta subset) in the human genome are active (at that time our conservative estimate of active L1s in the diploid genome was 30–60) and that elements possessing reverse transcriptase activity aren't necessarily active for retrotransposition. However, of greater importance was her finding that there is a very wide range of retrotransposition activity among active elements, roughly 100-fold (Figure 15.1). Her data had also increased the number of known active human L1s from 2 to 7. She finished her experiments, wrote up her thesis, defended back at Hopkins where she had started graduate work, and went off to medical school at Robert Wood Johnson in New Jersey.

But Sassaman was not finished. She needed to get her key paper published, and I needed her help to finish it up. I also wanted her to put all of the presently known active L1 sequences into her thesis as an appendix. Over Christmas vacation from medical school in 1996, Donna returned to the lab a number of days to complete these tasks. Her paper went off to *Nature Genetics* in early 1997, and it was published in the spring (Sassaman et al., 1997). Moreover, she did put all those sequences into a very valuable file at the back of her thesis. Donna Sassaman, M.D.-Ph.D., is presently in the practice of Internal Medicine and Pediatrics in Wilkes-Barre, PA, and the Geisinger Clinic in Danville, PA.

The other jock was Brook Brouha. Brook graduated at the very top of his class at Dartmouth College, also my alma mater, and before coming to Penn he had taught high school for a couple years in Hanover, New Hampshire. He entered the M.D.-Ph.D. program and the Genetics Graduate program in 1996. I remember Brook from lectures I gave to a graduate genetics course and how interested he seemed in everything. Brook was an avid skier, competitive mountain biker, and a gym rat. He put many of us on his monthly program of gym exercises, requiring just two sessions of 40 minutes each per week. Brook would make up a new program of lifting exercises each month, and within a few years he had about 30 people doing his program. We called it "Body by Brook." Very catchy! We told him he should write a book about it, including nutritional information, but he has yet to do it. To this day, I still get a monthly workout routine from Brouha.

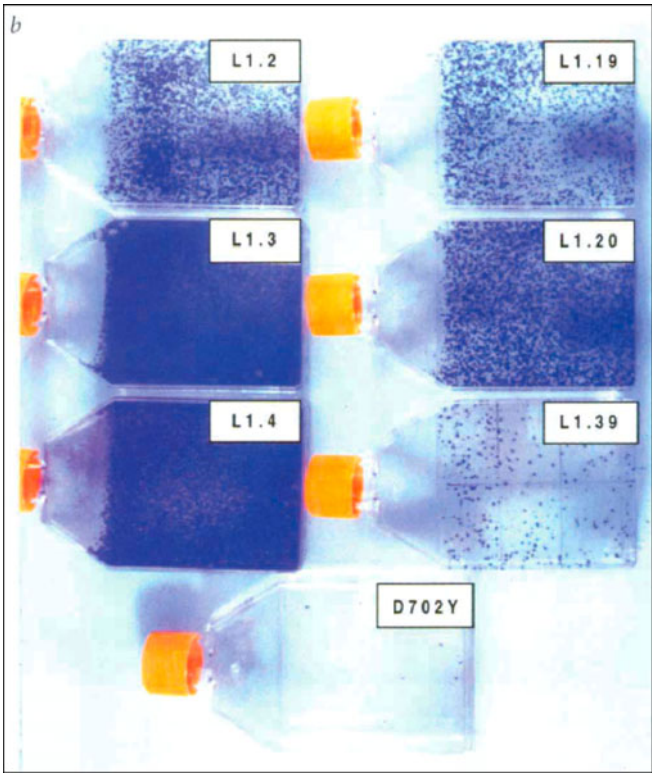


Figure 15.1 Human L1s have highly variable retrotransposition capability in cell culture. Assays of L1.2A, two other human L1s isolated from the mother of JH-27, L1.3, and L1.4, and three active L1s isolated by Sassaman, L1.19, L1.20, and L1.39. Each dot within the flasks represents a colony of cells with a retrotransposition event. Note the substantial variation in activity. The D702Y mutant of L1.2A affects reverse transcriptase and is nearly dead for activity.

In late 2000, Brouha was having difficulty. After over two years of graduate work in another lab, his project was floundering. He had just had a disastrous thesis committee meeting at which his project had been pretty much shredded by his committee. A very bad sign! Brook decided to change labs, but he went about it very methodically. He interviewed with perhaps ten or more potential mentors, including me, to determine which project might be feasible to complete within less than two years so that he could return to medical school in the late fall of 2002. I thought of a circumscribed project: Use the computer to find all the potentially active L1s in the human genome working draft (HGWD) that was due out in early 2001, isolate by PCR all of those

elements, clone them into the vector containing the Ostertag *EGFP* cassette (see Chapter 17), and assay them all for retrotransposition in cell culture. Brouha liked the project, and after two months of interviews and soul-searching, he decided to join the lab.

Brouha was amazing! He already knew the techniques needed for the project. He was good at searching the database. He was a quick learner and mastered the field rapidly. Then a young Penn undergraduate named Josh Shustak showed up looking for work for school credit, and I assigned him to Brouha. This was a great move because Josh did a fine job helping Brouha with all aspects of the project. In short order, with the help of Richard Badge of Moran's lab, Brook found that there were 90 full-length human-specific elements in HGWD that also contained two intact ORFs. These L1s were potentially active for retrotransposition. Using a long-range polymerase chain reaction (PCR), Brook and Josh were able to isolate 82 of these L1s. They then cloned them into the pCEP4 vector containing the *EGFP* indicator cassette. They would then transfect 143B osteosarcoma cells, the most active human cell line for retrotransposition, select the cells that had been transfected, wait a week, and then carry out fluorescence activated cell sorting (FACS) to determine the frequency of green cells, that is, retrotransposition.

After considerable effort, Brouha found that in the haploid genome of HGWD 40 of the 82 elements he had tested had retrotransposition activity. He then estimated that in a comparable diploid genome there would be 80–100 active L1s, a number somewhat greater than the earlier estimate of Sassaman. (Brouha's estimate of active L1s in the average diploid human genome from 2002 is still the best estimate available.) Brouha also found that many of the human-specific full-length L1s were polymorphic as to presence or absence in human genomes. The most active elements were generally the youngest, being present in less than 50% of human genomes tested. Brouha also showed that a small number of the 40 active elements, 6, were extremely active relative to the other 34 and contained 84% of the total retrotransposition activity in HGWD. Moreover, by 2002 there were 6 full-length L1s not found in HGWD that had been isolated as disease-causing elements that had recently retrotransposed in human beings. When these 6 L1s were tested in Brouha's assay, 5 of the 6 were in the very active category (5 are shown in Figure 15.2).

Although there was controversy in the lab concerning the use of the word "hot" in the context of very active L1s, we decided to call these exceptional elements "hot." So although "hot" L1s are infrequent in the human genome, they account for the bulk of retrotransposition activity in the human population (Figure 15.2).

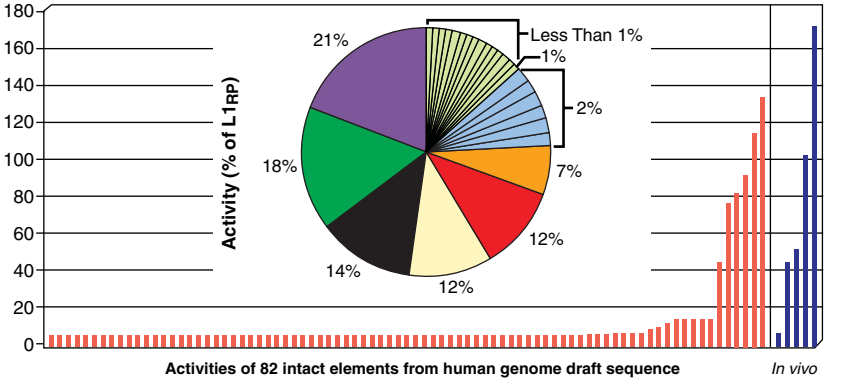


Figure 15.2 The distribution of L1 retrotransposition activity. The measured potential activity of L1s from both the human genome working draft (HGWD) and de novo human insertions is shown. The histogram depicts the activities of 82 intact L1s from the HGWD and five human L1s involved in recent disease-causing insertions. The entire pie in the pie chart represents the total of all of the activity of the 82 L1s from the HGWD. Each slice of the pie represents the activity of a single element. The six "hot" elements (large slices) represent 84% of the total measured potential activity in the HGWD. Dark (blue in e-book) bars at right represent activities of five L1s known to have retrotransposed in vivo. Four of the 5 shown are "hot." A sixth L1 involved in a recent disease-causing insertion is also "hot" in this assay (not shown).

When Brouha was writing his paper on these data, he wanted to have one summary figure that showed the chromosomal location of all 82 tested L1s, their allele frequency, their L1 subset, their ability to retrotranspose, and their relative activity as retrotransposons. Brouha and Shustak thought very hard about this figure and finally came up with a plan. They would show all the chromosomes individually and portray the L1s as human stick figures next to their chromosomal location. Allele frequencies were shown by the act of shading of the human figure. The extent of activity was shown by both the size of the figure and its state of recumbence. The "hot" L1s were shown as large and standing tall. Dead L1s were smaller figures lying flat on their backs. Different L1 subsets were shown as different shadings of

the figures. For example, an upright tall figure shaded in from the waist down next to the short arm of chromosome 6 is a highly active Ta subset L1 with an allele frequency of 0.5 on the short arm of chromosome 6. In my view, this figure is one of the most innovative I have ever seen, and all the credit for it goes to Brouha and Shustak (Figure 15.3). Brouha published this work in the *Proceedings of the National Academy of Science* (Brouha et al., 2003).

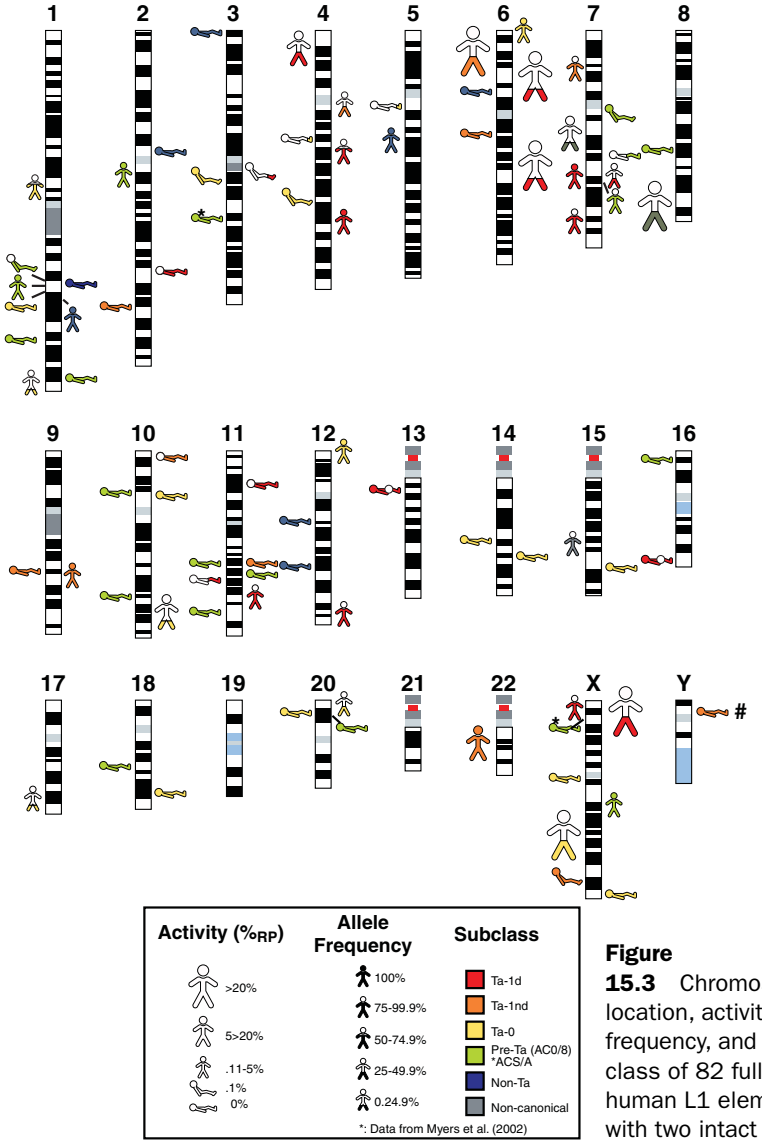


Figure 15.3 Chromosomal location, activity, allele frequency, and subclass of 82 full-length human L1 elements with two intact ORFs.

While Brouha was doing this work, he also solved another mysterious case. A Dutch boy with chronic granulomatous disease, an X-linked condition in which the immune system is defective, had an L1 insertion into the *CYBB* gene, causing his disease. The L1 insertion carried a 3' transduction of single copy sequence that allowed the Netherlands' group to find the precursor L1 on chromosome 2. However, the insertion had telltale sequence differences from the full-length potential precursor found on one of his chromosome 2s (Figure 15.4). His other chromosome 2 lacked the potential precursor L1. The mother of the patient had one chromosome 2 L1 identical to the patient's L1 on chromosome 2 but different from the patient's L1 insertion. She also had a second chromosome 2 that had the sequence of the insertion in the patient's X chromosome. Thus, the patient's insertion could have arisen from this L1 present in the mother, but we believed then that the insertion would have necessarily occurred before the end of maternal meiosis I because the patient did not receive the chromosome 2 bearing the same L1 as the insertion. The father was not available, but the chance that he had contributed the insertion-producing L1 was calculated after other studies as only 2%. Thus, the case for insertion from a chromosome 2 L1 of the mother was strong (Brouha et al., 2002). However, recent work of Hiroki Kano (discussed in Chapter 22) suggests that the insertion could have equally occurred prior to the end of maternal meiosis I as postulated in Brouha's published work or during early embryonic development of the patient. Either scenario now seems quite plausible.

With both of these studies completed, Brouha was able to finish his Ph.D. work in the lab within 20 months and get back to medical school in the fall of 2002. He decided that he wanted to continue to balance professional work after medical training with his outdoor sporting activities. On one of his ski trips out West, he found his soul mate, a female physician of Asian Indian–American origin. She liked many of the things he did, including the gym workouts. They were married after Brouha finished at Penn. He enjoyed Dermatology, and this specialty would afford him sufficient free time. Brouha has recently completed his Dermatology residency and a Dermatology Pathology fellowship. He's now residing in San Diego, a great place for outdoor activities, and he's setting up a Dermatology Pathology practice there. I'm sure he'll be very successful.

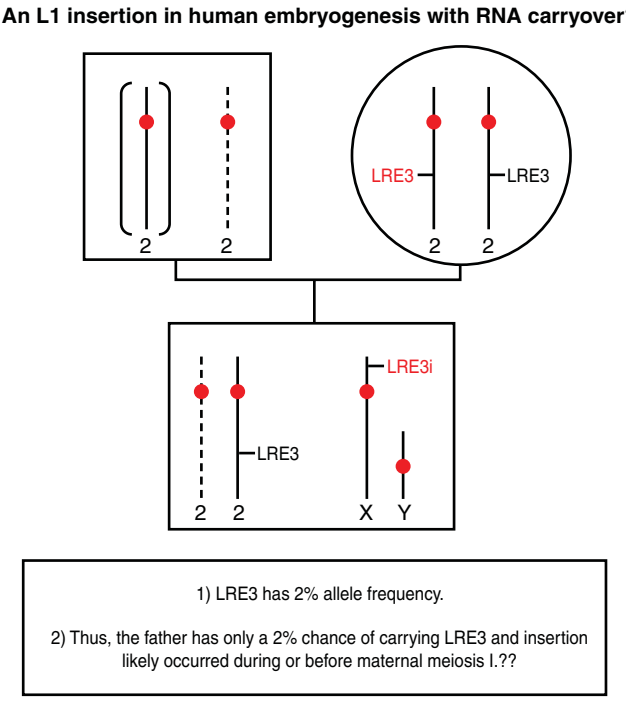


Figure 15.4 A disease-causing L1 insertion occurred either before the end of maternal meiosis I or in early embryonic development. The patient has chronic granulomatous disease, an X-linked disease, caused in this instance by insertion of LRE3, a very active L1 element. The precursor of LRE3 is on chromosome 2, but the patient does not carry the same LRE3 allele (colored black) as the insertion allele (colored light gray—or red in the e-book version). His mother has two LRE3 alleles, the allele present on chromosome 2 in the patient, and a second allele that is identical in sequence to the patient's insertion over the entire insertion. The father was unavailable, but the LRE3 allele identical to the insertion allele has only a 2% incidence in the population. Thus, it is highly likely that the insertion arose from a maternal LRE3. It could have equally occurred prior to maternal meiosis I when the chromosomes segregate to different developing germ cells or in early embryonic development. At the time this study was carried out, an insertion in early embryonic development was considered unlikely. However, in light of work over the past five years, it is now considered a strong possibility.

This page intentionally left blank

16

The mayor and the Frenchman

In 1996, two other trainees entered the lab. One was a graduate student, Ralph DeBerardinis, and the other was a postdoc, Thierry Naas. Ralph was my first student from the University of Pennsylvania. He had begun the M.D.-Ph.D. program at Penn in 1994, and after an eight-week rotation in the lab he decided to do his thesis work with us. Ralph was a *bona fide* Italian-American Philadelphian. He was raised in the region, went to St. Joseph's University outside Philadelphia, and now was studying at Penn close to Center City. Ralph was a gregarious young man. When the lab went out to lunch together, he seemed to know and greet almost everyone on the street. He soon had acquired the moniker, "The Mayor." Ralph was at ease with everyone and loved to discuss his science and the rest of the science going on in the lab.

DeBerardinis also liked to run. He and another graduate student friend would go for five-mile runs in the late afternoon before returning to work in the lab after dinner. After a few months in the lab, Ralph had a pair of old, very grubby, beat up, smelly running shoes that were ready for the trash but that he kept in the lab. After the first time he successfully obtained a tough DNA clone with those shoes nestled in his bottom desk drawer, they became known in the lab as "Ralph's lucky cloning shoes." When lab members wanted to get a difficult clone, they would invoke Ralph's lucky cloning shoes. If the shoes were still in the desk drawer, chances are the cloning would be successful! Sometimes the shoes were even placed on the PCR machine to get a dicey PCR to work.

Thierry was a Frenchman, but he came from the region where the borders of France, Germany, and Switzerland meet. He had finished his Ph.D. at the renowned Biozentrum in Basel, Switzerland, under Nobel laureate Werner Arber. He recounted that he had lived with his parents in a small town on the French side of the border and rode his motorbike into work at Basel every day. I met Thierry at a Transposable Element meeting in Toulon, France, in 1995, and he expressed his desire to come to my lab in Philadelphia for postdoctoral training the next year. He then obtained an EMBO fellowship that gave him two years of support. Thierry was a great addition to the lab but a relatively short-term one. After making several trips back to France to present his credentials for a faculty job in Paris, he was offered the job and took it in 1998. So Naas was a productive member of the group for only two years.

At about the time Thierry and Ralph came into the lab, there was breaking news on mouse L1s. Two disease-producing insertions of full-length L1 elements were reported in neurological disorders, one in the spastic mouse (Kingsmore et al., 1994) and the other in the *Orleans* reeler mouse (Takahara et al., 1996). Both of these L1s contained long 5' untranslated regions, suggesting excellent promoters, and two intact ORFs, suggesting that, like the active human L1s, they had retrotransposed through *cis*-preference and that they were likely still active for further retrotransposition. Thierry and Ralph made constructs of these two L1s with the *neo* retrotransposition cassette inserted in the 3' untranslated region downstream of the second ORF. They then carried out the cell culture assay with the marked $L1_{spa}$ and $L1_{Orl}$ elements and showed that these elements remained very active for retrotransposition. As expected, the elements also had reverse transcriptase activity in the yeast-based assay.

However, their DNA sequences were different from any previously discovered mouse L1s. The 5' untranslated region in mouse L1s is quite different from the same region in human L1s. Instead of a single stretch of sequence of around 900 nucleotides as observed in human L1, the mouse 5' untranslated region as shown by Edgell and Hutchison contains monomers of about 210 nucleotides. These monomers are of one type in a single element but are of different

types in the range of mouse L1 elements. $L1_{spa}$ and $L1_{Orl}$ had monomers of a known type, the F type, while their ORFs contained sequences common to the two elements but different from any mouse L1 observed previously. We called this new L1 variety T_F, T for transposable and F for the monomer type in the 5' untranslated region. Naas and DeBerardinis also found that mouse L1's T_F subfamily contained a large number of members and appeared to be expanding in the present-day mouse genome. They published this work as co-first authors in *EMBO Journal* in 1998 (Naas et al., 1998). It was the first paper from the lab on mouse L1 elements, and it dispelled the notion that L1s of the A type are the most active family of L1s in the mouse genome.

After Naas returned to his job in Paris where he has been quite successful doing bacterial genetics, DeBerardinis continued to study the T_F subfamily of L1s. He found that there were approximately 3,000 full-length T_F elements in one strain of mice, and he isolated 11 of them. After sequencing these 11, he found that they were very similar in sequence, averaging 99.8% identity. Moreover, 7 of the 11 elements were active in cell culture, suggesting that there were ~1800 active T_F elements in the mouse genome. DeBerardinis also found that there was considerable polymorphic variation among different strains of mice. Some of the 11 T_F elements were present in one or two strains but absent in all the others (Figure 16.1). These data suggested further that the T_F subfamily of active L1s was presently expanding in mouse strains. Ralph published this work at the end of 1998 in *Nature Genetics* (DeBerardinis et al., 1998).

DeBerardinis also had a major project that was not going well. During the Short Course at Bar Harbor, Maine, in 1996, he met Mario Capecchi, an inventor of the technique of homologous recombination in mice, and with Oliver Smithies, a 2007 Nobel laureate. DeBerardinis and Capecchi had discussed the possibility of making a mouse that would retrotranspose an active human L1 element. Capecchi had made some suggestions, and Ralph had independently decided that this was a great Ph.D. project. Although I knew it might be difficult, I gave him my enthusiastic support.

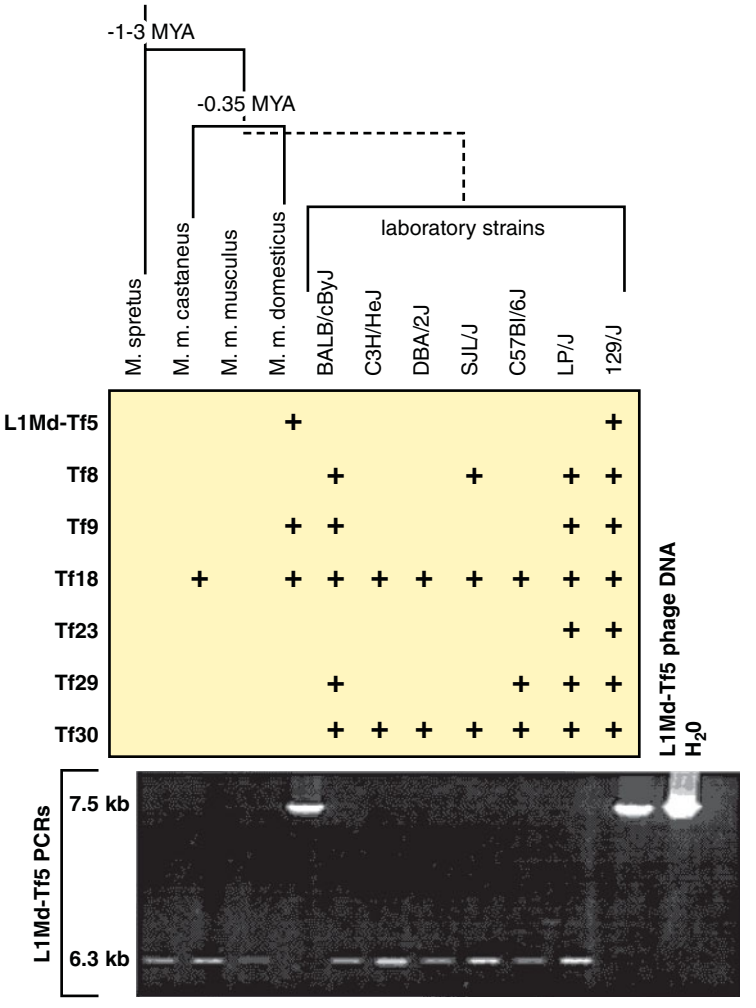


Figure 16.1 T_F elements are recent insertions in the mouse. The phylogenetic tree (top) shows evolutionary divergence points for M. Spretus and M. musculus species complex. The dashed line connecting the lab strains to the M. musculus complex indicates that the strains do not strictly belong to a particular M. musculus subspecies. Hybridization experiments revealed that all genomes in the tree contain many T_F sequences and that T_F is most abundant in M. spretus. Genomic PCR reactions (bottom) across the L1 T_F15 insertion site, which are aligned with the genomes shown in the top panel, verify presence of the element (7.5kb product) in strain 129/J and a wild-derived M. m. domesticus strain, and absence of the element (0.3kb product) in all other tested genomes. Middle: results of similar analyses, using genomic PCR to test polymorphism of six other T_F elements. Plus indicates presence of the L1, and lack of a plus indicates absence of the element. There is significant polymorphism for the seven elements tested among the mouse strains.

Ralph decided to use an enhanced green fluorescent protein (*EGFP*) reporter gene for retrotransposition *in vivo*. He went ahead and made a mouse carrying an *EGFP* transgene with a CMV promoter and an SV40 poly A signal. This transgene also contained an acrosin signal peptide sequence between the promoter and the *EGFP* gene to concentrate the EGFP in the acrosome of sperm cells. DeBerardinis reasoned that he could then place the sperm under fluorescent light and see fluorescence in the acrosome of the sperm head. Indeed, these positive control mice had the green fluorescence in every sperm head! Then Ralph disrupted the *EGFP* with an intron and put this new retrotransposition cassette into the SmaI site that Moran had created in the 3' untranslated region of a very active human L1. But when this new marked L1 was put into HeLa cells in the cell culture assay, it was a bust! There was never any retrotransposition. It gave the same result as a negative control L1. What could be the problem? It was two more years before Moran solved this problem for us. Meanwhile, DeBerardinis had done his analysis of T_F mouse elements, gotten his Ph.D. degree, and gone back to finish medical school.

DeBerardinis was a star in medical school and completed a combined pediatric-genetics residency at the Children's Hospital of Philadelphia on the Penn campus. During his residency, he worked successfully with Craig Thompson at Penn on cancer metabolism. Ralph is now an Assistant Professor of Pediatrics at Southwestern Medical School in Dallas.

This page intentionally left blank

17

Ostertag's coups

In 1998, Eric Ostertag, a soft-spoken, no-nonsense, precise Germanic type, came to the lab as a rotation student. Eric was a Wisconsin "cheesehead" who had graduated from the University of Wisconsin at Madison and was another M.D.-Ph.D. student at Penn. He had been a member of my small discussion group in the first-year Genetics course for medical students in 1996, so I knew he was very smart and a deep thinker. During his short rotation in the lab, it became clear that Eric would be an outstanding graduate student. I was very pleased when a few months later Eric told me that he wanted to do his Ph.D. in the lab. Although Eric obtained a lot of data on a number of projects, I discuss only four here.

First, Ostertag wanted to test the new retrotransposition cassette made by DeBerardinis in which enhanced green fluorescent protein, *EGFP*, a screenable marker for retrotransposition, was substituted for the *neo* gene. This switching to a new retrotransposition marker can be tricky because the backward intron must go into a site in the gene from which it can be spliced without difficulty, but Ralph had been able to find such a site in the *EGFP* gene. The only problem was the cassette just did not work in the cell culture assay.

In the meantime, John Moran had recently started up his lab at the University of Michigan. Because Moran loved to talk science, we had told him about DeBerardinis' trouble getting retrotransposition to work with the new cassette in cell culture, let alone the mouse. We had discussed DeBerardinis's retrotransposition cassette, so John knew that the cassette had been altered and now contained an SV40 poly A signal sequence instead of a thymidine kinase (TK) poly A signal after the marker gene. Moran discovered that the SV40 poly A

signal sequence contained a poly A signal (AATAAA) not only in the desired direction, but also in the other orientation. This news was crucial! Now we knew why Ralph could not obtain any sign of retrotransposition with the altered cassette. The RNA transcript from the L1 promoter would be cleaved after the SV40 poly A signal and before it reached the marker gene. No transcription through the marker gene meant no opportunity for reverse transcription and integration of the marker gene. Eric needed to replace the SV40 poly A signal with the original TK poly A signal sequence because it contained the signal (AATAAA) in the desired orientation only. This vignette points out that anybody, even a very smart person, can still make mistakes!

Eric then quickly found excellent retrotransposition with his new cassette that contained a CMV promoter driving the disrupted *EGFP* gene, followed by a TK poly A signal sequence. He carried out experiments on the kinetics of retrotransposition in cell culture in HeLa and human 143B osteosarcoma cells. Ostertag found that retrotransposition in cells took about 48 hours to get started and that it progressed in a linearly increasing fashion for about 10 days (Ostertag et al., 2000). Later, the EGFP cassette became popular in the field because the readout was easy. You looked for green cells under a fluorescent microscope, and you could count the green cells by fluorescent cell sorting, or FACS, after a few days or one week (Figures 17.1 and 17.2). The *EGFP* cassette with the TK poly A signal remains a popular readout for retrotransposition (Ostertag et al., 2000).

At the same time, Eric began work on a mouse model of L1 retrotransposition. He made three different transgenes using the most active human L1 available at the time, $L1_{RP}$. This was an element that had inserted as a full-length L1 into the RP2 gene and knocked out the activity of the gene, causing retinitis pigmentosa, an eye disease, in a young boy (Schwahn et al., 1998). In one transgene, L1 was transcribed using a mouse RNA polymerase II large subunit (pPolII) promoter. In the second, the L1 used its endogenous promoter without a heterologous promoter. In the third, the L1 contained two missense mutations in the ORF1p region that previously were shown to completely abolish activity (the negative control line). He found intronless *EGFP*, a clear sign of retrotransposition, in sperm fractions of both pPolII and endogenous promoter-driven lines of mice but no sign of retrotransposition in the negative control line.

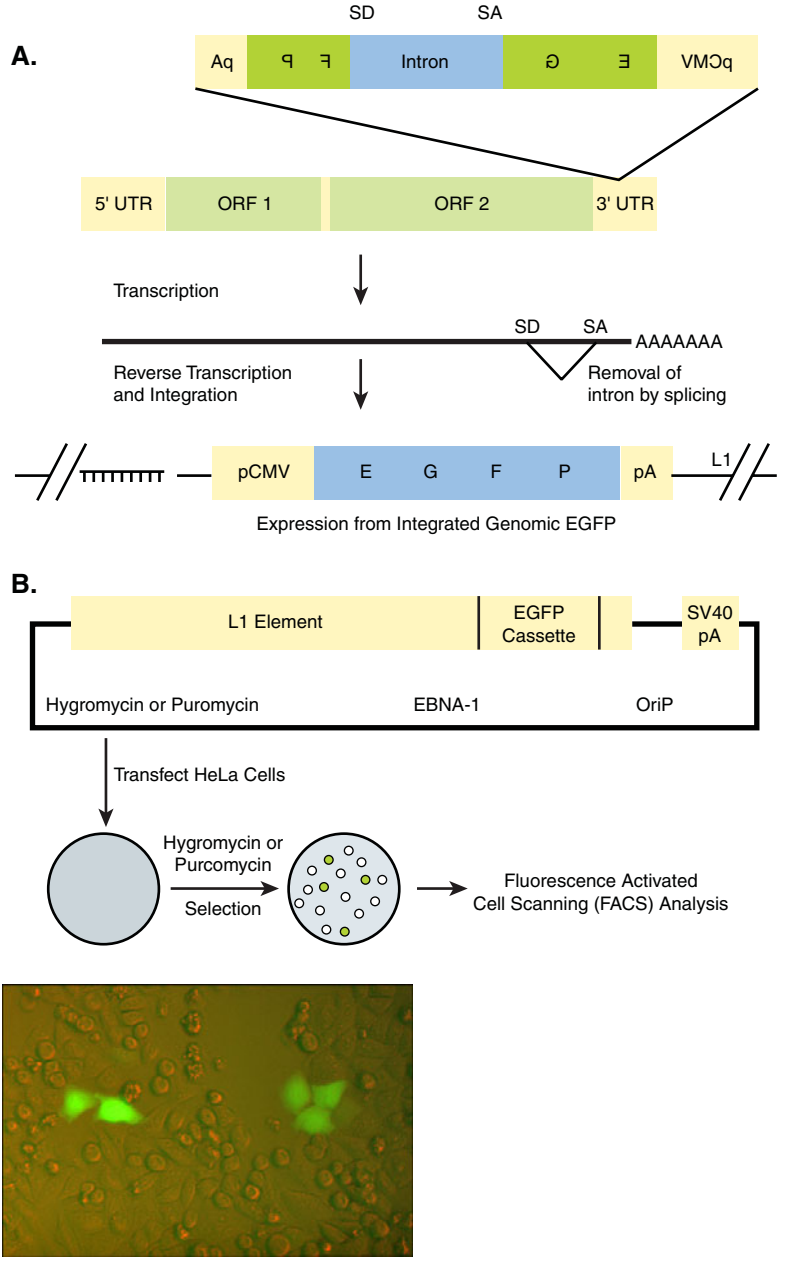


Figure 17.1 A retrotransposition assay using the screenable marker, enhanced green fluorescent protein, EGFP (A) An *EGFP* gene replaces *neo* in the retrotransposition cassette. (B) EGFP-positive cells with retrotransposition events are detected as green cells (light shaded cells in the figure) in fluorescent-activated cell sorting (FACS). (© 2000 Oxford University Press)

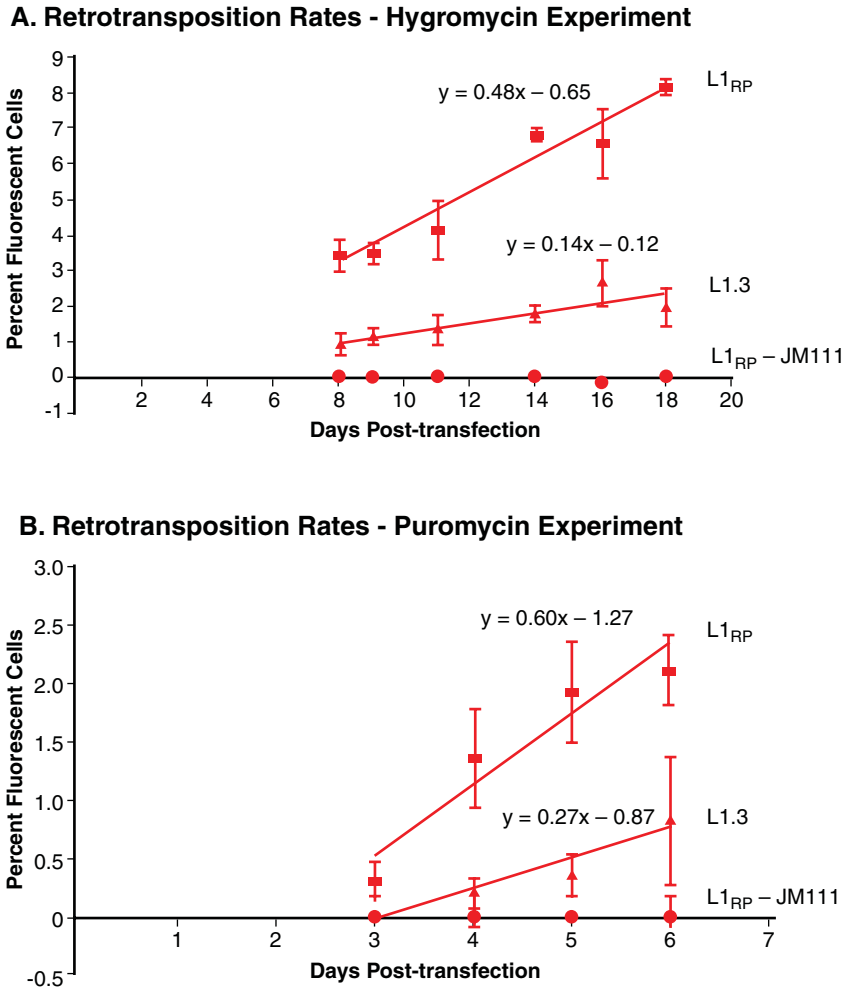


Figure 17.2 Characteristics of *EGFP*-based retrotransposition assay. With hygromycin selection of 143B human osteosarcoma cells, the assay begins on day eight because hygromycin killing takes roughly one week. Puromycin kills cells much more quickly, so the assay begins at day three. Note that $L1_{RP}$, a full-length L1 isolated from a patient with retinitis pigmentosa (RP), is roughly three times more active than L1.3, an L1 isolated from the mother of JH-27. JM111 mutant has two changes in the amino acid sequence of ORF1 that kill retrotransposition activity. (© 2000 Oxford University Press)

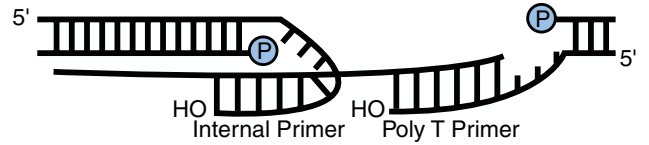
Retrotransposition events were present in roughly 1 in 70 mice carrying the pPolII promoter and in about 1 in 200 mice carrying the endogenous L1 promoter only. New insertions were characterized in

two mice from the pPolII promoter line, and they had all the characteristics of natural L1 insertions. Eric showed that L1 expression occurs predominantly in testis and ovary and that some insertions occur in male germ cell development. This was the first successful transgenic mouse model of L1 retrotransposition, and it was Eric's second big advance. The work was published in *Nature Genetics* in late 2002 (Ostertag et al., 2002).

One of his other two coups came a bit earlier and the other a bit later. They involved some impressive creative thinking on his part. In my mind, because of the creativity he demonstrated, they were even more impressive than his mouse model work. At the time, a major question in the L1 field was, "What is the mechanism that produces inversions of L1 sequence upon retrotransposition?" Since the early 1980s, it was known that about 25% of L1 insertions in humans and mice contained inversions of L1 sequence. These inversions often occurred near the 3' end of the L1 and always flipped whatever 5' end was present. Thus, starting from the 3' end, the L1 sequence might be 3'-G, F, D, E-5', flipping sequences D and E. The inversion never occurred on the 3' side of the L1 sequence, that is, 3'-F, G, E, D-5' was not seen, and no inverted L1 sequence was ever present in the middle of an insertion. Another aspect of L1 inversions was that they could also result in small deletions or duplications of L1 sequence. The mechanism producing an inversion was a big mystery that had eluded all investigators in the field for a long time. At nearly every seminar I gave on the L1 work of the lab, someone, usually a senior investigator, would ask about the mechanism of inversion formation, and I would be forced to confess my ignorance.

Then Ostertag cracked the case! In looking over insertion sequences on his computer, he found something very interesting! Sequence of the L1 RNA trailing the start of the inversion sequence was often complementary for 2–4 nucleotides with the sequence of endogenous DNA prior to the breakpoint of the inversion (Figure 17.3). Statistical analysis indicated that this complementarity was highly unlikely to occur by chance. And to top it off, this complementarity had a precedent. It was reminiscent of the complementarity of the RNA poly A tail at the 3' end of the L1 sequence with T residues adjacent to the endonuclease nick site.

A



Accession #	1	2	3	4	5	6	1	2	3	4	5	6
AC004053	G	T	T	T	T	T	T	T	C	T	T	A
Z84814	T	T	T	G	A	T	T	T	T	T	T	G
Z95325	A	A	C	A	A	A	T	T	T	T	T	G
AC004491	T	A	T	A	T	A	T	T	T	C	T	T
AC002122	G	T	A	A	T	T	T	T	T	A	A	A
AC006131	T	T	T	C	C	T	T	G	T	T	T	T
Z70758	A	C	C	T	A	A	T	C	T	T	C	A
AC004883	T	T	G	T	A	T	C	T	A	T	T	A
AC007486	G	G	T	T	T	T	T	T	T	T	T	T
AL034425	C	T	G	T	C	G	T	T	T	T	G	A
AL030998	A	G	T	A	A	C	T	T	T	T	C	T
AC004220	A	G	G	G	G	A	T	T	C	T	T	C
AF036938	C	T	T	A	A	T	T	T	T	T	T	T
AL031117	C	T	G	G	G	A	T	T	T	T	A	C
A023284	T	A	T	G	A	T	T	T	T	T	T	C
U09115 (LRE2)	T	C	G	T	C	G	T	T	T	A	G	T
AL022153	C	C	T	T	A	G	T	T	T	T	T	T
Total/17	14	15	13	11	2	6	16	15	14	14	11	7

B

Internal Primer Poly T Primer

Position	r	p-value	Position	r	p-value
1	14	1.14E-06	1	16	3.02E-09
2	15	7.43E-08	2	15	7.43E-08
3	13	1.24E-05	3	14	1.14E-06
4	11	6.25E-04	4	14	1.14E-06
5	2	.95	5	11	6.25E-04
6	6	.23	6	7	.11

Figure 17.3 Twin-priming hypothesis: Complementarity of the primers. (A) The postulated internal primer and the poly T primer were analyzed for complementarity to their predicted binding sites on the L1 RNA. The first six nucleotides, numbered from the 3'-hydroxyl, are listed. Nucleotides are highlighted in gray (blue in e-book) if they are complementary to the corresponding nucleotide on the L1 RNA. The last row lists the number of complementary nucleotides at each position, out of a possible total of seventeen. (B) The number of matches (r) at each position and the corresponding P-values, representing the likelihood of obtaining r matches or greater by chance alone. (© 2001 Cold Spring Harbor Laboratory Press)

Ostertag then postulated that a second target primed reverse transcription (TPRT) reaction was occurring on the top strand of DNA. He reasoned that when this second strand reverse transcript was resolved, it would lead to an inversion that would fit all the characteristics of those observed in nature. He postulated that sometime during first strand reverse transcription on the bottom strand of genomic DNA, a nick is made in the top strand of DNA, and a second

molecule of ORF2p begins reverse transcribing the L1 RNA just in front of the first ORF2p (Figure 17.4). He also found something interesting about the resolution of this process.

There was usually complementarity of the inversion end (now internal in the final L1 insertion) and the 5' end of the uninverted sequence. This complementarity suggested that non-homologous end joining (NHEJ) was responsible for the completion of the process. When this complementarity occurred somewhere within the growing reverse transcription products, sequence was likely chewed back to that key point and ligation occurred (Figure 17.4). If the reverse transcription of the bottom or first strand continued beyond the start of the reverse transcription of the second strand, then the result would be a duplication of some of the L1 sequence. Thus, Eric's "twin-priming" hypothesis fit all the observed data, and it is now the generally accepted mechanism for inversion formation (Ostertag and Kazazian, 2001). However, this mechanism was a surprise to me at the time, and it was probably a surprise to the rest of the field!

Eric's second intellectual coup came when he solved a mysterious human insertion in the alpha-spectrin gene that caused a hereditary elliptocytosis, an autosomal dominant red cell abnormality, in a family. His analysis clearly showed that the SVA element that had been described a few years earlier was another non-autonomous retrotransposon, like Alu, that was mobilized through a *trans*-effect of L1.

Eight years earlier, an insertion of single-copy, 700-nucleotide sequence into the alpha-spectrin gene had been reported (Hassoun et al., 1994). This sequence was clearly a retrotransposition event because it was surrounded by target site duplications. The origin of this insertion had remained a mystery. How could an insertion of non-repetitive sequence occur? All known retrotransposons were repetitive sequence. However, by the time Ostertag began to study this phenomenon, the human genome draft sequence (HGWD) had been published and was online. Upon searching the database with the insertion sequence, Eric found that it was located downstream or 3' to a full-length SVA sequence on another chromosome. He immediately thought—aha, this looks like a typical 3' transduction event, similar to the 3' transductions seen in L1 insertions! SVA must have a weak poly A signal just like L1. In addition, there were two other interesting features of this insertion. First, the insertion actually

contained an inversion of a portion of the single-copy 3' transduced sequence, and the 5' truncation had been so severe that no SVA sequence was included in the insertion. The inversion sequences fit Ostertag's "twin-priming" model, and all of the evidence strongly suggested that this was an SVA insertion mediated by the L1 reverse transcriptase in which 1) 3' transduction, 2) inversion, and 3) severe 5' truncation had occurred (Figure 17.5). Eric went on to track the lineage of the precursor SVA in the genome and to characterize other SVA elements in the genome. However, this retrotransposition event is particularly important because 1) it is illustrative of three different aspects of L1 retrotransposition occurring in one event, and 2) the retrotransposed sequence is not obviously related to any retrotransposon (Ostertag et al., 2003).

I don't know how Ostertag was able to maintain his high level of productivity from 1999 through 2002. He was going through a difficult divorce that must have weighed on him. Yet in the lab, he always maintained an upbeat attitude. He even talked me into joining him and Brouha in twice a week workouts at the University gymnasium. These were arduous, and Eric kept getting stronger and stronger. Perhaps it was the combination of the lab work and the gym workouts that kept him going with so much optimism.

An important aspect of Ostertag's work on these last two projects is that he initiated them and was responsible for their seminal ideas. Yes, we had discussed potential mechanisms to account for L1 inversions, but none of those ideas included priming of reverse transcription on top-strand DNA by a second ORF2p molecule. Many individuals in the field had considered the problem of L1 inversions, but when Ostertag described his model and the data backing it up, it was the first time I had heard a solution that fit all the data. Likewise, Eric solved the problem of the origin of the single-copy insertion in the alpha-spectrin gene independently. This one was simpler because the offending SVA was present in HGWD. However, he still made the connections to 3' transduction, inversion, and severe truncation without my intervention. Needless to say, I was very pleased with Ostertag's independence and creativity.

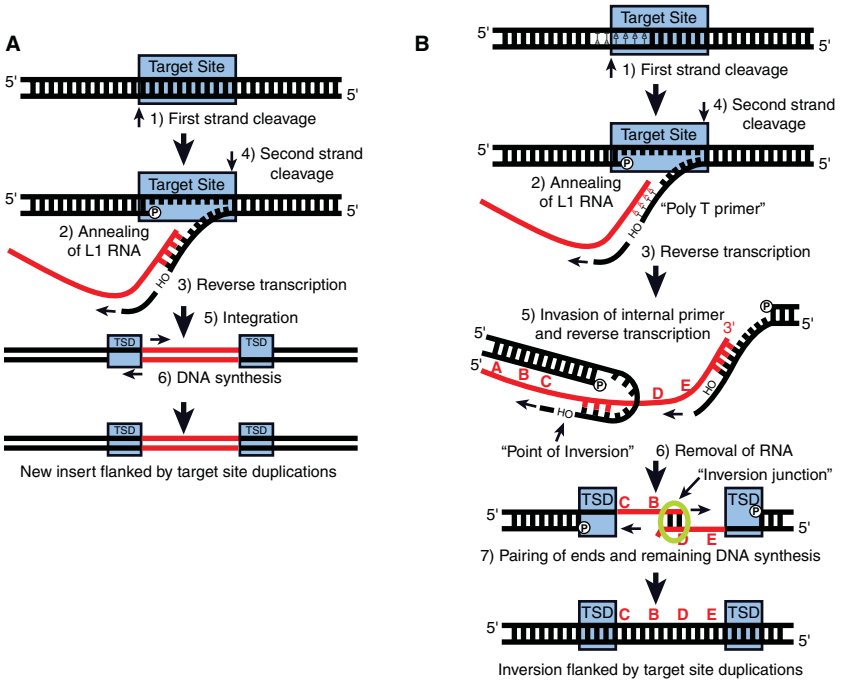


Figure 17.4 Target primed reverse transcription and twin priming. (A) Schematic of target primed reverse transcription (TPRT), based on in vitro studies of the R2 element from Bombyx mori (Luan et al. 1993). TPRT involves (1) Cleavage of the first DNA strand at the target site by the retrotransposon endonuclease (EN). (2) Annealing of retrotransposon RNA at the nick. (3) Reverse transcription from the free 3'-hydroxyl by the retrotransposon reverse transcriptase (RT). (4) Cleavage of the second DNA strand. (5) Integration at the double-strand break. (6) Removal of RNA and completion of DNA synthesis. The TPRT process produces target site duplications (TSDs) at the flanks of the newly integrated retrotransposon. (B) Twin priming is a modification of TPRT with the following steps: (1) The L1 EN cleaves one strand of its DNA target site, producing the poly T primer. (2) The poly(A) tail of the L1 RNA anneals on the poly T primer. (3) L1 RT uses the L1 RNA as a template and the poly T primer to initiate reverse transcription. (4) The L1 EN cleaves the second DNA strand before reverse transcription has been completed, producing the internal primer. (5) The internal primer invades the L1 RNA and primes reverse transcription, likely by a second ORF2p molecule. (6) The RNA is removed from the RNA/cDNA structure. (7) The single-stranded cDNAs pair at a region of limited complementarity, and the remaining DNA synthesis is completed. The entire process results in an L1 inversion flanked by perfect target site duplications. The L1 RNA sequence is represented by 5'-A-B-C-D-E-3'. After the inversion, the insertion sequence is 5'-C-B-D-E-3'. (© 2001 Cold Spring Harbor Laboratory Press)

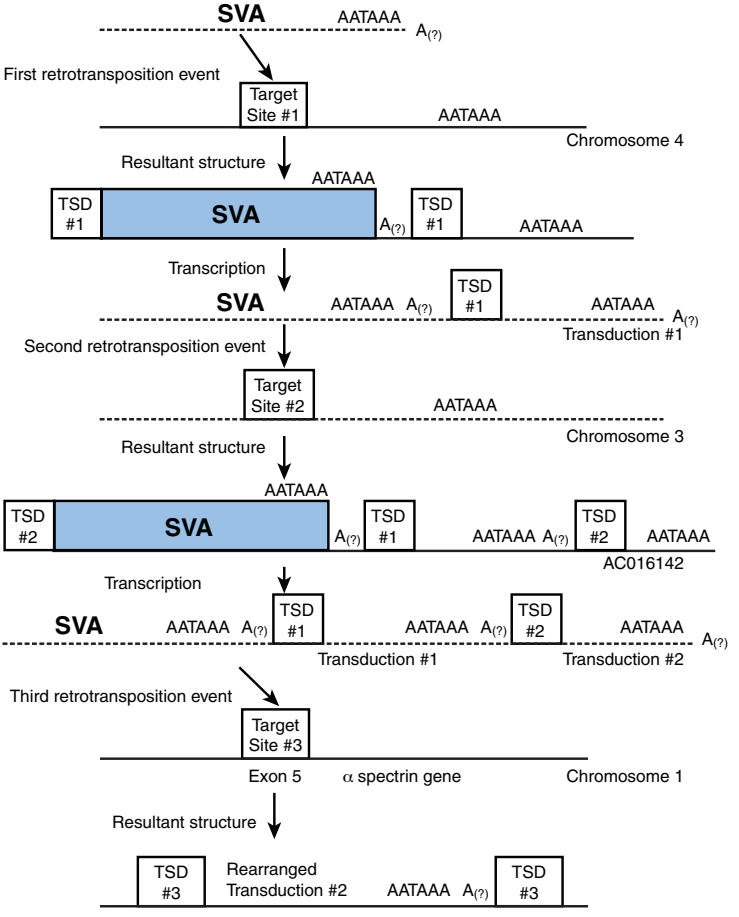


Figure 17.5 The sequence of SVA retrotransposition events. A full-length SVA element of unknown origin retrotransposed into the empty-site sequence on chromosome 4, represented on sequences AC037439 and AC068352 of HGWD. During subsequent transcription of the new SVA at this site, the SVA poly A was bypassed in favor of a downstream poly A, producing the first transduction event. The full-length SVA and transduction 1 inserted into a new genomic location at target site 2 on chromosome 3 to produce the sequence found in genomic sequence AC016142. During subsequent transcription of the new SVA on chromosome 3, both the SVA poly A and the poly A from transduction 1 were bypassed in favor of a poly A farther downstream, producing the second transduction event. The transcript containing a full-length SVA element and both transduction events produced an insertion into target site 3 on chromosome 1, which is in exon 5 of the α-spectrin gene. However, the integration process resulted in a structure that was 5' truncated and inverted compared with the precursor, a common process in L1-mediated retrotransposition. RNA is represented by a dashed line; DNA is represented by a solid line. (© 2003 Elsevier)

Since I've known him, Ostertag has always had an entrepreneurial spirit. So it is no surprise to me that he has gone on to cofound several biotechnology companies, one of which has a core technology related to mobile DNA. Indeed, that company is called Transposagen.

This page intentionally left blank

18

The independent Canadian

John Goodier entered the lab in 1997. John was a native of Toronto, Canada, who had done his Ph.D. in Newfoundland with Willie Davidson, an evolutionary biologist, spent some time in marine biology in Japan, and did a short postdoc on Alu expression with Rich Maraia at the NIH. Goodier was fiercely independent, to the point of discussing his work sparingly at lab meetings until a particular project was nearly finished. Then he would take up to two hours in lab meeting to thoroughly present the work that was always complete and very interesting. It became a standing joke in the lab that no one knew what exactly John was up to at any given time. John was generally quiet, but he could express his displeasure on occasion. He also did not like to have his work discussed at meetings before it was ready for submission because of concern that he would be "scooped." Thus, Goodier's views on how one does science and communicates it were a bit different from the somewhat more open approach I prefer. Over the years, one could say that we've agreed to disagree occasionally because we respect each other's science.

John began working on the biochemistry of retrotransposition, but as I alluded to earlier, he liked to carry a number of projects at the same time. Soon, he and Ostertag were analyzing 3' transduction events using the human and mouse genome databases. Holmes had shown back in 1994 in her work on the insertion causing Duchenne muscular dystrophy that because of the weak poly A signal at the 3' end of L1, the transcript was often not cleaved until it reached a strong poly A signal downstream of the L1. This latter signal could be hundreds or thousands of nucleotides 3' of the L1. In 1999, John Moran had done cell culture assays in which he moved the retrotransposition cassette from the 3' untranslated region of an active human

L1 to the 3' flank of the element. Moran found that there was still active retrotransposition in cultured cells, demonstrating that 3' transduction was not infrequent. His work showed that if an active L1 sits just upstream of a gene or an exon of a gene, that gene or exon could be shuffled by L1 retrotransposition to a new genomic site (Moran et al., 1999).

Indeed, in 2006 Mark Batzer and colleagues reported that 3' transduction of SVA had led to three extra copies of the AMAC gene (two exons and ~1500 nucleotides) in humans and African great apes, and that at least two of these extra copies retained activity in different human tissues (Xing et al., 2006). Because L1 reverse transcriptase is likely required for SVA retrotransposition as shown by Ostertag and others, and Ostertag also showed that 3' transduction of SVA occurred in nature, Batzer's findings were not a huge surprise, but they were still exciting. They remain among the best examples of retrotransposition of 3' transduced sequence leading to new genes. Batzer also showed in that work that SVA 3' transduction events have occurred frequently (in roughly 10% of SVA insertions), and they've led to 53 kilobases of duplicated sequence in the human genome relative to that of the great ape. Thus, SVA transduction events are another mechanism for exon shuffling.

In 2000, Goodier along with Ostertag collected a large number of L1 insertions that had brought along 3' flanking sequences. It seemed that roughly 20–25% of all L1s in the genome had 3' transductions. However, because natural 3' transductions were easy to find in the genome databases, we knew that this time we needed to publish rapidly. Goodier wrote up the paper and sent it to *Nature Genetics*, but although the reviews were not really critical, it was turned down because the reviewers perceived that it lacked novelty and importance. I immediately called Hunt Willard, a good friend and eminent scientist who was the Editor of *Human Molecular Genetics*. Willard said to send him the paper with the *Nature Genetics* reviews. We did, he accepted the paper within a few days, and it was published within a month (Goodier et al., 2000). At the same time, a very similar paper appeared from Jef Boeke's lab by Oksana Pickeral in another journal (Pickeral et al., 2000). We were fortunate that the "old boy network" had worked so well!

Goodier then turned his attention to searching the L1s in the mouse genome, and he immediately found another subfamily of mouse L1s. This new subfamily was distinct in its 5' untranslated monomer sequences from the previously known F, A, and T_F subfamily monomers. His analysis found that this subfamily contained some 1,500 full-length elements of which about 500 had two intact ORFs. Goodier performed the retrotransposition assay on three members of this new mouse subfamily and found that all three were active. One of these is the most active natural mouse L1 known to date. He also showed after studying a number of these new elements in different mouse strains that this subfamily is also expanding rapidly in present day mouse strains. After Goodier analyzed a number of L1s from the A subfamily he could estimate that the number of active L1s in the diploid mouse genome (the combination of active L1s in the A, T_F, and new subfamily) is ~3000, a significantly greater number that the 30–60 estimate of Sassaman and the later 80–100 estimate of Brouha for the number of active L1s in the diploid human genome (Table 18.1).

Table 18.1 The estimated number of active human and mouse L1s. The mouse has three active subfamilies of L1s, while humans have only one. Note that estimates of the number of active mouse L1s relied on assays of a relatively small number of T_F, A, and G_F L1s.

	Human	**Mouse**
Active subfamily	Ta	T_F, A, G_F
# full-length subfamily L1s/diploid genome	160–240	3000, 6500, 500
Estimated % capable of retrotransposition	45–50%	64% (7/11) T_F 14% A, 80% G_F
Estimated # active L1s in diploid genome	80–100	2000 T_F, 800 A, 400 G_F

When it came time to name this subfamily, Goodier suggested G_F, stating that the monomers had about 70% sequence similarity to the F monomers, and G was the next letter after F in the alphabet. Of course, we all thought that he was naming the new subfamily G for Goodier. We still kid him about it! Goodier published this work in 2001 in *Genome Research* without the pressure he had felt with the 3' transduction paper (Goodier et al., 2001).

After his postdoc, Goodier continued in the lab as a Senior Research Investigator and recently a Research Assistant Professor. He obtained considerable independent research support through a two-year research grant from the NIH and a three-year grant from the Department of Defense. He writes very well, and he continues to apply for independent support on the biochemical and cell biological analysis of the L1 retrotransposon life cycle.

In 1996, after Moran had developed the cell culture assay for retrotransposition, I thought the assay would be a bonanza for figuring out the biochemistry of the L1 life cycle. All we'd have to do is transfect cells and look for where the L1 RNA went, what happened to the ORF1p and ORF2p proteins, and so on. We'd have the whole L1 story solved in 5–7 years. Well, at this writing it's 14 years later, and although there has been progress on the biochemistry, it has been slow and halting. L1 RNA is very difficult to find even in transfected cells, and the ORF2p has been nigh onto impossible. In 2002 and 2003, Goodier finally succeeded in seeing ORF2p but in a very artificial system.

Everyone in the field knew that ORF2p encoded endonuclease and reverse transcriptase activities, but no one had seen the protein. Kurt Engelke, a friend of Goodier's, had a vaccinia virus/T7 RNA polymerase system that could greatly amplify production of any protein in cultured cells. Using Kurt's system and with Kurt's help, Goodier finally saw ORF2p as a predominantly cytoplasmic protein, with a nucleolar distribution in the nucleus of a small subset of cells. ORF2p present in the nucleolus! That fit with Buzdin's very recent surprising finding. Anton Buzdin, a young Moscovite, had done an analysis of the HGWD and found a number of chimeric DNA sequences in which the 5' end usually encoded a small nucleolar RNA, and the 3' end was the 3' end of L1 sequence, as shown in Figure 18.1 (Buzdin et al., 2003). He postulated that the L1 reverse transcriptase was copying the L1 RNA when it jumped or switched templates to a small nucleolar RNA and copied it. Wow! That was an unexpected surprise! These whole DNA copies were then inserted back into the genome. However, given the added DNA was derived from a nucleolar RNA, it wasn't clear how the L1 reverse transcriptase found its way into nucleoli. Now Goodier had found that in some cells L1 ORF2p with its reverse transcriptase could enter nucleoli. If these ORF2 proteins were in the process of reverse transcribing L1 RNA, they could then

occasionally switch to a small nucleolar RNA template (Goodier et al., 2004). A potential solution for a thorny problem!

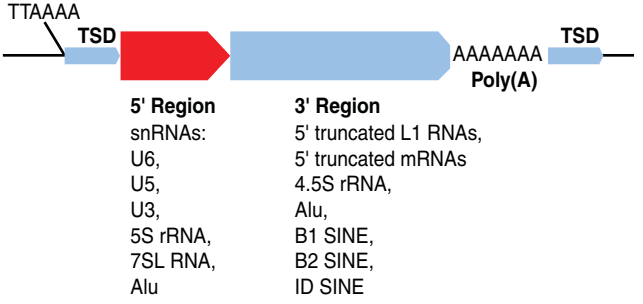


Figure 18.1 Schematic representation of the chimeric retrogenes of humans and mice found in public databases. Note that the 3' ends are L1s, Alus, or 3' ends of mRNAs in humans and L1 and various SINEs in mice. The 5' ends are the DNAs of snRNAs (small nuclear RNAs), various U RNAs involved in splicing, 5S ribosomal RNA, 7SL RNA, and Alu RNA. The chimeric retrogenes end in a poly A tail and are surrounded by target site duplications (TSDs). TT/AAAA at the 5' end represents the insertion site mediated by L1 endonuclease.

Goodier then began to study the location of the L1 proteins in human cells. I've told you something about ORF2p, but I need to provide some further information about ORF1p, a 40kDa protein found mainly in large cytoplasmic foci. Sandy Martin at the University of Colorado Health Sciences Center has extensively studied ORF1p and shown that it exists as a trimer (three molecules joined together) with a dumbbell-like shape (Martin et al., 2003). She and others also had earlier provided evidence that it had RNA binding activity (Hohjoh and Singer, 1996; Kolosha and Martin, 1997; Kolosha and Martin, 2003; Kulpa and Moran, 2005). Martin and Bushman found that ORF1p acted as a nucleic acid chaperone protein, aiding nucleic acid strand transfer steps during reverse transcription (Martin and Bushman, 2001). Recently, the structure of ORF1p has been solved by X-ray crystallography, and its trimer, dumbbell configuration has been verified (Khazina and Weichenrieder, 2009; Weichenreider, 2010, personal communication). Goodier had antibodies to both ORF1p and ORF2p, but outside of the artificial vaccinia virus system ORF2p was still very tough to visualize. Concentrating on ORF1p, he found it in large cytoplasmic foci, as had been seen many times previously, but

he was able to localize these sites to stress granules, recently described granules that increase in frequency when cells are stressed. Some proteins known to localize to stress granules also interacted with ORF1p. John suggested in his paper published in *Molecular and Cellular Biology* that targeting ORF1p, and possibly the L1 ribonucleoprotein particle (RNP), to stress granules was a mechanism for controlling retrotransposition and its associated genetic and cellular damage (Goodier et al., 2007). However, in this work, he was unable to localize L1 RNA, derived either from a transfected plasmid or from endogenous L1s. He still suggested that the L1 RNA, ORF1p, and other components of the L1 RNP were being trapped in stress granules and removed from the retrotransposition cycle.

Very recently, Goodier and the group of Nicolas Gilbert, a trainee of Moran and now a principal investigator at the CNRS in Montpellier, France, have obtained the important result that the L1 RNA can be localized in the cytoplasm of transfected cells with ORF1p and ORF2p. Gilbert sees ORF2p by tagging ORF2 and applying an antibody to the tag (Doucet et al., 2010). Goodier visualizes ORF2p using an antibody to native ORF2p. Gilbert sees ORF2p signal in a majority of transfected cells, while Goodier is only able to visualize the protein in a small minority of those cells, as shown in Figure 18.2 (Goodier et al., 2010). These are very important results because Deanna Kulpa with John Moran had earlier shown that ORF2p reverse transcriptase activity is present in isolated L1 RNPs (Kulpa and Moran, 2006). Now, the ORF2p has been seen in human cells localized with L1 RNA and ORF1p, strongly suggesting that these sites are RNP particles. However, it has not yet been possible to find ORF2p co-localized with L1 RNA and possibly ORF1p in the nucleus where the protein(s) and RNA are thought to be carrying out the retrotransposition process. Thus, there has been recent progress on the biochemistry and cell biology of L1 retrotransposition, but there is still much work to be done. Considering these studies are quite new, it will be interesting to see where they lead in the future.

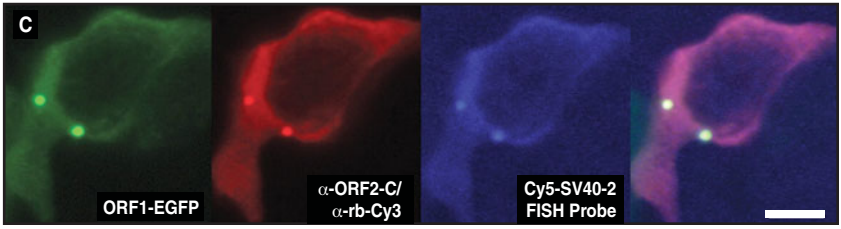


Figure 18.2 Localization of ORF1p, ORF2p, and L1 RNA together in the cytoplasm of human embryonic kidney 293T cells. Confocal micrograph of *EGFP*-tagged ORF1p, ORF2p detected by an antibody to the C-terminus of ORF2p, and L1 RNA detected by Cy5-SV40-2 FISH (fluorescent in situ hybridization) probe. (Used with permission from Goodier et al; © 2007, American Society for Microbiology.)

This page intentionally left blank

19

The musician scientist

Every year since joining Penn, I was asked to lecture to the Clinical Pathology residents on hemoglobin disorders. In 1999, one of the students was Nina Luning Prak, another very smart M.D.-Ph.D. from Penn who had done her Ph.D. in Immunology with Martin Weigert at Fox Chase. In early 2000, Nina came to me with the proposition that she would do the two-year research stint that was part of her residency training in my lab. She was very lively and highly motivated, so I agreed. Nina also impressed me with her enthusiasm for classical music, about which I am also passionate. She played the violin in a local string quartet that got together informally once a week. It was also great to hear some good classical music in the lab after all the pop and other music played by previous lab members. I could also count on Nina to give me wonderful classical CDs as presents on special occasions.

As I expected, Nina was a hard worker. She made a valiant though unsuccessful effort to obtain retrotransposition of a marked L1 from one chromosome location to another. Note that Moran's assay had demonstrated retrotransposition from an episome (an extrachromosomal piece of DNA) into chromosomal DNA but not from one chromosome to another. (Let me point out how many projects initiated in my lab have failed—perhaps 50% or more. Luckily, my trainees have usually picked up new ones that were successful. Perseverance in science needs to be tempered by flexibility.)

She also worked with Ostertag in the attempt to make a retrotransposing, transgenic mouse. As her two years in the lab were ending, she was getting close to examining one such transgenic carrying a human L1 driven by a pPolII (RNA polymerase II) promoter. The L1 was marked with an *EGFP*-cassette. However, before she had the data she moved into her own lab in the Department of Pathology and

Lab Medicine at Penn. Soon after her move, she called me and was very excited. She had evidence of retrotransposition in the founder mouse carrying the transgene. His testes, specifically his seminiferous tubules, were glowing green with *EGFP* under fluorescent light (Figure 19.1). "You've got to see this," she exclaimed. Indeed, the testes were green with *EGFP*, and the insertion was passed on to roughly 30% of the founder's offspring. Thus, here was evidence for a very early embryonic insertion event that was heritable. Nina's work presaged the work of Babushok and Kano in the lab, showing that embryonic insertions are not uncommon, but this was the first demonstration of an embryonic event, and it had occurred very early in development from an L1 that was driven by a heterologous promoter. Nina published her paper on tracking an embryonic retrotransposition event in the *Proceedings of the National Academy of Science* (Prak et al., 2003).

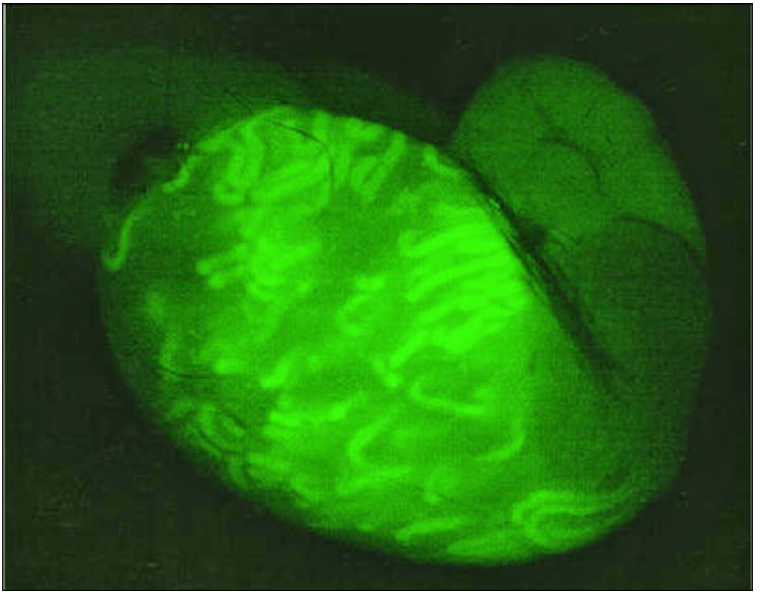


Figure 19.1 *EGFP* fluorescence in the seminiferous tubules (light shading of tubules) of a founder mouse carrying an active human L1 marked with an *EGFP* retrotransposition cassette driven by a CMV (cytomegalovirus) promoter.

Nina also loved to teach, and she mentored an undergraduate, Alex Farley. Alex did a nice piece of work suggesting that more active L1s produced insertions that were longer (less 5' truncated) than less active ones (Farley et al., 2004). At this writing, Nina is on the faculty at Penn in the Department of Pathology and Lab Medicine, and she continues her research and highly-regarded teaching of medical and graduate students.

This page intentionally left blank

20

Young ladies in the back bay

This chapter deals with two inseparable young ladies who shared the bay at the far end of the lab. One was Bolivian, Maria del Carmen Seleme, called Marie, and the other was Russian, Daria Babushok, called Dasha. Let me begin with Marie Seleme.

I first met Seleme at the railway station in Avignon, France. In 1999, Marie was finishing her Ph.D. with Alain Bucheton, a noted expert on non-LTR retrotransposons in Drosophila, and applied to me for a postdoc position at Penn. Seleme had an interesting background. She was born and raised in La Paz, Bolivia, but attended university in Barcelona, Spain, and then continued her graduate studies with Bucheton in Paris. When Bucheton moved to the Institut de Genetique Humaine in Montpellier, Marie moved with his group. Bucheton wrote her a strong recommendation, but I still wanted to meet her, if possible. After I attended an unrelated meeting in Paris, my wife and I decided to travel to Provence by train for a few days of vacation. We would stay in a nice rural hotel just outside Avignon, rent a car, and tour the Luberon. I checked the map and found that Montpellier was not far from Avignon. Seleme and I arranged to meet for an interview at the train station in Avignon; her train was on time, and we had a delightful time over coffee in the station. I immediately accepted her for the position, and she returned to Montpellier.

Seleme was a black-haired, intense young lady who worked very hard, but for a long while it seemed that each of her efforts would meet a wall or yield unpublishable results. She did collaborate on a couple projects that were published by other groups, but those weren't fully satisfying either to her or to me. Finally, in 2003 after she had been in the lab for over three years, she came to me for a

heart-to-heart discussion. She needed a successful project, but we both wanted it to have real significance for the field.

I was interested in polymorphisms of human L1 elements (differences in the L1s found in different people)—not only presence/absence polymorphisms, but also nucleotide polymorphisms. Both types of polymorphism had been observed previously. Presence/absence polymorphisms of the full-length L1s that hybridized with the JH-27 oligomer had been found by Dombroski and were also well known from the work of Tony Furano and Mark Batzer. We had known since 1990 that nucleotide polymorphisms occurred in human L1s. In trying to find the precursor of the JH-27 insertion into the factor VIII gene, we had found one allele in the commercial phage library, L1.2A, and a second allele, L1.2B, as the actual precursor in the phage library from the mother of JH-27. In Moran's original cell culture assay and all the subsequent assays in our lab in Philly, we had used L1.2A, with two nucleotide substitutions from the JH-27 precursor, L12B. Then in 2002, Sheila Lutz of the Moran lab tested L1.2B in cell culture and found that it was roughly ten times more active than L1.2A (Lutz et al., 2003). First Lutz and then Alex Farley of our lab showed that most of this difference in retrotransposition activity was due to a single nucleotide substitution (Lutz et al., 2003; Farley et al., 2004).

This was the backdrop for Seleme's study. Because of the L1.2 alleles, we knew that small changes in the DNA sequence of an L1 could alter its activity greatly. Of Brouha's six "hot" full-length elements, there were three that could readily be studied in a number of human beings. Marie was game to take on this project, but there were serious obstacles to overcome. She needed a reliable PCR across the 6-kb elements that would produce very few errors in the sequence of the element due to the PCR procedure. Seleme tried a number of different DNA polymerases for this long range PCR and finally hit on a polymerase called Phusion that gave the 6+kb products with very few sequence errors. When she resequenced a number of cloned Phusion 6kb L1 products, she found only three sequence errors in 100,000 nucleotides amplified by PCR and sequenced. Then she had to carry out the PCR of the three "hot" L1s from 160 diploid genomes representing four human populations. After that all 480 products needed to be cloned into the pCEP4 plasmid. The retrotransposition assay was carried out in triplicate using the *EGFP* assay used by

Brouha. That meant she would need to carry out almost 1,500 assays. This was very arduous work, not only because of the large numbers of samples, but also because a number of samples needed to be repeated, and some just would not yield any usable data. After two more long years, Seleme had the data and wrote up her paper, "Extensive Individual Variation in Human L1 Retrotransposition Activity Leads to Significant Genetic Diversity," for the *Proceedings of the National Academy of Science.* It was published in early 2006 (Seleme MdelC et al., 2006).

For each of the three "hot" L1s, she found one previously uncharacterized allele in every three to five genomes, including some with nonsense and insertion/deletion mutations. We expected some variation in specific L1s but certainly not this much. As usual, another surprise! These mutations would eliminate production of ORF1p or ORF2p. Single or multiple nucleotide substitutions that altered amino acids in one or another of the proteins drastically affected the retrotransposition efficiency of some alleles. One-third of elements were no longer hot, and these so-called cool alleles substantially increased the range of individual susceptibility to retrotransposition events. Adding the activity of the three elements in each individual resulted in a surprising degree of variation in mobilization capability, ranging from 0% to 390% of a reference L1. Marie's data suggested that individual variation in retrotransposition potential makes an important contribution to human genetic diversity (Figure 20.1, also see Figure 20.2 on the fate of full-length L1s upon entering the genome). This was work that made both her and me proud.

But this project took a toll. Soon thereafter, Seleme decided to take a job at the University of Pittsburgh. Her husband, Jean-Hugo, had been unable to alter his visa status so that he could work in the U.S. Pittsburgh promised to help with his visa, and Marie's permanent residency in the U.S. It was appropriate for her to leave Penn, but I still miss her smiling face. I treasure a gift that she gave me after one of her trips home to Bolivia, a very interesting carved wooden chess set from Peru. It featured the Spaniards versus the Incas. The Spaniards had castles, and the Incas had mud huts. The Spaniards had horses, and the Incas had llamas. Although I don't play much chess, this set remains prominently displayed on my desk. Whenever I see it, I smile and think of Marie Seleme.

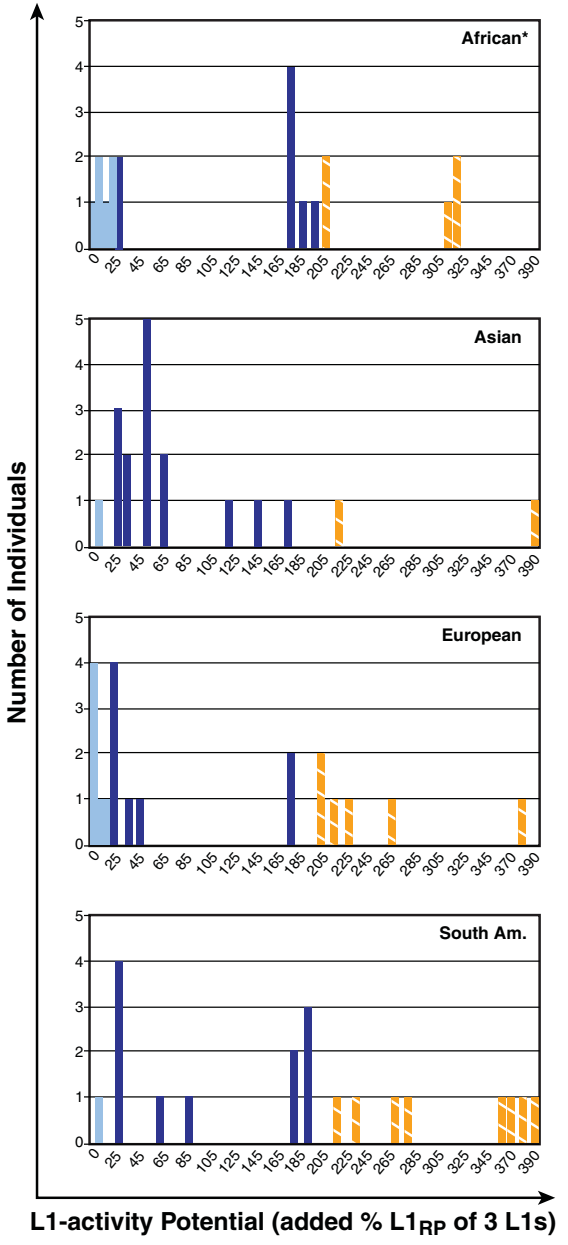


Figure 20.1 Combined retrotransposition potential of three hot L1s per individual in four populations. From 26% (African) to 55% (South American) of individuals per population have a unique L1 activity potential. Light gray (light blue in e-book version), black (dark blue), and hatched (orange) bars represent individuals lacking a hot L1 phenotype (<25%), having an intermediate L1 activity, and having a high L1 activity (>200%), respectively.

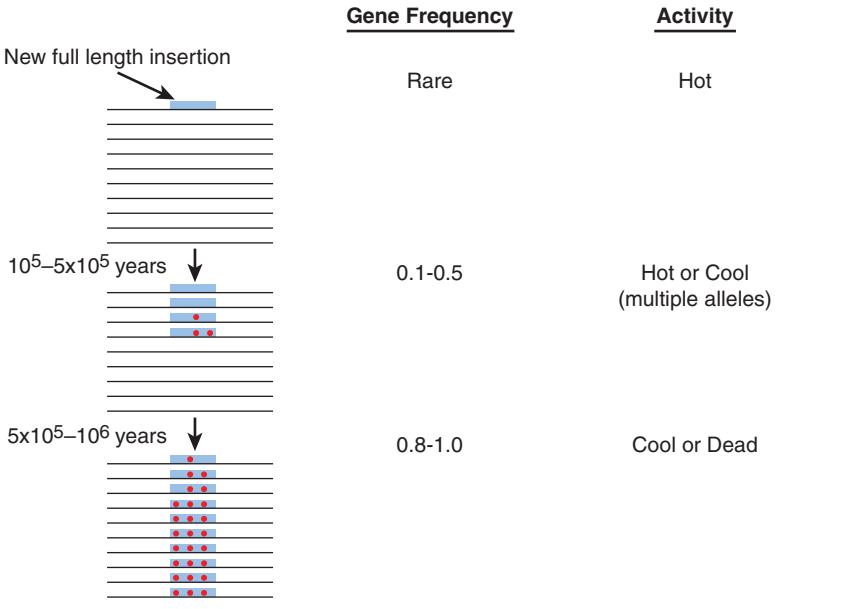


Figure 20.2 Model of the evolution of a full-length L1 insertion in a population. Data presented in (Seleme MdelC et al., 2006) and evidence that hot L1s account for most new insertions (Brouha et al., 2003) suggest that new insertions are derived from hot L1s. Data on alleles of the "hot L1s studied in (Seleme MdelC et al., 2006) indicate that, after a hot L1 reaches an intermediate gene frequency in the population, it has a significant proportion of cool alleles. As an L1 approaches fixation, mutations produce cool alleles and dead alleles. Shaded box, L1 insertion in chromosomes (lines); black dots (red in e-book), mutations.

Then there was Marie's buddy, Dasha Babushok. She too had an interesting background. She was born and raised in Novosibirsk in far Eastern Russia where her father was a physical scientist. Dasha came to Maryland as a young teenager. She attended Bryn Mawr College in a Philadelphia suburb, did very well, and then was admitted into the M.D.-Ph.D. program at Penn in 2000. In 2002, I was acting as an advisor for the Genetics graduate program at Penn, and Babushok came to speak with me. I guess that she enjoyed our meeting because she decided to do a lab rotation with me. I was delighted to have her! In 2003, she came back to the lab for her thesis work. In a word, Dasha was brilliant!

She quickly grasped the L1 field and decided to come up with an independent project. She would find an active L1 in a new yeast species

whose genome was being sequenced, demonstrate its retrotransposition activity in cell culture, and then use the power of yeast genetics to find host factors important for retrotransposition. Granted, this was a risky project, but she gave it a good effort. She found a few full-length non-LTR retrotransposons in the species, but they all either had stop codons in the ORF regions or were otherwise inactivated. After nearly a year of work, she decided to throw in the towel on this project. Another disappointing failure! In the meantime, she was working on a very important aspect of L1 retrotransposition in the transgenic mouse.

Ostertag had made new mice carrying a very active human L1 driven by a heat shock protein-70 (HSP-70) promoter. These transgenic mice had frequent insertions, but disappointingly nearly all were somatic, that is, they had occurred after fertilization and were present in much less than one insertion per cell in the adult mouse. Dasha perfected a published technique called TAIL-PCR to characterize the L1 insertions. At this time, although we knew that Jef Boeke was characterizing insertions in L1 transgenic mice that he had made, only the two characterized L1 insertions from our 2002 *Nature Genetics* paper had been published. Thus, it was important to find out whether new insertions *in vivo* would mimic the L1 insertions found naturally in the genome. TAIL-PCR (Thermal Asymmetric Interlaced PCR) is a complicated series of PCR reactions designed to isolate the unknown DNA sequence flanking a known sequence. In this case, human L1 sequence was the known sequence, and Dasha wanted to find the sequence downstream of the L1 sequence. The problem was that the human L1 insertions in the mice were somatic and present in only one copy per 10 to 100 cells. In other words, they were rare in genomic DNA! But Dasha persevered, like so many of my other students and postdocs. She characterized the flanking sequence of 51 insertions and showed that, as expected, they had a broad genomic distribution without any real sign of preferential insertion sites (Figure 20.3). She also characterized the 5' ends and the complete structures of 33 of the 51 *de novo* events, finding a large number of highly truncated L1s, as over half (27/51) were <1/3 the length of a full-length element. New integrants carried all the structural characteristics typical of genomic L1s, including a number with inversions and deletions. Notably, 13% (7/51) of all insertions contained a short stretch of extra nucleotides at their 5' end, which

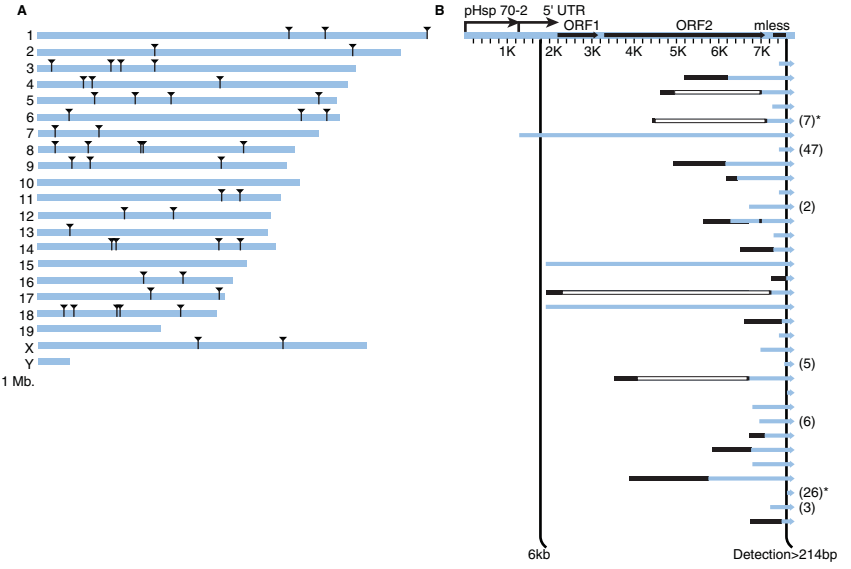


Figure 20.3 Genomic distribution and structural features of de novo insertions. (A) Using TAIL-PCR, 51 de novo integration sites were determined, 48 of which were uniquely mapped to the mm6 assembly of the mouse genome. Depicted are the locations of 47 inserts (one mapped to an unplaced location and is not shown). 1-Mb scale reference is shown at the bottom. (B) The structures of 33 fully characterized de novo inserts. A hypothetical full-length (FL) human L1 insert is shown at the top, with vertical lines indicating the 6-kb endogenous FL element and the 214-bp detection limit in our study. De novo integrants are shown on separate lines, aligned to the FL element. (Direct fragment) Gray (blue in e-book) rightward arrow; (inverted fragment) black rectangle; (deletion of sequence in inverted elements) white rectangle; (extra 5' nt) numbers in parentheses; (dual inversions) asterisk. Three elements mobilized ~6kb of sequence; one is FL, and two are nearly FL. (© 2006 CSHLP)

she postulated were the result of template-jumping by the L1-encoded reverse transcriptase to genomic DNA flanking the insertion site. Here, Dasha showed her brilliance. She came up with a unified model of L1 integration that explains all of the characteristic features of L1 retrotransposition, such as 5' truncations, inversions, extra nucleotide additions, and 5' boundary and inversion point micro-homologies (Figure 20.4). I still like this model of integration very much because it also fits the data on integration that Shawn Christiansen and Tom Eickbush proposed from their biochemical studies of integration of the R2 retrotransposon of *D. melanogaster*

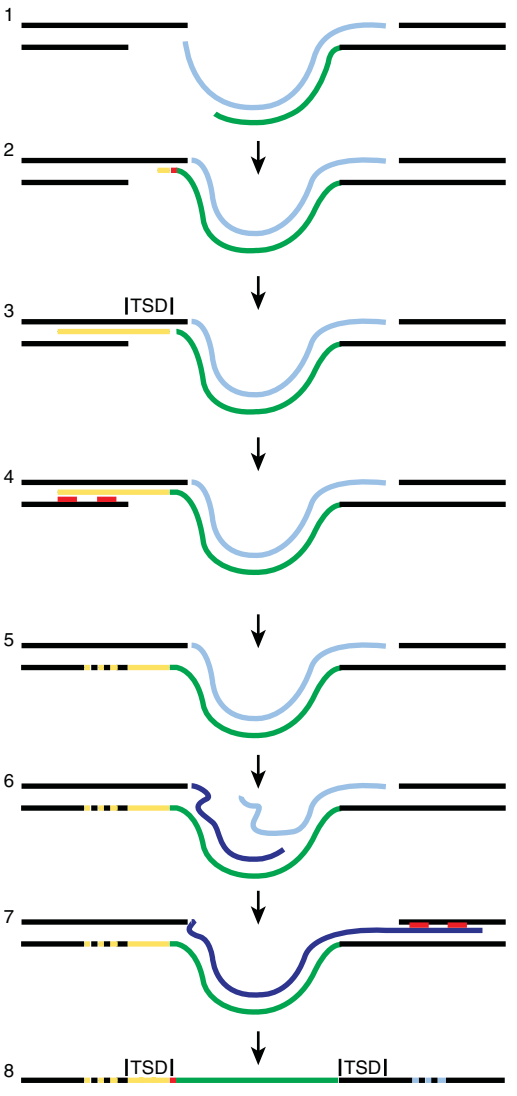


Figure 20.4 Model for L1 integration. (Flanking host DNA) black lines; (L1 RNA) top gray line (light blue in e-book); (nascent cDNA strand) bottom gray line (green in e-book); (nascent 2nd strand) dark gray line (dark blue); (homologous recombinational repair) dark crossed lines (yellow); [recombination products (no sequence changes expected)] stippled lines; (5' microhomology-guided base pairing) dotted line between nascent DNA and top-strand DNA in (2); (TSD) target site duplication. First, a host DNA strand is cleaved by the L1 EN, reverse transcription of the L1 RNA is initiated in a standard TPRT reaction and is followed by the downstream cleavage of the second strand (1). Upon reaching the end of L1 RNA, the L1 RT attempts template jumping from the 5' end of the RNA template onto the upstream overhang of host DNA (2). This template jump may be facilitated by annealing of one or more nucleotides at the 3' end of the nascent cDNA with those in the overhang of host DNA, or by the addition of several untemplated bases (Bibillo and Eickbush, 2004). The former results in an apparent 5' microhomology; the latter creates unexplained nucleotides at the element's 5' boundary. After the jump, L1 RT likely continues copying the host DNA region, adding a stretch of DNA complementary to the host's top strand to the L1 cDNA (3). Depending on the length of added homologous DNA, the nascent strand can be joined to the host's bottom strand by simple nick ligation or by the host's homologous recombinational repair machinery (4 and 5). After the attachment of the 5' end of L1 cDNA to the host's bottom strand, a second molecule of L1 RT likely completes the synthesis of the second strand (6), either displacing RNA from the DNA:RNA duplex during the reaction or relying on RNA degradation by host enzymes. Finally, the nick is repaired by simple ligation or by the host's homologous recombination machinery (7), creating a typical L1 insertion (8). (© 2006 CSHLP))

(Christensen and Eickbush, 2004; Christensen and Eickbush, 2005). Key aspects of the model are reverse transcription of the first DNA strand past the insertion site with subsequent recombination between that growing DNA strand and the top strand of genomic DNA, followed by DNA synthesis of the second strand by the ORF2p reverse transcriptase using first-strand DNA as a template. Dasha wrote up this work and quickly published it in *Genome Research* in 2006 (Babushok et al., 2006). Other reasonable models for the L1 retrotransposition mechanism have been presented (Gilbert et al., 2005; Symer et al., 2002).

For her second act, Babushok had to get some help. Very early on in the lab, she was carrying out cell culture assays of human L1 retrotransposition and found something strange; a messenger RNA sequence of another gene had been spliced into the L1 RNA, and the hybrid RNA had been reverse transcribed and inserted into the genome. Was this an example of a rare phenomenon, trans-splicing, splicing between two different messenger RNAs? We wondered whether there were examples of this in nature. I knew that Nori Okada, a colleague in Tokyo, was doing genomic computational analyses of processed pseudogenes, messenger RNAs that were retrotransposed back into the genome by L1. I contacted Okada and asked him to look for events in the HGWD and the mouse genome database in which parts of two genes had been spliced together to make a new gene. Okada asked Koichi Ohshima, his computational expert, to look for such hybrid genes. Ohshima found none in the mouse genome but did find one such gene in the human genome. Here again was an unexpected finding!

We know that most new genes arise by duplication of existing gene structures, after which relaxed selection on the new copy frequently leads to mutational inactivation of the duplicate; only rarely does a new gene with modified function emerge. Ohshima had found a novel gene that represented a unique mechanism of gene creation. In this new mechanism, a new combination of functional domains was assembled at the RNA level from distinct genes, and the resulting chimera was then reverse transcribed and integrated into the genome by the L1 retrotransposon (Figure 20.5). Dasha characterized this novel gene, which she called PIPSL. It was created from an intergenic transcript between the phosphatidylinositol-4-phosphate

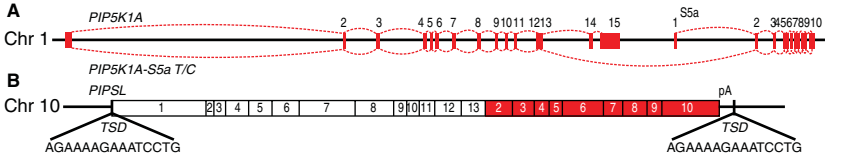


Figure 20.5 Formation of a new gene, PIPSL. Neighboring 15-exon PIP5K1A and 10-exon S5a genes on Chr 1 are spliced to form PIP5K1A, S5A, and PIP5K1A-S5a TIC (transcription-induced chimeric) mRNAs. (Shaded rectangles—red in e-book version) Exons; (curved lines) splicing. (B) PIP5K1A-S5a TIC was retrotransposed by L1 to create the PIPSL gene on Chr 10. (TSD) Target site duplication; (pA) A-rich repeat. (White rectangles) Regions corresponding to PIP5K1A exons; (shaded rectangles—red in e-book) regions corresponding to S5a exons. (© 2007 CSHLP)

5-kinase (PIP5K1A) and the 26S proteasome subunit (S5A) genes in a hominoid ancestor. These two genes are just 6kb apart, and a single RNA transcript containing both genes occurs *in vivo*. In other words, transcription often goes beyond the poly A signal of the PIP5K1A gene through to the S5A gene. Then this elongated transcript undergoes splicing from the penultimate, or next-to-last, exon of the PIP5K1A gene to an early exon of the S5A gene. This new hybrid gene, PIPSL, is transcribed specifically in the testis both in humans and chimpanzees and is repressed after transcription by independent mechanisms in these primate lineages. The PIPSL gene encodes a chimeric protein combining the lipid kinase domain of PIP5K1A and the ubiquitin-binding motifs of S5A. Strong positive selection on PIPSL led to its rapid divergence from the parental genes PIP5K1A and S5A, forming a chimeric protein with a distinct cellular localization and minimal lipid kinase activity but significant affinity for cellular ubiquitinated proteins. Babushok enlisted Charles Abrams of the Department of Medicine at Penn for the kinase studies, but she carried out the studies on affinity of the PIPSL protein to ubiquitinated proteins on her own. PIPSL is a tightly regulated, testis-specific novel ubiquitin-binding protein formed by an unusual exon-shuffling mechanism in hominoid primates and represents a key example of rapid evolution of a testis-specific gene.

In 2006, Babushok had submitted the work to *Genome Research* before obtaining functional data on the hybrid gene, but her paper was rejected because the reviewers required the functional data. Now that she had obtained those data, the paper was accepted for publication in

August, 2007. This time, *Genome Research* wanted to feature it with a cover picture (Babushok et al., 2007). Dasha designed a nice cover figure of the science, but she wanted to add some extra character. Given the work was on hominoid evolution, and the retrotransposition event she was describing was ~17 million years old, Dasha took her camera to the Philadelphia Zoo. Using her zoo photos, she made a great cover that featured one orangutan in the foreground swinging on a vine that looked awfully like a DNA double helix and another orangutan in the background looking out at the viewer (Figure 20.6).

Figure 20.6 *Genome Research* cover showing creation of a new gene. In addition to its role as an insertional mutagen and a potent substrate for homologous recombination, L1 has produced a number of new genes by reverse transcribing cellular mRNAs and integrating the resultant cDNAs back into the genome. An unusual product of such an L1-mediated retrotransposition process is a new primate gene PIPSL, which originated from a read-through, chimeric transcript between the neighboring phosphatidylinositol-4-phosphate 5-kinase (PIP5K1A) and the 26S proteasome subunit S5A (called PSMD4 in the figure) genes. (© 2007 CSHLP)

I have high hopes that Babushok will continue in research after her medical training. She has finished medical school and is presently a resident in Internal Medicine at the Massachusetts General Hospital in Boston. She plans to begin a fellowship in Hematology-Oncology at Penn in July, 2011. However, I can't leave Babushok without mentioning her main activity outside the lab, orienteering. Dasha was a short, trim young woman who looked like she might be able to run all day. Indeed, that's what she often did! Orienteering is a sport that is a little like cross-country running but is running point-to-point using a compass for direction. The sport is very popular in Europe but is somewhat under the radar in the U.S. However, there are orienteering clubs, and national meets in the U.S. Dasha's husband James is very much into orienteering, so Dasha joined him and became, as expected, quite good at it. James has become one of the top orienteerers in the U.S. and a member of the national team. Dasha and James travel to national meets and to Europe for international competition. However, I suspect that Dasha's medical training at Mass General has now cut into her orienteering activities.

21

The brilliant young lady from China

In the fall of 2000, a small, highly intelligent, young Chinese lady came to discuss a potential lab rotation with me. Nuo Yang had recently entered graduate school at Penn in the Gene Therapy program, having just completed medical school in Beijing. I was surprised that Nuo's command of the English language was quite good for someone who had recently come to the U.S. Later, I learned from her that her GRE scores were outstanding. After I told her about our retrotransposition projects, she said the field was too complicated and difficult for a first rotation, but she would likely return in six months after she had gained some experience. Indeed she did, and after a successful lab rotation, Nuo decided to join the lab in 2001. After a short while, she independently picked a project.

Yang decided to make mutations in the promoter region (the first 600 nucleotides at the 5' end) of a highly active L1 using a mutation-producing PCR protocol and then cloning the PCR products back into the L1 to test both promoter activity and ability to retrotranspose in the cell culture assay. In 1990, Gary Swergold had shown in an important paper that L1 had an internal promoter and that the first 100 nucleotides of the element were important for transcription (Swergold, 1990). However, he also showed the importance for transcription of nucleotides up to position 668 or two-thirds of the L1 5' UTR. In 1993, Becker et al. showed that there was an important YY1 transcription factor-binding site at nts. +21 to +13 on the antisense strand of L1 (Becker et al., 1993), and in 2004 Athanikar et al. found that this site was important for setting the location of the transcription start site at +1 (Athanikar et al., 2004). Meanwhile, in 2000, Heidmann's lab had shown that SRY binding sites (SOX factor binding) between nucleotides +472 and +477 and between nucleotides

+572 and +577 were also important in L1 transcriptional activity (Tchenio et al., 2000). Yet the number of known transcription factors that were required for L1 expression was very small. It seemed highly likely that many more transcription factors were critically important for L1 expression. Thus, it was quite reasonable for Yang to search for new transcription factor binding sites in the 5' UTR of L1.

She tested a large number of single nucleotide mutations and found one that was particularly interesting. The single change at nucleotide +100 from the 5' end reduced retrotransposition to 8% of control. When Yang searched the database for potential transcription factor binding sites in the region, she found a RUNX3 site at nucleotides +83 to +101 (Figure 21.1). Yang went on to show that this binding site was functional in the transcription of L1 by carrying out further mutation analysis of the site and studies of the binding of RUNX3 to the site. This was an important study that added RUNX3 to the short list of transcription factors that have a large effect on the expression of L1 elements (Yang et al., 2003).

However, Yang needed another project to finish her Ph.D. At the time, micro RNAs and small interfering RNAs (siRNAs) were *au courant* in biology. Following on the work of Fire and Mello, they were seemingly being found everywhere. Speek had recently reported that there was an antisense promoter that started transcription in the middle of the 5' untranslated region of L1 and drove transcription backwards toward the 5' end. This promoter could serve as another start site for genes upstream of L1 in the opposite orientation as the L1 (Speek, 2001). Both Moran's lab and mine had done experiments showing that deletion of L1 sequence just downstream of the middle of the 5' untranslated region led to increased transcription and retrotransposition of active L1s in cell culture. So Yang reasoned that perhaps the production of antisense RNA along with sense strand RNA resulted in small amounts of double-strand RNA that would be cleaved by the RNAi machinery to produce small RNAs that then would cleave L1 RNA. If you deleted the antisense promoter, you would reduce the double-strand RNA and in turn reduce the amount of small RNAs produced, leading to increased levels of L1 RNA. This antisense mechanism could be a cellular response to reduce retrotransposition by active L1s.

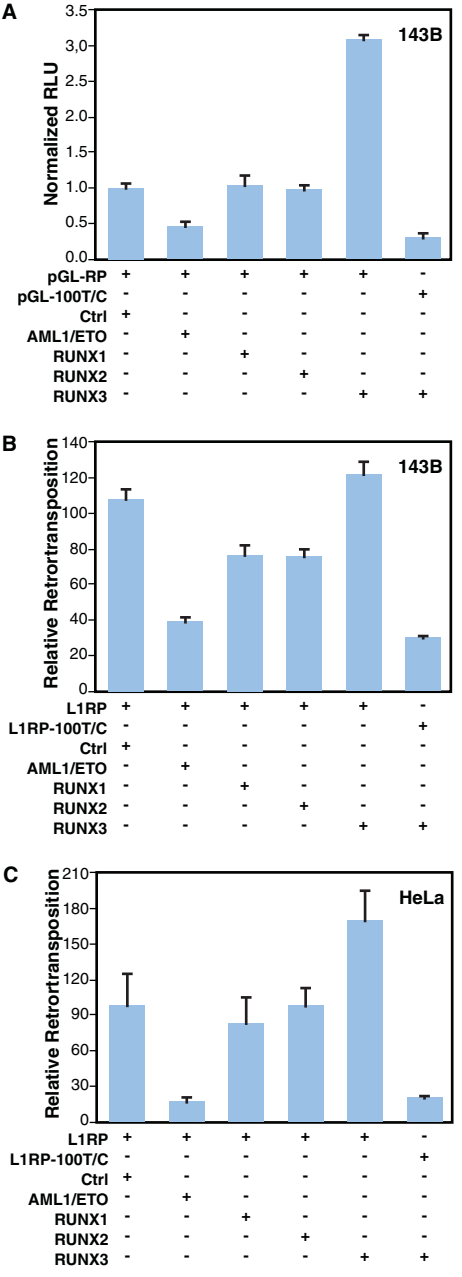


Figure 21.1 RUNX3 addition increases L1 retrotransposition. Earlier, Yang showed that a single nucleotide substitution at position +100 within a RUNX3 binding site in the L1 promoter region reduced retrotransposition of an active human L1 by >90%. Here she demonstrates in (A) that exogenous RUNX3 greatly increases L1 promoter activity in 143B cells compared with empty vector. Retrotransposition increases in 143B cells (B) and HeLa cells (C) upon transfection of a RUNX3-containing plasmid along with human L1. Transfection of a dominant negative (dn) RUNX3 reduces retrotransposition in both cell types by 85–90%. (Yang et al., 2003. Used with permission by Oxford University Press.)

Yang first traced the transcription start sites of the antisense transcript to around +500 nucleotides into the L1 and found that antisense transcription was only about 10% as active as sense transcription (Figure 21.2). She also confirmed that deletion of nucleotides +600 to +900 of the L1 5' region led to an increase of 1.5–2 fold in L1 transcription and retrotransposition, and a stabilization (increased half-life) of L1 RNA. Most importantly, she found that there were small RNAs of about 20 nucleotides in length derived from the first 500 nucleotides of L1 in some human cultured cells but not others. When she carried out experiments in which she knocked down Dicer1, a key component of the siRNA pathway, she did find a small, roughly 1.5–2-fold increase in L1 transcription and retrotransposition. This work was published in *Nature Structural and Molecular Biology* in 2006 after Yang had obtained her Ph.D. degree in 2005 (Yang and Kazazian, 2006).

However, some of Yang's findings, including the interpretation that L1 RNA stability is affected by a siRNA mechanism, remain controversial. Prabhat Mandal, a postdoc in the lab, has confirmed the finding of the small RNAs coming from the +400 to +500 region of the 5' end of L1, but he cannot find other features of an siRNA mechanism affecting L1 RNA stability. It may be that the L1 RNA is stabilized by deletion of nucleotides +600 to +900 through an altered secondary structure that is unrelated to a cleavage effect of small RNAs. Supporting this view, Mandal now believes that small RNAs derived from L1 do not have the structure seen in other siRNAs. This story of Yang's work points out that convincing data may have interpretations other than those postulated. This is called, "being led down the garden path," and is another example of the dictum, "expect the unexpected." This time, the advice is...don't be so convinced of the expected explanation that you overlook the unexpected one.

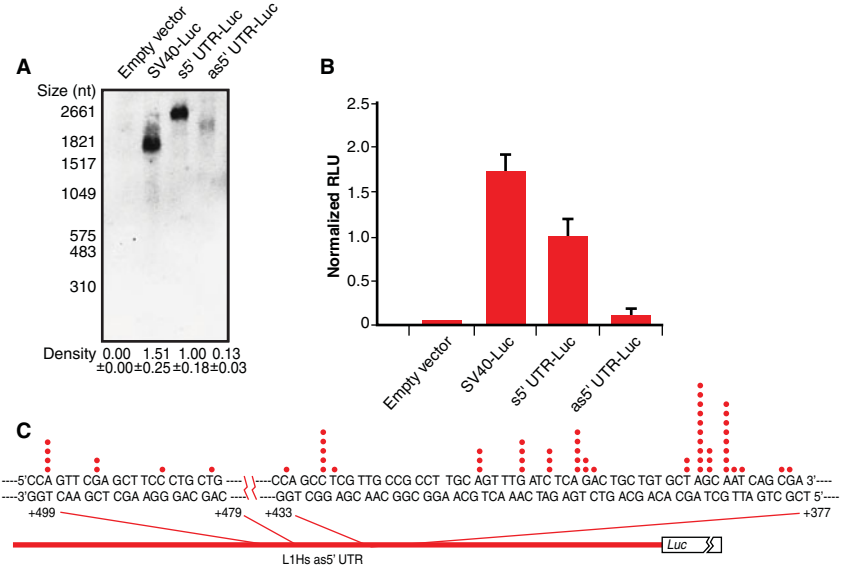


Figure 21.2 Characterization of the antisense promoter (ASP) in the human L1 5' UTR. (a) Detection of the ASP-derived transcript by Northern blotting. Total RNA was extracted from HeLa cells transfected with pGL3-Basic vector containing either no promoter (empty vector), SV40 control promoter (SV40-Luc) or human L1 5' UTR in the sense (s5' UTR-Luc) or antisense direction (as5' UTR-Luc). Density quantification of the signal (n = 3) is below gel. (b) Luciferase reporter assay of expression from SV40 control promoter (SV40-Luc), sense human L1 5' UTR (s5' UTR-Luc) and ASP (as5' UTR-Luc). pGL3-Basic vector containing no promoter (empty vector) was used as a negative control. RLU, relative luciferase units. Error bars show s.d. (n = 9). (c) Transcription start sites of ASP identified by 5' RACE. Dot, one occurrence of transcription initiation at the specific nucleotide. (Yang et al., 2006. Used with permission by Oxford University Press.)

This page intentionally left blank

22

Hiroki's big surprises

In 2003, Hiroki Kano from Osaka University approached me at a meeting in Japan with a desire to join the lab as a postdoc. He was just finishing his Ph.D. degree with Toda in Medical Genetics. I learned later after Hiroki joined the lab in early 2005 that Kano had an M.D. degree and was a trained orthopedic surgeon who had carried out his Ph.D. lab research in the evenings and weekends. He brought with him to the U.S. his wife and two early school-age children.

Kano's English was serviceable. He understood the language reasonably well and spoke haltingly. However, he read and wrote the language very well. Hiroki was soft-spoken but knew his own mind and his opinions were always very much respected. His work was impeccable, and his work ethic set a good example for the grad students and other postdocs. He was also very helpful to his lab mates, making suggestions and aiding experiments.

Upon his arrival in 2005, we talked about his project. I felt that we needed to do a transgenic experiment using a human L1 without a heterologous promoter on the L1. At that point, we had made many transgenic mice carrying human L1s with a heterologous promoter such as pPolII or Hsp70. Similarly, Boeke had only made transgenic mice carrying an L1 driven by a heterologous promoter, the CAG (beta-actin) promoter (An et al., 2006). Ostertag had made one transgenic line with L1 driven by its own endogenous promoter, but the level of retrotransposition in that line was roughly one event in every 200–300 sperm, and no insertion-positive adults had been recovered. Perhaps five years earlier when I had given a seminar at NIH, I had met with Gary Felsenfeld. Gary had suggested that we might get excellent retrotransposition in transgenic mice if at both

ends of the L1 we placed chicken beta-globin insulator sequences that he had discovered (Chung et al., 1997). Now five years later, I was finally ready to take Felsenfeld up on this suggestion. These insulator sequences should block modifiers of chromatin that would shut off transcription. A transgene L1 between the insulators should be transcribed into RNA no matter where it landed in the mouse genome. Kano cloned the insulators at both sides of a human L1. He also made the retrotransposition cassette much shorter by removing the *EGFP* gene. Now the cassette was limited to the γ-globin intron only. We made mice carrying this human transgene and called the mice the insulator line.

Meanwhile, Alysson Muotri in Rusty Gage's lab at the Salk Institute had made an amazing, very surprising observation. He got the human $L1_{RP}$ containing an *EGFP* retrotransposition cassette from us and made a transgenic mouse line with it. This mouse line was very active for retrotransposition in adult animals. The result was surprising because the L1 promoter in the transgene was derived from the L1 itself, and we had never obtained highly active retrotransposition of a human L1 in a mouse line without using a heterologous promoter on the L1. More surprisingly, Muotri found significant retrotransposition in neural progenitor cells and in various regions of the brain of the transgenic mice. Muotri, Moran, and Gage had published this surprising result in *Nature*, along with the suggestion that L1 retrotransposition might be a significant contributor to human behavioral diversity (Muotri et al., 2005).

Obviously, we were interested in determining whether the retrotransposition events from this line were inherited, and if they were, whether the line, that we called the Gage line, would be useful in making mice that had insertional mutations knocking out genes. For the previous five years, our applied goal was to use L1 as an insertional mutagen, find mice with an insertion that had developed a pathologic condition, and then find the genomic site of the L1 insertion in order to determine the affected gene. We would then have evidence that the knocked-out gene had potentially caused the disease. In this way, we

hoped to find genes involved in various conditions, such as cancer and diabetes. But first we needed to obtain a mouse line in which there was a high frequency of inherited L1 retrotransposition events. Our L1 transgenics in which an Hsp70 promoter drove the human L1 gave rise only to insertions that were somatic and not inherited. But surely the transgenic animals carrying a native L1 without a promoter from another source (a heterologous promoter) would give rise to inherited events, even if their frequency were low!

Kano bred his transgenic mice with mice that were not carrying the transgene. Both the insulator line and the Gage line had a high frequency of insertions—about two-thirds of the adult offspring carrying the transgene had insertions detected in tail DNA. However, the intensity of the PCR band demonstrating the presence of an insertion was weaker than would have been expected if the insertion were present in every cell and transmitted through a germ cell. Moreover, between 5% and 10% of offspring had an insertion even though they had *not* inherited the transgene (Figure 22.1). We had seen this result in a few of Ostertag's mice way back in 2001. Those mice had a heterologous promoter, the pPolII (RNA polymerase II) promoter driving the L1, and those insertions were clearly inherited. They must have occurred before the end of meiosis I because that was the time when the two chromosomes of a pair separate. If an insertion occurred before the end of meiosis I from, say, a chromosome 2 carrying the transgene, the mouse we were testing could be derived from a germ cell containing the insertion but lacking the chromosome 2 with the transgene. In this case, the insertion would be present in every cell, and 50% of the offspring of that mouse would carry the insertion. So Kano proceeded to breed the animals carrying an insertion and lacking the transgene, but—surprise, surprise—none of their offspring contained the insertion. He did the same experiment breeding a number of mice that had insertions but lacked the transgene, which we called insertion +, transgene - mice, and always got the same result. The insertions were not heritable and had not occurred during meiosis I but must have occurred in early embryogenesis.

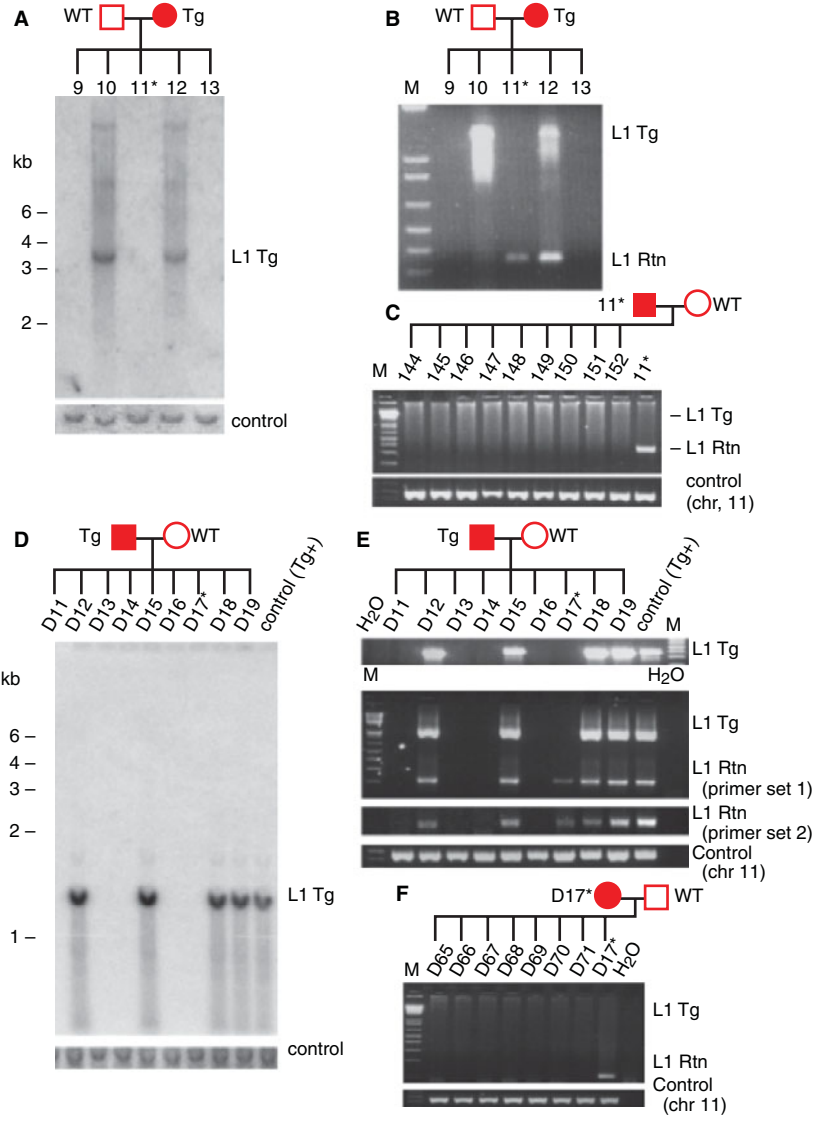


Figure 22.1 L1 retrotransposition caused by L1 RNA carried over through meiosis, fertilization, and embryogenesis in the L1RP mouse (A–C) and in the L1LRE3 mouse (D–F). (Kano et al., 2009. Used with permission by CSHL Press.)

(A) Southern blot analysis on tail DNA isolated from offspring of an L1RP transgenic female mouse. A 1.4-kb DNA probe generated from the retrotransposition cassette of the L1RP transgene was expected to hybridize to both transgene and retrotransposition insertion, but it demonstrated only transgene bands despite the presence of retrotransposon amplicons by PCR. The membrane was rehybridized with an unrelated DNA probe generated from mouse chromosome 11 as a DNA loading control. (B) Genotyping PCR on tail DNA indicates an L1 retrotransposition event in a mouse lacking the transgene (mouse 11). (C) Mouse 11 (transgene-negative, retrotransposition event-positive) was bred with a wild-type mouse, and its offspring were genotyped. No offspring of this mouse inherited the retrotransposition insertion, indicating mosaicism of the L1 retrotransposition event in mouse 11. A control PCR on mouse chromosome 11 was performed to confirm the amount and quality of DNA. (D–F) Similar data to those in A–C are shown for the offspring of an L1LRE3 transgenic male mouse using tail DNA. The transgenic male mouse was bred with a wild-type female mouse, and its offspring were genotyped by Southern blot using a 503-bp probe generated from the L1 3'UTR and SV40 poly(A) signal sequence of the L1LRE3 transgene (D) and by PCR (E). Two independent PCR primer sets were used to confirm the presence of retrotransposition events. (F) The single offspring (D17) that had a retrotransposition event while lacking the L1 transgene was bred with a wild-type mouse. As shown in C, none of its offspring inherited the retrotransposition event. Note that RNA carry over has occurred from both the female transgene carrier in (A-C) and the male transgene carrier in (D–F). Asterisk denotes a transgene-negative, retrotransposition event-positive mouse. (Tg) transgene; (Rtn) retrotransposition event; (WT) wild-type animal; (M) 1-kb plus DNA Ladder (Invitrogen).

We were incredulous! How could it be that most insertions coming from an L1 containing only the endogenous promoter would occur in early embryonic development? Look at mammalian genomes! The human and mouse genomes are loaded with over 500,000 copies of

L1s that have been inherited from generation to generation. Those insertions must have either occurred in germ cells or in very early development. That made it very difficult to believe that most insertions occur late enough in development so that they are not heritable.

OK, so Hiroki, if your result is correct, you should be able to find the L1 RNA from the transgene in embryos lacking transgene DNA. Moreover, there should be more retrotransposition in the morula and blastocyst stages of embryonic development than in male or female germ cells. Indeed, when Kano analyzed the L1 RNA in spermatogenic fractions of transgenic males and in single blastocysts derived from mating transgenic males with non-transgenic females, he found about the same RNA amounts in single blastocysts as in the germ cell fractions. He also found blastocysts lacking the L1 transgene that contained a significant amount of L1 RNA (Figure 22.2). This was another big surprise because it showed that L1 RNA could be carried over through spermatogenesis, through fertilization, and into the embryo. It could then remain intact until the blastocyst stage, that is, day four after fertilization. A very long time indeed!

Then Hiroki looked at the insertions present in spermatogenic fractions of male transgenic animals and in their blastocyst offspring. As we now expected, there was much greater retrotransposition in the blastocysts than in spermatogenic fractions. Spermatogenic fractions contained barely detectable retrotransposition, while in blastocysts retrotransposition was perhaps 100 to 200 times greater (Figure 22.3). Hiroki also did the same study using transgenic rats carrying a similar human L1 transgene but without insulators. Interestingly, all the lines of these rats had high levels of retrotransposition, but, as in the mouse, essentially all the retrotransposition occurred in early development, not in germ cells. All the data from the mouse transgenic animals were repeated in the rat transgenics. Now Kano wrote up his paper and sent it to *Nature*. It was promptly rejected because the reviewers did not believe the data!

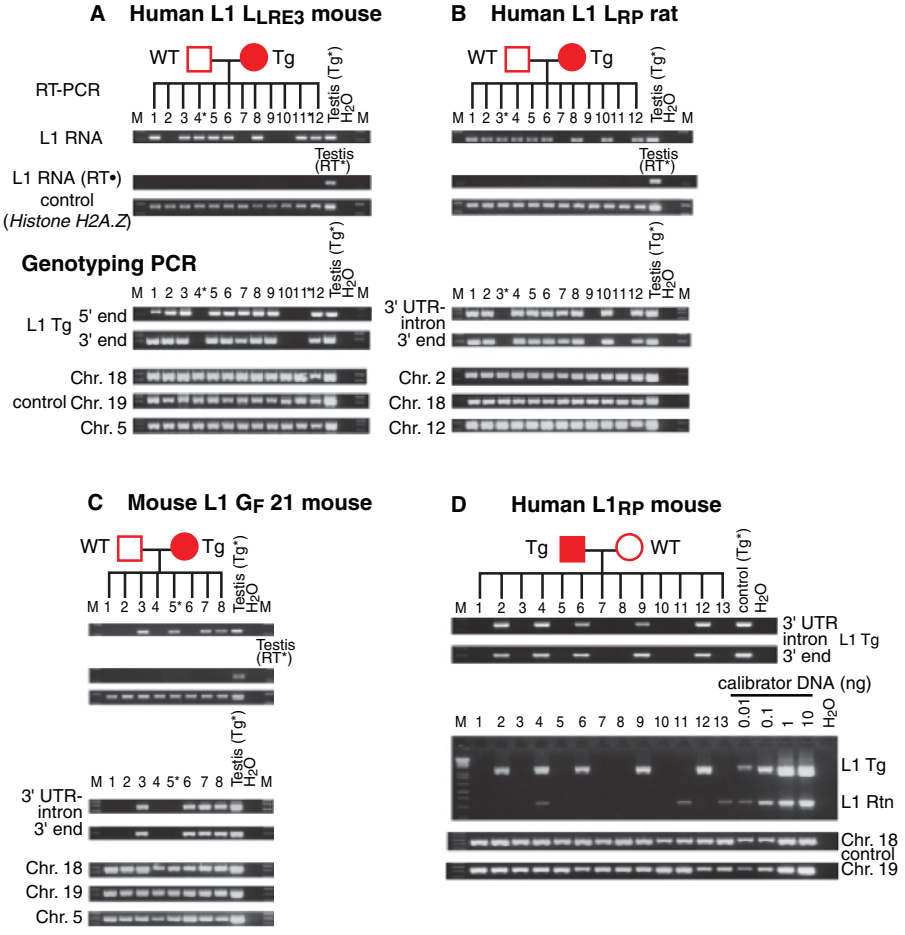


Figure 22.2 Retrotransposition in embryos lacking L1 transgene. Single preimplantation embryos lacking transgene contain L1 RNA (A–C) and L1 retrotransposition events (D). RT–PCR and genotyping PCR on offspring of an $L1_{LRE3}$ mouse (A), an $L1_{RP}$ rat (B), and an L1 G_F 21 mouse (C). L1RNA and L1 DNA of single morulae or blastocysts were detected by RT–PCR and genotyping PCR, respectively. To exclude a false negative genotype for transgene, each embryo was genotyped by two different primer sets for L1 transgene, and three control loci. In A–C, an asterisk denotes a transgene-negative, L1 RNA-positive embryo. (D) Retrotransposition in individual blastocysts. Genotyping PCR was done on single blastocysts of the $L1_{RP}$ mouse. For semi-quantification, mouse DNA carrying 1 retrotransposition event/diploid genome (Ostertag et al., 2002) was used as calibrator DNA. The DNA amount of each blastocyst used in the intron-flanking PCR was ~0.1–0.5 ng, suggesting that retrotransposition events in blastocysts 4, 11, and 13 are present in << 1 copy/cell. (RT) reverse transcriptase; (Tg) transgene; (Rtn) retrotrans-position event; (WT) wild-type animal; (M) 1-kb plus DNA Ladder (Invitrogen).

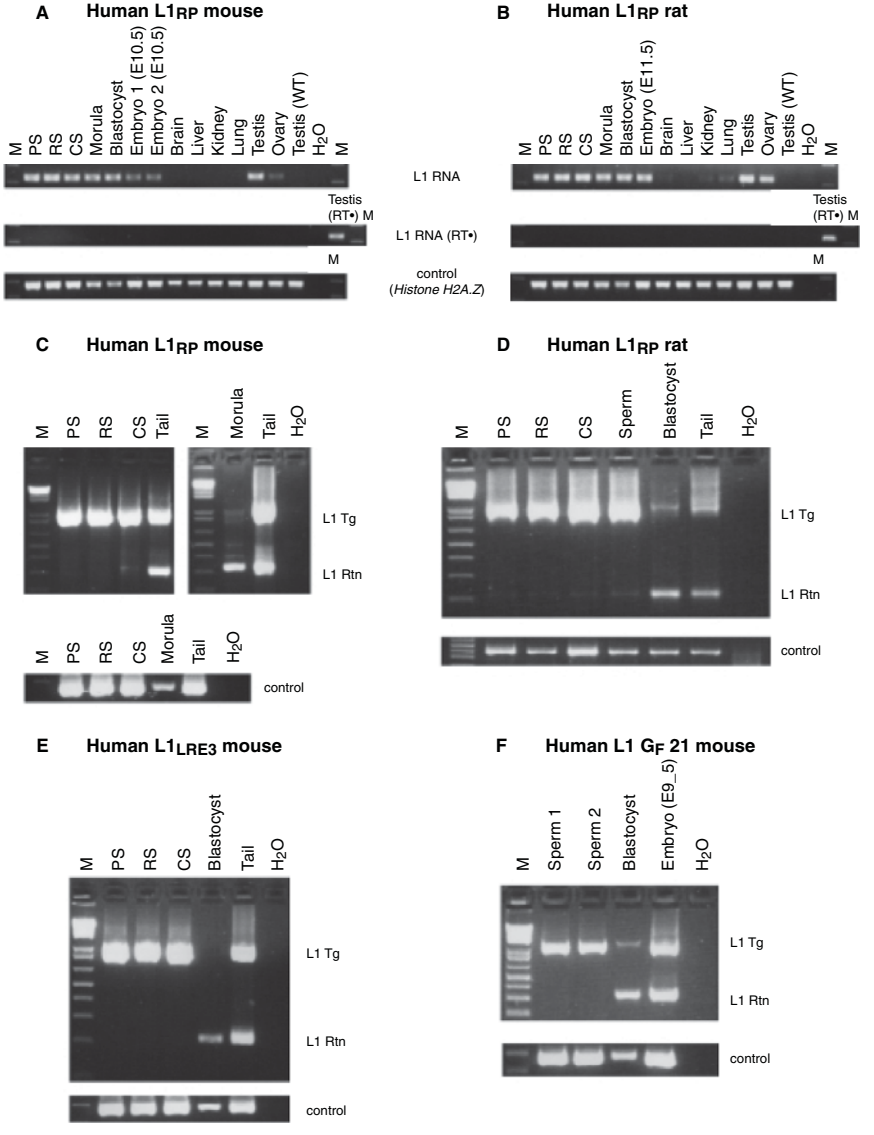


Figure 22.3 L1 transcripts and retrotransposition events in various developmental stages and adult tissues. (Kano et al., 2009. Used with permission by CSHL Press.)

(A,B) RT–PCR on transgenic L1RP mouse (A) and L1RP rat (B) spermatogenic cell fractions, preimplantation embryos (morulae and blastocysts), E10.5–E11.5 embryos, and adult tissues. Only a head portion of E10.5–E11.5 embryos was subjected to RT–PCR in order to eliminate contamination of germ cells in the embryonic developmental stages. Testis from wild-type adult animals was used as a negative control for L1 RNA from the transgene. Histone H2A.Z gene was used as an endogenous control. (C–F) Genotyping PCR on L1RP mouse (C), L1RP rat (D), L1LRE3 mouse (E), and L1 GF 21 mouse (F) spermatogenic cell fractions and pooled preimplantation embryos (L1RP mouse line, 10 morulae; L1RP rat line, 25 blastocysts; L1LRE3 mouse line, 9 blastocysts; L1 GF 21 mouse line, 12 blastocysts). Spermatogenic cell fractions were prepared from transgene-positive, retrotransposition event-negative mice (C,E,F) and transgene-positive, retrotransposition event-positive rats (D). Nested PCR was performed on each sample, which was optimized to amplify small products preferentially. In D, similar amounts of DNA (5 ng) from rat spermatogenic cell fractions, pooled blastocysts, and tail were subjected to PCR. Genomic DNA of the Actb region was amplified to confirm the amount of DNA. (M) 1-kb plus DNA Ladder (Invitrogen); (RT) reverse transcriptase; (Tg) transgene; (Rtn) retrotransposition event; (WT) wild type.

In the meantime, I thought that we needed to redo the experiments using a mouse L1 because there might be species-specific effects that could alter the timing of L1 insertion. I also thought that we should make a human L1 transgene without any cassette and determine whether the intron was leading to RNA carry over. During the summer of 2007, we made these new transgenes and began the experiment. In 2008, we got the data. The active mouse L1 behaved just as the human L1 had. We saw L1 RNA carry over from germ cells, and most retrotransposition of the mouse L1 occurred in early embryonic development and not in germ cells (Figures 22.2 and 22.3). Although we could not study retrotransposition in the offspring of transgenic animals bearing the native human L1 without the intron, we could show that they had L1 RNA carry over just as we had observed in all other transgenic lines. Some mouse blastocysts carrying the native human L1 had L1 RNA but lacked the transgene.

So we rewrote the paper with all the new data and sent it to *Science*. One reviewer was strongly in favor of publication, and one again just didn't believe the data and made a few factual errors in the review. The third reviewer killed the paper. He or she said we needed to show retrotransposition timing from the intronless human transgene, an experiment that was undoable. We made format changes and sent the paper to *Genes and Development*, a highly regarded developmental biology journal. What a different result! The paper was accepted within three weeks and published online within one month of submission (Kano et al., 2009). The lesson from this experience is that a paper with very surprising, almost unbelievable, results is very difficult to publish, even though the data are overwhelming! Kano had worked patiently and tirelessly for four years on this one important piece of work. I think that in the end he will be rewarded.

His work correlates nicely with the work of the Nicole Coufal in the Gage lab on somatic insertions. She has recently shown in human cadaver tissues that L1 retrotransposition events are more frequent in hippocampus than in the heart or liver. Her estimate was roughly 80 extra copies of L1 per cell in hippocampus as compared to heart or liver. These data show that there is ongoing L1 retrotransposition after early embryonic development in neural progenitor cells destined to form parts of the human brain (Coufal et al., 2009).

In other recent work, Belancio et al. have shown that both full-length and processed L1 RNA copies are produced in a wide variety of somatic tissues and transformed cell lines (Belancio et al., 2010). Because they also have evidence that L1s can produce DNA double strand breaks (see later), they suggest that L1 may be an endogenous mutagen in many somatic tissues. Another piece of evidence supporting L1 insertion in embryogenesis is the case of choroideremia in a Dutch family in which the mother was clearly a germinal and somatic mosaic for the insertion. In this mother, it is clear that the insertion occurred during her embryonic development (van den Hurk et al., 2007). In addition, Garcia-Perez et al. have shown that human ES cells can support retrotransposition of a transfected active L1 (Garcia-Perez et al., 2007b). Thus, the evidence for an important role of somatic L1 retrotransposition in addition to germ line insertion seems to be accumulating rapidly.

23

A young man with a purpose

The story in this chapter is still ongoing. In the spring of 2006, I met Dustin Hancks for the first time. On first glance, Dustin appeared to be quite a character, but later I learned that he was the real deal. Dustin was applying to the Gene Therapy graduate program at Penn. He was a student at Southern Illinois University and had worked with David Duvernell, who had found a retrotransposon similar to L1 in fish that he called *swimmer*. During his experience with Duvernell, Dustin had become hooked on mobile DNA. While still an undergrad, he decided that he wanted to do his graduate work on human retrotransposons. So he applied only to the University of Michigan to work with Moran, Johns Hopkins to work with Jef Boeke, and Penn to work in my lab. Fortunately for me, he was accepted at Penn and joined our graduate program.

Hancks had many interesting qualities. First was his appearance! He had bushy black hair and large facial features. He also was a bit chunky and was somewhat obsessed with listening to tunes on his iPod. This latter aspect meant that he didn't hear your first effort to communicate because he was wearing earphones. On top of all this, Dustin counted as a minority student because his mother was Mexican. The upshot of this status was that because he was such a good graduate student, one could be confident that his applications for support would be successful.

So Hancks dutifully showed up in early September 2006, to set up a lab rotation. I thought it best for both him and me that he take his third and last lab rotation with us so that he could get a running start on his thesis project. Thus, I suggested other labs for his first two

rotations, and he took the suggestions, did two rotations with excellent investigators, and, true to his original decision to work on retrotransposition, returned to my lab in the early spring of 2007.

Hancks wanted to find a project on his own. I said fine, and for a while he worked on a few things unsuccessfully. Then, as so often happens if one is patient, a nice project fell into his lap. John Goodier and I were writing a review on retrotransposition when John discovered a paper that neither of us had seen even though it had come out a year earlier. The paper described an SVA retrotransposition event in which the entire HLA-A gene of 14kb had been deleted in three unrelated patients with leukemia (Takasu et al., 2007).

In addition to Alus and processed pseudogenes, in evolutionarily recent times (15–20 million years), L1s have also been retrotransposing SVA elements. As mentioned earlier, SVAs are composed of a poly A tail, a SINE-R sequence representing part of the *env* gene and the 3' long terminal repeat (LTR) of a human endogenous retrovirus, a *V*ariable *N*umber *T*andem *R*epeat (VNTR) region, two portions of a backward Alu, and a number of hexamer repeats. SVA is misnamed because the name is backwards—the Alu is near the 5' end, and the SINE-R is at the 3' end. Because of the 3' poly A tail and SVA's length, it is very likely that SVA is transcribed by RNA polymerase II, the same polymerase that transcribes most genes and likely transcribes L1. There are roughly 2700 SVAs in the human genome, most of which are full-length or about 3,000 nucleotides long. SVAs are expanding in primate genomes and have caused eight known cases of human disease through retrotransposition.

Goodier and I showed the SVA paper to Hancks, and he and another graduate student, Adam Ewing, analyzed the inserted SVA and found something very interesting. The insertion started at its 5' end well beyond the usual hexamers, and indeed the hexamers were absent. When this extra sequence was analyzed, it turned out to be the first exon (the 5' end) of another gene, the MAST2 gene. Further analysis showed that the 3' end of MAST2 sequence ended precisely at the 3' end of the first exon of MAST2 and was joined to a sequence internal to the SVA. The sequence in the SVA was an excellent 3' splice site, so it was very likely that at some point an SVA was sitting in the first intron of MAST2, and after MAST2 transcription into RNA,

a splicing event had occurred between the 3' end of exon 1 of MAST2 and a splice site within the SVA. RNA of this type containing a MAST2-SVA fusion had then been retrotransposed into the HLA locus causing a 14kb deletion of HLA-A. Then Ewing looked further for MAST2-SVA fusion sequences in the human genome and found a sizable number, about 80, of them. Now Hancks had a very interesting and worthwhile project, to determine the impact of this type of exon trapping by SVA on the human genome. Hancks carried out both computational analysis of the genome and molecular experiments and determined that when an SVA was located in the intron of a gene in the same orientation as the gene, it could disrupt gene expression through splicing into the SVA. It could also lead to making part of SVA a new exon in the gene ("exonization") because of splicing into the SVA, followed by splicing out of the SVA into the next exon of the gene (Figure 23.1).

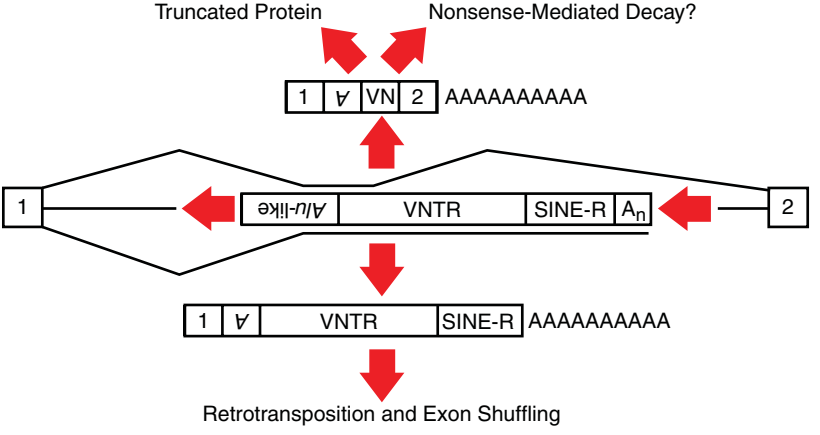


Figure 23.1 SVA alternative splicing outcomes. An intronic truncated SVA is shown (middle). The SVA is 5' truncated because such SVAs may still be spliced. If SVAs are exonized, they will either generate a truncated protein or subject the mRNA to nonsense-mediated decay due to inclusion of SVA nonsense codons (top). If SVAs mimic an endogenous gene-trap, that is, provide a 3' splice site followed by termination at the SVA or downstream poly A signal, this may result in truncated proteins, but more importantly the retrotransposition of exons (bottom). (Hancks et al., 2009; Damert et al.,2009. Used with permission by CSHL Press.)

Dustin obtained a model splicing construct from Russ Carstens, an RNA maven working upstairs from our lab at Penn, put full-length SVAs into the construct, and showed frequent splicing into the SVA and exonization of SVA in cell culture (Figure 23.2). In addition, Hancks found that much of the time transcription of SVA began not at the hexamers as previously thought, but upstream at other promoter sites in the genome.

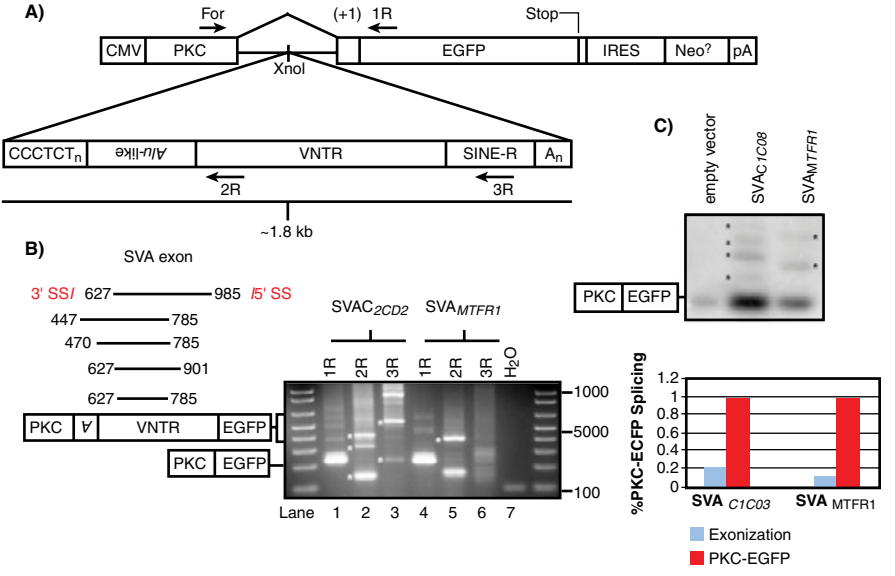


Figure 23.2 SVA gene-trapping and exonization are not rare. (A) Two SVAs were cloned into PKC-EGFP (Newman et al., 2006) to test the mutagenic potential of SVA splicing. Primers used for RT-PCR are marked. (B) RT-PCR was performed on total RNA extracted from 293T cells transiently transfected with pPKC-EGFP containing one of two different SVAs cloned into the intron. SVA exonization events (left panel) are annotated with the first and last nucleotide of the SVA exon, all of which occur within the Alu-like and VNTR domains. A representative agarose gel displaying SVA alternative splicing events is shown (right panel). (*) Indicates bands verified by DNA sequencing to be SVA splicing events. (C) Semi-quantitative PCR to determine the frequency of SVA exonization. Ten cycles of PCR on cDNA from individual pPKC-EGFP, pPKC-SVAC2CD3, and pPKC-SVAMTFR1 transfections were carried out using PKC For and 1R. PCR products were resolved on a 2% agarose gel, followed by overnight transfer to a membrane and subsequent probing using a DNA probe targeting the PKC exon (top panel). (*) Indicates bands quantified by a phosphorimager. Total SVA exonization was normalized to PKC-EGFP splicing within each respective lane and graphed (bottom panel). (Hancks et al., 2009; Damert et al., 2009. Used with permission by CSHL Press.)

It appeared that Hancks and Ewing had the data for a very good paper, but then we heard that Mark Batzer had an SVA paper in the works that described the MAST2-SVA fusion family of sequences. I was worried that this other paper would partially scoop Dustin's work. Batzer was a respected and friendly colleague, so one Saturday morning I called him. He told me that Gerald Schumann was the senior author of their paper that had been submitted to *Genome Research*, and it had been rejected without review. I told Mark and Gerald that I would contact the editor, tell her about our data that corroborated much of theirs, and ask her to consider both papers together for publication. The editor agreed, and both Dustin Hancks and Gerald Schumann improved their data. The two papers ended up being quite complementary. Gerald's concentrated on the upstream transcription start sites of SVA, while Dustin's concentrated on exon trapping and exonization of SVA. The two papers were accepted and appeared together in the journal (Hancks et al., 2009; Damert et al., 2009). The upshot was that Gerald and I were both very happy. By contacting him, his rejected paper had been improved and published in the same journal that had rejected it, and my worries about publication priority had been put to rest. Now in late-2010, Dustin Hancks continues to work toward his Ph.D. degree, but he has proven his mettle and is sure to get it soon.

Now it's time to leave my lab for the time being and discuss further the current state of knowledge in mobile DNA gained through the work of many other labs. I first discuss mammalian elements other than L1, with a discussion of both SINEs, or short interspersed elements, and other LINEs. I then talk about the effects of mammalian retrotransposons on genome structure and plasticity, followed by a discussion of the many ways that the host has devised to protect itself from genome invasion by mobile DNA. A brief discussion of the question, "Why mobile DNA?" is followed by a recounting of my best guess as to what the future holds in the mobile DNA field—where are we going from here and what the next big surprises might be.

This page intentionally left blank

24

Other mobile DNA in mammalian genomes

Alu elements

In primate genomes, a major SINE, the Alu element, maintains a prominent position. In the human genome, there are some 1.1 million Alus, accounting for about 11% of the genome mass. Alu elements evolved from 7SL RNA of the signal recognition particle. Alus are dimers of roughly 140 non-identical nucleotide monomers that are separated by an A-rich region. The left monomer contains an internal RNA polymerase III promoter composed of short A and B boxes. At the 3' end is a poly A tail, similar to that seen in the L1 and SVA elements. Alus acquired their name because these elements contain an Alu restriction endonuclease site.

Alus started to amplify in genomes about 65 million years ago and reached a peak of amplification about 40 million years ago. Because of their fairly recent amplification, Alus are not present in non-primate mammalian genomes, although there is an element with sequence similarity to Alu, B1, in the mouse genome. Alu elements have been classified into subfamilies based on their sequence variation that can also be used to age them in genomes. The great majority of Alus, the J and S subfamilies, are generally very old and inactive, while about 200,000 of the 1.1 million Alus in the human genome belong to the Y subfamily and its younger subtypes. The most active Alu subtype is AluYa5, which, although present in only about 3,000 copies per genome, has accounted for more than 15 cases of disease through insertional mutation (Deininger and Batzer,

1999). Because Alus do not encode any protein, they are nonautonomous retrotransposons requiring the L1 machinery *in trans* for their mobility.

Dewannieux et al. provided the formal proof of Alu *trans* mobilization by an active L1 in the cell culture assay (Dewannieux et al., 2003) (see Chapter 25, "Effects of Retrotransposons on Mammalian Genomes"). Later, Bennett and colleagues carried out retrotransposition assays in cell culture of a number of Alu elements from the Ya5, Ya8, and other Y, S and J subfamilies using an active human L1 to drive Alu retrotransposition. Although the AluJ elements tested were all dead for retrotransposition, the authors found that some AluS elements (4 of 16 tested) remain active for retrotransposition in cell culture. An even larger proportion of all AluY subfamily members tested were active in the assay. Bennett et al. found that the ability of Alu to interact with the SRP9/14 proteins of the signal recognition particle was highly correlated with the ability of the Alu to retrotranspose in cell culture. Surprisingly, these authors estimated that there are at least 850 Alus in the human genome that are capable of retrotransposition, and there may actually be thousands of active Alus, a number that dwarfs the number of active L1s in the human genome (Bennett et al., 2008).

Other LINE elements

The human genome contains hundreds of thousands of LINE elements other than L1 that are called L2 and L3. None of these elements contain intact ORFs, and none have been known to retrotranspose in humans. However, Okada and his colleagues found that many L2-SINE pairs from various animals share similar 3' tails. They hypothesized that, unlike the reverse transcriptase of L1 that recognizes only the poly A tail, the reverse transcriptase of L2 elements interacts with and requires sequences at the 3' end of the element. Similar sequences at the 3' end of the SINE and the 3' end of the L2 aid the L2 enzyme to reverse transcribe the SINE *in trans* along with the L2 *in cis*. These LINE-SINE pairs are found in a number of clades in the animal kingdom. Okada calls these LINEs stringent because they must recognize DNA sequence at their 3' ends (or the similar sequence at the 3' end of SINEs) to carry out reverse transcription (Figure 24.1) (Ohshima et al., 1996). L1 is called relaxed

because its reverse transcriptase does not require a specific DNA sequence. L1 reverse transcriptase only recognizes the poly A tail sequence.

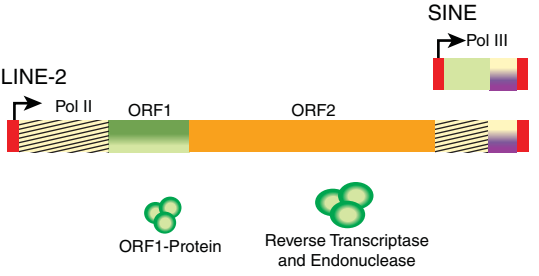


Figure 24.1 Schematic presentation of a LINE and SINE that have the same 3' tail sequence. There are ~20 examples of LINEs and SINEs with similar 3' tails.

Okada's group found that members of the LINE (UnaL2) and SINE (UnaSINE1 and UnaSINE2) families from the eel genome share similar 3' tails. They then adapted the L1 retrotransposition assay in HeLa cells to the eel elements and showed that the 3' conserved tail of UnaL2 is necessary for its retrotransposition. As hypothesized, the UnaSINE1 3' tail (and later UnaSINE2) was recognized *in trans* by the UnaL2 reverse transcriptase at a surprisingly high rate, at the time providing the first experimental evidence that a SINE can be mobilized by the retrotransposition machinery of a partner LINE (Kajikawa and Okada, 2002). The work of Heidmann's group demonstrating mobilization of Alus by the L1 machinery followed soon thereafter (Dewannieux et al., 2003).

Endogenous retroviruses

HERVs, or human endogenous retroviruses, make up some 8% of the human genome, yet up to the present time there have been no active HERVs discovered. HERVs are LTR-retrotransposons that look very similar to retroviruses. They are composed of long terminal repeats (LTRs), *gag*, *pol*, *prt*, and *env* genes encoding a core protein, protease, polymerase (reverse transcriptase) and envelope gene, respectively. Each HERV is given an added letter that denotes the tRNA

that is used to prime reverse transcription of the element, for example, HERV-K is primed by a $tRNA_{lys}$. Most HERVs have defective genes, and the *env* gene is defective in nearly all HERVs. A number of HERV-Ks, the youngest subfamily, have one or more of their ORFs intact. One natural HERV-K has all of its ORFs intact but has a mutation in the highly conserved (YXDD) motif of its reverse transcriptase domain (Mayer et al., 1999). HERVs that are polymorphic as to presence have been found in both chimpanzee and human genomes. This fact suggests that HERVs have been active retrotransposons in the very recent past <3 million years ago. However, no presently active HERVs have yet been isolated from the human genome, and no *de novo* or very recent insertions (<100 years) have been observed in humans.

At one point, my lab made an effort to demonstrate retrotransposition by a HERV in tissue culture. Jens Mayer, the graduate student at the University of the Saar in Germany who reported the HERV-K with intact ORFs that lacked an intact YXDD motif, decided to do postdoctoral training in my lab. This turned out to be the lab's only foray into HERV biology. Jens's goal was to obtain retrotransposition of his nearly intact HERV-K in tissue culture. He inserted the active *neo* retrotransposition cassette into the *env* gene, fixed the defective YXDD motif, and made a number of other modifications in the assay. However, all of his efforts were to no avail. After two years of trying, he could not retrotranspose this HERV-K in cultured cells. How frustrating! He returned to his university and continued his work in other areas of HERV biology and its evolution. He has recently received tenure at the University of the Saar. A great honor in the German system!

On the other hand, two groups have been able to reconstruct replication-competent HERV proviruses from consensus sequences of HERV elements. These reconstructed elements are able to retrotranspose in cell culture (Dewannieux et al. 2006; Lee and Bieniasz 2007) and can also infect human cells. Beyond these full-length active HERVs produced in the test tube, there is the possibility that two partially functional HERVs might be able to complement one another *in trans* within cells to produce an active HERV. Whether the production of an active HERV in the laboratory that has the potential to invade and propagate in human cells is ethical is an important question for debate. Some scientists question the appropriateness of these experiments.

In contrast to the human genome, both chimp and gorilla genomes have been sites for a considerable number of endogenous retrovirus insertions over the past 3–4 million years (postdating the separation of the great ape and hominoid lineages). Thus it is possible that the chimp and/or gorilla genomes contain active HERVs that are still capable of retrotransposing.

In addition, there are a few DNA examples of potentially important host sequences derived from HERVs. For example, the two *syncytin* genes in both humans and mice are likely derived from the *env* gene of an endogenous retrovirus. The mouse *syncytin-A* gene has recently been knocked out, leading to death of homozygotes *in utero* between embryonic days 11.5 and 13.5. Studies of the placentas of these null embryos demonstrate that *syncytin-A* is essential for trophoblast cell differentiation and syncytiotrophoblast morphogenesis during placenta development (Dupressoir et al., 2009). Thus some genes captured from ancestral retroviruses have provided important new and apparently indispensible functions in mammals.

Mice also have a large number of endogenous retroviral sequences that are inactive and non-infectious largely due to an inactivated *env* gene. However, when mice are infected with an exogenous retrovirus, the exogenous retrovirus can recombine with an endogenous retrovirus, creating an active *env* gene and thereby producing an active, infectious agent. In addition, among murine retroviruses members of at least three different families are competent to form infectious viral particles and have an extracellular life cycle. Recently, an active rat endogenous retrovirus was isolated. This element is active for retrotransposition in cell culture, has an active *env* gene, and is polymorphic for presence in inbred rat strains, suggesting ongoing retrotransposition in the rat genome (Wang et al., 2010).

The human genome also contains a wide variety of DNA transposon relics that make up about 3% of the genome's mass. These elements include transposons called tigger, pogo, Charlie, and mariner that have been thought to be dead for many millions of years because no presence/absence polymorphisms of them in human beings or other primates have been described (Lander et al., 2001). (Note the enormous contribution of Arian Smit to the analysis of mobile DNA in this human genome paper.)

LTR-retrotransposons in the mouse

Intracisternal A-particle (IAP)

In primates, active retrotransposons are limited in type to non-LTR elements (L1s) and SINEs (Alus and SVAs). While in mice, active retrotransposons are not only L1s and SINEs, but also a number of different kinds of LTR-retrotransposons. Because of the numerous types of active retrotransposons in mice, insertion events account for 10% of all mutations in mice. This compares to about 0.1% of mutations in human beings.

The mouse genome contains about 1000 intracisternal A particles (IAPs) that were named because they have been seen in viral-like particles within cells. IAPs have a structure very similar to that of retroviruses with LTRs at both ends of the element and *gag*, *pol*, *prt* (protease), and *env* genes internal to the LTRs. The great majority of IAP elements have defective ORFs. In particular, all but a few IAPs have defective or deleted *env* gene regions. However, these defective elements have been the ones that have retrotransposed into the mouse genome, causing a number of isolated cases of disease. It turns out that these events have occurred *in trans* using the activities of an essentially intact IAP element. A small number of IAPs exist that contain intact *gag*, *prt*, and *pol* genes without active *env*. These "intact" IAPs are able to mobilize defective IAPs in a cell culture assay, and they are presumed to be responsible for the retrotransposition of the "defective" IAPs *in vivo* (Ribet et al., 2008; Saito et al., 2008). Some of these progenitor IAPs have intact *gag, pol, prt, and env* genes. One of them was shown to have all the characteristics of a retrovirus, producing particles at the cell membrane and releasing infectious virions. Non-autonomous, defective IAPs can be derived from this progenitor in the laboratory (Ribet et al., 2008). Note that IAP mobility *in trans* differs from L1 mobility that mainly occurs *in cis*.

Early transposon (Etn)

Another element that has been mobilized *in trans* is the mouse early transposon (Etn). This element is repeated hundreds of times in the mouse genome and has given rise to at least eight different cases of recent disease-causing events in the mouse. The Etn contains LTRs and other features of LTR-retrotransposons, plus sequence unrelated

to that of LTR-retrotransposons and retroviruses. These elements also lack intact ORFs. Mager used computer searches to detect a small region of previously unrecognized type D retroviral *pol* homology within ETn elements. She then used this small region of homology to isolate a family of mouse endogenous proviral elements with *gag*, *prt*, and *pol* genes similar to simian type D viruses. This new family of mouse endogenous proviruses, called MusD, is present in several hundred copies in the mouse genome. Interestingly, the MusD LTRs and other regions are closely related to ETn subfamily members that have recently transposed. MusD elements predate the ETns, indicating that ETns were likely created via recombination events resulting in a near complete substitution of MusD coding sequences with unrelated DNA (Mager and Freeman, 2000). The Heidmann lab has gone on to isolate three intact MusD elements from the mouse genome and demonstrated retrotransposition *in trans* of Etn in tissue culture driven by an active MusD element (Ribet et al., 2004; Ribet et al., 2008). Thus, mouse ETns use the proteins supplied by MusD proviruses for retrotransposition *in trans*.

Mammalian apparent LTR-retrotransposon (MaLR)

So what about the other disease-producing mouse retrotransposon, MaLR? Mammalian apparent LTR-retrotransposons (MaLRs) are a superfamily of perhaps 40,000–100,000 members that continue to retrotranspose and cause disease in mice. These retrotransposons have structural similarities to retroviruses, but the putative product of a 1350 nucleotide ORF found in the consensus sequence does not resemble any retroviral protein. This internal sequence is usually excised in inserted MaLRs. These elements are present in rodents, primates, and other species, suggesting that their origin dates back more than 80–100 million years. Together, disease-producing insertions of LTR-retrotransposons (IAP, Etn, and MaLR) outnumber those of L1s in the mouse by roughly five fold.

Other SINEs

Like their Alu homologue in primates, the ~350,000 B1 elements in rodents evolved from the 7SL RNA gene, a small non-coding RNA species. Also like Alu, they are transcribed through an internal RNA

polymerase III promoter. However, a B1 element regulates the mouse and rat gene, *Nkg2d*, and represents a novel source of RNA polymerase II promoter activity (Lai et al., 2009). B2 elements, close relatives of B1 that instead evolved from tRNA, have also been found occasionally to provide a polymerase II promoter to a mouse gene. Both B1 and B2 elements are non-autonomous retrotransposons that are a few hundred nucleotides long and do not encode any protein. One retrotransposition event involving B1 has been found that caused a mouse disease (Gilbert et al., 2004), but no B2 disease-producing insertion has been observed to date. Both of these elements can retrotranspose in tissue culture using the reverse transcriptase of mouse L1 *in trans*. Interestingly, B2, the element that seems not to retrotranspose in nature, retrotransposes in cell culture at a rate that is 20 to100 times greater than that of B1 (Dewannieux and Heidmann, 2005).

Ultraconserved SINEs

Then there are the mysterious ultraconserved SINEs in the genome. These SINEs, present in mammalian genomes, are likely derived from ancient retrotransposons. They are about 200 nucleotides in length, are present in a few hundred to 1000 copies in the genome depending on the species, and are extremely conserved (100% identity among all mammals). One such SINE originated some 410 million years ago and is still likely active in the "living fossil" fish, the Indonesian coelacanth. This SINE has one copy that acts as a long-range enhancer for an important neuro-developmental gene, ISL1, in the mouse, and other copies of this SINE are likely functional as enhancers (Bejerano et al., 2004).

Another ultraconserved SINE, AmnSINE1, has a very similar story. It too is about 200 nucleotides long and is present in about 125 copies in the mouse. Similar to the "living fossil" SINE, two individual AmnSINE1s are long-range enhancers (over 150kb distances) of neuro-developmental genes (Sasaki et al., 2008). The sequences of both of these types of SINEs provide little clue as to how they entered the genome so long ago and replicated to so many copies. Of course, the other mystery is, how and why are they so ultraconserved? A few copies have acquired a known function, but what about the remainder?

25

Effects of retrotransposons on mammalian genomes

Mammalian genome evolution has in large part been driven by retrotransposons. In previous chapters, I've discussed our work on the 17–20% of mammalian genomes that are L1 sequence, mostly L1 remnants. Another 10–12% of those genomes are repeat sequences, such as Alu elements, that have been inserted into the genome by retrotransposition using the L1 endonuclease and reverse transcriptase. Both L1 and Alu elements have a substantial effect on the evolution of genomes. I've mentioned the L1 insertions in humans that occur at a rate that is still unknown, but recent estimates place it between 1 in 100 and 1 in 150 individuals. Alu insertions are more frequent, occurring at a rate of roughly 1 in 50 meioses. Some fraction of insertions is inherited, but a potentially much larger number are somatic and not inherited. How the somatic insertions are distributed among tissues and at various stages of development are still unknown. As previously mentioned, L1 and SVA insertions can themselves contain sequence inversions that change the inserted sequence in an unpredictable way. Previously, I discussed 3' sequence transductions mediated by L1 and SVA along with some effects of the L1 antisense promoter. There are many other effects produced by mammalian non-LTR elements on genome evolution, gene expression, and possibly on human behavioral diversity.

Effects of purifying selection on the distribution of retrotransposons in human genomes

Although many retrotransposon insertions are likely neutral, a number must be mildly detrimental even if they do not cause overt

disease. These insertions should be subjected to purifying or negative selection over time. In 1988, Korenberg and Rykowski found by *in situ* hybridization that Alus are concentrated at Giemsa negative chromosomal bands while L1s were mainly found at Giemsa positive bands (Korenberg and Rykowski, 1988). Giemsa negative bands contain gene-rich regions, while Giemsa positive bands are generally gene poor. However, we now know that since Alu elements are mobilized into the genome by L1 reverse transcriptase and the two elements have the same insertion site sequences, the genome-wide distribution of Alus and L1s should be identical at the moment of their insertion. Because the insertion site sequence, 5'-TTTT/AA-3' (where / denotes the cleavage site), is very common in the genome, both Alu and L1 should have very wide genomic distributions. Thus at the time of their insertion, both Alu and L1 should enter the genome at similar, nearly random sites. Any changes in the distribution of these elements should occur due to selection over thousands to millions of years after their insertion. Jumping ahead to 2001 and the human genome reference sequence (Lander et al., 2001), we find that Korenberg and Rykowski's observations were indeed correct and at this time: Alus and L1s do have different genomic distributions. However, young human-specific L1s are present in a broad distribution that is not skewed away from genes. Young L1s are distributed within introns of genes as expected, but their orientation is skewed with a marked deficiency of L1s in the same orientation as the gene (sense orientation) and the expected number of L1s in the antisense orientation to the gene (Ewing and Kazazian, 2010b; Symer et al., 2002). Thus, there appears to be fairly rapid, purifying selection against both intergenic L1s and intronic L1s inserted in the sense orientation. Presumably since Alus are very small (~300 bp), they are tolerated in gene-rich and intronic regions.

To look at an even broader effect of L1 insertion on the genome, Boissinot et al. compared sex chromosomal and autosomal regions of similar GC contents and found that both the human X and Y chromosomes contain many times as many full-length (FL) old L1 elements per megabase as the autosomes (Boissinot et al., 2004). Also both sex chromosomes contain more of the longer, but not quite full-length, L1s than the autosomes. However, the autosomes are not deficient in short L1s relative to the sex chromosomes. Because the X and Y chromosomes in males can't use recombination to remove deleterious

sequences, they concluded that most full-length L1s were deleterious and subject to purifying selection. Thus it appears that there exists negative selection for any L1 in gene-rich genomic regions and for full-length L1s throughout the genome. In addition, if longer, perhaps full-length, L1s are used as "booster stations" of the Xist signal for X chromosome inactivation, then there may also be positive selection for full-length L1s on the X (see discussion in a later section of this chapter).

Retrotransposition of Alu, SVA, and mRNA by L1 *in trans*

Our ability to isolate L1 precursors of insertions that were identical in sequence to the insertion over its length was good evidence for "*cis* preference" as the usual mode of L1 retrotransposition (Dombroski et al., 1991). *Cis* preference means that the ORF1p and ORF2p translated from a particular L1 RNA stay with that RNA to provide the activities important for its insertion. This initial evidence was strengthened by finding that full-length insertions in both human beings and mice always contained intact ORFs, in spite of the fact that only very few full-length L1s in these species have intact ORFs. Later, both the Heidmann and Moran groups demonstrated *cis* preference in cell culture experiments (Esnault et al., 2000; Wei et al., 2001), and Kulpa and Moran showed it biochemically (Kulpa and Moran, 2006). Now it is regarded as proven.

However, this leads to an incompletely resolved mystery in the field. If L1s obey *cis* preference, why do L1s retrotranspose Alus *in trans* at roughly twice the rate that they mobilize themselves? Yes, we do know that Alu RNAs are aided in getting to the ribosomes by the signal recognition particle (SRP) proteins 9/14 (Sarrowa et al.,1997), but it seems like that can't be the whole answer. Something further must favor Alu retrotransposition *in trans*. How do Alus hijack the L1 proteins when other L1s can't seem to pull it off?

Retrotransposition of Alu by L1 *in trans* was beautifully shown in cell culture by Thierry Heidmann's lab in 2003 (Dewannieux et al., 2003). Heidmann put a 7SL enhancer on the Alu to aid Alu transcription by RNA polymerase III. He also changed the intron in the retrotransposition cassette from the γ-globin intron to a self-splicing intron and mutated the cassette to eliminate potential terminators of

RNA polymerase III transcription. He then co-transfected a retrotransposition-cassette marked Alu that had previously retrotransposed in a human patient with an unmarked active L1.2B or an $L1_{RP}$. Interestingly, although my lab had not tested L1.2B in cell culture up until this time, we had sent him this element five years earlier instead of L1.2A, the element we were testing. Fortuitously for Heidmann, L1.2B was many times more active than L1.2A. He obtained retrotransposition of Alu that was dependent on ORF2 of L1.2B but not on ORF1 (Figure 25.1). This result was interesting because ORF1 is required for retrotransposition of L1 itself. As mentioned earlier, Alu is thought to interact with the L1 reverse transcriptase as the L1 ORF2p is being synthesized on the ribosome because Alu associates with SRP9/14, protein components of the signal recognition particle that in turn associate with ribosomes (Boeke, 1997).

Earlier, Heidmann along with Moran had shown that processed pseudogene formation is also dependent upon L1 reverse transcriptase (Esnault et al., 2000; Wei et al., 2001). There are some 8,000 copies of processed pseudogenes in the human genome. Processed pseudogenes are copies of messenger RNAs that have been retrotransposed. They are intronless, end in a poly A tail, and are usually surrounded by short target site duplications, like those seen for L1 and Alu elements. Processed pseudogenes do not arrive at their new genomic sites containing their required RNA polymerase II promoters. This is because these promoters are external to the messenger RNA sequence at their original site. Thus, processed pseudogenes are usually inactive because they are not transcribed.

However, interesting examples of reactivated genes derived from processed pseudogenes have been found. These include a number of examples of the retrotransposition of a cyclophilin A messenger RNA into the TRIM5 gene during primate evolution. In at least one of these, the expressed TRIM5-cyclophilin A hybrid protein confers resistance to HIV-1 in the owl monkey (Sayah et al., 2004). A second example of a reactivated processed pseudogene is the fibroblast growth factor 4 (fgf4) retrotransposed copy in the dog. This so-called retrogene has accumulated mutations that have changed its function. Now certain dog breeds, such as dachsund, corgi, and basset hound, have a short-legged appearance (chondrodysplasia) on the basis of the activity of this fgf4 retrogene in their genomes (Parker et al., 2009).

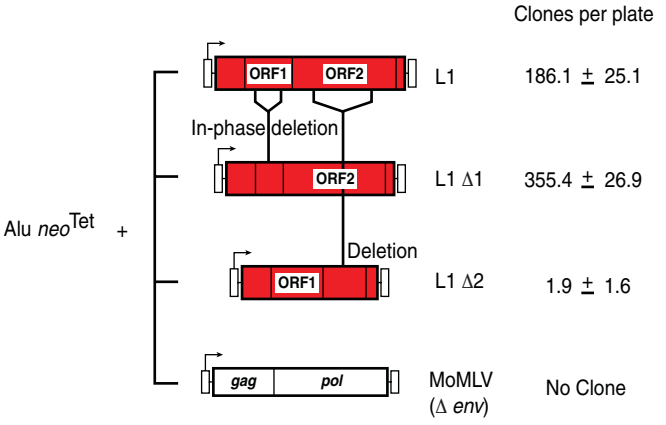


Figure 25.1 Elements required *in trans* for Alu retrotransposition. Assays for Alu retrotransposition were carried out with an Alu that had previously retrotransposed in a human patient marked with a retrotransposition cassette containing a self-splicing intron. Transfected *in trans* was an active L1 with expression vectors rendered defective for ORF1 (423-bp in-phase deletion) (L1Δ1) or ORF2 (2,137-bp deletion) (L1Δ2) and with an expression vector for retroviral Gag-Pol proteins (from the Moloney Murine Leukemia Virus). The constructs used *in trans* were derived from the L1.2B element.

As mentioned earlier, SVA elements are also likely retrotransposed *in trans* using the L1 reverse transcriptase. All the signs of L1 action are present in natural retrotranspositions of SVA in humans. The alpha-spectrin insertion of single copy DNA previously mentioned demonstrates that retrotransposition events containing only single-copy sequence can occur. These insertions happen following a 3' transduction of an SVA (or L1) and severe truncation of the reverse transcribed product. It is also possible that many poly A stretches in the human genome are derived from retrotransposition of a reverse transcript of an SVA, Alu, or L1 element in which reverse transcription aborts in the poly A tail.

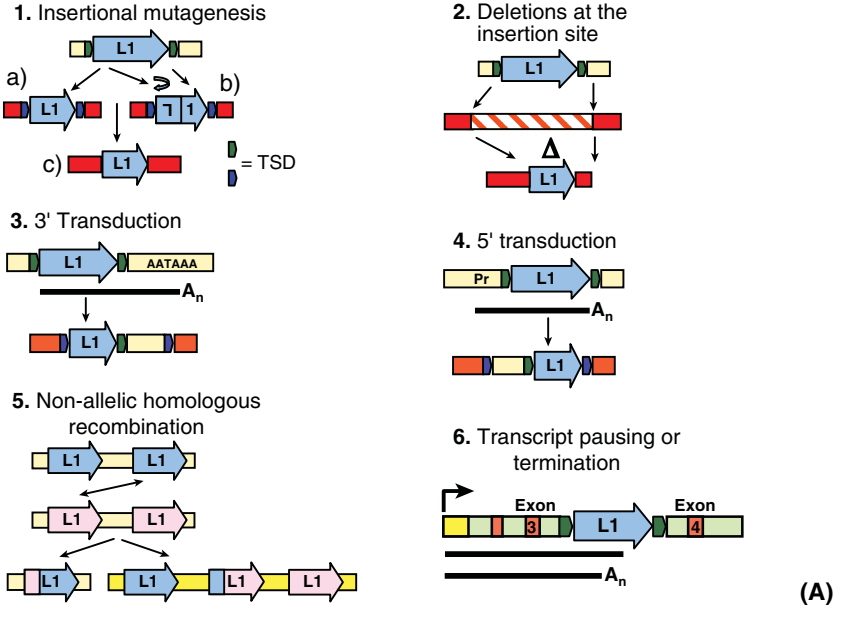
So what sets up SVA sequences for *trans* retrotransposition by L1s? Certainly, SVA sequences are much more retrotransposable than the average mRNA, which can produce a processed pseudogene. Perhaps SVAs get to the ribosomal neighborhood like Alus do. Yet it is unlikely that the antisense oriented Alu sequences in the SVA sequence would produce this effect. Perhaps SVA RNA interacts with the L1 RNP in the cell nucleus. The bottom line is that the mechanism

by which SVA has become favored for retrotransposition by the L1 machinery remains unknown.

Although one can show that one third of the human and mouse genomes is derived from L1 retrotransposition, some investigators have tried to find DNA sequences that inserted into the genome more than 150 million years ago. These sequences have acquired so many mutations from the moment that they inserted until the present that it is difficult to discover their origins. However, some attempts to do just that have had some success, and now it is estimated that >50% of the genome is composed of retrotransposed sequences.

Non-allelic homologous recombination

Retrotransposed sequences in the mammalian genome occur in so many copies that their sheer copy number leads to the possibility of mispairing and homologous recombination. Mispairing of similar sequences and unequal crossing over (homologous recombination), results in duplications and deletions. These can occur between two Alu sequences or two L1 sequences (see Figure 25.2A and B on how retrotransposons affect the cell).



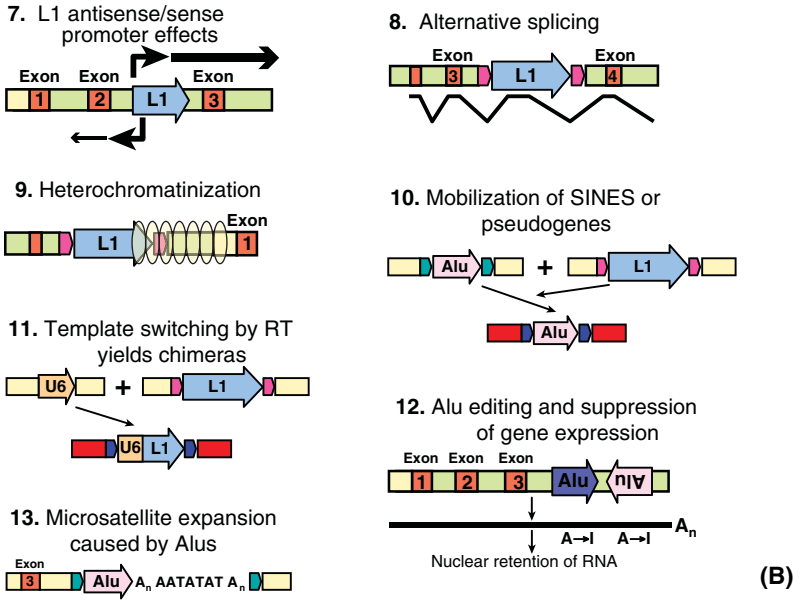


Figure 25.2 How retrotransposons affect the cell. Insertions may be full length or 5' truncated or contain inversions (1a and b). EN-independent insertions also occur at low frequency (1c). Deletions may accompany insertions (2). Flanking sequence, either 5' or 3', may be carried along in a retrotransposition event (3 and 4). Mispairing and crossing over between LINEs or SINEs can lead to deletions or duplications (5). Transcriptional pausing can occur in retrotransposons, and poly A signals within an L1 can lead to premature termination of transcription (6). The antisense promoter in the L1 5' UTR can produce new transcription start sites for genes upstream of the L1 on the opposite strand (7). Splice sites within L1s residing in introns can lead to new exons within genes (8). L1s can alter the chromatin state, thereby altering gene expression (9). L1 reverse transcriptase can mobilize Alu, SVA, and mRNA, leading to further genome expansion (10). Template switching of L1 reverse transcriptase from L1 RNA to other sequences, like U6 RNA or Alu RNA, can produce chimeric insertions in the genome (11). Editing by ADAR of inverted Alus can suppress gene expression by nuclear retention of the mRNA (12). Alu elements seed formation and expansion of microsatellites that have been occasionally associated with disease (13).

There have been at least 50 examples of Alu-Alu homologous recombination events causing cases of human disease (Batzer and Deininger, 2002; Lehrman et al., 1985). On the other hand, although L1s offer the possibility for longer stretches of homology than do Alus, there have been only a handful of known occurrences of L1-L1

homologous recombination causing human disease. On an evolutionary scale, the story is a little different. Batzer's group has compared the chimp and human genomes for deletions due to homologous recombination between repeat sequences. Roughly 660 Alu-Alu recombination events have occurred in the chimp lineages deleting about 700kb of DNA, and ~500 recombination events have deleted ~400kb from the human lineage (Han et al., 2007; Han et al., 2008). Similarly, there have been 70+ L1-L1 recombination events in the human lineage, deleting about 500kb of DNA (Han et al., 2008). Thus these homologous recombination events have deleted about 10^6 basepairs from the DNA of each species, including some gene coding regions. In addition, one notable L1-L1 homologous recombination in present day humans deleted 520kb, including the EvC loci, thereby causing Ellis van Creveld syndrome in a consanguineous family (Temtamy et al., 2008).

Both L1 insertions and L1-mediated insertions can occasionally result in deletion

In cell culture, about 10% of L1 insertion events are accompanied by deletion of genomic DNA at the insertion site. These insertions are distinctive in that they lack target site duplications (Gilbert et al., 2002; Symer et al., 2002). In human beings, a few insertions, perhaps 10% of those observed, are accompanied by deletions, including notably one deletion of 46kb causing deficiency of the pyruvate dehydrogenase complex that accompanied the insertion of a full-length L1 (Mine et al., 2007). Deletions can also rarely be seen associated with insertions mediated by L1. As mentioned earlier, a 14kb deletion of the HLA-A gene associated with an SVA insertion was likely mediated by an active L1 acting *in trans* (Takasu et al., 2007).

3' and 5' transductions of genomic sequence associated with L1 and L1-mediated retrotransposition

As mentioned in Chapter 11, "A quirk of L1 elements..." 3' transduction of sequence occurs in at least 10% of human L1 retrotranspositions (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000). These are due to failure of cleavage of the L1 RNA at the

3' end of the element. Subsequently, cleavage of these elongated transcripts occurs just downstream of the next poly A signal sequence. Ostertag's and Batzer's work showed that 3' transduction can also occur in association with SVA retrotransposition (Ostertag et al., 2003; Xing et al., 2006).

In cell culture, Moran had shown that 5' transduction during retrotransposition is also possible, but 5' transduction requires a transcription start site upstream of the 5' end of L1 and continued reverse transcription beyond the L1 itself (Moran et al., 1996). The best example of a natural 5' transduction occurred in association with a mouse L1 insertion when a full-length L1 retrotransposed into the *Nr2e3* gene on mouse chromosome 9, causing retinal degeneration. This L1 was, as expected, a T_F family member and contained a strong 6-monomer promoter at its 5' end. In addition, it contained 28 nucleotides between the end of the monomers and the 5' target site duplication. These nucleotides were single copy, and they allowed the unequivocal identification of the precursor L1 on mouse chromosome 4. This L1 had sequence identity with the insertion and was situated just downstream of the 28-nucleotide transduction sequence. Thus this precursor L1 was identified through the 5' transduced sequence. Because of the sequence identity of the precursor to the inserted L1, the insertion also demonstrated the importance of *cis* preference, not only for human L1s, but also for mouse L1s (Chen et al., 2006).

Effects on gene expression

L1s residing in genes appear to have a number of effects on gene expression. There is evidence from the Boeke lab that because of the high A content of L1, RNA polymerase II tends to pause as it reads through L1 DNA. When L1 lies in an intron in the same orientation as the gene, this pausing may affect transcription of the gene (Han et al., 2004). On the other hand, L1s in the antisense orientation to the gene should not have this effect. In support of the existence of this effect in nature, L1s in introns of genes tend to be much more frequent in the antisense orientation than the sense one, suggesting selection against L1s in the sense orientation.

Another effect of L1s in introns of genes is due to the presence of many premature polyadenylation sites within the L1 body. These sites

can lead to cleavage within the L1 RNA after any of these sites. In cell culture, only about 10% of the L1 RNA is full length, while the great majority is truncated due to this effect. Although premature polyadenylation occurs in both L1 orientations, it predominates in the antisense orientation (Perepelitsa-Belancio and Deininger, 2003). Wheelan et al. have found 15 examples in the human genome of a phenomenon that they call "gene breaking" of the RNA transcript. In these events, a full-length L1 sitting in a gene intron in the antisense orientation can lead to breakage of the transcript into two parts. First, the 5' end of the transcript ends with a premature polyadenylation site in the body of the L1. Second, the 3' portion of the transcript restarts with the antisense promoter (mentioned earlier in Chapter 21, "The brilliant young lady from China,") in the 5' untranslated region of the L1. Both of these breakage events can lead to the creation of new genes if the new RNA is retrotransposed back into the genome (Wheelan et al., 2005).

In addition, remnants of transposable elements have been found to encode enhancers for expression of genes located many kilobases away from them. One such repetitive element enhancer was in the long terminal repeat (LTR) of an endogenous retrovirus (ERV) present upstream of the locus control region (LCR) of the β-globin gene cluster (Pi et al., 2010). The LTR is responsible for stimulating β-globin synthesis and reducing fetal globin production. When the LTR is deleted γ-globin production increases substantially. Other repetitive elements, Alu and L1 fragments, affect gene expression within the 5' UTR of the gene. An example is the repetitive elements in the 5' UTR of a human zinc-finger gene (Landry et al., 2001).

Antisense promoter effects

The human L1 has not only the expected sense promoter within the sequence at the 5' end of the element, but also a relatively weak antisense promoter that begins RNA synthesis around nucleotide +500 in the 5' untranslated region (see Chapter 21). The antisense promoter has roughly 10% the activity of the sense promoter. This antisense promoter, originally described by Speek, supplies an alternative start site from within many full-length L1s for genes lying in the opposite orientation upstream to the L1s. Thus this antisense promoter has an

effect on the expression of around 100 genes (Speek, 2001; Nigumann et al., 2002).

Mouse L1 also has an antisense promoter activity, but it lies in the ORF1 region and not in the monomer promoter region. Moreover, the antisense sequence of active mouse L1s contains a number of splice sites, both donor and acceptor sites, that are used to add exons to mRNAs (Zemojtel et al., 2007).

Template switching produces L1 and other chimeras

The human genome contains over 150 copies of L1 chimeras in which a small nuclear RNA, usually the RNA splicing factor U6, is attached at its 3' end to a 5' truncated L1 copy (Buzdin et al., 2003). These U6-L1 chimeras are scattered throughout the human genome, and they began to be formed in the primate lineage as long ago as 50 million years. A much smaller number of chimeras contain other U RNAs, 5s rRNA, and 7SL RNA at their 5' ends, and Alu or mRNAs (processed pseudogenes) at their 3' ends (Gilbert et al., 2005). Because the bulk of these chimeras end in a poly A tail and are flanked by target site duplications, they must be formed through the use of L1 reverse transcriptase, either *in cis* or *in trans*. In the most common case, the reverse transcriptase must begin synthesis of L1, but then switch templates to the multiple Ts at the 3' end of U6 RNA. Template switching accounts for essentially all of these chimeric events, including those rare events in which Alu or mRNA forms the 3' portion of the chimera (Garcia-Perez et al., 2007a). In mammalian genomes, a small number of chimeras that formed from fusion of three different RNAs have also been identified (Gogvadze et al., 2005). These sequences must result from double template switches of the L1 reverse transcriptase.

U6-L1 chimeras have been readily identified in the retrotransposition cell culture assay, In fact, in HeLa cells roughly 1 in 15 G418 (*neo* resistant) retrotransposition events is a U6-L1 chimera, suggesting that this chimera occurs at least 10 times more frequently in cell culture than it has in genome evolution (Gilbert et al., 2005). How the small nuclear RNA located in the nucleolus contacts the L1RNA and L1 reverse transcriptase remains a mystery, although there is evidence that the L1 RNP may spend a portion of its life in the nucleolus.

A role in X chromosome inactivation?

A role for L1s located on the X chromosome was first proposed by Gartler and Riggs in 1983 (Gartler and Riggs, 1983) and elaborated by Lyon in 2000 (Lyon, 2000). The hypothesis stated that L1s acted as booster stations to spread the inactivation signal mediated by the RNA of the *Xist* gene on the inactive X chromosome in mammals. This hypothesis has remained controversial with evidence presented both in its favor and against it. In its favor, the X chromosome contains about 2 to 3 times as many L1s per megabase as the autosomes, and the distribution of L1s on the X is non-random with clusters of L1s around the X-inactivation center and genes that are preferentially inactivated. On the negative side, a South American rodent that has lacked L1s for at least eight million years retains appropriate X inactivation (Cantrell et al., 2009). Likewise, another rodent, the spiny rat has an XO sex chromosome constitution, so it does not require X inactivation. However, it does have an excess of L1 elements on its lone X chromosome relative to its autosomes (Scott et al., 2006).

Very recently, a study from the Heard lab has presented further evidence in favor of a role for L1s in promoting heterochromatization of X chromosome regions (Chow et al., 2010). They show that LINEs participate in creating a silent nuclear compartment into which genes become recruited. A subset of young L1 elements in the mouse is expressed during X-chromosome inactivation rather than being silenced. They demonstrate that this L1 expression requires the specific heterochromatic state induced by *Xist*. These young expressed L1s often lie within regions of the X chromosome that escape inactivation, even though they are close to genes that are inactivated. Thus it is possible that L1s may facilitate X chromosome inactivation at different levels, with silent L1s involved in assembly of the heterochromatic nuclear compartment induced by *Xist*, while active L1s participate in local propagation of the X chromosome inactivation signal into regions that would otherwise escape it.

Endonuclease-independent L1 retrotransposition

L1 can enter the genome not only after a DNA nick made by its endonuclease, but also by an endonuclease-independent mechanism. These latter insertions have unusual features, including integration at

atypical target sites that are not 5'-TTTT/AA-3', deletions at the target site, incorporation of other DNA sequences at the insertion site, and initiation of reverse transcription of L1 RNA internal to its 3' end. How these unusual features are produced is poorly understood, but they are believed to be the result of reverse transcription priming from naturally occurring DNA nicks/breaks in the chromosome, followed by resolution using host DNA repair pathways. Thus this mechanism has been thought of as applying a "genome band aid" and is another way in which L1 has affected genome evolution. Endonuclease-independent insertions have also been rarely observed among naturally occurring events. These insertions validate the hypothesis put forward by Edgell and Huchison in 1984 (Voliva et al., 1984).

In cell culture, endonuclease-independent insertions are only observed with any frequency in cells deficient in DNA repair after transfection with an endonuclease-defective L1 (Morrish et al., 2002). Interestingly, in this circumstance they occur mainly at telomere sequences by a mechanism very similar to that of telomerase reverse transcriptase (Morrish et al., 2007). For about ten years, the sequence similarities between non-LTR retrotransposon reverse transcriptase and telomerase reverse transcriptase have led scientists to speculate as to which one is the evolutionary precursor of the other (Eickbush, 1997). Although the concentration of endonuclease-independent retrotransposition events at telomeres does not resolve this question, it certainly underlines the relationship between these ancient reverse transcriptases.

Effects of somatic insertions

As I mentioned in Chapter 22, "Hiroki's Big Surprises," Alysson Muotri in the lab of Fred Gage showed that L1 retrotransposition events occur in neuronal precursor cells, and the insertions tend to be concentrated in genes active in neurons. Most impressively, they showed that retrotransposition events occur in developing neuronal cells in many parts of the brain in transgenic mice (Muotri et al., 2005). This is further evidence that mammals are somatic mosaics for L1 insertions. Recently Coufal, working with Gage and Moran, has found in human cadaver samples that human hippocampus contains more L1 insertions than human heart and liver (Coufal et al., 2009).

They estimated that the hippocampal cells contain ~80 more L1 copies than heart or liver cells. Whether these copies are all new retrotransposition events remains to be determined. This is quite an astonishingly high number because, depending upon the timing of the insertions in development, many cells will have different complements of new somatic insertions. Muotri has also recently shown that *de novo* L1 retrotransposition is greater in the hippocampus of transgenic mice subjected to exercise as compared to sedentary transgenic mice (Muotri et al., 2009). These data suggest that environmental changes could alter the frequency of somatic retrotransposition in parts of the mammalian brain, leading to the speculation that L1 and L1-mediated retrotransposition may play a significant role in the diversity of human behavior. Perhaps a portion of the difference between the behavior and/or psychopathology of identical twins is due to retrotransposition events affecting the expression of key neuronal genes. Because these insertions are somatic and not heritable, it is unclear whether the ability of retrotransposons to produce human behavioral diversity would be subjected to positive selection. Meanwhile, the possibility of behavioral modification due to L1 retrotransposition is fascinating—but still controversial.

26

Host factors involved in L1 retrotransposition

Over evolutionary time, there has clearly been a battle between mobile DNA and host organisms. Mobile DNA is continuously expanding genomes, altering them, and affecting the expression of many genes. Thus, mobile DNA has added to the plasticity and diversity needed to continue the process of the evolution of species. Although this mutation process is usually detrimental, that is, most mutations are either bad or at best neutral for an organism, it can be occasionally favorable and worthy of genetic selection. On the other hand, organisms can't allow mobile DNA to take over their genomes. Organisms need the means to control the spread and activity of mobile DNA. This is the continuous battle that rages. Mobile DNA expands the genome. The genome fights back with controls on this expansion.

Although we can predict that the human genome produces a very large number of host factors and strategies to thwart L1 retrotransposition, only a few are known today. However, there is a new experimental system that holds promise for discovery of a number of such factors in the near future. I describe this new system first because of its potential and then go on to discuss what is now known about host factors.

A few years ago, Russell Poulter made a fascinating discovery in *C. albicans*, the common yeast that infects humans and is very far removed evolutionarily from *S. cerevisiae*, the budding yeast that is so well known to biologists and geneticists. *S. cerevisiae* has about 30 Ty1 elements and a few other Ty elements, and that's it (Kim et al., 1998). Ty elements are LTR-retrotransposons, so this yeast completely lacks

non-LTR or L1-like retrotransposons. Poulter found three similar L1-like elements in *C. albicans* that he called *zorros* (Goodwin et al., 2001). One of these, *zorro3*, appeared to be full length and contained two intact ORFs with sequence similarities to human and mammalian L1 elements. He then marked the element with a retrotransposition cassette and showed that it could retrotranspose in *C. albicans* cells (Goodwin et al., 2007).

Han has recently shown that *zorro3* can be reengineered for use in *S. cerevisiae.* He changed the codons of the two ORFs to fit the codon usage of *S. cerevisiae*, and then he added a retrotransposition cassette used in budding yeast, a backward gene for histidine production disrupted by an artificial forward intron into the 3' untranslated region of *zorro3.* [This cassette is the same one used previously by Garfinkel and Curcio in many studies of Ty1 retrotransposition in budding yeast (Curcio and Garfinkel, 1991).] Interestingly, this L1-like element from *C. albicans* was now able to retrotranspose remarkably well in *S. cerevisiae*. Analysis of mutants and the structures of the insertions demonstrated an amazing resemblance to the retrotransposition events mediated by mammalian L1s. The data suggest that *S. cerevisiae*, an excellent system for genetic analysis, has retained the basal host machinery for L1 retrotransposition (Dong et al., 2009). Thus this experimental system should be extremely useful in identifying and characterizing cellular factors involved in mammalian L1 retrotransposition. Because there are now available gene knockouts for every yeast gene, it should be possible with mass screening to identify a number of host genes that when knocked out either increase or decrease *zorro3* retrotransposition.

APOBEC3 proteins affect reverse transcription of L1

APOBEC3 (apoprotein B-editing catalytic polypeptide 3) proteins are a novel group of proteins involved in innate immunity that act against retroviral infection. In humans, there are seven APOBEC3s—A, B, C, D, F, G, and H—that act as cytidine deaminases. They deaminate cytidine to uracil in the growing first strand of DNA during reverse transcription, causing many mutations and genome instability. However, the Vif-proteins of HIV-1 and other

retroviruses are protective against the action of APOBEC3s. Yet because retroviral reverse transcription occurs by a very different mechanism from that of non-LTR retrotransposons (in the cytoplasm within viral particles as opposed to the nucleus on chromosomal DNA), it is surprising that some APOBEC3s also affect L1 retrotransposition. A number of groups have shown that APOBEC3s A and B can enter the nucleus and inhibit both L1 and Alu retrotransposition in cell culture (Bogerd et al., 2006; Muckenfuss et al., 2006; Stenglein and Harris 2006; Schumann 2007). Interestingly, this inhibitory effect is not related to a cytidine deaminase activity because the inserted DNA does not have G-A changes (the complementary nucleotides to C-T) in its sense strand sequence. Another APOBEC3, APOBEC3G (A3G), greatly inhibits L1-dependent retrotransposition of marked Alus (Chiu et al., 2006; Hulme et al., 2007). This effect is not through inhibition of L1 function in retrotransposition of Alus, but by sequestration of Alu RNAs in cytoplasmic, high-molecular-weight A3G complexes away from the nuclear L1 machinery (Chiu et al., 2006). So it now appears that the different APOBEC3s act to suppress HIV-1, L1 and Alu by three different mechanisms, cytidine deamination for HIV-1, an unknown effect on retrotransposition for L1 and Alu, and, in the case of APOBEC3G, a sequestering of Alu in cytoplasmic complexes.

Inhibition of non-LTR retrotransposons by small RNAs

The field of small RNAs is moving very rapidly with new information appearing from one week to the next. I briefly discuss the three classes of small RNAs in *Drosophila melanogaster* to provide an overview. Then I discuss the role of the mammalian Piwi homologs in control of mammalian retrotransposons.

The three types of small RNAs are classified according to their mechanisms of biogenesis (Zhou et al., 2009). MicroRNAs (miRNAs) are approximately 21–23 nucleotides, ubiquitously expressed, and are processed from hairpin-like precursors first by Drosha/Pasha and then by Dcr-1/Loquacious complexes. These RNAs usually associate with AGO1 and regulate the expression of protein-coding genes.

Piwi-interacting RNAs (piRNAs) are approximately 24–28 nt, associate with Piwi-family proteins, and can arise from single-stranded precursors. piRNAs function in transposon silencing (to be discussed) and are mainly restricted to gonadal tissues.

Endo-siRNAs are approximately 21-nt and are found in both germline and somatic tissues. They are produced by a different Dicer, Dcr-2 and do not depend on Drosha/Pasha complexes for processing. They predominantly bind to AGO2 and target both mobile elements and protein-coding genes. Surprisingly, some endo-siRNAs depend for their synthesis on the dsRNA-binding protein Loquacious (Loqs), which is thought to be a partner for Dcr-1 and a cofactor for miRNA biogenesis. However, endo-siRNA production depends on a specific Loqs isoform, Loqs-PD, which is distinct from Loqs-PB, which is required for the production of microRNAs. Paralleling their roles in the biogenesis of distinct small RNA classes, Loqs-PD and Loqs-PB bind to different Dicer proteins, with Dcr-2/Loqs-PD complexes driving endo-siRNA biogenesis and Dcr-1/Loqs-PB complexes driving microRNA biogenesis (Zhou et al., 2009)

piRNAs, the second class just mentioned, appear to inhibit the accumulation of L1, IAP, and other retrotransposon RNAs in male germ cells by stimulating *de novo* methylation of retrotransposon regulatory sequences. These small RNAs appear to act specifically on retrotransposon RNA and not the expression of "single-copy" genes. Piwi proteins interact with piRNAs and have been implicated also in control of transposon RNA accumulation and in methylation of transposable elements in mammals (Nakano et al., 2008).

In the restricted window of development in which methylation patterns on DNA are set during embryogenesis, there is robust expression of two Piwi protein homologues, MILI and MIWI2 in the mouse. In that species, dispersed copies of transposable elements initiate the pathway, producing primary piRNAs, which mostly join MILI in the cytoplasm. MIWI2 has nuclear localization, and its association with piRNAs depends upon MILI. MIWI2 complexes are enriched for secondary piRNAs antisense to the elements that it controls. Loss of MILI or MIWI2 proteins leads to increased production of L1 and IAP retrotransposon RNA (Aravin et al., 2007; Carmell et al., 2007), presumably due to the defective DNA methylation that

occurs in these mutant mice. Male mice carrying these mutants have meiotic catastrophe of sperm cell development and are infertile. Another knockout, affecting the Maelstrom protein, also has increased production of L1 RNA and meiotic collapse, leading to male infertility (Soper et al., 2008), Because piRNAs are still produced in dnmt3L mutants, which fail to methylate transposons (Bourc'his and Bestor, 2004), the Piwi pathway lies upstream of known mediators of DNA methylation. Thus piRNAs are involved in determining the specificity of DNA methylation in germ cells. Although understanding of Piwi proteins and piRNA biology has increased substantially over recent years, major gaps still exist in our understanding of these enigmatic RNAs and how they affect mobilization of retrotransposons.

Epigenetic effects on L1 retrotransposition

Mutants that reduce DNA methylation in the mouse appear to increase remarkably the expression of L1s and IAPs (Bourc'his and Bestor, 2004; Yoder et al., 1997). This and other evidence suggest that DNA methylation reduces L1 expression and probably retrotransposition and that DNA hypomethylation increases L1 activity. The correlation of L1 retrotransposition in developing germ cells with relative hypomethylation in these cells supports this notion. Yet the strength of this effect relative to other effects is still an open question.

The protein that binds to methyl groups on DNA and is involved in global DNA methylation, MeCP2, appears to be important in L1 retrotransposition in neural tissues. Muotri et al have found that L1 neuronal transcription and retrotransposition in mice are increased in the absence of MeCP2. Using neuronal progenitor cells derived from human induced pluripotent stem cells and human tissues, they showed that patients with Rett syndrome, a model of autism spectrum disorder caused by mutations in MeCP2, have increased susceptibility for L1 retrotransposition. These data add retrotransposition to the molecular events in human neurological disease (Muotri et al., 2010).

Yet another epigenetic phenomenon has been found that could affect the activity of retrotransposed sequences. Garcia-Perez et al. have discovered that new insertions produced by transfected, marked L1s in embryonic carcinoma cells are shutdown upon their

insertion (Garcia-Perez et al., 2010). The shutdown appears mediated by histone deacetylases (HDACs) considering that inhibitors of HDAC reverse the shutdown. The data suggest that perhaps when new insertions occur, the host may actively eliminate or reduce its ability to re-retrotranspose by modifying the chromatin into which they have landed. Yes, most insertions will be dead on arrival due to 5' truncation of the L1 (it then lacks its internal promoter) or inversion of L1 sequence. Yet here is another potential mechanism to shut down the expression of new full-length retrotransposition events—chromatin modification by HDACs.

27

Why mobile DNA?

Why mobile DNA has attained such a prominent fraction of so many genomes, particularly those of plants and mammals, remains a puzzling question. It seems that in those genomes, a constant struggle between the host and mobile DNA is continually present. We know that transposable elements have been important drivers of genome evolution. But what evidence is there that transposable elements have contributed function to the individual that could have led to its selection and increased proportion in many genomes?

The opossum genome has a very high fraction (52%) of recognizable transposable element sequence, much of which consists of evolutionarily new elements (Mikkelsen et al., 2007). This fact suggests that creation of new transposable elements may be ongoing. The same observation holds for the SVA elements of primates. These elements, now 2700 strong in the human genome, are also a recent creation of the past 20 million years. Yet SVAs are "jumping" now at a substantial frequency, given the number of disease cases caused by SVA insertions that have been identified. In addition, over the past 40 million years, only one L1 subfamily has emerged at any one time in the human lineage. There has been plenty of opportunity for L1 extinction, but it hasn't happened even though the number of active L1s in any individual genome at any time has remained relatively small. To my knowledge there is no living organism that lacks mobile DNA in its genome. Thus transposable elements appear to be a creative force for genome change.

And how have non-autonomous transposable elements survived to become so prominent in genomes? Specifically, why have the single L1 subfamilies continued over millions of years to facilitate the

mobilization of Alu sequences to the extent that Alus now substantially outnumber L1s in the human genome? Why has the genome not eliminated active L1 completely and at the same time eliminated Alu expansion and likely SVA mobility? These questions remain unanswered, but they again suggest that mobile DNA is important to the very existence of organisms.

Mobile DNA has likely affected individual diversity through continuous insertion in neuronal precursor cells in individual humans. Some evolutionary biologists believe that increasing behavioral variation and diversity would be under positive selection even though the somatic insertions producing the effect are not inherited from one generation to the next. Other evolutionary biologists do not believe that such effects would be under positive selection.

Positive effects of mobile DNA on the organism that are inherited from one generation to the next are only hypothetical. Perhaps reverse transcriptase is required at some developmental stage. Perhaps L1 endonuclease has a cellular function. Perhaps a stress response of mobile DNA is crucial in development. These are all unproven suggestions.

Recently, Lu and Clark espoused another interesting view in *Genome Research* (Lu and Clark, 2010). They believe that piRNAs can severely repress the activity of retrotransposons, but some retrotransposons are required to generate the piRNAs. They carry out computer simulations that show that piRNAs can reduce the number of segregating retrotransposons by >50% and increase the fitness of individuals by >2%. They find that retrotransposons that generate piRNAs can easily attain high gene frequencies, but, paradoxically, so can retrotransposons that are targeted by piRNAs because their deleterious effects are reduced. Lu and Clark suggest that piRNAs may shelter retrotransposons by "shielding the host from the deleterious consequences of retrotransposition." Later, when piRNAs attain a higher frequency, host fitness relies on piRNA expression to repress the retrotransposons. This makes piRNAs generated by retrotransposons crucial to the host.

28

The future of mobile DNA research

Genome-wide analysis of recent retrotransposition events

The work of Adam Ewing, a graduate student in my lab and a number of members of the Moran, Devine, Burns and Boeke labs points to one productive future line of research (Beck et al., 2010; Huang et al., 2010; Iskow et al., 2010; Ewing and Kazazian, 2010b). The beginnings of this line of research have been published in mid-2010. Next generation sequencing has led to the ability to obtain billions of nucleotide sequence at one time. Using one company's technology (Solexa), as recently as 2008 these sequences were 35 nucleotides long. In 2009, they increased to 76 and then 100 nucleotides. In the near future, each read will likely provide 200 nucleotides of accurate sequence.

Again, it took an outside source to push me into this project to find the location of all human-specific L1s in any genome. It was Vivian Cheung, a good friend and colleague at Penn, who did the prodding and Adam Ewing who enthusiastically put in the effort. Cheung was working on whole genome approaches to gene expression analysis, and she wondered aloud to me one day why I didn't look at all recent L1 insertions in the human genome. That nudge led me to remember the pioneering work by Gary Swergold (Ovchinnikov et al., 2001) and Richard Badge (Badge et al., 2003) on finding human-specific L1s (L1Hs) in the genome. Of course, that work was pre-next generation sequencing, but both these groups of investigators took advantage of the short, specific sequences near the 3' end of L1Hs that distinguished them from other older L1s. I immediately realized that

Cheung had the nub of a good idea. If we could locate all the L1Hs in any genome, a lot of good biology would follow. Using primers specific for the sequences at the 3' end of L1Hs elements and carrying out PCR into 3' flanking sequence, Ewing developed a technique to find essentially all of the roughly 1000 L1Hs elements in any human genome (Ewing and Kazazian, 2010b). In studying 15 unrelated individuals, he found that, on average, the genome of any two individuals differs at 289 L1Hs sites. That is, in comparing any two genomes, there were nearly 150 different places in which one genome has an inserted L1 not present in the other and vice-versa. He also calculated that the retrotransposition rate for L1 elements is 1 in 140 meioses with 95% confidence limits of 1 in 90 to 1 in 240 meioses. He estimated that the number of relatively common polymorphic L1s in the world population of 6 billion people with a gene frequency >.05 is between 3,000 and 10,000 (Ewing and Kazazian, 2010b). Of course, it is likely from the data of Beck et al., 2010 and Iskow et al., 2010 that the number of very rare or "private" L1Hs in the world population is very much greater, perhaps hundreds of thousands to millions. Hopefully, a better estimate of this number will be forthcoming shortly.

Huang et al. from Kathy Burns' and Jef Boeke's labs used the different technique of hybridization to microarrays of closely spaced oligonucleotides (TIP-ChIP) to detect L1Hs locations (Huang et al., 2010). These labs also carried out PCR off the 3' ends of the elements to restriction sites onto which known oligonucleotides were ligated and then hybridized the PCR products to the array, looking for hybridization to three or more successive oligonucleotides. Although most of their work concentrated on L1s located on the X chromosome, they did carry out genome-wide analyses. This technique also appears to work well. In the genome-wide analysis, among potential L1Hs sites not present in the human reference genome, roughly 60% turned out to be true non-reference L1s. Huang et al. also found a candidate L1 insertion that may have caused an X-linked mental retardation. In screening some 60 patients, they found a rare insertion in a gene expressed in neurons that was not present in any other patient studied or in a panel of control individuals. However, the mentally normal parents of the patient were unavailable, so whether the insertion was *de novo* in the patient or present in one parent could not be demonstrated.

Iskow et al. from the lab of Scott Devine also used a PCR technique to determine the 3' flanks of L1Hs elements (Iskow et al., 2010). This group used the 454 technology for their next-generation sequencing. Their most interesting observation related to a study of paired normal and lung cancer tissue from patients. In studying 20 normal-tumor pairs, they found 9 *de novo* L1Hs insertions in 6 different tumors that were not present in the normal tissue. Moreover, this group studied global methylation in 59 specific genomic regions subject to DNA methylation and found that the tumors in which the new insertions occurred had relative hypomethylation of these regions, while the remainder of the tumors had significantly more DNA methylation in the regions tested. The finding of *de novo* L1 insertions in tumors is striking and suggests that retrotransposition may play a role, either in some general sense, or in the specific etiology of some cancers. Previously, although the cell culture assay of retrotransposition has been carried out in transformed cells, there has been almost no observation of insertions that might play a role in cancer etiology. The finding of Miki et al. in the early 1990s of an L1 insertion disrupting the adenomatous polyposis coli (APC) gene in colorectal cancer that was absent from normal colon tissue of the patient stands as the only example (Miki et al., 1992).

Similar work from Witherspoon et al. of the Jorde lab used high-throughput techniques to detect young Alu elements in human genomes (Witherspoon et al., 2010). They used primers specific to Alu sequence, carried out PCR off the 3' ends of Alus into the flanking DNA, and then sequenced the PCR products by next-generation sequencing. They applied their technique (ME-scan for mobile element scan) to human AluYb8 and AluYb9 subfamilies, the youngest Alu subfamilies. In four individuals, they found 2,758 AluYb8 and AluYb9 insertions, including nearly all those that are present in the human reference genome, as well as 487 that are not and presumably polymorphic in the human population. At a sequencing depth of 355,000 paired reads per sample, the sensitivity and specificity of ME-Scan were both approximately 95%, very high indeed. They concluded that in light of continuing improvements to high-throughput sequencing technology, it should be possible to employ their technique to genotype insertions of almost any mobile element family in many individuals from any species. Ewing in our lab has already

shown that his technique can be successfully modified to detect human SVA polymorphisms.

In the meantime, a large number of individual genomes are being sequenced in the 1000 genomes project. However, nearly all of the genomes in this 1000 genomes study are being sequenced to very low coverage, meaning that many L1Hs locations are missed until sequence coverage is increased substantially. Yet if one combines the sequences from many or all of the individuals for a single analysis, it is possible to detect a large number of non-reference mobile elements, particularly L1s, and because many of the sequences include both the 3' poly A tail and abutting flanking sequence, the insertion sites of many of the non-reference L1s can be located to single nucleotide resolution. The combination of 1000 genomes data along with the data of Iskow et al. (Iskow et al., 2010; Ewing and Kazazian, 2011) have increased the number of non-reference L1Hs elements verified by two or more methods to 1016, bringing the total of known polymorphic L1Hs, including those in the human reference genome to 1419. Unverified L1Hs should, after verification, bring this total to over 1700. Combining these 1700 L1 polymorphisms with an expected total of 2500 Alu polymorphisms and perhaps 200 SVA polymorphisms should yield nearly 4500 mobile DNA variants for study in population genetic and disease susceptibility studies. We plan to use a high-throughput genotyping method on a microarray chip to study all of these polymorphisms simultaneously. These markers should add significant variants for analysis in genome-wide association studies (GWAS).

So let's say that one or more of these techniques is highly successful. What biological facts can we learn? First, even as individuals in the 1000 genomes project are sequenced at high coverage, the project will only give us information about the individuals being sequenced. Yet we can learn from those sequences the extent of presence/absence polymorphism in L1Hs sequence, young Alu sequence, and SVAs. How different are individuals and ethnic groups in their retrotransposon content? Is one population, say Africans because they are earliest humans, more likely to have their own set of mobile DNA elements that are specific to that population? Ewing's analysis of 1000 Genomes Project individuals has found over 100 L1Hs elements specific to 3 African population groups as opposed to only 4 or 5 specific to either

3 European or 3 Asian groups (Ewing and Kazazian, 2011). At present, we know that any two individuals differ in their L1Hs content by at least 100 L1s out of roughly 1000 total. Yet as more and more individuals are sequenced, the number of new L1Hs not seen previously decreases dramatically so that after 17 individuals the number of novel L1Hs in individual 18 is less than 10. Will we find that any two Yorubans (West Africans from Nigeria) are more similar or different in L1Hs content than say any two Japanese?

Much more new biological information should be obtainable with the detection techniques geared to any specific genome. As costs of next generation sequencing drop and microarray techniques improve, it should be possible to study retrotransposon insertions in a large number of genomes. Study of trios (or quads) containing the DNA of both parents and one or more children aimed at finding new insertions in the children that are *de novo* events, that is, absent from both parents, will provide a substantial improvement in the frequency estimate of new retrotransposition events for all three human retrotransposons. Do the *de novo* insertions of all retrotransposons occur at a frequency of 1 in every 10–20 meiotic events as previously estimated or at the frequency of the more recent estimate of 1 in 40–60 meioses (1 in 100–150 for L1s and 1 in 50 for Alus)?

Another key biological question is this: What is the frequency of somatic insertions detectable in lymphocyte DNA? This question can be addressed by analysis of the DNA of identical twins. In the twins, any difference in retrotransposon content would be due to insertions occurring after fertilization. The work of Kano and Ostertag using transgenic animals (Kano et al., 2009) and Muotri and the Gage lab on transgenic mouse and human cadavers (Muotri et al., 2005; Coufal et al., 2009; Muotri et al., 2010) strongly suggests that significant retrotransposition occurs in somatic cells early in development. At this juncture, we don't know the extent to which DNA of a somatic insertion can be diluted by cellular DNA and remain detectable by sequencing or microarray hybridization. Can an insertion be detected if it is present in only 1 cell in 100? Mixing experiment should tell us whether such a rare somatic insertion can be detected. Thus we should learn soon the extent to which somatic retrotransposon insertions are important for human biology and diversity.

What about somatic insertions in tissues other than lymphocytes? If retrotransposon insertions occur sufficiently early in development, it may be possible to find them in other cell types, such as neural precursor cells. If insertions are occurring at the frequency suggested by Coufal et al. (Coufal et al., 2009) and the same insertions are present in a reasonable proportion of cells, they should be detectable using cadaver tissue samples. Similarly, one should be able to use a modified technique to detect new L1 or other retrotransposon insertions in mouse tissues and embryos. These studies should add to the information on somatic retrotransposon activity. Our lab has had the opportunity, in collaboration with the Moran lab, to analyze clones of cells started from the same culture but grown separately. In two ovarian teratocarcinoma clones, Ewing obtained tantalizing data that several L1Hs insertions are present in one clone but not the other. Some of these insertions appear to have occurred early in passage while at least one appears to be a later event. He has also found a few L1Hs insertions present in one human embryonic stem cell clone but absent from a second clone of cells derived from the same cells as the first. Again, these data are further evidence of the extent of somatic retrotransposition events. In another application of this new technology, we have begun to analyze trios (mother, father, fetus) including a fetus that died late in gestation. In only two such trios studied to date, one fetal placenta had an L1Hs insertion discovered by the high-throughput technique that was absent in both parents. Paternity was confirmed by other studies, and the insertion was verified by PCR. Although the insertion was in an intron of a neuronally expressed gene, we don't know whether the *de novo* insertion was related to the fetal demise.

Another interesting question is this: How many highly active or "hot" L1s are there in the human population? And does the number of "hot" L1s in the genome of each individual determine their susceptibility to retrotransposition? Brouha had shown that among 6 retrotransposition events in which the full-length precursor L1 was available, 5 were due to "hot" L1s that he defined as having at least 30% the activity of a "very, very active" human L1 control (Brouha et al., 2003). Beck et al. recently analyzed the full-length L1s from fosmid libraries of 6 individuals of different ethnic groups. Fosmids contain ~40kb of random genomic DNA and are similar to cosmids

except they are based on the bacterial F-plasmid. They specifically studied fosmids whose length was roughly 6kb (the length of a full-length human L1) larger than that expected from the human reference genome sequence. They found that each of the 6 individual genomes had 2 of the 6 "hot" L1s found by Brouha et al. Amazingly, they also found that each of the 6 individuals had 3 to 9 additional "hot" L1s, which they defined as having 10% or greater the activity of a very active human element. Their criteria for a "hot" L1 turn out to be similar to those of Brouha et al., and they also had only 4 elements in the group with activity between 10% and 30% of their very active control element. In sum, the 6 individuals had 37 new "hot" L1s, bringing the total for "hot" L1s to 43 in 7 genomes (Beck et al., 2010)—take a look at Table 28.1. In addition, 4 of the new "hot" L1s were found in one or a very small number of many individuals tested, suggesting that there may be many rare, essentially private (present in only one family) L1s in the human population.

Table 28.1 Novel Ta L1s found in 6 different individual fosmid libraries. From the top, the libraries are derived from an unknown, Japanese, Yoruban, Chinese, European, and Yoruban. (Beck et al., 2010. Used with permission by Elsevier.)

Dimorphic Elements	Novel (Not in dbRIP)	Active	Hot	HGR "Hot" Elements
5	5	4	4	2
16	16	9	8	2
20	18	11	9	2
13	12	9	8	2
8	7	4	3	2
7	7	6	5	2
69 Total	65 Total	43 Total	37 Total	

From the 1000 Genomes Project, Ewing has found 180 full-length L1s not in the Beck et al. dataset. Of these, over 120 have been verified by at least one other group and have not been previously analyzed for retrotransposition activity. From the >50% yield of "hot" L1s from full-length L1s in Beck et al. (Beck et al., 2010), we can expect that 60 or more of the 1000 Genomes Project L1s will be

scored as "hot." Thus it is likely that the number of "hot" L1s in the world population is in the thousands and may well be between 100,000 and 1,000,000. As stated earlier, these are the important L1s because they are believed to be the elements that are actively retrotransposing at the present time. It is also likely that these are the elements providing reverse transcriptase for the *trans* mobilization of Alus and SVAs.

So we seem to be closing in on an estimate of the number of new insertions of L1s, Alus, and SVAs in the world population per year. It now appears, contrary to the prevailing view, that this number may be quite large. If we consider that the best estimate of the rate of L1 retrotransposition per live birth is ~1 in 100–140, the similar estimate for Alu is ~1 in 50, and the number of live births per year according to world statistics is 130 million, then there would be roughly 1 million L1 events and 2.5 million Alu insertions per year.

From these numbers, we can make a rough estimate of the number of Mendelian disease-causing insertions per year in the world population. A conservative estimate of exon sequence is 1% of the genome; then 1% of 3.5 million L1 and Alu insertions per year is 35,000. Roughly 700 of the 25,000 human genes are haploinsufficient, meaning that knocking out the function of one of the two copies will produce disease. Then 2.8% (700/25,000) of 35,000 is ~1000 insertions into exons of haploinsufficient genes per year worldwide. If 50% of these insertions are highly deleterious to gene function, then ~500 insertions would cause single gene disease in the world per year. Because only 70 deleterious insertions have been found over the past 20 years, there must be a major underascertainment of these disease-producing retrotransposition events.

Moreover, the estimate of 3.5 million total retrotransposition events per year above only includes germ line and very early embryonic insertions. Somatic L1 insertions in transgenic animals occur much more frequently than do germ line and very early heritable embryonic ones. Well, you say, transgenic animals may not accurately reflect the *in vivo* situation. To which I answer that all retrotransposition studies using cultured cells and transgenic animals to date have mirrored natural *in vivo* phenomena. If we assume then that the transgenic data are also reflecting real life, then somatic insertions

may add another 10–100 million to the 1 million germ line L1 insertions worldwide per year without even counting potential somatic Alu and SVA insertions. Thus the total number of ongoing retrotransposition events is potentially astronomical.

The role of retrotransposition in disease

Is the role of retrotransposon insertion in disease etiology significantly greater than is now believed? At present, retrotransposon insertions are believed to have a very limited role in causing disease. Only about 70 insertions of various retrotransposons have been discovered in isolated cases of human disease. Yet for most disorders whose etiology lies in mutations affecting a single gene, mutations covering only about 70–80% of cases are discovered. Perhaps some fraction of the remaining cases is due to retrotransposon insertions. And what about common diseases thought to be caused by mutation in multiple genes? Could retrotransposon insertions play a significant role in these disorders?

For studies of disease etiology, two types of samples could be used. One sample type is discordant identical twins. For many conditions, identical twins are frequently concordant for the disease, that is, they both acquire it, but less than 100% of the time. For schizophrenia, identical twins are concordant about 60% and discordant about 40% of the time. For autism spectrum disorder, the concordance rate is about 90%. We hypothesize that in some cases one identical twin acquired the disease while the other did not because the affected twin received a somatic retrotransposon insertion into an important neuro-developmental gene. Ewing is presently studying this possibility using DNA from such discordant identical twins. As mentioned earlier, Burns and Boeke are studying X-linked mental retardation. Because there are many unexplained cases of X-linked mental retardation, it is reasonable to believe that a single retrotransposition event into a key gene on the X chromosome could cause the condition. However, it will be critical to study families because any potentially causative insertion should be either *de novo* or present in an unaffected mother and/or her female relatives. It will also be critical to demonstrate that a putatively causal insertion affects the function of any suspected gene. The possible role of retrotransposon

insertions in other conditions, including fetal wastage, bipolar disorder, and other diseases will also be explored by these new techniques in the near future.

What about the role of retrotransposition in various cancers? If insertions are mainly occurring in early development and in cells early in their differentiation, then their role in oncogenesis should be greater than we now suspect. As mentioned, Iskow et al. have found evidence for *de novo* L1Hs insertions in lung tumors (Iskow et al., 2010). Before this recent report, only one insertion had been discovered in a tumor that was not present in normal tissue of the individual. As previously mentioned in this chapter, this was the L1 insertion in a colon cancer that knocked out an APC gene, a known predisposing lesion in colorectal cancer (Miki et al., 1992). Other constitutional L1 and Alu insertions have been found in cancer-susceptibility genes, such as BRCA1, in both the cancer and normal tissue. The new techniques for finding *de novo* insertions are being used to study paired tumor-normal tissue samples in various cancers. Thus it is likely that we will learn soon whether retrotransposition of L1, Alu, and SVA elements have a more significant role in disease etiology than is presently believed.

Biochemical characterization of retrotransposition intermediates

Another large area that is becoming ripe for exploration is the biochemistry and cell biology of retrotransposons. Recently, ORF2p was finally detected in L1 RNPs by both specific antibodies and a PCR assay called LEAP that detects reverse transcriptase activity (Kulpa and Moran, 2006). ORF2p was detected in the cell cytoplasm both by an antibody to the native protein and by an antibody to an epitope tag on the protein (Doucet et al., 2010; Goodier et al., 2010). Some important questions that will be answered soon include: What non-L1 derived proteins and RNAs are present in the L1 RNP? What is the life cycle in the cell of the L1 RNP? What fraction of the L1 RNP enters the nucleus? Does it ever associate with the nucleolus? Because the L1 RNP appears to associate with stress granules in the cell cytoplasm, does this constitute putting the L1 RNP into a "waste

basket" for removal, or is the stress granule merely a storage site? Is there a role for the P (or processing) body in L1 RNA processing? Are PIWI-like proteins involved? What are the other host proteins besides those associated with the L1 RNP that are critical for L1 retrotransposition? These are just a few of the important questions relating to L1 biology that will be resolved soon.

Along with answers to many of these questions, we should start to see the three-dimensional structures of the L1 proteins, their RNPs, and the various individual host components of the L1 RNP. We already have a key structure of ORF1p worked out by Oliver Weichenrieder's lab (Khazina and Weichenrieder, 2009). A much tougher structure to obtain will be that of ORF2p, a much larger protein, present in very, very small amounts both in transformed cells and after transfection of cells with L1 plasmids. I expect that methods to obtain and purify larger amounts of ORF2p, perhaps from isolated L1 RNPs, will be available in the foreseeable future. Down the road a bit, techniques should be developed so that the three-dimensional structure of the entire L1 RNP can be solved, providing new insights into L1 biology.

We also look forward to getting a clearer picture of the hierarchy of controls on mobility of mammalian retrotransposons. At present, the picture is quite murky. Much is made of DNA methylation repressing transposons, but the evidence for this assertion is incomplete. Likewise, the relationship of mobile DNA to small RNAs and the control of retrotransposition in lower mammals and human beings is also supported by data showing increased retrotransposon RNA in knockouts of proteins associated with piRNAs, but we await evidence of piRNAs effects on retrotransposition itself. As far as the effects of these proteins on L1 retrotransposition itself in mammals, we should learn soon how the gene knockout of these proteins, MILI, MIWI2, Maelstrom, and others, affect retrotransposition of a mouse L1 transgene in the mouse. It should also be possible to determine whether knockout of these proteins leads to an increase in *de novo* retrotransposition events in developing testes of mice carrying the knockouts. Here is another application for the new high-throughput technologies to detect L1 insertions in any genome, albeit that the technology needs to be adapted to detect active mouse L1s.

Another large area ripe for some definitive answers is the effect of cellular stress on retrotransposition of Alu and L1. There have been suggestions that various stressors increase L1 and Alu expression and potentially retrotransposition. Stress effects on retrotransposon expression and retrotransposition need further work both in cell culture and in the animal. In addition, what global changes in cell machinery accompany and are caused by increases in L1 or Alu expression? There is evidence that increased L1 expression causes an increase in double strand breaks in DNA (Gasior et al., 2006). These data need to be confirmed and solidified. Further data regarding genome-wide effects of increased L1 and Alu expression should be forthcoming soon.

Beyond these expectations, I am still hopeful that one or more real functions in the individual will be demonstrated for mobile DNA. Much of "junk DNA" has now turned out to have some function as the producer of important small RNA molecules and as enhancers for distant gene expression. It is about time that function is found for much of the transposable element sequences present in the mammalian genome. Can it be that most of these many, many sequences are really neutral to the genetic fitness of the organism and not performing any function, whether positive or negative, for the individual? We know about their usefulness in driving genome evolution, but what about a function in the individual that might be under positive selection? I keep coming back to the possibility that reverse transcriptase has some required function during embryonic development in mammals. In human beings, the only dependable source of reverse transcriptase is L1 elements. Because reverse transcriptase was crucial at the beginning of life forms on the planet, it is tempting to think that it still retains some important function today. Alternatively, perhaps the piRNAs derived from mammalian transposable elements carry out some critical function for the organism beyond reducing the expression of transposable elements. I believe that a nice surprise awaits the scientific community concerning function for mobile DNA in the biology of the individual organism, human being, or other mammal. We shall see, but whatever the future holds, it's sure to be at least as exciting as the past! Many surprises and unexpected observations await us!

29

Predictions for mobile DNA

It's always fun to go out on a limb and predict what the future holds for the field, realizing that because of all the unexpected surprises the odds are great that most of one's predictions will be inaccurate. But here goes anyway. The following is a list of top ten predictions, not in any specific order. The list focuses on mammalian mobile DNA because that is the area with which I am most familiar. Some of these predictions are on basic stuff, while others have medical significance.

1. Mammalian mobile DNA, specifically the non-LTR retrotransposon, does have an important function in the individual that provides positive selection for the element. I consider this prediction somewhat iffy because it's hard to find function for mobile elements in individuals of any species. There seems to be a balancing act, an arms race, between the element's efforts to take over a genome and the host's attempts to restrain it.
2. Most L1 retrotransposition occurs in early development, not in the germ cell. There is already data on this one, but much of it comes from transgenic animals and may not reflect the endogenous situation. The large number of somatic events suggests that the number of new retrotransposon insertions in the world population in every generation is well into the millions of events.
3. L1 RNPs are functional retrotransposition intermediates and they contain 20–50 complexed proteins and 5–10 RNAs in addition to L1 RNA. Some more rare RNAs in L1 RNPs are Alu RNA, SVA RNA, and mRNAs on their way to insertion into the genome. Determining the functions and relative importance of the proteins and non-retrotransposing RNAs will be

critical to learning how retrotransposition intermediates interact with other parts of the cellular machinery.

4. The complete mechanism of non-LTR retrotransposition, including the mystery of second strand DNA synthesis, will be solved in the next five years. At the moment, the best data on this point comes from Eickbush's studies on the insect retrotransposon, R2. At this point, there is not a good *in vitro* system for studying the late steps of mammalian retrotransposition, but this should come.
5. L1 5' truncation is due to cleavage of L1 RNA by a host enzyme and not poor processivity of L1 reverse transcriptase. An *in vitro* system should help here also.
6. Three-dimensional structures for the L1 proteins and the L1 RNP will be solved and lead to new insights. Isolating sufficient ORF2p to carry out X-ray crystallography will be very difficult, but it will get done, although it may take some time. I'd love to see a 3D structure of ORF2p in the process of reverse transcription of L1 RNA.
7. Individual human beings differ significantly in their frequency of retrotransposition, varying from 1 in 10 meioses in some individuals to 1 in 200 meioses in others. This makes some people much more susceptible to retrotransposition than others. This variation in retrotransposition frequency also holds for somatic insertions. The new sequencing technology should give us information on this prediction within the next few years.
8. Retrotransposons have a small but significant role in the etiology of many complex diseases, such as autism spectrum disorder and schizophrenia. Again, we'll know the role of retrotransposons in complex diseases very soon.
9. Retrotransposition occurs frequently in early, poorly differentiated cancer cells and is important in the etiology of a small minority of human cancers.
10. Genome studies of ancient human remains shows that the human genome is continuing to expand at between 1 and 10 million base pairs per million years. This expansion is heavily influenced by retrotransposition. It should be possible to obtain information on this prediction in the near future.

We may need to come back 50 years from now to determine which, if any, of these predictions is correct. I'd be happy if five out of ten made it to prime time. For sure, we now know that what used to be called "junk DNA," the 98+% of the genome that lies between and outside of exon sequences (introns and intergenic sequences) is not junk at all. Mobile DNA and its remnants make up a large fraction of that "junk." This is the DNA that we in the mobile DNA field find so interesting and exciting to study. This is the DNA we treasure. I'm reminded of the PBS television show "Antiques Roadshow" in which old curios, nicknacks, and other items found in people's attics or purchased at flea markets are appraised by experts. Many of the items remain "junk." Some are moderately valuable, while others are worth $25,000 to $200,000. These last items are true treasures. It will be interesting to see how much "junk" DNA remains "junk" and how much takes on real value in human biology. I'm betting on much more treasure than we now imagine.

This page intentionally left blank

References

Adams, J.W., Kaufman, R.E., Kretschmer, P.J., Harrison, M., and Nienhuis, A.W. (1980). "A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene." *Nucleic Acids Res* 8, 6113–6128.

An, W., Han, J.S., Wheelan, S.J., Davis, E.S., Coombes, C.E., Ye, P., Triplett, C., and Boeke, J.D. (2006). "Active retrotransposition by a synthetic L1 element in mice." *Proc Natl Acad Sci U S A* 103, 18662–18667.

Antonarakis, S.E., Boehm, C.D., Giardina, P.J., and Kazazian, H.H., Jr. (1982). "Nonrandom association of polymorphic restriction sites in the beta-globin gene cluster." *Proc Natl Acad Sci U S A* 79, 137–141.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). "Developmentally regulated piRNA clusters implicate MILI in transposon control." *Science* 316, 744–747.

Athanikar, J.N., Badge, R.M., and Moran, J.V. (2004). "A YY1-binding site is required for accurate human LINE-1 transcription initiation." *Nucleic Acids Res* 32, 3846–3855.

Babushok, D.V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Mandal, P.K., Okada, N., Abrams, C.S., and Kazazian, H.H., Jr. (2007). "A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids." *Genome Res* 17, 1129–1138.

Babushok, D.V., Ostertag, E.M., Courtney, C.E., Choi, J.M., and Kazazian, H.H., Jr. (2006). "L1 integration in a transgenic mouse model." *Genome Res* 16, 240–250.

Badge, R.M., Alisch, R.S., and Moran, J.V. (2003). "ATLAS: a system to selectively identify human-specific L1 insertions." *Am J Hum Genet* 72, 823–838.

Bainton, R.J., Kubo, K.M., Feng, J.N., and Craig, N.L. (1993). "Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system." *Cell* 72, 931–943.

Baltimore, D. (1995). "Discovery of the reverse transcriptase." *FASEB Journal* 9, 1660–1663.

Bao, W., Jurka, M.G., Kapitonov, V.V., and Jurka, J. (2009). "New superfamilies of eukaryotic DNA transposons and their internal divisions." *Mol Biol Evol* 26, 983–993.

Batzer, M.A., and Deininger, P.L. (2002). "Alu repeats and human genomic diversity." *Nat Rev Genet* 3, 370–379.

Beauregard, A., Curcio, M.J., and Belfort, M. (2008). "The take and give between retrotransposable elements and their hosts." *Annu Rev Genet* 42, 587–617.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). "LINE-1 retrotransposition activity in human genomes." *Cell* 141, 1159–1170.

Becker, K.G., Swergold, G.D., Ozato, K., and Thayer, R.E. (1993). "Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element." *Hum Mol Genet* 2, 1697–1702.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). "Ultraconserved elements in the human genome." *Science* 304, 1321–1325.

Belancio, V.P., Roy-Engel, A.M., Pochampally, R.R., and Deininger, P. (2010). "Somatic expression of LINE-1 elements in human tissues." *Nucleic Acids Res* 38, 3909–3922.

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). "Active Alu retrotransposons in the human genome." *Genome Res* 18, 1875–1883.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." *Proc Natl Acad Sci U S A* 74, 3171–3175.

Bingham, P.M., Kidwell, M.G., and Rubin, G.M. (1982). "The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family." *Cell* 29, 995–1004.

Boeke, J.D. (1997). "LINEs and Alus—the polyA connection." *Nat Genet* 16, 6–7.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). "Ty elements transpose through an RNA intermediate." *Cell* 40, 491–500.

Bogerd, H.P., Wiegand, H.L., Hulme, A.E., Garcia-Perez, J.L., O'Shea, K.S., Moran, J.V., and Cullen, B.R. (2006). "Cellular inhibitors of long interspersed element 1 and Alu retrotransposition." *Proc Natl Acad Sci U S A* 103, 8780–8785.

Boissinot, S., Entezam, A., Young, L., Munson, P.J., and Furano, A.V. (2004). "The insertional history of an active family of L1 retrotransposons in humans." *Genome Res* 14, 1221–1231.

Boissinot, S., and Furano, A.V. (2005). "The recent evolution of human L1 retrotransposons." *Cytogenet Genome Res* 110, 402–406.

Bourc'his, D., and Bestor, T.H. (2004). "Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L." *Nature* 431, 96–99.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G.J. (2008). "An epigenetic role for maternally inherited piRNAs in transposon silencing." *Science* 322, 1387–1392.

Britten, R.J., and Kohne, D.E. (1968). "Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms." *Science* 161, 529–540.

Brouha, B., Meischl, C., Ostertag, E., de Boer, M., Zhang, Y., Neijens, H., Roos, D., and Kazazian, H.H., Jr. (2002). "Evidence consistent with human L1 retrotransposition in maternal meiosis I." *Am J Hum Genet* 71, 327–336.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). "Hot L1s account for the bulk of retrotransposition in the human population." *Proc Natl Acad Sci U S A* 100, 5280–5285.

Brown, P.O., Bowerman, B., Varmus, H.E., and Bishop, J.M. (1987). "Correct integration of retroviral DNA in vitro." *Cell* 49, 347–356.

Burke, W.D., Calalang, C.C., and Eickbush, T.H. (1987). "The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme." *Mol Cell Biol* 7, 2221–2230.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003). "The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination." *Nucleic Acids Res* 31, 4385–4390.

Cantrell, M.A., Carstens, B.C., and Wichman, H.A. (2009). "X chromosome inactivation and Xist evolution in a rodent lacking LINE-1 activity." *PLoS One* 4, e6252.

Carlson, C.M., Dupuy, A.J., Fritz, S., Roberg-Perez, K.J., Fletcher, C.F., and Largaespada, D.A. (2003). "Transposon mutagenesis of the mouse germline." *Genetics* 165, 243–256.

Carmell, M.A., Girard, A., van de Kant, H.J., Bourc'his, D., Bestor, T.H., de Rooij, D.G., and Hannon, G.J. (2007). "MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline." *Dev Cell* 12, 503–514.

Chambeyron, S., Popkova, A., Payen-Groschene, G., Brun, C., Laouini, D., Pelisson, A., and Bucheton, A. (2008). "piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline." *Proc Natl Acad Sci U S A* 105, 14964–14969.

Chatterji, M., Tsai, C.L., and Schatz, D.G. (2004). "New concepts in the regulation of an ancient reaction: transposition by RAG1/RAG2." *Immunol Rev* 200, 261–271.

Checkley, M.A., Nagashima, K., Lockett, S.J., Nyswaner, K.M., and Garfinkel, D.J. "P-body components are required for Ty1 retrotransposition during assembly of retrotransposition-competent virus-like particles." *Mol Cell Biol* 30, 382–398.

Chen, J., Rattner, A., and Nathans, J. (2006). "Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements." *Hum Mol Genet* 15, 2146–2156.

Chen, J.M., Stenson, P.D., Cooper, D.N., and Ferec, C. (2005). "A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease." *Hum Genet* 117, 411–427.

Chiu, Y.L., Witkowska, H.E., Hall, S.C., Santiago, M., Soros, V.B., Esnault, C., Heidmann, T., and Greene, W.C. (2006). "High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition." *Proc Natl Acad Sci U S A* 103, 15588–15593.

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass, J.L., Attreed, M., Avner, P., Wutz, A., Barillot, E., et al. "LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation." *Cell* 141, 956–969.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." *Cell* 12, 1–8.

Christensen, S., and Eickbush, T.H. (2004). "Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage." *J Mol Biol* 336, 1035–1045.

Christensen, S.M., and Eickbush, T.H. (2005). "R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA." *Mol Cell Biol* 25, 6617–6628.

Chung, J.H., Bell, A.C., and Felsenfeld, G. (1997). "Characterization of the chicken beta-globin insulator." *Proc Natl Acad Sci U S A* 94, 575–580.

Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. (2006). "Birth of a chimeric primate gene by capture of the transposase gene from a mobile element." *Proc Natl Acad Sci U S A* 103, 8101–8106.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. (2002). "Human L1 element target-primed reverse transcription in vitro." *Embo* J 21, 5899–5910.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). "L1 retrotransposition in human neural progenitor cells." *Nature* 460, 1127–1131.

Craig, N.L. (1995). "Unity in transposition reactions." *Science* 270, 253–254.

Craig, N.L. (1997). "Target site selection in transposition." *Annu Rev Biochem* 66, 437–474.

Curcio, M.J., and Derbyshire, K.M. (2003). "The outs and ins of transposition: from mu to kangaroo." *Nat Rev Mol Cell Biol* 4, 865–877.

Curcio, M.J., and Garfinkel, D.J. (1991). "Single-step selection for Ty1 element retrotransposition." *Proc Natl Acad Sci U S A* 88, 936–940.

Dai, J., Xie, W., Brady, T.L., Gao, J., and Voytas, D.F. (2007). "Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin." *Mol Cell* 27, 289–299.

Damert, A., Raiz, J., Horn, A.V., Lower, J., Wang, H., Xing, J., Batzer, M.A., Lower, R., and Schumann, G.G. (2009). "5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome." *Genome Res* 19, 1992–2008.

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G., and Chovnick, A. (1990). "Evidence for horizontal transmission of the P transposable element between Drosophila species." *Genetics* 124, 339–355.

Davies, D.R., Goryshin, I.Y., Reznikoff, W.S., and Rayment, I. (2000). "Three-dimensional structure of the Tn5 synaptic complex transposition intermediate." *Science* 289, 77–85.

DeBerardinis, R.J., Goodier, J.L., Ostertag, E.M., and Kazazian, H.H., Jr. (1998). "Rapid amplification of a retrotransposon subfamily is evolving the mouse genome." *Nat Genet* 20, 288–290.

Deininger, P.L., and Batzer, M.A. (1999). "Alu repeats and human disease." *Mol Genet Metab* 67, 183–193.

Devine, S.E., and Boeke, J.D. (1996). "Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III." *Genes Dev* 10, 620–633.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). "LINE-mediated retrotransposition of marked Alu sequences." *Nat Genet* 35, 41–48.

Dewannieux, M., and Heidmann, T. (2005). "L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells." *J Mol Biol* 349, 241–247.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). "Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice." *Cell* 122, 473–483.

Dombroski, B.A., Feng, Q., Mathias, S.L., Sassaman, D.M., Scott, A.F., Kazazian, H.H., Jr., and Boeke, J.D. (1994). "An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae." *Mol Cell Biol* 14, 4485–4492.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). "Isolation of an active human transposable element." *Science* 254, 1805–1808.

Dombroski, B.A., Scott, A.F., and Kazazian, H.H., Jr. (1993). "Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element." *Proc Natl Acad Sci U S A* 90, 6513–6517.

Dong, C., Poulter, R.T., and Han, J.S. (2009). "LINE-like retrotransposition in Saccharomyces cerevisiae." *Genetics* 181, 301–311.

Doucet, A.J., Hulme, A.E., Sahinovic, E., Kulpa, D.A., Moldovan, J.B., Kopera, H.C., Athanikar, J.N., Hasnaoui, M., Bucheton, A., Moran, J.V., et al. (2010). "Characterization of LINE-1 ribonucleoprotein particles." *PLoS Genet* 6, in press. www.plosgenetics.org

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., and Heidmann, T. (2009). "Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene." *Proc Natl Acad Sci U S A* 106, 12127–12132.

Eichinger, D.J., and Boeke, J.D. (1988). "The DNA intermediate in yeast Ty1 element transposition copurifies with virus-like particles: cell-free Ty1 transposition." *Cell* 54, 955–966.

Eickbush, T.H. (1997). "Telomerase and retrotransposons: which came first?" *Science* 277, 911–912.

Esnault, C., Maestre, J., and Heidmann, T. (2000). "Human LINE retrotransposons generate processed pseudogenes." *Nat Genet* 24, 363–367.

Ewing, A.D., and Kazazian, H.H. (2011). "Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans." *Genome Res* 21, March.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). "High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes." *Genome Res* 20, 1262–1270.

Farley, A.H., Luning Prak, E.T., and Kazazian, H.H., Jr. (2004). "More active human L1 retrotransposons produce longer insertions." *Nucleic Acids Res* 32, 502–510.

Fedoroff, N., Wessler, S., and Shure, M. (1983). "Isolation of the transposable maize controlling elements Ac and Ds." *Cell* 35, 235–242.

Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D. (1996). "Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition." *Cell* 87, 905–916.

Fire, A.Z. (2007). "Gene silencing by double-stranded RNA." *Cell Death Differ* 14, 1998–2012.

Freeman, J.D., Goodchild, N.L., and Mager, D.L. (1994). "A modified indicator gene for selection of retrotransposition events in mammalian cells." *Biotechniques* 17, 46, 48–49, 52.

Gabriel, A., and Boeke, J.D. (1991). "Reverse transcriptase encoded by a retrotransposon from the trypanosomatid Crithidia fasciculata." *Proc Natl Acad Sci U S A* 88, 9794–9798.

Garcia-Perez, J.L., Doucet, A.J., Bucheton, A., Moran, J.V., and Gilbert, N. (2007a). "Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase." *Genome Res* 17, 602–611.

Garcia-Perez, J.L., Marchetto, M.C., Muotri, A.R., Coufal, N.G., Gage, F.H., O'Shea, K.S., and Moran, J.V. (2007b). "LINE-1 retrotransposition in human embryonic stem cells." *Hum Mol Genet* 16, 1569–1577.

Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P., et al. (2010). "Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells." *Nature* 466, 769–773.

Garfinkel, D.J., Boeke, J.D., and Fink, G.R. (1985). "Ty element transposition: reverse transcriptase and virus-like particles." *Cell* 42, 507–517.

Garfinkel, D.J., Nyswaner, K.M., Stefanisko, K.M., Chang, C., and Moore, S.P. (2005). "Ty1 copy number dynamics in Saccharomyces." *Genetics* 169, 1845–1857.

Gartler, S.M., and Riggs, A.D. (1983). "Mammalian X-chromosome inactivation." *Annu Rev Genet* 17, 155–190.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). "The human LINE-1 retrotransposon creates DNA double-strand breaks." *J Mol Biol* 357, 1383–1393.

George, J.A., Traverse, K.L., Debaryshe, P.G., Kelley, K.J., and Pardue, M.L. (2010). "Evolution of diverse mechanisms for protecting chromosome ends by Drosophila TART telomere retrotransposons." *Proc Natl Acad Sci U S A* 107, 21052-21057

Gilbert, C., Schaack, S., Pace, J.K., 2nd, Brindley, P.J., and Feschotte, C. (2010). "A role for host-parasite interactions in the horizontal transfer of transposons across phyla." *Nature* 464, 1347–1350.

Gilbert, N., Bomar, J.M., Burmeister, M., and Moran, J.V. (2004). "Characterization of a mutagenic B1 retrotransposon insertion in the jittery mouse." *Hum Mutat* 24, 9–13.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). "Multiple fates of L1 retrotransposition intermediates in cultured human cells." *Mol Cell Biol* 25, 7780–7795.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). "Genomic deletions created upon LINE-1 retrotransposition." *Cell* 110, 315–325.

Gogvadze, E.V., Buzdin, A.A., and Sverdlov, E.D. (2005). "Multiple template switches on LINE-directed reverse transcription: the most probable formation mechanism for the double and triple chimeric retroelements in mammals." *Bioorg Khim* 31, 82–89.

Goodier, J.L., and Kazazian, H.H., Jr. (2008). "Retrotransposons revisited: the restraint and rehabilitation of parasites." *Cell* 135, 23–35.

Goodier, J.L., Mandal, P.K., Zhang, L., and Kazazian, H.H., Jr. (2010). "Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion." *Hum Mol Genet* 19, 1712–1725.

Goodier, J.L., Ostertag, E.M., Du, K., and Kazazian, H.H., Jr. (2001). "A novel active L1 retrotransposon subfamily in the mouse." *Genome Res* 11, 1677–1685.

Goodier, J.L., Ostertag, E.M., Engleka, K.A., Seleme, M.C., and Kazazian, H.H., Jr. (2004). "A potential role for the nucleolus in L1 retrotransposition." *Hum Mol Genet* 13, 1041–1048.

Goodier, J.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2000). "Transduction of 3'-flanking sequences is common in L1 retrotransposition." *Hum Mol Genet* 9, 653–657.

Goodier, J.L., Zhang, L., Vetter, M.R., and Kazazian, H.H., Jr. (2007). "LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex." *Mol Cell Biol* 27, 6469–6483.

Goodwin, T.J., Busby, J.N., and Poulter, R.T. (2007). "A yeast model for target-primed (non-LTR) retrotransposition." *BMC Genomics* 8, 263.

Goodwin, T.J., Ormandy, J.E., and Poulter, R.T. (2001). "L1-like non-LTR retrotransposons in the yeast Candida albicans." *Curr Genet* 39, 83–91.

Grimaldi, G., Skowronski, J., and Singer, M.F. (1984). "Defining the beginning and end of KpnI family segments." *EMBO* J 3, 1753–1759.

Grindley, N.D., Whiteson, K.L., and Rice, P.A. (2006). "Mechanisms of site-specific recombination." *Annu Rev Biochem* 75, 567–605.

Grundy, G.J., Hesse, J.E., and Gellert, M. (2007). "Requirements for DNA hairpin formation by RAG1/2." *Proc Natl Acad Sci U S A* 104, 3078–3083.

Haldane J. B. S. (1935). "The rate of spontaneous mutation of a human gene." *J. Genet.* 31, 317–326.

Han, J.S., Szak, S.T., and Boeke, J.D. (2004). "Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes." *Nature* 429, 268–274.

Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., and Batzer, M.A. (2008). "L1 recombination-associated deletions generate human genomic variation." *Proc Natl Acad Sci U S A* 105, 19366–19371.

Han, K., Lee, J., Meyer, T.J., Wang, J., Sen, S.K., Srikanta, D., Liang, P., and Batzer, M.A. (2007). "Alu recombination-mediated structural deletions in the chimpanzee genome." *PLoS Genet* 3, 1939–1949.

Han, Y.W., and Mizuuchi, K. (2010). "Phage Mu transposition immunity: protein pattern formation along DNA by a diffusion-ratchet mechanism." *Mol Cell* 39, 48–58.

Hancks, D.C., Ewing, A.D., Chen, J.E., Tokunaga, K., and Kazazian, H.H., Jr. (2009). "Exon-trapping mediated by the human retrotransposon SVA." *Genome Res* 19, 1983–1991.

Hassoun, H., Coetzer, T.L., Vassiliadis, J.N., Sahr, K.E., Maalouf, G.J., Saad, S.T., Catanzariti, L., and Palek, J. (1994). "A novel mobile element inserted in the alpha spectrin gene: spectrin dayton. A truncated alpha spectrin associated with hereditary elliptocytosis." *J Clin Invest* 94, 643–648.

Hattori, M., Kuhara, S., Takenaka, O., and Sakaki, Y. (1986). "L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein." *Nature* 321, 625–628.

Heidmann, T., Heidmann, O., and Nicolas, J.F. (1988). "An indicator gene to demonstrate intracellular transposition of defective retroviruses." *Proc Natl Acad Sci U S A* 85, 2219–2223.

Hiom, K., Melek, M., and Gellert, M. (1998). "DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations." *Cell* 94, 463–470.

Hohjoh, H., and Singer, M.F. (1996). "Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA." *Embo J* 15, 630–639.

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., and Kazazian, H.H., Jr. (1994). "A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion." *Nat Genet* 7, 143–148.

Holmes, S.E., Singer, M.F., and Swergold, G.D. (1992). "Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element." *J Biol Chem* 267, 19765–19768.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T., et al. (2010). "Mobile interspersed repeats are major structural variants in the human genome." *Cell* 141, 1171–1182.

Hulme, A.E., Bogerd, H.P., Cullen, B.R., and Moran, J.V. (2007). "Selective inhibition of Alu retrotransposition by APOBEC3G." *Gene* 390, 199–205.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). "Natural mutagenesis of human genomes by endogenous retrotransposons." *Cell* 141, 1253–1261.

Itakura, K., Rossi, J.J., and Wallace, R.B. (1984). "Synthesis and use of synthetic oligonucleotides." *Annu Rev Biochem* 53, 323–356.

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvak, Z. (1997). "Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells." *Cell* 91, 501–510.

Jensen, S., Gassama, M.P., and Heidmann, T. (1999). "Taming of transposable elements by homology-dependent gene silencing." *Nat Genet* 21, 209–212.

Ji, Y., Resch, W., Corbett, E., Yamane, A., Casellas, R., and Schatz, D.G. (2010). "The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci." *Cell* 141, 419–431.

Jiang, N., Feschotte, C., Zhang, X., and Wessler, S.R. (2004). "Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs)." *Curr Opin Plant Biol* 7, 115–119.

Jones, J.M., and Gellert, M. (2004). "The taming of a transposon: V(D)J recombination and the immune system." *Immunol Rev* 200, 233–248.

Jurka, J. (1997). "Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons." *Proc Natl Acad Sci U S A* 94, 1872–1877.

Kajikawa, M., and Okada, N. (2002). "LINEs mobilize SINEs in the eel through a shared 3' sequence." *Cell* 111, 433–444.

Kano, H., Godoy, I., Courtney, C., Vetter, M.R., Gerton, G.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2009). "L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism." *Genes Dev* 23, 1303–1312.

Kay, M.A., Manno, C.S., Ragni, M.V., Larson, P.J., Couto, L.B., McClelland, A., Glader, B., Chew, A.J., Tai, S.J., Herzog, R.W., et al. (2000). "Evidence for gene transfer and expression of factor IX in haemophilia B patients treated with an AAV vector." *Nat Genet* 24, 257–261.

Kazazian, H.H., Jr. (2004). "Mobile elements: drivers of genome evolution." *Science* 303, 1626–1632.

Kazazian, H.H., Jr., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988). "Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man." *Nature* 332, 164–166.

Keng, V.W., Ryan, B.J., Wangensteen, K.J., Balciunas, D., Schmedt, C., Ekker, S.C., and Largaespada, D.A. (2009). "Efficient transposition of Tol2 in the mouse germline." *Genetics* 183, 1565–1573.

Kennedy, A.K., Guhathakurta, A., Kleckner, N., and Haniford, D.B. (1998). "Tn10 transposition via a DNA hairpin intermediate." *Cell* 95, 125–134.

Khan, H., Smit, A., and Boissinot, S. (2006). "Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates." *Genome Res* 16, 78–87.

Khazina, E., and Weichenrieder, O. (2009). "Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame." *Proc Natl Acad Sci U S A* 106, 731–736.

Kidwell, M.G., Kidwell, J.F., and Nei, M. (1973). "A case of high rate of spontaneous mutation affecting viability in Drosophila melanogaster." *Genetics* 75, 133–153.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. (1998). "Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence." *Genome Res* 8, 464–478.

Kingsmore, S.F., Giros, B., Suh, D., Bieniarz, M., Caron, M.G., and Seldin, M.F. (1994). "Glycine receptor beta-subunit gene mutation in spastic mouse associated with LINE-1 element insertion." *Nat Genet* 7, 136–141.

Kirchner, J., Connolly, C.M., and Sandmeyer, S.B. (1995). "Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element." *Science* 267, 1488–1491.

Kolosha, V.O., and Martin, S.L. (1997). "In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition." *Proc Natl Acad Sci U S A* 94, 10155–10160.

Kolosha, V.O., and Martin, S.L. (2003). "High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1)." *J Biol Chem* 278, 8112–8117.

Korenberg, J.R., and Rykowski, M.C. (1988). "Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands." *Cell* 53, 391–400.

Kuduvalli, P.N., Mitra, R., and Craig, N.L. (2005). "Site-specific Tn7 transposition into the human genome." *Nucleic Acids Res* 33, 857–863.

Kuiper, M.T., and Lambowitz, A.M. (1988). "A novel reverse transcriptase activity associated with mitochondrial plasmids of Neurospora." *Cell* 55, 693–704.

Kulpa, D.A., and Moran, J.V. (2005). "Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition." *Hum Mol Genet* 14, 3237–3248.

Kulpa, D.A., and Moran, J.V. (2006). "Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles." *Nat Struct Mol Biol* 13, 655–660.

Kumar, A., Seringhaus, M., Biery, M.C., Sarnovsky, R.J., Umansky, L., Piccirillo, S., Heidtman, M., Cheung, K.H., Dobry, C.J., Gerstein, M.B., et al. (2004). "Large-scale mutagenesis of the yeast genome using a Tn7-derived multipurpose transposon." *Genome Res* 14, 1975–1986.

Kunkel, L.M. (1989). The Wellcome lecture, 1988. "Muscular dystrophy: a time of hope." *Proc R Soc Lond B Biol Sci* 237, 1–9.

Lai, C.B., Zhang, Y., Rogers, S.L., and Mager, D.L. (2009). "Creation of the two isoforms of rodent NKG2D was driven by a B1 retrotransposon insertion." *Nucleic Acids Res* 37, 3032–3043.

Lambowitz, A.M., and Zimmerly, S. (2010). "Group II introns: mobile ribozymes that invade DNA." *Cold Spring Harb Perspect Biol*., in press.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* 409, 860–921.

Lee, Y.N., and Bieniasz, P.D. (2007). "Reconstitution of an infectious human endogenous retrovirus." *PLoS Pathog* 3, e10.

Lee, J., Cordaux, R., Han, K., Wang, J., Hedges, D.J., Liang, P., and Batzer, M.A. (2007). "Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons." *Gene* 390, 18–27.

Leem, Y.E., Ripmaster, T.L., Kelly, F.D., Ebina, H., Heincelman, M.E., Zhang, K., Grewal, S.I., Hoffman, C.S., and Levin, H.L. (2008). "Retrotransposon Tf1 is targeted to Pol II promoters by transcription activators." *Mol Cell* 30, 98–107.

Lehrman, M.A., Schneider, W.J., Sudhof, T.C., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1985). "Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains." *Science* 227, 140–146.

Li, J., Han, K., Xing, J., Kim, H.S., Rogers, J., Ryder, O.A., Disotell, T., Yue, B., and Batzer, M.A. (2009). "Phylogeny of the macaques (Cercopithecidae: Macaca) based on Alu elements." Gene 448, 242–249.

Lu, J., and Clark, A.G. (2010). "Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila." *Genome Res* 20, 212–227.

Luan, D.D., and Eickbush, T.H. (1995). "RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element." *Mol Cell Biol* 15, 3882–3891.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). "Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition." *Cell* 72, 595–605.

Lutz, S.M., Vincent, B.J., Kazazian, H.H., Jr., Batzer, M.A., and Moran, J.V. (2003). "Allelic heterogeneity in LINE-1 retrotransposition activity." *Am J Hum Genet* 73, 1431–1437.

Lyon, M.F. (2000). "LINE-1 elements and X chromosome inactivation: a function for "junk" DNA?" *Proc Natl Acad Sci U S A* 97, 6248–6249.

Mager, D.L., and Freeman, J.D. (2000). "Novel mouse type D endogenous proviruses and ETn elements share long terminal repeat and internal sequences." *J Virol* 74, 7221–7229.

Martienssen, R.A. (2010). "Heterochromatin, small RNA and post-fertilization dysgenesis in allopolyploid and interploid hybrids of Arabidopsis." *New Phytol* 186, 46–53.

Martin, S.L. (1991). "Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells." *Mol Cell Biol* 11, 4804–4807.

Martin, S.L., Branciforte, D., Keller, D., and Bain, D.L. (2003). "Trimeric structure for an essential protein in L1 retrotransposition." *Proc Natl Acad Sci U S A* 100, 13815–13820.

Martin, S.L., and Bushman, F.D. (2001). "Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon." *Mol Cell Biol* 21, 467–475.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., Boeke, J.D., and Gabriel, A. (1991). "Reverse transcriptase encoded by a human transposable element." *Science* 254, 1808–1810.

Mayer, J., Sauter, M., Racz, A., Scherer, D., Mueller-Lantzsch, N., and Meese, E. (1999). "An almost-intact human endogenous retrovirus K on human chromosome 7." *Nat Genet* 21, 257–258.

McBlane, J.F., van Gent, D.C., Ramsden, D.A., Romeo, C., Cuomo, C.A., Gellert, M., and Oettinger, M.A. (1995). "Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps." *Cell* 83, 387–395.

McClintock, C.B. (1950). "The origin and behavior of mutable loci in maize." *Proc Natl Acad Sci U S A* 36, 344–355.

Mello, C.C. (2007). "Return to the RNAi world: rethinking gene expression and evolution." *Cell Death Differ* 14, 2013–2020.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. (1992). "Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer." *Cancer Res* 52, 643–645.

Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. (2007). "Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences." *Nature* 447, 167–177.

Mine, M., Chen, J.M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Ferec, C., Abitbol, M., Ricquier, D., et al. (2007). "A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element." *Hum Mutat* 28, 137–142.

Mitra, R., McKenzie, G.J., Yi, L., Lee, C.A., and Craig, N.L. (2010). "Characterization of the TnsD-attTn7 complex that promotes site-specific insertion of Tn7." *Mob DNA* 1, 18.

Mizrokhi, L.J., Georgieva, S.G., and Ilyin, Y.V. (1988). "*jockey*, a mobile Drosophila element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II." *Cell* 54, 685–691.

Mizuuchi, K. (1983). "In vitro transposition of bacteriophage Mu: a biochemical approach to a novel replication reaction." *Cell* 35, 785–794.

Mizuuchi, K. (1997). "Polynucleotidyl transfer reactions in site-specific DNA recombination." *Genes Cells* 2, 1–12.

Mol, C.D., Kuo, C.F., Thayer, M.M., Cunningham, R.P., and Tainer, J.A. (1995). "Structure and function of the multifunctional DNA-repair enzyme exonuclease III." *Nature* 374, 381–386.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). "Exon shuffling by L1 retrotransposition." *Science* 283, 1530–1534.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). "High frequency retrotransposition in cultured mammalian cells." *Cell* 87, 917–927.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). "Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres." *Nature* 446, 208–212.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. (2002). "DNA repair mediated by endonuclease-independent LINE-1 retrotransposition." *Nat Genet* 31, 159–165.

Mouw, K.W., Rowland, S.J., Gajjar, M.M., Boocock, M.R., Stark, W.M., and Rice, P.A. (2008). "Architecture of a serine recombinase-DNA regulatory complex." *Mol Cell* 30, 145–155.

Muckenfuss, H., Hamdorf, M., Held, U., Perkovic, M., Lower, J., Cichutek, K., Flory, E., Schumann, G.G., and Munk, C. (2006). "APOBEC3 proteins inhibit human LINE-1 retrotransposition." *J Biol Chem* 281, 22161–22172.

Muotri, A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., and Gage, F.H. (2005). "Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition." *Nature* 435, 903–910.

Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). "L1 retrotransposition in neurons is modulated by MeCP2." *Nature* 468, 443–446.

Muotri AR, Zhao C, Marchetto MC, Gage FH. (2009) "Environmental influence on L1 retrotransposons in the adult hippocampus." *Hippocampus*.19, 1002–1007.

Naas, T.P., DeBerardinis, R.J., Moran, J.V., Ostertag, E.M., Kingsmore, S.F., Seldin, M.F., Hayashizaki, Y., Martin, S.L., and Kazazian, H.H. (1998). "An actively retrotransposing, novel subfamily of mouse L1 elements." *Embo* J 17, 590–597.

Nakano, M., Cardinale, S., Noskov, V.N., Gassmann, R., Vagnarelli, P., Kandels-Lewis, S., Larionov, V., Earnshaw, W.C., and Masumoto, H. (2008). "Inactivation of a human kinetochore by specific targeting of chromatin modifiers." *Dev Cell* 14, 507–522.

Nigumann, P., Redik, K., Matlik, K., and Speek, M. (2002). "Many human genes are transcribed from the antisense promoter of L1 retrotransposon." *Genomics* 79, 628–634.

Nikaido, M., Rooney, A.P., and Okada, N. (1999). "Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales." *Proc Natl Acad Sci U S A* 96, 10261–10266.

Ohshima, K., Hamada, M., Terai, Y., and Okada, N. (1996). "The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements." *Mol Cell Biol* 16, 3756–3764.

Orgel, L.E., and Crick, F.H. (1980). "Selfish DNA: the ultimate parasite." *Nature* 284, 604–607.

Orkin, S.H., and Kazazian, H.H., Jr. (1984). "The mutation and polymorphism of the human beta-globin gene and its surrounding DNA." *Annu Rev Genet* 18, 131–171.

Orkin, S.H., Kazazian, H.H., Jr., Antonarakis, S.E., Goff, S.C., Boehm, C.D., Sexton, J.P., Waber, P.G., and Giardina, P.J. (1982). "Linkage of beta-thalassaemia mutations and beta-globin gene polymorphisms with DNA polymorphisms in human beta-globin gene cluster." *Nature* 296, 627–631.

Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL, Kazazian HH Jr.(2002) "A mouse model of human L1 retrotransposition." *Nat Genet* 32, 655–60.

Ostertag, E.M., Goodier, J.L., Zhang, Y., and Kazazian, H.H., Jr. (2003). "SVA elements are nonautonomous retrotransposons that cause disease in humans." *Am J Hum Genet* 73, 1444–1451.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001). "Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition." *Genome Res* 11, 2059–2065.

Ostertag, E.M., Prak, E.T., DeBerardinis, R.J., Moran, J.V., and Kazazian, H.H., Jr. (2000). "Determination of L1 retrotransposition kinetics in cultured cells." *Nucleic Acids Res* 28, 1418–1423.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). "Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion." *Genome Res* 11, 2050–2058.

Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G., et al. (2009). "An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs." *Science* 325, 995–998.

Perepelitsa-Belancio, V., and Deininger, P. (2003). "RNA truncation by premature polyadenylation attenuates human mobile element activity." *Nat Genet* 35, 363–366.

Pi, W., Zhu, X., Wu, M., Wang, Y., Fulzele, S., Eroglu, A., Ling, J., and Tuan, D. (2010). "Long-range function of an intergenic retrotransposon." *Proc Natl Acad Sci U S A* 107, 12992–12997.

Pickeral, O.K., Makalowski, W., Boguski, M.S., and Boeke, J.D. (2000). "Frequent human genomic DNA transduction driven by LINE-1 retrotransposition." *Genome Res* 10, 411–415.

Pothof, J., van Haaften, G., Thijssen, K., Kamath, R.S., Fraser, A.G., Ahringer, J., Plasterk, R.H., and Tijsterman, M. (2003). "Identification of genes that protect the C. elegans genome against mutations by genome-wide RNAi." *Genes Dev* 17, 443–448.

Prak, E.T., Dodson, A.W., Farkash, E.A., and Kazazian, H.H., Jr. (2003). "Tracking an embryonic L1 retrotransposition event." *Proc Natl Acad Sci U S A* 100, 1832–1837.

Ray, D.A., Feschotte, C., Pagan, H.J., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W., and Craig, N.L. (2008). "Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus." *Genome Res* 18, 717–728.

Ray, D.A., Walker, J.A., and Batzer, M.A. (2007). "Mobile element-based forensic genomics." *Mutat Res* 616, 24–33.

Ribet, D., Dewannieux, M., and Heidmann, T. (2004). "An active murine transposon family pair: retrotransposition of "master" MusD copies and ETn trans-mobilization." *Genome Res* 14, 2261–2267.

Ribet, D., Harper, F., Dupressoir, A., Dewannieux, M., Pierron, G., and Heidmann, T. (2008). "An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus." *Genome Res* 18, 597–609.

Robert, V.J., Katic, I., and Bessereau, J.L. (2009). "Mos1 transposition as a tool to engineer the Caenorhabditis elegans genome by homologous recombination." *Methods* 49, 263–269.

Rubin, G.M., Kidwell, M.G., and Bingham, P.M. (1982). "The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations." *Cell* 29, 987–994.

Saito, E.S., Keng, V.W., Takeda, J., and Horie, K. (2008). "Translation from nonautonomous type IAP retrotransposon is a critical determinant of transposition activity: implication for retrotransposon-mediated genome evolution." *Genome Res* 18, 859–868.

Sakai, J., and Kleckner, N. (1997). "The Tn10 synaptic complex can capture a target DNA only after transposon excision." *Cell* 89, 205–214.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). "The paleontology of intergene retrotransposons of maize." *Nat Genet* 20, 43–45.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. (1996). "Nested retrotransposons in the intergenic regions of the maize genome." *Science* 274, 765–768.

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., Kimura-Yoshida, C., Matsuo, I., Sumiyama, K., Saitou, N., et al. (2008). "Possible involvement of SINEs in mammalian-specific brain formation." *Proc Natl Acad Sci U S A* 105, 4220–4225.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, H.H., Jr. (1997). "Many human L1 elements are capable of retrotransposition." *Nat Genet* 16, 37–43.

Sayah, D.M., Sokolskaja, E., Berthoux, L., and Luban, J. (2004). "Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1." *Nature* 430, 569–573.

Schumann, G.G. (2007). "APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition." *Biochem Soc Trans* 35, 637–642.

Schwahn, U., Lenzner, S., Dong, J., Feil, S., Hinzmann, B., van Duijnhoven, G., Kirschner, R., Hemberger, M., Bergen, A.A., Rosenberg, T., et al. (1998). "Positional cloning of the gene for X-linked retinitis pigmentosa 2." *Nat Genet* 19, 327–332.

Schweidenback, C.T., and Baker, T.A. (2008). "Dissecting the roles of MuB in Mu transposition: ATP regulation of DNA binding is not essential for target delivery." *Proc Natl Acad Sci U S A* 105, 12101–12107.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). "Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence." *Genomics* 1, 113–125.

Scott, L.A., Kuroiwa, A., Matsuda, Y., and Wichman, H.A. (2006). "X accumulation of LINE-1 retrotransposons in Tokudaia osimensis, a spiny rat with the karyotype XO." *Cytogenet Genome Res* 112, 261–269.

Seleme Mdel, C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H., Jr. (2006). "Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity." *Proc Natl Acad Sci U S A* 103, 6611–6616.

Shehee, W.R., Loeb, D.D., Adey, N.B., Burton, F.H., Casavant, N.C., Cole, P., Davies, C.J., McGraw, R.A., Schichman, S.A., Severynse, D.M., et al. (1989). "Nucleotide sequence of the BALB/c mouse beta-globin complex." *J Mol Biol* 205, 41–62.

Sijen, T., and Plasterk, R.H. (2003). "Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi." *Nature* 426, 310–314.

Skowronski, J., Fanning, T.G., and Singer, M.F. (1988). "Unit-length line-1 transcripts in human teratocarcinoma cells." *Mol Cell Biol* 8, 1385–1397.

Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D., and Corces, V.G. (1994). "An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus." *Genes Dev* 8, 2046–2057.

Soper, S.F., van der Heijden, G.W., Hardiman, T.C., Goodheart, M., Martin, S.L., de Boer, P., and Bortvin, A. (2008). "Mouse maelstrom, a component of nuage, is essential for spermatogenesis and transposon repression in meiosis." *Dev Cell* 15, 285–297.

Speek, M. (2001). "Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes." *Mol Cell Biol* 21, 1973–1985.

Spradling, A.C., and Rubin, G.M. (1982). "Transposition of cloned P elements into Drosophila germ line chromosomes." *Science* 218, 341–347.

Stenglein, M.D., and Harris, R.S. (2006). "APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism." *J Biol Chem* 281, 16837–16841.

Swergold, G.D. (1990). "Identification, characterization, and cell specificity of a human LINE-1 promoter." *Mol Cell Biol* 10, 6718–6729.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). "Human L1 retrotransposition is associated with genetic instability in vivo." *Cell* 110, 327–338.

Takahara, T., Ohsumi, T., Kuromitsu, J., Shibata, K., Sasaki, N., Okazaki, Y., Shibata, H., Sato, S., Yoshiki, A., Kusakabe, M., et al. (1996). "Dysfunction of the Orleans reeler gene arising from exon skipping due to transposition of a full-length copy of an active L1 sequence into the skipped exon." *Hum Mol Genet* 5, 989–993.

Takasu, M., Hayashi, R., Maruya, E., Ota, M., Imura, K., Kougo, K., Kobayashi, C., Saji, H., Ishikawa, Y., Asai, T., et al. (2007). "Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon." *Tissue Antigens* 70, 144–150.

Tchenio, T., Casella, J.F., and Heidmann, T. (2000). "Members of the SRY family regulate the human LINE retrotransposons." *Nucleic Acids Res* 28, 411–415.

Temin, H.M. (1976). "The DNA provirus hypothesis." *Science* 192, 1075–1080.

Temtamy, S.A., Aglan, M.S., Valencia, M., Cocchi, G., Pacheco, M., Ashour, A.M., Amr, K.S., Helmy, S.M., El-Gammal, M.A., Wright, M., et al. (2008). "Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis-van Creveld syndrome with borderline intelligence." *Hum Mutat* 29, 931–938.

Ton-Hoang, B., Pasternak, C., Siguier, P., Guynet, C., Hickman, A.B., Dyda, F., Sommer, S., and Chandler, M. (2010). "Single-stranded DNA transposition is coupled to host replication." *Cell* 142, 398–408.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). "Bursts of retrotransposition reproduced in Arabidopsis." *Nature* 461, 423–426.

Turlan, C., and Chandler, M. (2000). "Playing second fiddle: second-strand processing and liberation of transposable elements from donor DNA." *Trends Microbiol* 8, 268–274.

van den Hurk, J.A., Meij, I.C., Seleme, M.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S., et al. (2007). "L1 retrotransposition can occur early in human embryonic development." *Hum Mol Genet* 16, 1587–1592.

Voliva, C.F., Martin, S.L., Hutchison, C.A., 3rd, and Edgell, M.H. (1984). "Dispersal process associated with the L1 family of interspersed repetitive DNA sequences." *J Mol Biol* 178, 795–813.

Wang, Y., Liska, F., Gosele, C., Sedova, L., Kren, V., Krenova, D., Ivics, Z., Hubner, N., and Izsvak, Z. (2010). "A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains." *Genome Res* 20, 19–27.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Boeke, J.D., and Moran, J.V. (2001). "Human L1 retrotransposition: cis preference versus trans complementation." *Mol Cell Biol* 21, 1429–1439.

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). "Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution." *Genome Res* 15, 1073–1078.

Witherspoon, D.J., Xing, J., Zhang, Y., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2010). "Mobile element scanning (ME-Scan) by targeted high-throughput sequencing." *BMC Genomics* 11, 410.

Woods-Samuels, P., Wong, C., Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., and Antonarakis, S.E. (1989). "Characterization of a nondeleterious L1 insertion in an intron of the human factor VIII gene and further evidence of open reading frames in functional L1 elements." *Genomics* 4, 290–296.

Worton, R.G., and Thompson, M.W. (1988). "Genetics of Duchenne muscular dystrophy." *Annu Rev Genet* 22, 601–629.

Wu, S.C., Meir, Y.J., Coates, C.J., Handler, A.M., Pelczar, P., Moisyadi, S., and Kaminski, J.M. (2006). "piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells." *Proc Natl Acad Sci U S A* 103, 15008–15013.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., and Batzer, M.A. (2006). "Emergence of primate genes by retrotransposon-mediated sequence transduction." *Proc Natl Acad Sci U S A* 103, 17608–17613.

Xiong, Y., and Eickbush, T.H. (1988a). "The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons." *Mol Cell Biol* 8, 114–123.

Xiong, Y.E., and Eickbush, T.H. (1988b). "Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm." *Cell* 55, 235–246.

Yanagihara, K., and Mizuuchi, K. (2003). "Progressive structural transitions within Mu transpositional complexes." *Mol Cell* 11, 215–224.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. (2009). "Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE." *Science* 325, 1391–1394.

Yang, N., and Kazazian, H.H., Jr. (2006). "L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells." *Nat Struct Mol Biol* 13, 763–771.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). "Cytosine methylation and the ecology of intragenomic parasites." *Trends Genet* 13, 335–340.

Zemojtel, T., Penzkofer, T., Schultz, J., Dandekar, T., Badge, R., and Vingron, M. (2007). "Exonization of active mouse L1s: a driver of transcriptome evolution?" *BMC Genomics* 8, 392.

Zhao, F., Qi, J., and Schuster, S.C. (2009). "Tracking the past: interspersed repeats in an extinct Afrotherian mammal, *Mammuthus primigenius*." *Genome Res* 19, 1384–1392.

Zhou, L., Mitra, R., Atkinson, P.W., Hickman, A.B., Dyda, F., and Craig, N.L. (2004). "Transposition of hAT elements links transposable elements and V(D)J recombination." *Nature* 432, 995–1001.

Zhou, R., Czech, B., Brennecke, J., Sachidanandam, R., Wohlschlegel, J.A., Perrimon, N., and Hannon, G.J. (2009). "Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform." *RNA* 15, 1886–1895.

Glossary

Alleles
Different DNA sequences at the same site or locus in genomic DNA. For example, we now know that an L1 from a specific chromosomal locus may have a number of alleles in the human population. These alleles may differ from each other at only 1-5 nucleotides out of 6000 in a full-length L1.

Consensus sequence
A consensus or majority-rule sequence is a sequence of a repeat in which each nucleotide is determined by nucleotide in the majority of sequences at that position. For example, if among 10 sequences of L1 elements, position 5000 contains an A in 6 and a G in 4, the consensus sequence for position 5000 is A.

Direct repeats
DNA sequences at the ends of mobile DNA that are directly repeated. For example, sequence at the 5' end is 5'-ATCG while sequence at the 3' end is 5'-ATCG-3'.

DNA transposons
Segments of DNA that contain a transposase enzymatic activity that allow them to be removed from one genomic site and placed in another genomic location. The two ends of these DNA transposons are also very important for their mobility.

Endonuclease
Non-LTR retrotransposons encode an endonuclease activity that makes a nick in one strand of DNA at the target. The nick site serves as the site for beginning or priming of reverse transcription. Mammalian endonucleases of retrotransposons have sequence similarities to the apurinic/apyrimidinic endonucleases and exonuclease III of *E. coli*.

Enhancers
DNA sequences that stimulate gene transcription into RNA. These sequences are usually less than 100 base pairs in size. These sequences can be located either upstream or downstream of the gene on which they act. They can be either in the forward or reverse orientation to that gene. They can be at great distance (up to one million base pairs) from that gene.

Envelope gene
The envelope gene is encoded by retroviruses and aids the retrovirus to migrate from one cell to another. Many LTR retrotransposons contain envelope gene sequences, but these are defective and prevent the retrotransposon from migrating to another cell.

Exons
Regions of the gene that generally can be translated into protein. However, 5' and 3' untranslated regions are also included in exon sequence.

Gene frequency
The frequency of a particular gene in a human population. Note that most genes are on autosomal chromosomes so they exist in two copies in each individual. So for example, the gene frequency of an autosomal sickle cell gene, β^s, present in one copy in 10% (10/100 people) of a particular population is 10/200 genes or .05.

Genome
All DNA in the chromosomes in the nucleus of the cell. The human genome is about $3x\ 10^9$ base pairs in size. Cytoplasmic mitochondria have small mitochondrial genomes of about 16,000 base pairs.

Heterochromatin
Tightly compacted chromatin containing genes that are inactive. Heterochromatin in some tissues and developmental states may become euchromatin with activated genes in other developmental states.

Integrase
This enzyme catalyzes the integration reaction of an LTR retrotransposon into genomic DNA.

Intron
DNA sequence between gene exons. They are removed from the premessenger RNA to make gene exons contiguous. The average gene contains 8 introns and introns comprlse ~30% of the genome.

Inverted repeats
Inverted DNA sequences at the ends of DNA transposons. For example, sequence at the 5' end (left end) of a transposon is 5'-ATCG-3' while sequence at its 3' end (right end) is 5'-GCTA-3'.

LTR retrotransposons
DNA segments of a particular structure that have the ability to be transcribed, reverse transcribed, and integrated into a new genomic site. These are copy and paste mobile elements. LTR stands for long terminal repeat. LTRs are direct repeats of 300–1000 base pairs of DNA sequence at the two ends of the element. These elements generally encode a protein coat for a cytoplasmic viral-like particle, a protease, a reverse transcriptase, and an integrase.

messenger RNA (mRNA)
RNA transcribed from gene DNA that carries protein-coding information out of the nucleus to the protein-synthesizing machinery in the cytoplasm for translation.

Non-LTR retrotransposons
DNA segments that are mobilized by a copy and paste mechanism. They are transcribed into RNA, the RNA is then reverse transcribed and integrated into a new DNA site in a single step. These elements do not contain long terminal repeats (LTRs) but have a poly(A) tail at their 3' ends.

Nucleophilic attack
Donation of one or more electrons in a chemical reaction involving covalent catalysis in which the donated electron(s) bond other chemical groups.

Poly(A) tail
A sequence of numerous adenylate residues located at the 3' end of messenger RNAs. Long stretches of As are also present at the 3' ends of L1s, Alus, and SVA elements at the time of their insertion into genomes. After a few generations in the genome, the A tails shorten from 40–120 As to 10–20 As.

Pre-messenger RNA
This RNA contains both the exons and introns of genes. After the introns are removed and the RNA is processed further at its ends and transported to the cytoplasm from the nucleus, pre-messenger RNA becomes messenger RNA.

Processed pseudogenes
Retrotransposed copies of mature messenger RNAs (lacking introns) that have been reverse transcribed by L1 reverse transcriptase and their DNA inserted into a new genomic site by target-primed reverse transcription.

Promoter
A region of DNA usually upstream of a gene that facilitates the gene's transcription into RNA. For L1, the promoter is within the L1 sequence itself at its 5' end.

Repetitive DNA
DNA present in many copies in any genome. For example, in the human and other primate genomes, repetitive DNA accounts for about 70% of the genome. Single-copy DNA is DNA present in one or very few copies and makes up ~30% of the genome.

Reverse transcriptase
The enzyme that catalyzes the conversion of RNA into DNA, the reverse of the canonial conversion of DNA into RNA. The only known endogenous source of reverse transcriptase in the human genome is encoded by L1 elements. Reverse transcriptase is also encoded by exogenous sequences not present in the human genome, most notably retroviruses like HIV.

Segmental duplication
Duplications of DNA sequence that can be 200–500 kilobase pairs in length. These duplications can be on the same chromosome within a few hundred kilobase pairs of each other or on different chromosomes.

Target-primed reverse transcription (TPRT)
The process of first strand DNA synthesis of non-LTR retrotransposons. TPRT begins with a nick of one DNA strand by retrotransposon endonuclease followed by reverse transcription using the 3' terminal OH at the nick as primer and the retrotransposon RNA as template.

Target-site duplication
A duplication of target DNA at the site of mobile DNA insertion. The length of a target site duplication depends greatly on the mobile element being inserted. For some elements, it is fixed, while for others, it is variable. For example, the target site duplications for mammalian L1 elements are usually between 6 and 20 base pairs in length.

Telomeres
The ends of chromosomes are called telomeres. In mammals, these are formed using a telomere RNA guide template and a telomerase, a reverse transcriptase similar in sequence to the reverse transcriptase of non-LTR retrotransposons.

Template
A sequence of nucleic acid, either DNA or RNA, that is copied by different enzymes. DNA polymerases copy the strands of the DNA double helix. RNA polymerases copy DNA strands into RNA. Reverse transcriptases copy RNA strands into DNA. The latter enzymes can also copy DNA strands into DNA.

Transcription
The process of synthesis of RNA from a DNA template. Reverse transcription is the process of synthesis of DNA from an RNA template.

Transposable elements or mobile DNA
Sequences of DNA that can move from one genomic location to another, either by a cut-and-paste mechanism (DNA transposons) or a copy-and-paste mechanism (retrotransposons)

V(D)J recombination
The process of forming intact immunoglobulin genes from disparate genes separated by great distances in genomic DNA. The RagI/RagII enzymes do the recombining using specific sequences at the junctions of the variable (V) and (D)J regions. The Rag enzymes are evolutionarily derived from a DNA transposon.

Viral-like particles (VLPs)
Particles synthesized by retroviruses and LTR-retrotransposons composed of protein coats (encoded by the *gag* gene) surrounding retroviral or retrotransposon RNA and reverse transcriptase. For retroviruses and LTR-retrotransposons, reverse transcription occurs in these particles.

This page intentionally left blank

Index

NUMBERS

A

B

C

D

E

F

G

H

I

J–K

L

M

N

O

P

Q–R

S

T

U–V

W

X–Y–Z

This page intentionally left blank

In an increasingly competitive world, it is quality of thinking that gives an edge—an idea that opens new doors, a technique that solves a problem, or an insight that simply helps make sense of it all.

We work with leading authors in the various arenas of business and finance to bring cutting-edge thinking and best-learning practices to a global market.

It is our goal to create world-class print publications and electronic products that give readers knowledge and understanding that can then be applied, whether studying or at work.

To find out more about our business products, you can visit us at www.ftpress.com.