Palindromes on the human X chromosome: Testis-biased transcription, gene conversion and evolution

by

Jennifer R. Saionz

B.S., Biology
The George Washington University, 1998

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2005

Signature of Author: _____
Department of Biology
January 14, 2005

Certified by: _____
David C. Page
Professor of Biology
Thesis Supervisor

Accepted by: _____
Stephen Bell
Professor of Biology
Chair, Biology Graduate Student Committee

PALINDROMES ON THE HUMAN X CHROMOSOME: TESTIS-BIASED TRANSCRIPTION, GENE CONVERSION AND EVOLUTION

by

JENNIFER R. SAIONZ

Submitted to the Department of Biology on January 14, 2005
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biology

ABSTRACT

Recent genomic studies of the Y chromosome revealed massive, testis-specific palindromes that span 30% of the chromosome and are subject to gene conversion. We conducted studies to determine whether similar palindromes exist on the human X chromosome and, if they exist, to what degree they share the features of the Y chromosome palindromes.

We performed an electronic search for palindromes on the human X chromosome resulting in the identification of 24 palindromes comprising 1.8% of the chromosome. The palindromes consist of sequences 9.5 to more than 140 kilobases long duplicated in inverted orientation separated by a 0.2 to 164 kilobase spacer. The paired palindrome arms display greater than 99 percent nucleotide identity.

We determined the palindrome associated gene content and experimentally evaluated their transcription range. All the genes residing in palindrome arms and spacers are transcribed in the testis, with almost two thirds predominantly testis-transcribed. To determine if the testis-transcription bias is due to a chromosome-wide enrichment for testis-transcribed genes, we used publicly available expression data to compare the ratio of palindrome-associated X-linked testis genes with non-palindrome-associated X-linked testis genes. We confirmed that the proportion of testis genes in palindromes is significantly different than that of testis genes on the entire X chromosome.

We pursued a comparative sequencing strategy to trace the evolution of the X chromosome palindromes. We sequenced bacterial artificial chromosomes (BACs) from chimpanzee, orangutan and rhesus monkey genomic libraries containing sequence orthologous to several of the human X chromosome palindromes. We found some of the palindromes conserved in all species studied, demonstrating the origins of these palindromes before the rhesus monkey and human lineages split 25 million years ago.

Despite their ancient origin, all of the palindromes studied display greater than 99 percent nucleotide identity between paired arms, suggesting that gene conversion between palindrome arms maintains the arm to arm similarity. We also uncovered insertions and deletions between orthologous palindrome arms that had been

subsequently homogenized to the opposite arm of the palindrome. The largest deletion of 14.5 kilobases is the largest known example of a gene conversion homogenized indel in mammals.

Thesis Supervisor:	David C. Page
Title:	Professor of Biology Member, Whitehead Institute for Biomedical Research Investigator, Howard Hughes Medical Institute

In memory of

Moritz Saionz & Albert Benjamin Gerber

ACKNOWLEDGEMENTS

My fellow Page lab members have been a wonderful source of friendship, collegial advice, constructive criticism and moral support. Yanfeng Lim, my fellow conspirator from day 1; Julian Lange, Andy Baltus, Jesse Potash, Jana Koubova, Jessica Alfoldi, Jenn Hughes, Alex Bortvin, Mark Gill, Janet Marszelak. Tatyana Pyntikova, the perfect baymate. Doug Menke, the best "senior grad student" ever. Laura Brown, equally generous with time, expertise and friendship. In a convoluted way, I owe the genesis of my thesis project to Jeremy Wang. I would like to thank Steve Rozen whose assistance and advice were crucial to this project. I would especially like to thank Helen Skaletsky for so much: advice, help, discussion and mentoring.

David Page, my advisor and mentor. Thank you for the opportunity to learn to be a scientist. Thank you for allowing me to make my own mistakes and encouraging me to follow a path of my own choosing. Thank you for introducing me to the intellectually exciting world at the crossroads of sex chromosome biology and germ cell biology.

I am indebted to my collaborators. I would like to thank Bob Blakesley, Eric Green and colleagues at the NIH Intramural Sequencing Center at the National Human Genome Research Institute for their "heroic efforts" to sequence orangutan and rhesus monkey BACs. I would like to thank Tina Graves, Pat Minx, Rick Wilson and colleagues at the Genome Sequencing Center at the Washington University School of Medicine for their great skill in sequencing the chimpanzee BACs.

I would especially like to thank my wonderful friends Allison, Natalia and Alex who kept me grounded and made me laugh. Sarah Banana, I promise to be there when you finish. Family is more than genetics: thank-you Mom, Dad, Sarah (again), Libs, Bec, David; Grammom & Grampop, Mom-mom & Pop-pop.

Table of Contents

CHAPTER 1

Introduction

Introduction

Much effort has been devoted to discerning patterns in the organization of the human genome. The duplicated genes and genomic regions comprise a widespread genomic structural pattern. Duplications may provide fodder for the evolution of new gene functions or may serve other purposes, such as increased gene dosage, by remaining highly similar. Functional patterning of the genome can be observed in a non-random distribution of genes within the genome. Biased distributions can be seen in the close linkage of similarly transcribed genes or the preferential accumulation of genes providing a male fitness advantage onto the sex chromosomes. The union of structural patterning with functional patterning is an exciting intersection where sequence structure may play a role in the function of embedded genes.

I. Structural organization of the genome by gene duplication

Gene duplication is an important source of material for evolution to utilize in generating novelty (Ohno 1970). Novelty may take the form of new function, better partitioning of function or enhanced function of one or more of the duplicates. There are instances of duplication where the selective advantage in preserving the duplicated sequence has not yet been inferred. Duplications of various ages and degrees of divergence have been studied. For many ancient duplications, evidence of a common ancestor is limited to the similarity of genes embedded in a small fraction of the

duplicated sequence. Yet many recent human or primate specific duplications retain the marks of the duplication mechanism.

Types of duplications

Duplicated sequence can be generated by whole chromosome or genome duplication (polyploidization), by duplication of segments of sequence or by duplication of single genes. Whole genome duplications can occur by means of errors in mitotic or meiotic reductive cell divisions in gametic cell lineages or by fertilization by multiple sperm (Otto and Yong 2002). Regional duplications can result from unequal breakage and reunion of non-homologous sequences (Maeda and Smithies 1986). Also, unequal, or non-allelic, homologous crossover between two copies of short similar sequences on sister or homologous chromatids can duplicate sequence of varying lengths residing between them (Smith 1976). Transduction of small pieces of sequence along with replicating transposable elements has been demonstrated to duplicate up to 3 kb of sequence (Goodier et al. 2000; Pickeral et al. 2000). Transposition of genes by way of RNA intermediates can duplicate single genes, but does not duplicate untranscribed regulatory elements such as promoters (Wang 2004). All of these mechanisms produce duplicated sequence. However, whether duplicated sequence is retained is a separate matter.

Classical model of gene duplication

The classical model of gene duplication, as put forth by Ohno (1970), predicts that duplicated genes can undergo two potential fates. Most likely to occur is that one of

the two will be lost. Loss of function occurs because only one gene is required to provide the function originally specified for the ancestral gene. The gene that supplies function is preserved by selection, while the other gene may accumulate mutations. Most of the time those mutations will be neutral or deleterious, in which case the gene is lost either due to genetic drift or, in the case of deleterious mutation, by counter selection. The road to loss is termed non-functionalization. In rare instances, the gene will acquire a beneficial mutation that leads to a novel function, neo-functionalization. When one gene acquires the novel function, the remaining gene is predicted to maintain the ancestral function.

Duplication/Degeneration/Complementation model

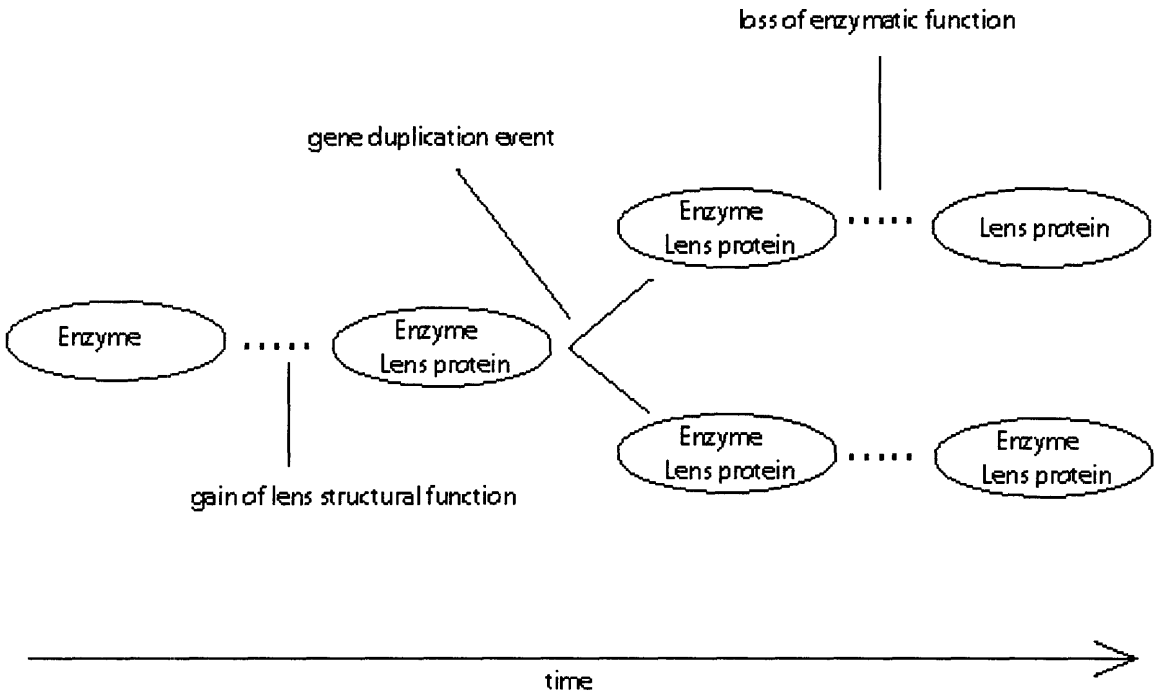
Because of the rarity of beneficial mutations and thus neo-functionalization, the classical model predicts that retention of duplicated genes does not occur frequently. To explain the prevalence of duplicated genes in many genomes, Force and colleagues proposed another model. The duplication-degeneration-complementation (DDC) model invokes the concept of subfunctionalization (Force et al. 1999; Lynch and Force 2000). Subfunctionalization predicts that duplicate genes are preserved when each is required for distinct and separate functions originally provided by the single ancestral gene. Loss-of-function mutations will occur in both gene duplicates; however, the mutations are most likely to generate defects in different, complementary functions, such that both genes are required in order to fulfill the function of the ancestral gene. The original model proposed that the loss-of-function mutations would occur in the regulatory regions of the genes. In this manner, while an ancestral gene may be expressed in several tissues, mutations in different regulatory binding sites in the daughter genes would cause of loss

of expression for the two in different tissues. Together, the two daughter genes recapitulate the range of expression in the ancestral gene. Similarly, if both genes accumulate mutations that reduce expression levels, but together they can provide the necessary dosage of gene product, then both will be retained.

While the DDC model is the latest model to explain the retention of duplicated genes, the idea of subfunctionalization of gene duplicates was first proposed by Piatigorsky and Wistow (1991) in a model termed 'gene sharing.' The 'gene sharing' model was brought about by their work on lens crystallins, structural proteins in the eye, where they observed that novel tissue-specific function was gained by a gene before the duplication event, rather than after, as predicted by the classical model (Figure 1). The lens crystallins include members that also act as various different types of enzymes outside of the eye. Ancestrally, those genes exclusively encoded enzymatic proteins that later developed additional roles as lens structural proteins. When present in the lens, these proteins do not act in their enzymatic roles, although they may still fulfill enzymatic roles in other tissues. The 'gene sharing' concept comes into play in some species-specific duplications where these multifunctional proteins duplicated, giving rise to one daughter gene that lost the enzymatic role and another daughter gene that has maintained both roles.

Subfunctionalization models are designed to account for the maintenance of gene duplicates soon after the duplication event; however, it is still compatible with later neo-functionalization (Force et al. 1999; Lynch and Force 2000). The requirement of complementary duplicates extends the time during which mutations can arise. The lifting of constraints that existed in the ancestor due to pleiotropic effects

Figure 1. Gene Sharing model of lens crystallins genes. A gene encoding an enzyme that was expressed in many tissues gained a function as a structural protein in the lens of the eye. After this multifunctional gene was duplicated, one of the daughter genes lost the widespread enzymatic role, but retained the role as a lens protein.


loss of enzymatic function
gene duplication event
Enzyme
Lens protein
Lens protein
Enzyme
Enzyme
Lens protein
Enzyme
Lens protein
Enzyme
Lens protein
gain of lens structural function
time

might permit mutations in formerly constrained regulatory or coding regions in the functionally partitioned daughter genes. Novel function may arise or better adaptation to the particular subfunction can occur.

Degenerative complementation in mammalian Hox3 genes

The Hox genes encode transcription factors that play a role in patterning segmental identity along the anterior to posterior body axis (McGinnis and Krumlauf 1992). In a vertebrate ancestor, a single cluster of tandemly duplicated Hox genes was itself twice duplicated such that today, the genes in vertebrates are organized into four clusters on separate chromosomes. A Hox gene cluster is comprised of 13 different paralogs (homologs related by duplication) lying in tandem, although not all paralogs have been retained in each cluster (Amores et al. 1998). The loss of paralogs is presumed to have resulted from functional redundancy. In zebrafish, the four clusters were again duplicated, with one of the clusters later being lost (Amores et al. 1998).

McClintock and colleagues used the DDC subfunctionalization model to explain their observations on the functions and expression patterns of the zebrafish *hoxb1* duplicates (McClintock et al. 2002). Because the mouse *Hoxb1* regulatory sequences are conserved between mouse, chick and pufferfish, it is hypothesized that they were present in the ancestral *Hoxb1* gene before the entire cluster was duplicated in the zebrafish lineage. In the mouse, the *Hoxb1* gene is expressed early in gastrulation in the hindbrain, with later expression limited to a single segment of the developing hindbrain. The zebrafish *hoxb1b* and *hoxb1a* genes together recapitulate the expression pattern of the mouse *Hoxb1*. Zebrafish *hoxb1b* is expressed early in the hindbrain, while zebrafish

hoxb1a is expressed later in the same single segment of the hindbrain as mouse *Hoxb1*. A retinoic acid response element downstream of the coding sequence is responsible for the early expression in the mouse, and is also present in the zebrafish *hoxb1b*, but not the later expressed zebrafish *hoxb1a*. An autoregulatory control element upstream of the coding sequence in the mouse is required for later expression. Zebrafish *hoxb1a* also has an upstream autoregulatory element and shares with the mouse *Hoxb1* the later autoregulated expression. The zebrafish *hoxb1b* gene has neither an autoregulatory element nor later expression. In this manner, the function of the ancestral *Hoxb1* gene was partitioned between *hoxb1a* and *hoxb1b*. Interestingly, *hoxb1b* coding sequence cannot be used to rescue *hoxb1a* function, leading McClintock and colleagues to hypothesize coding sequence differentiation due to a lack of selective constraints for maintaining the functions required in the later expression period.

Use of global gene expression profiles to explore the DDC model

Two studies of gene expression profiles in mammals have attempted to test the DDC model. Both tested the hypothesis that duplicated genes should diverge in their expression profile across tissues using microarray data from the same data set (Su et al. 2002). Neither, however, compared expression levels between duplicates. The first study, performed by Makova and Li (2003), looked at duplicate genes present in the human genome. They found that the expression range of duplicated genes diverged in a linear relationship with the time elapsed since duplication. Huminiecki and Wolfe (2004) improved on the earlier study by controlling for changes in expression that occur in any gene over time. They compared differences between species-specific paralogs (homologs

by duplication) in human or mouse to the differences between single copy orthologs (homologs by descent) in different species. They found that paralogs' coding sequence and expression profile diverged faster than orthologs' of the same age of divergence. They also predicted that subfunctionalization should be detected as a decrease in expression breadth in paralogs relative to orthologs. Indeed, Huminiecki and Wolfe found that expression range narrows as the number of duplicates rose. Furthermore, tissue-specific genes were more likely to belong to large gene families.

Dosage and retention of gene duplications

The effects of increased gene dosage after duplication may be an important determinant of whether a duplication is retained. Kondrashov and Koonin (2004) looked into the types of genes that are maintained after duplication. They divided genes into dosage sensitive and dosage insensitive genes. A gene yielding a phenotype when heterozygous for a loss-of-function allele is dosage sensitive, whereas a gene that does not yield a phenotype unless homozygous for a loss-of-function allele is dosage insensitive. Based on these definitions, they predict that genes associated with dominant genetic disorders should be dosage sensitive, as change in the dosage of a single allele is associated with disease. Similarly, they predict that genes associated with recessive genetic disorders should be dosage insensitive, as change in the dosage of both alleles is required for disease. Kondrashov and Koonin compared these two sets of genes and they found that without regard to category of gene function, the dosage sensitive genes were more likely to be members of larger gene families than were dosage insensitive genes. They interpret their results to confirm earlier predictions (Wright 1934) that duplications

of dosage-sensitive genes will be retained because of advantages related to increased dosage. Importantly, these predictions consider the immediate aftermaths of duplication and do not preclude later divergence of the dosage sensitive genes.

Another side to the question of duplication of dosage-sensitive genes is the prediction that only when the duplication is due to a polyploidization event will stabilizing selection maintain duplications of genes whose function requires precise stoichiometric relationships between their encoded proteins (Lynch and Conery 2000). If a function requires the precise balancing of dosage between two or more genes, the duplication of just one of them should be selected against. These two theories concerning dosage-sensitive genes are not mutually exclusive, and may each explain a portion of duplicated genes.

Segmental duplications represent recent sequence duplications

Segmental duplications, also called low copy repeats, are blocks of genomic DNA greater than 1 kb that are nearly identical and are present in at least two locations in the genome. Roughly 5% of the human genome is part of a 1 kb or greater duplication block, as compared to just over 1% of the fly genome and 4.25% of the worm genome (Lander et al. 2001). Most studies of segmental duplications limit their analyses to those with greater than 90% identity. Segmental duplications are assumed to be the result of recent duplication events as they exhibit high nucleotide sequence identity over long stretches of non-coding DNA. Because it is generally not subject to selection, non-genic sequence is assumed to accumulate mutations at a steady rate reflecting its age (Li 1997). Segmental duplications, by and large, represent duplication events of the kind that may have given

rise to divergent duplicate genes, yet are young enough that the non-coding sequence surrounding the duplicated genes is still recognizably similar. Under the assumption of constant divergence, these 90% identical sequences are presumed to have duplicated 40 million years ago, around the time of the divergence of prosimian (lemurs, lorises, galagos) and anthropoid (Old World and New World monkeys, apes) primate lineages (Bailey et al. 2001). The study of such segmental duplications has shed light on gene duplication events that occurred in the recent past.

Ancient duplications contain only the sequence that is selectively maintained, such as coding sequence and regulatory sequence. Sequence that is not selectively maintained in ancient duplications will have accumulated a sufficient amount of mutation that it can no longer be recognized as homologous. Thus, the fingerprints of the duplication mechanism generally cannot be discerned in the extant sequence of ancient duplications. Recent duplications, in many cases, still exhibit the marks of the duplication mechanism. Study of the boundaries of segmental duplications show an enrichment for retrotransposons of the *Alu* family, suggesting that unequal homologous recombination between *Alu* elements generated those sequence duplications (Bailey et al. 2003). Interestingly, the *Alu* enrichment was most significant for duplications separated by more than 1 Mb of sequence.

Segmental duplications, because they are regions of highly homologous sequence, are prone to non-allelic homologous recombination that may generate large scale rearrangements, deletions or duplications of sequence often associated with genomic disease (Lupski 1998). Studying those non-allelic homologous recombination events further aids our understanding of the mechanisms of gene duplication. Non-allelic

homologous recombination between segmental duplications on chromosome 17p11.2 is associated with Charcot-Marie-Tooth disease (CMT1A; OMIM 118220) and hereditary neuropathy with liability to pressure palsies (HNPP; OMIM 162500), diseases of the peripheral nervous system. These segmental duplications, termed the CMT1A-REPs, are two 24 kb long, 98.7% identical direct repeats spanning a 1.5 Mb unique sequence (Reiter et al. 1997). Duplication or deletion of the unique sequence between the CMT1A-REPs affects dosage of the *PMP22* gene, a component of the peripheral nervous system myelin, which causes the CMT1A or HNPP diseases (Boerkoel et al. 1999). Interestingly, a *mariner* transposase-like element was mapped to one of the CMT1A-REPs in both human and chimpanzee and shown to be transcribed in the testis, raising the possibility that double-strand breaks might be enhanced by the *mariner* element, thereby predisposing the region to recombination (Kiyosawa and Chance 1996; Reiter et al. 1996).

Bailey and colleagues (2002) examined recent duplications of the human genome, looking at genes that were duplicated in their entirety, to find what functions are enriched within sequence duplicated recently. In regions duplicated within the last 40 million years, they found significant numbers of genes associated with immunity and defense, membrane surface interaction, drug detoxification and growth and development. Immune response genes were also overrepresented among the paralogous genes found by Makova and Li (2003) that were recently duplicated, but highly diverged. Both studies suggest that the retention of duplicates involved with these functions is consistent with selection for adaptation or positive selection.

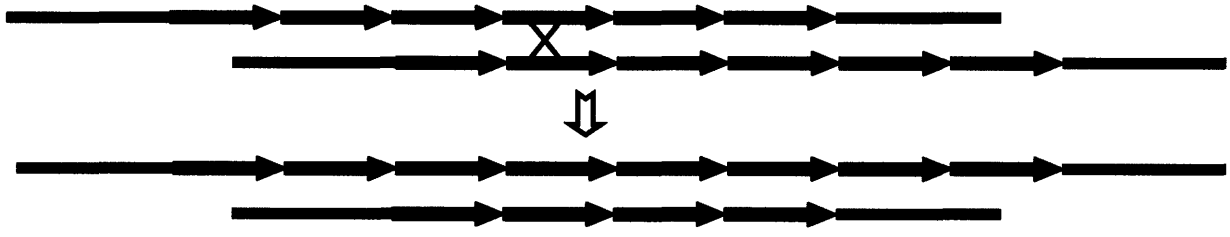
Gene duplications that do not diverge

As stated above, in some cases duplicated genes are maintained because multiple copies of the gene provide benefit through increased gene dosage and high expression levels. Purifying selection prevents fixation of mutations that modify gene function, but allows mutation at silent sites. Purifying selection is diagnosed by examining the ratio of synonymous substitutions to nonsynonymous substitutions. Purifying selection is evident among the large gene families of ubiquitin genes and histone H4 genes where high rates of silent substitutions reveal ancient duplication times, yet the low rates of nonsynonymous substitutions demonstrate strong selection toward amino acid conservation (Nei et al. 2000; Piontkivska et al. 2002). Concerted evolution will maintain sequence similarity at all sites (Hurst and Smith 1998). Studies of duplicated genes can be confused when high levels of nucleotide similarity due to concerted evolution are mistaken for high levels of nucleotide similarity due to a recent divergence. When DNA sequence analysis of genes was first possible, it was presumed that duplicated sequences always accumulate mutations and diverge at sites that are not under selection. Duplicated sequences that maintain sequence identity at unselected sites can be mistaken for recent duplications.

Two mechanisms have been shown to be involved in the concerted evolution of homologous sequences: unequal crossing over and gene conversion (Figure 2). Unequal crossing over occurs when two homologous, but not allelic sequences on sister or homologous chromatids pair and exchange DNA (Smith 1976). Unequal crossing over

Figure 2. Mechanisms of concerted evolution. A. Unequal chromatid exchange. Non-allelic sequences on sister or homologous chromatids align and recombine. When unequal chromatid exchange occurs between tandem arrays of sequence, contraction and expansion of the array can homogenize the sequence repeats. B. Gene conversion. Gene conversion is the non-reciprocal transfer of genetic information from one DNA molecule to an homologous DNA molecule (Paques and Haber 1999).

A. Unequal chromatid exchange (sister or homolog)



B. Gene conversion

operates in tandem duplicates where contraction and expansion of arrays can generate homogenization of the duplicates. Gene conversion is the non-reciprocal transfer of genetic information by means of homologous recombination intermediates (Paques and Haber 1999). Gene conversion was known to occur in eukaryotes such as yeast, but not shown to occur in mammals until almost 25 years ago. Slightom and colleagues (1980) cloned the two human fetal globin chain genes, $^{A}\gamma$ and $^{G}\gamma$, from both allelic chromosomes in a single individual. They discovered that on one chromosome, part of the $^{A}\gamma$ had been overwritten with sequence from the $^{G}\gamma$ such that $^{A}\gamma$ was more similar to $^{G}\gamma$ on the same chromosome than $^{A}\gamma$ on the allelic chromosome. They proposed that instances where genes exhibited high levels of identity suggesting more recent duplication than implied by the surrounding sequence might represent occurrences of gene conversion. The gene conversion event in the fetal globin genes represents a rare event; however, there are many instances of tandem arrays of genes that preserve their high identity through regular concerted evolution.

U2 snRNA genes tandem array provides high dosage of a splicing factor

The U2 small nuclear RNA (snRNA) genes function in splicing of pre-mRNAs. In humans, the U2 snRNA is encoded by a multigene family arranged in a single tandem array at chromosome 17q21 (Hammarstrom et al. 1985; Lindgren et al. 1985). Individuals may inherit between 6 to over 30 copies of the 6 kilobase (kb) repeat unit (Liao et al. 1997). The many copies of the U2 snRNA gene may serve to increase gene dosage. Interestingly, the repeat unit found in old world monkeys is 5 kb longer than the repeat unit in apes. The difference is due to the deletion of a retrovirus sequence in the

ape lineage that was subsequently homogenized between the repeat units of the array (Pavelitz et al. 1995).

The high level of sequence identity between repeat copies is maintained by concerted evolution, such that natural selection acts upon the entire array as a single genetic unit, as opposed to acting on the individual gene copies (Van Arsdell and Weiner 1984). Indeed, the few polymorphisms that do exist create variation in arrays in different individuals; only very seldom can sequence variants be detected in different repeat units within the same array (Liao et al. 1997). The rarity of within array heterogeneity implies rapid homogenization of the tandem repeats through intrachromosomal genetic exchange. While unequal crossover can explain the variability in repeat number, intrachromatid gene conversion probably also plays a role in the concerted evolution of the U2 snRNA genes (Liao et al. 1997).

Color vision genes exhibit concerted evolution, but maintain divergence at key sites

The genes required for mediation of color vision encode three different membrane proteins. The blue pigment gene lies on chromosome 7 in humans, while the red and green pigment genes are on the X chromosome. The X-linked red and green pigment genes are organized in a tandem array where a single red pigment gene lies at the head of the array and a variable number of green pigment genes are duplicated behind it (Nathans et al. 1986a; Nathans et al. 1986b; Vollrath et al. 1988). This tandem array evolved after the divergence of old world and new world primates. In New World monkeys, there is a single autosomal blue pigment gene and a single X-linked pigment gene. The X-linked pigment gene in New World monkeys is polymorphic for color perception, that is, any

particular allele may be sensitive to one of a few different visual wavelengths. The X-linked polymorphisms generate sex-linked variation in color vision such that males are always dichromats and females either dichromats or trichromats (reviewed in Jacobs 1996).

In humans, the coding regions of the red and green pigment genes are both 1092 basepairs (bp), yet there are only 25 nucleotide differences between the two genes and the difference in color perception is due to 3 amino acid changes (Nathans et al. 1986b). All in all, the red and green gene repeats are 98% identical over the approximately 38 kb repeat unit, with over 99% identity between green gene repeats (Feil et al. 1990; Vollrath et al. 1988). Introns of the two genes show extraordinarily high sequence identity, suggesting that frequent or recent gene conversion preserves the similarity (Shyue et al. 1994). The polymorphic number of green pigment gene copies points to unequal crossover as at least a partial mechanism for concerted evolution of the repeat units. Red-green color blindness is caused by non-allelic homologous recombination between red and green repeats that results in either deletion of all but the first red gene or in hybrid 5'red-3'green or 5'green-3'red genes in the first or second position of the array, respectively (Figure 3). Individuals may inherit a gene array with a hybrid gene or genes in positions downstream of the second gene, but do not display an altered color vision phenotype because only the first two genes in the array are expressed (Winderickx et al. 1992; Yamaguchi et al. 1997). Females carrying one wildtype and one mutant copy of the array will have normal color vision, except in cases of highly skewed X-inactivation (Jorgensen et al. 1992).

Figure 3. Red-green color blindness is caused by non-allelic homologous recombination between the red and green pigment genes. A. Normal array of red and green pigment genes. B. Deletion of all but the 5' red pigment gene. C. Hybrid 5'red 3'green pigment gene in the first position of the array. D. Hybrid 5'green 3'red pigment gene in the second position of the array.

A. 

B. 

C. 

D. 

Gene duplication provided the necessary substrates for the evolution of color vision. Key divergent nucleotide substitutions generated difference in the absorption of light by the red and green pigment genes thereby allowing trichromatic color vision. Selection prevents the genes from diverging markedly and thus altering color perception. Gene conversion and unequal crossing over are mechanisms for preventing divergence. However, the tandem repeat organization of the genes is inherently unstable, thereby putting males, who carry only one copy of the array, at risk for inherited or de novo gene conversion or gene deletion causing color blindness.

Segmental duplications that do not diverge.

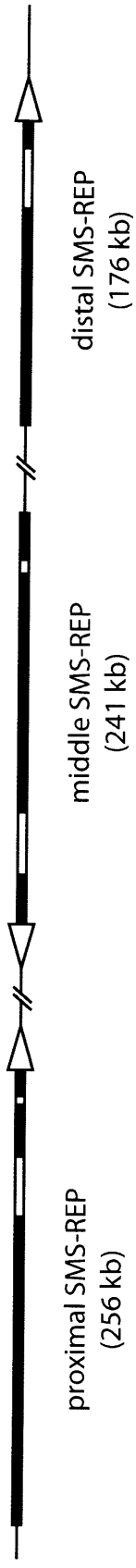
There are several examples of large segmental duplications that by virtue of their high nucleotide identity should be recent duplications. However, through hybridization experiments on genomic DNA in other primates, it has been demonstrated that the duplications are more ancient than their high sequence identity implies. Aside from a study of the Y chromosome palindromes in chimpanzee and human, to be discussed later, no comparative sequence studies have been completed on these duplications. The observation that some segmental duplications are older than expected is remarkable because segmental duplications are largely non-coding, non-genic sequence. That tens of kilobases of duplicated non-coding sequence maintain high sequence identity cannot be readily explained under current models of gene duplications. Most of these observations have been phrased as interesting side notes and with no speculation on possible selective advantages for concerted evolution of these duplications. Again, the palindromes on the Y chromosome are the primary exception and will be discussed later.

Smith-Magenis Syndrome repeats on 17p11.2

The phenotype of patients with Smith-Magenis syndrome includes distinctive facial characteristics, developmental delay, cognitive impairment and behavioral abnormalities (OMIM 182290). More than 90% of cases are caused by a deletion of about 4 Mb at chromosome 17p11.2, but the gene or genes underlying the syndrome have not yet been definitively ascertained (Park et al. 2002). The deletion is generated by non-allelic homologous recombination between two low copy repeats, termed SMS-REPs (Chen et al. 1997). There are three SMS-REPs; the proximal repeat is 256 kb in length, the middle 240 kb and the distal 176 kb (Figure 4 and 5a). The proximal and distal repeats are oriented in tandem, while the middle repeat is inverted relative to the others. About 170 kb of the repeat is over 98% identical between SMS-REP copies, with the largest block of homology 126 kb (Park et al. 2002).

Based on the greater than 98% nucleotide identity of the repeats, the duplication would be expected to be present only in the chimpanzee and human genomes. However, using fluorescent in situ hybridization (FISH) with human SMS-REP-specific probes, Park and colleagues (Park et al. 2002) demonstrated the presence of three SMS-REP copies in apes, old world monkeys and new world monkeys. New world monkeys diverged from prosimians over 40 million years ago (Kumar and Hedges 1998). The high nucleotide identity despite ancient duplication suggests that concerted evolution between the repeats prevented more significant divergence, but no selective advantage for concerted evolution has been proposed. Clues for a selective advantage based on gene content of the duplicated sequences have not been uncovered; of the 14 apparent transcription units

Figure 4. The Smith-Magenis Syndrome repeats (SMS-REPs) lie on chromosome 17p11.2. A common 4 Mb deletion is generated by recombination between the proximal and distal repeats. The red, black and yellow sequence regions display >98 percent identity between proximal and distal repeats and the green regions display >95 percent identity. White represents unique sequence. Open arrowheads denote direction of the repeats relative to each other. (Figure colors used as in Park et al. 2002)


distal SMS-REP
(176 kb)
middle SMS-REP
(241 kb)
proximal SMS-REP
(256 kb)

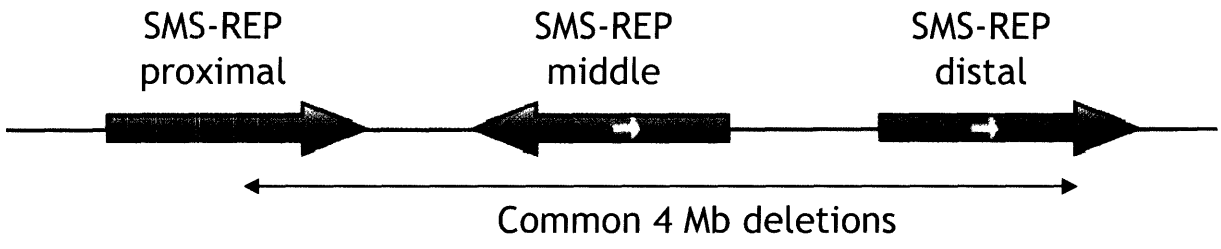
within the repeats, only two appear to be functional genes; neither is characterized (Park et al. 2002).

Other segmental duplications

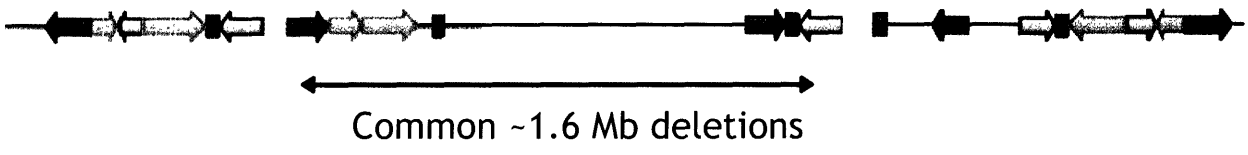
Studies of other genomic diseases have revealed additional segmental duplications that are inferred to undergo concerted evolution based upon detection of the duplications in species further diverged than the levels of nucleotide identities between the duplicated sequences imply. Williams-Beuren syndrome has features that include cardiovascular defects, facial dysmorphia, infantile hypercalcemia and cognitive and personality disorder (OMIM 194050). The most common form of Williams-Beuren syndrome (Figure 5b) is linked to a 1.5-2 Mb deletion generated by non-allelic homologous recombination between ~300 kb, 98% similar segmental duplications on chromosome 7q11.23 (Baumer et al. 1998; Valero et al. 2000). The clinical phenotype of the syndrome is likely due to haploinsufficiency of multiple genes within the critical region. While the region is not duplicated in rodents, comparative FISH analysis demonstrated that the segmental duplications responsible for the genomic deletions exist in multiple copies in chimpanzee, gorilla, orangutan and gibbon, suggesting the duplication event occurred before the hominoid diversification 20 million years ago (DeSilva et al. 1999). Similarly, the segmental duplications that engage in non-allelic recombination to generate the 3 Mb deletion on 22q11.2 linked to DiGeorge and velocardiofacial syndromes (DGS/VCFS, Figure 5c) are comprised of sequences that are 97-98% similar, but, again, are revealed by comparative FISH to be the result of

Figure 5. Genomic structure of segmental duplications that do not diverge. A. Duplications in the Smith-Magenis syndrome region. B. Duplications within the critical region for Williams-Beuren syndrome. C. Duplications associated with DiGeorge and velocardiofacial syndromes. (Figure adapted from Stankiewicz and Lupski 2002)

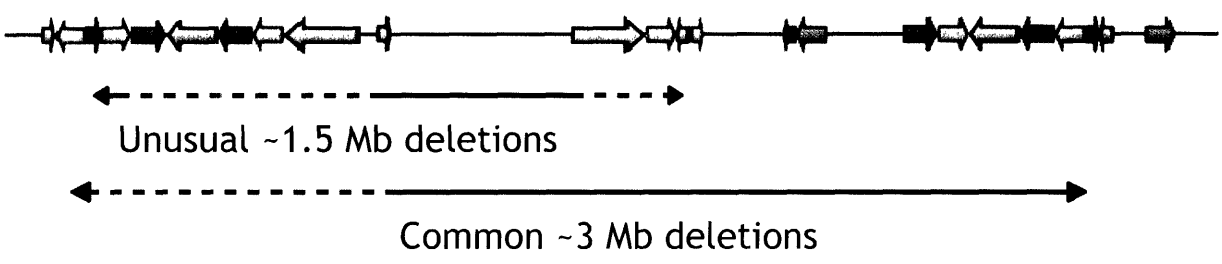
A. Smith-Magenis Syndrome



B. Williams-Beuren Syndrome



C. DiGeorge/velocardiofacial Syndromes



duplications before the divergence of Old World monkeys and hominoids 25 million years ago (Shaikh et al. 2000). Like the SMS-REPs, there is no hypothesized selective advantage that would explain the persistence of the Williams-Beuren or DGS/VCFS associated segmental duplications.

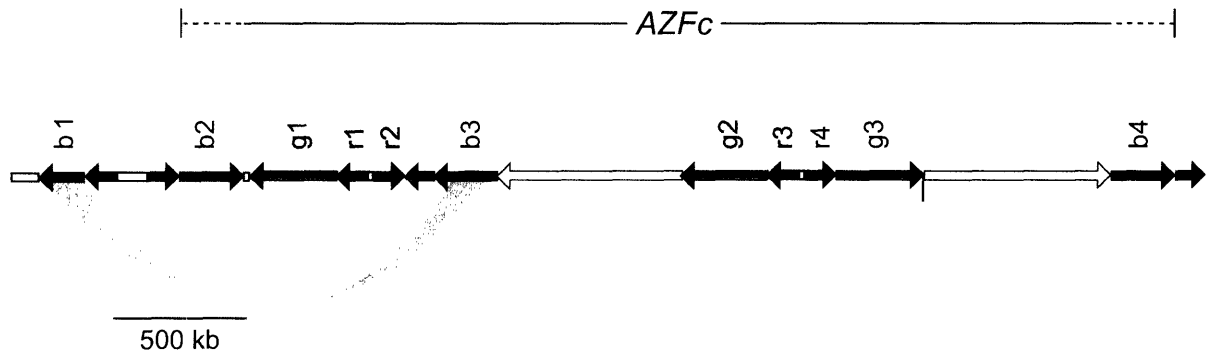
Palindromes on the Y chromosome

The complete sequence of the human Y chromosome revealed that 30% of the euchromatin is organized as inverted duplications, or massive imperfect palindromes. The arms of the palindromes range in size from 8 kb to 1.5 Mb with paired arm nucleotide identity greater than 99.9% (Rozen et al. 2003; Skaletsky et al. 2003). Like the segmental duplications described above, the Y palindromes, too, are subject to non-allelic recombination generating genomic disease (Figure 6). Recombination between tandemly repeated sequences created by the complex internal structures of the palindromes can generate deletions associated with male infertility or predisposition to male infertility (Kuroda-Kawaguchi et al. 2001; Repping et al. 2003; Repping et al. 2002). Also, unequal crossover between palindrome arms in separate chromatids generates iso-dicentric chromosomes (Vollrath et al. 1992, J. Lange, personnal communication)

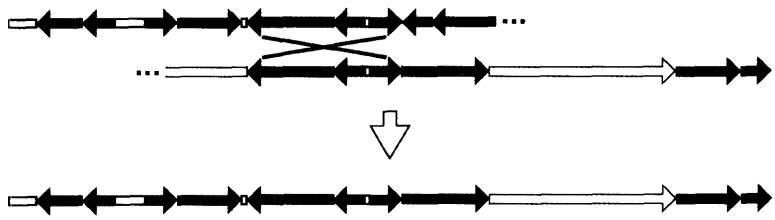
Not only did evidence from PCR amplification of palindrome boundary sequence reveal that orthologous palindromes are present in chimpanzee and gorilla, but comparative analysis was undertaken using sequenced orthologous palindromes in the chimpanzee (Rozen et al. 2003). Thus, it was discovered that not only is there concerted evolution between palindrome arms, but the nucleotide substitution rate between

Figure 6. Non-allelic recombination of the human Y chromosome. A. The central bar shows the organization of the repeats within AZFc region of the Y chromosome (Kuroda-Kawaguchi et al. 2001). The colored arcs show the repeats involved in recurrent deletions in AZFc. B. Model of homologous recombination generating the gr/gr deletion. C. Model of homologous recombination generating the b1/b3 deletion. D. Model of non-allelic homologous recombination between palindrome arms generating an isodicentric Y chromosome. (A, B and C adapted from Repping et al. 2003, D courtesy of Julian Lange)

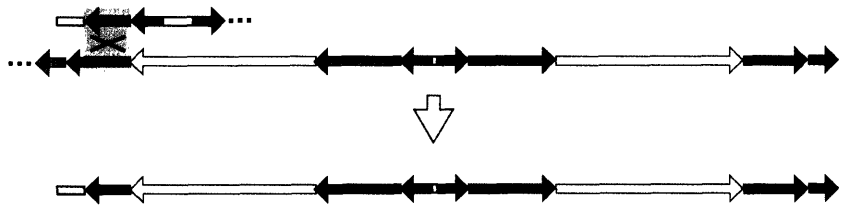
A.


AZFc
b1
b2
g1
r1
r2
b3
g2
r3
r4
g3
b4
500 kb

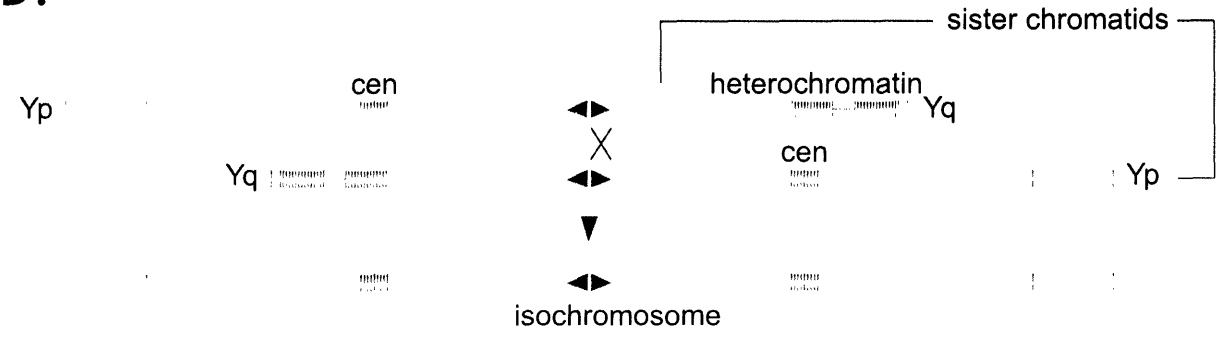
B.



C.



D.


sister chromatids
Yp
cen
heterochromatin
Yq
Yq
cen
Yp
isochromosome

orthologous palindrome arms is reduced relative to the surrounding single-copy sequence (Rozen et al. 2003). As palindromes are inverted duplications, and therefore are unable to undergo unequal crossover without generating chromosomal rearrangements, gene conversion is the likely mechanism preserving sequence identity. Human population studies show that gene conversion between palindrome arms is ongoing and frequent (Rozen et al. 2003). A genome-wide electronic search revealed that palindromes with spacers under 100 kb can be found elsewhere in the human genome; however, there is a clear enrichment for palindromes on the Y and the X chromosomes (See Chapter 2 and Warburton et al. 2004). Still, the Y is the clear outlier in the genome with an unprecedented amount of the chromosome organized in conserved palindromes. Hypotheses on function of Y palindromes as duplications that have not diverged can be inferred from the expression pattern of the associated genes; the expression patterns of all of the genes associated with Y palindromes are restricted to the testis. These hypotheses will be discussed below.

II. Functional organization of the genome by gene clustering

There are a rising number of reports that eukaryotic genomes are functionally organized. Genes may not be randomly distributed through the genome, but instead exhibit biases in their location. On the most basic level of functional clustering of genes, many duplicated genes are closely linked and still retain regulatory elements producing similar expression patterns. Any correlations between functions of linked duplicated genes are best studied in the context of duplication rather than functional organization of the genome. However, there are also many studies demonstrating co-expression of

closely linked non-homologous genes in the same tissue or the same pathway in yeast, worms, flies and mammals. Exclusion of genes expressed in a certain tissue from a chromosome, or bias of genes expressed in a certain tissue to a chromosome, have also been observed in both vertebrates and invertebrates. The primary hypothesis for the organization of co-expressed genes into clusters is that entire regions can be maintained in an open chromatin conformation to enhance transcriptional coregulation or coordination (Hurst et al. 2004).

Clustering of co-expressed genes

The first whole-genome study of gene clustering in eukaryotes was published in 1998 in a study of genome-wide transcriptional profiling of the mitotic cell cycle (Cho et al. 1998). In that study, it was found that of the ~6.7% of the genes in the yeast genome that exhibit cell-cycle dependent expression, 25% of those are present as adjacent pairs. A second group, using the same data set along with transcriptional profiling data sets from sporulation and mating-type pheromone response, included nearby non-adjacent genes in determining gene clustering (Cohen et al. 2000). While most of the significant clustering of co-expressed genes were limited to pairs and trios, they also found groups of clustered genes that span up to 26 kb. Aside from co-expression inferred from transcriptional profiling, clustering of genes related by common pathways and functions has also been examined. Of ~2000 adjacent pairs of genes, just under 400 pairs were in the same functional class (Cohen et al. 2000). Pathway information for non-duplicated genes in the yeast genome showed that the vast majority of genes in the same pathway were located in clusters in the genome (Lee and Sonnhammer 2003).

The study of gene clusters in the worm *Caenorhabditis elegans* has been confused by the existence of operons, single transcription units of multiple genes, which are prevalent in the *C. elegans* genome (Blumenthal et al. 2002). Although many dismiss the operons of *C. elegans* as a trivial means of clustering genes, operons are rare in other eukaryotes, and are an efficient method of co-regulating gene expression. About 15% of *C. elegans* genes are present in operons of 2-8 genes (Blumenthal et al. 2002), but the distribution of function and transcription is not random. Tissue and cell-type specific genes are rarely located in operons, whereas genes that code for factors involved in transcriptional regulation, translation and RNA degradation are much more likely to be encoded in operons (Blumenthal and Gleason 2003). Despite the large portion of genes located in operons, clusters of 2-5 co-expressed genes in regions up to 25 kb that are neither in the same operon nor paralogous have also been observed (Lercher et al. 2003a; Roy et al. 2002). The worm genome appears to have two methods of increasing transcriptional efficiency, both of which simplify transcription: operons at the promoter level and gene clusters at the chromatin level. A comparison between the types of genes that are in clusters relative to those in operons and any overlap between the two categories might provide insight into which genes in higher eukaryotes should be predicted to lie in clusters as well. It has been found that operons are conserved within nematodes (Blumenthal and Gleason 2003). A comparable study testing conservation of co-expressed gene clusters would be valuable in assessing the importance of gene clusters, as well.

In *Drosophila melanogaster,* gene clusters are on average significantly larger than in worms or yeast. A study by Spellman and Rubin (2002) used microarray data to find

clusters of similarly expressed genes. They found that 20% of genes in the fly genome exist in clusters of 10-30 genes over 20-200 kb regions. They did not find correlation between the clusters and any known chromosomal structures involved in transcriptional control. Another study by Boutanaev and colleagues (2002) looked at genes solely transcribed in the testis based on EST profiling. They found over a third of all testis-specific genes were organized in clusters containing four or more genes. As more genomic data becomes available for other *Drosophila* species, any conservation in clusters will assist in determining the significance of the gene clustering.

There have been a multitude of genome-wide studies performed to identify clusters of co-expressed genes in humans (Table 1). Many of those studies were intended to reveal information about chromosomal domains containing genes that might be relevant to medical conditions and human disease. Most of these studies entailed mapping of cDNAs from relevant tissue libraries to the human genome (Bortoluzzi et al. 1998; Ko et al. 1998; Qiu et al. 2002; Yager et al. 2004), while a few used serial analysis of gene expression (SAGE) (Caron et al. 2001; Lercher et al. 2002; Yamashita et al. 2004), or array hybridization (Gabrielsson et al. 2000; Yager et al. 2004). The clusters in these studies extend to over a megabase and often explicitly include genes within the clusters that are not expressed in the tissue of interest (Dempsey et al. 2001). The strongest correlation between clusters of genes and their transcriptional profile is among housekeeping genes, that is genes that are widely expressed (Lercher et al. 2002). For the studies identifying clusters of genes expressed in particular tissues, most looked at all genes expressed in the tissue of interest without excluding duplicated genes or housekeeping genes. By including housekeeping genes, many tissue cluster studies may

Table 1. Studies of clustering of co-expressed genes in mammals. Breadth of analysis indicates whether the study was genome-wide or restricted to particular chromosomes. Data source indicates the type of data used to determine tissue expression: RT-PCR of genes from a panel of tissues, cDNA library sequencing, Serial Analysis of Gene Expression (SAGE), cDNA array hybridization, EST database searches or Affymetrix microarray hybridization. Controlled for duplicates indicates whether the study removed duplicated genes from the analysis, controlled for housekeeping indicates whether the study removed ubiquitously expressed genes from the analysis.

Organism	Breadth of analysis	Data source	Controlled for duplicates	Controlled for housekeeping	Results	Study
human	Chromo-somes 21, 22	ESTs	no	no	2 hotspots, 1 coldspot for cardiovascular genes; all 3 corresponded to high gene density regions	Dempsey et al. 2001
human/ mouse	Chromo-some 21	RT-PCR in mouse	yes	yes	brain hotspot; cold spots for heart, lung, testis and muscle; clusters syntenic to mouse clusters	Reymond et al. 2002
human	genome	cDNA	no	no	muscle genes clustered on 17, 19 & X	Bortuluzzi et al. 1998
mouse	genome	cDNA	no	no	extra-embryonic clusters on 2, 7, 9 & 17; X is cold spot; 17 cluster is part of the t-complex	Ko et al. 1998
human	genome	cDNA array	no	no	adipose tissue clusters	Gabrielsson et al. 2000
human	genome	SAGE	no	no	highly expressed genes cluster	Caron et al. 2001
human	genome	SAGE	yes	NA	housekeeping clusters, no tissue-specific clusters, high expression clusters due to housekeeping genes	Lercher et al. 2002
human	genome	ESTs	no	no	brain clusters; brain cold spots; clusters are in avg gene density regions	Qiu et al. 2002
human	genome	SAGE	no	yes	6 liver-related, 5 colon-related, no breast- or brain-related clusters	Yamashita et al. 2004
human	genome	ESTs & affymetrix	no	yes	clusters of 4-10 cartilage genes, also cold spots; both correspond to regions of high gene density.	Yager et al. 2004

merely recapitulate the pattern set by the housekeeping genes. A study looking at the genes on human chromosome 21 used RT-PCR to define tissue distributions of transcription of mouse orthologs and excluded widely expressed housekeeping genes. (Reymond et al. 2002). They found clusters containing tissue co-expressed or tissue co-silenced genes and furthermore found that these clusters were syntenic in the mouse. Another study that controlled for both duplicates and housekeeping genes found that liver-related and colon- related clusters were syntenic to clusters in the mouse and rat genomes (Yamashita et al. 2004)

Perhaps because there are fewer studies, the data appears clearer for worms and flies that the phenomena of co-expressed gene clustering is real. For mammals, the number of studies that examined the question of identifying co-expressed clusters suggests that a meta-analysis of all of the different data sets would be useful. Overlapping results from different studies should add significance to the cluster predictions. Conflicting results should point out flaws in experimental designs. While the relevance of studies that did not take duplicated genes or housekeeping genes into consideration when predicting clusters, the studies performed with proper controls (Lercher et al. 2002; Reymond et al. 2002) make it clear that the human genome contains clusters of co-expressed genes. Studies demonstrating the conservation of clusters between humans and rodents (Reymond et al. 2002; Yamashita et al. 2004) lend further credence to co-expressed gene clustering because it demonstrates a selective advantage to having genes clusters rather than clusters being mere statistical anomalies.

Sexual antagonism and biases of gene content on the sex chromosomes

The X and Y chromosomes each differ in gene content when compared to the autosomes. These differences are the result of the presence of the male determining locus on the Y chromosome and because both the X chromosome and the Y chromosome are haploid in males. Close linkage to the male determining locus increases the likelihood a particular allele will pass through the male germ-line. Haploidy allows selection to act quickly on recessive alleles. Because of these two phenomena, sexually antagonistic genes are predicted to be found more frequently on sex chromosomes (Fisher 1931; Rice 1984). A sexually antagonistic gene or allele provides a benefit to one sex, yet is neutral or deleterious when expressed in the other sex. The classic example of a sexually antagonistic trait is the sexually dimorphic coloration of guppies (Winge 1927). Dun-colored female guppies are well camouflaged from predation, but the bright colors typical of male guppies greatly enhance mate attraction. Males are benefited by increased sexual attractiveness and thus an enhanced likelihood of reproductive success. The same traits appearing in a female would only increase her likelihood of predation. Of the 18 genes shown to affect color, 17 of them are on the sex chromosomes (Winge 1927).

Male-benefit genes are predicted to accumulate on the X chromosome because of the haploid nature of the X chromosome during its passage through the male germline. For recessive alleles increasing male fitness, linkage to the X chromosome enhances rate of allele fixation in the population. While the allele is still rare, it is hidden by the wildtype allele during its passage through heterozygous females; however, in the hemizygous male, the allele is fully exposed to selection. Thus, even when the disadvantage to one sex is greater than the advantage to the other sex, the allele will still

increase in frequency. Alternatively, the evolution of modifiers to restrict the expression pattern of the sexually antagonistic gene to the sex where it is beneficial will also suffice to bring the sexually antagonistic allele to fixation (Rice 1984).

Sexual antagonism can be invoked to explain the apparent bias of male genes on the human and mouse X chromosomes. Wang and colleagues (2001), seeking to catalog genes expressed in spermatogonia, the spermatogenic stem cells, found an enrichment on the sex chromosomes. Of the 25 genes identified, 3 were on the Y chromosome, and 10 on the X chromosome. A random distribution of 25 genes in the genome would generate an expectation of zero on the Y chromosome and two on the X chromosome. A separate study examining expression in pre-meiotic cells, including spermatogonia and using a mouse mutant with a pre-meiotic arrest phenotype, also found an enrichment on the sex chromosomes (Khil et al. 2004). Another study looking at male specific genes that are not expressed in the germline also discovered a propensity for X-linkage among genes expressed exclusively in the prostate (Lercher et al. 2003b). Saifi and Chandra (1999) demonstrated a bias of sex-related genes of on the X chromosome. Sex-related genes were identified by their association, when mutated, with sex-determination, reproductive or sexual differentiation pathologies and abnormalities. Despite the demonstrated statistical biases of male benefit genes on the X chromosome, it should be remembered that testis genes comprise a small fraction of the total number of genes on the X chromosome (see Chapter 2).

Despite the selection for male fitness enhancing genes on the X chromosome, a stronger selective pressure prevents accumulation of gene transcribed during meiosis on the X chromosome. During the meiotic stages of mammalian spermatogenesis, the X and

Y chromosomes are transcriptionally silenced (Wang 2004), thereby generating selection against genes required during these stages to reside on the X chromosome. In mice, while the X chromosome is enriched for testis genes expressed before meiotic X chromosome silencing, it has a deficit of genes expressed during meiosis (Khil et al. 2004). In the fly and worm, the X chromosome is also silenced (Kelly et al. 2002; Lifschytz and Lindsley 1972) and both species have an underrepresentation of genes involved in spermatogenesis on the X chromosome (Parisi et al. 2003; Reinke et al. 2000). In the mouse, the expression of X-linked genes subsequent to meiosis has been noted (Khil et al. 2004), but a genome-wide survey of post-meiotic testis is lacking. An interesting corollary of the exclusion of spermatogenesis genes is the finding that many autosomal retroposed genes with testis function originated from genes on the X chromosome in both flies and mammals (Betran et al. 2002; Wang 2004).

Like the X chromosome, the Y chromosome is also predicted to harbor male-benefit genes. Indeed, in mammals and flies, the Y chromosome is necessary for male fertility (Carvalho 2002; Lahn et al. 2001). Unlike the X chromosome, the Y chromosome never passes through the female germ-line, so counter-selection due to diminished female fitness never affects Y-linked genes. Experimental work in flies generating synthetic Y chromosomes demonstrated the accumulation of male-benefit, female-detriment genes on the Y chromosome (Rice 1998). There is some evidence that the genes on the human Y chromosome would be detrimental to females because females carrying a piece of the Y chromosome are predisposed to develop gonadoblastoma, an otherwise rare tumor, in their ovaries (Tsuchiya et al. 1995). The Y chromosome in humans is sequenced and the full content of its genes known (Skaletsky et al. 2003),

revealing a huge enrichment for testis genes; 62 of the 78 protein coding genes on the Y chromosome are testis-specific.

III. A functional role for structural organization

The intersection of functional organization and structural organization is not well-explored in mammalian genomes. At the present time, the only example known can be seen in the palindromes of the Y chromosome. Here, the obvious structural pattern of massive palindromes is associated with the functional pattern of clustered testis genes. Of the 62 Y-linked testis genes, 60 are compartmentalized within duplicated sequence on the Y chromosome. Duplicated sequence displaying over 92% nucleotide identity comprises almost half of the human Y chromosome. Massive palindromes, inverted repeats displaying greater than 99.9% nucleotide identity, comprise 30% of the Y chromosome and house only testis genes. The arms of the palindromes are as similar to each other as allelic chromosomes are to each other (Skaletsky et al. 2003). As discussed above, the palindrome arms, including genic and non-genic sequence, serve as substrates for ongoing gene conversion and homologous recombination. The genic sequences in palindrome arms are all expressed in the testis, and specifically within the germ cells of the testis (Skaletsky et al. 2003), thereby establishing the association between genomic structure and function.

The association between conserved palindromes and testis transcription bias is not unique to the Y chromosome, but is also seen on the X chromosome and, to a lesser extent, potentially among the autosomes as well (See Chapter 2 and Warburton et al.

2004). The Y chromosome is necessary for male fertility, making preservation of functional testis genes important for male reproductive fitness. Except for the two pseudoautosomal regions at either tip of the chromosome that pair with homologous regions on the X chromosome, the Y chromosome does not have a pairing partner during meiosis with which to recombine. Similarly, the X chromosome does not have a pairing partner during male meiosis. Compartmentalization of testis genes into recombination-capable genomic structures may allow homologous recombination between palindrome arms to provide the benefits of recombination in purging deleterious mutations. Indeed, the degeneration that is a hallmark of nonrecombining chromosomes (Charlesworth 1991; Rice 1994) is observed outside of the palindromes on the Y chromosome, but a decrease in the rate of evolution between orthologous palindrome arms relative to orthologous single-copy surrounding sequence (Rozen et al. 2003) suggests that degeneration may be slowed within the Y palindromes. Alternatively or additionally, pairing between palindrome arms may provide a scaffold for a germ cell specific chromatin conformation (Skaletsky et al. 2003; Warburton et al. 2004).

The functional role for palindrome genomic structure deserves further study. Despite the unique characteristic of the Y chromosome in never pairing with an allelic chromosome, other palindromes in the genome have been found, suggesting that the benefit of palindromes may be a broad one. As the searches for different functional and structural patterns in finished genomes progress, more intersections of function and structure will likely emerge. These intersections represent a further dimension to the information coded by genome sequences.

REFERENCES

Amores, A., A. Force, Y.L. Yan, L. Joly, C. Amemiya, A. Fritz, R.K. Ho, J. Langeland, V. Prince, Y.L. Wang, M. Westerfield, M. Ekker, and J.H. Postlethwait. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282:** 1711-1714.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Bailey, J.A., G. Liu, and E.E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Bailey, J.A., A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005-1017.

Baumer, A., F. Dutly, D. Balmer, M. Riegel, T. Tukel, M. Krajewska-Walasek, and A.A. Schinzel. 1998. High level of unequal meiotic crossovers at the origin of the 22q11. 2 and 7q11.23 deletions. *Hum Mol Genet* **7:** 887-894.

Betran, E., K. Thornton, and M. Long. 2002. Retroposed new genes out of the X in Drosophila. *Genome Res* **12:** 1854-1859.

Blumenthal, T., D. Evans, C.D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W.L. Chiu, K. Duke, M. Kiraly, and S.K. Kim. 2002. A global analysis of Caenorhabditis elegans operons. *Nature* **417:** 851-854.

Blumenthal, T. and K.S. Gleason. 2003. Caenorhabditis elegans operons: form and function. *Nat Rev Genet* **4:** 112-120.

Boerkoel, C.F., K. Inoue, L.T. Reiter, L.E. Warner, and J.R. Lupski. 1999. Molecular mechanisms for CMT1A duplication and HNPP deletion. *Ann N Y Acad Sci* **883:** 22-35.

Bortoluzzi, S., L. Rampoldi, B. Simionati, R. Zimbello, A. Barbon, F. d'Alessi, N. Tiso, A. Pallavicini, S. Toppo, N. Cannata, G. Valle, G. Lanfranchi, and G.A. Danieli. 1998. A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res* **8:** 817-825.

Boutanaev, A.M., A.I. Kalmykova, Y.Y. Shevelyov, and D.I. Nurminsky. 2002. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420:** 666-669.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.C. Hermus, R. van Asperen, K. Boon, P.A. Voute, S. Heisterkamp, A. van Kampen, and R. Versteeg. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289-1292.

Carvalho, A.B. 2002. Origin and evolution of the Drosophila Y chromosome. *Curr Opin Genet Dev* **12:** 664-668.

Charlesworth, B. 1991. The evolution of sex chromosomes. *Science* **251:** 1030-1033.

Chen, K.S., P. Manian, T. Koeuth, L. Potocki, Q. Zhao, A.C. Chinault, C.C. Lee, and J.R. Lupski. 1997. Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat Genet* **17:** 154-163.

Cho, R.J., M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis. 1998.

A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2:** 65-73.

Cohen, B.A., R.D. Mitra, J.D. Hughes, and G.M. Church. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26:** 183-186.

Dempsey, A.A., N. Pabalan, H.C. Tang, and C.C. Liew. 2001. Organization of human cardiovascular-expressed genes on chromosomes 21 and 22. *J Mol Cell Cardiol* **33:** 587-591.

DeSilva, U., H. Massa, B.J. Trask, and E.D. Green. 1999. Comparative mapping of the region of human chromosome 7 deleted in williams syndrome. *Genome Res* **9:** 428-436.

Feil, R., P. Aubourg, R. Heilig, and J.L. Mandel. 1990. A 195-kb cosmid walk encompassing the human Xq28 color vision pigment genes. *Genomics* **6:** 367-373.

Fisher, R.A. 1931. The evolution of dominance. *Biological Reviews* **6:** 345-368.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Gabrielsson, B.L., B. Carlsson, and L.M. Carlsson. 2000. Partial genome scale analysis of gene expression in human adipose tissue using DNA array. *Obes Res* **8:** 374-384.

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

Hammarstrom, K., B. Santesson, G. Westin, and U. Pettersson. 1985. The gene cluster for human U2 RNA is located on chromosome 17q21. *Exp Cell Res* **159:** 473-478.

Huminiecki, L. and K.H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14:** 1870-1879.

Hurst, L.D., C. Pal, and M.J. Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5:** 299-310.

Hurst, L.D. and N.G.C. Smith. 1998. The evolution of concerted evolution. *Proc R Soc Lond B Biol Sci* **265:** 121-127.

Jacobs, G.H. 1996. Primate photopigments and primate color vision. *Proc Natl Acad Sci U S A* **93:** 577-581.

Jorgensen, A.L., J. Philip, W.H. Raskind, M. Matsushita, B. Christensen, V. Dreyer, and A.G. Motulsky. 1992. Different patterns of X inactivation in MZ twins discordant for red-green color-vision deficiency. *Am J Hum Genet* **51:** 291-298.

Kelly, W.G., C.E. Schaner, A.F. Dernburg, M.H. Lee, S.K. Kim, A.M. Villeneuve, and V. Reinke. 2002. X-chromosome silencing in the germline of C. elegans. *Development* **129:** 479-492.

Khil, P.P., N.A. Smirnova, P.J. Romanienko, and R.D. Camerini-Otero. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat Genet* **36:** 642-646.

Kiyosawa, H. and P.F. Chance. 1996. Primate origin of the CMT1A-REP repeat and analysis of a putative transposon-associated recombinational hotspot. *Hum Mol Genet* **5:** 745-753.

Ko, M.S., T.A. Threat, X. Wang, J.H. Horton, Y. Cui, E. Pryor, J. Paris, J. Wells-Smith, J.R. Kitchen, L.B. Rowe, J. Eppig, T. Satoh, L. Brant, H. Fujiwara, S. Yotsumoto, and H. Nakashima. 1998. Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. *Hum Mol Genet* **7:** 1967-1978.

Kondrashov, F.A. and E.V. Koonin. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20:** 287-290.

Kumar, S. and S.B. Hedges. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917-920.

Kuroda-Kawaguchi, T., H. Skaletsky, L.G. Brown, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, S. Silber, R. Oates, S. Rozen, and D.C. Page. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat Genet* **29:** 279-286.

Lahn, B.T., N.M. Pearson, and K. Jegalian. 2001. The human Y chromosome, in the light of evolution. *Nat Rev Genet* **2:** 207-216.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M.

Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan J. Szustakowki P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lee, J.M. and E.L. Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13:** 875-882.

Lercher, M.J., T. Blumenthal, and L.D. Hurst. 2003a. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Res* **13:** 238-243.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31:** 180-183.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2003b. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol Biol Evol* **20:** 1113-1116.

Li, W.H. 1997. *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

Liao, D., T. Pavelitz, J.R. Kidd, K.K. Kidd, and A.M. Weiner. 1997. Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. *Embo J* **16:** 588-598.

Lifschytz, E. and D.L. Lindsley. 1972. The role of X-chromosome inactivation during spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-dosage compensation). *Proc Natl Acad Sci U S A* **69:** 182-186.

Lindgren, V., M. Ares, Jr., A.M. Weiner, and U. Francke. 1985. Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. *Nature* **314:** 115-116.

Lupski, J.R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14:** 417-422.

Lynch, M. and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151-1155.

Lynch, M. and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459-473.

Maeda, N. and O. Smithies. 1986. The evolution of multigene families: human haptoglobin genes. *Annu Rev Genet* **20:** 81-108.

Makova, K.D. and W.H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* **13:** 1638-1645.

McClintock, J.M., M.A. Kheirbek, and V.E. Prince. 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Development* **129:** 2339-2354.

McGinnis, W. and R. Krumlauf. 1992. Homeobox genes and axial patterning. *Cell* **68:** 283-302.

Nathans, J., T.P. Piantanida, R.L. Eddy, T.B. Shows, and D.S. Hogness. 1986a. Molecular genetics of inherited variation in human color vision. *Science* **232:** 203-210.

Nathans, J., D. Thomas, and D.S. Hogness. 1986b. Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science* **232:** 193-202.

Nei, M., I.B. Rogozin, and H. Piontkivska. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A* **97:** 10866-10871.

Ohno, S. 1970. *Evolution by gene duplication.* Springer-Verlag, New York.

Otto, S.P. and P. Yong. 2002. The evolution of gene duplicates. *Adv Genet* **46:** 451-483.

Paques, F. and J.E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* **63:** 349-404.

Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley, J. Andrews, S. Eastman, and B. Oliver. 2003. Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science* **299:** 697-700.

Park, S.S., P. Stankiewicz, W. Bi, C. Shaw, J. Lehoczky, K. Dewar, B. Birren, and J.R. Lupski. 2002. Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs. *Genome Res* **12:** 729-738.

Pavelitz, T., L. Rusche, A.G. Matera, J.M. Scharf, and A.M. Weiner. 1995. Concerted evolution of the tandem array encoding primate U2 snRNA occurs in situ, without changing the cytological context of the RNU2 locus. *Embo J* **14:** 169-177.

Piatigorsky, J. and G. Wistow. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science* **252:** 1078-1079.

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

Piontkivska, H., A.P. Rooney, and M. Nei. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol Biol Evol* **19:** 689-697.

Qiu, P., L. Benbow, S. Liu, J.R. Greene, and L. Wang. 2002. Analysis of a human brain transcriptome map. *BMC Genomics* **3:** 10.

Reinke, V., H.E. Smith, J. Nance, J. Wang, C. Van Doren, R. Begley, S.J. Jones, E.B. Davis, S. Scherer, S. Ward, and S.K. Kim. 2000. A global profile of germline gene expression in C. elegans. *Mol Cell* **6:** 605-616.

Reiter, L.T., T. Murakami, T. Koeuth, R.A. Gibbs, and J.R. Lupski. 1997. The human COX10 gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. *Hum Mol Genet* **6:** 1595-1603.

Reiter, L.T., T. Murakami, T. Koeuth, L. Pentao, D.M. Muzny, R.A. Gibbs, and J.R. Lupski. 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat Genet* **12:** 288-297.

Repping, S., H. Skaletsky, L. Brown, S.K. van Daalen, C.M. Korver, T. Pyntikova, T. Kuroda-Kawaguchi, J.W. de Vries, R.D. Oates, S. Silber, F. van der Veen, D.C. Page, and S. Rozen. 2003. Polymorphism for a 1.6-Mb deletion of the human Y

chromosome persists through balance between recurrent mutation and haploid selection. *Nat Genet* **35:** 247-251.

Repping, S., H. Skaletsky, J. Lange, S. Silber, F. Van Der Veen, R.D. Oates, D.C. Page, and S. Rozen. 2002. Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am J Hum Genet* **71:** 906-922.

Reymond, A., V. Marigo, M.B. Yaylaoglu, A. Leoni, C. Ucla, N. Scamuffa, C. Caccioppoli, E.T. Dermitzakis, R. Lyle, S. Banfi, G. Eichele, S.E. Antonarakis, and A. Ballabio. 2002. Human chromosome 21 gene expression atlas in the mouse. *Nature* **420:** 582-586.

Rice, W.R. 1984. Sex-chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735-742.

Rice, W.R. 1994. Degeneration of a nonrecombining chromosome. *Science* **263:** 230-232.

Rice, W.R. 1998. Male fitness increases when females are eliminated from gene pool: implications for the Y chromosome. *Proc Natl Acad Sci U S A* **95:** 6217-6221.

Roy, P.J., J.M. Stuart, J. Lund, and S.K. Kim. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418:** 975-979.

Rozen, S., H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, and D.C. Page. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423:** 873-876.

Saifi, G.M. and H.S. Chandra. 1999. An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proc R Soc Lond B Biol Sci* **266:** 203-209.

Shaikh, T.H., H. Kurahashi, S.C. Saitta, A.M. O'Hare, P. Hu, B.A. Roe, D.A. Driscoll, D.M. McDonald-McGinn, E.H. Zackai, M.L. Budarf, and B.S. Emanuel. 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* **9:** 489-501.

Shyue, S.K., L. Li, B.H. Chang, and W.H. Li. 1994. Intronic gene conversion in the evolution of human X-linked color vision genes. *Mol Biol Evol* **11:** 548-551.

Skaletsky, H., T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825-837.

Slightom, J.L., A.E. Blechl, and O. Smithies. 1980. Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21:** 627-638.

Smith, G.P. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science* **191:** 528-535.

Spellman, P.T. and G.M. Rubin. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* **1:** 5.

Stankiewicz, P. and J.R. Lupski. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18:** 74-82.

Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, and J.B. Hogenesch. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99:** 4465-4470.

Tsuchiya, K., R. Reijo, D.C. Page, and C.M. Disteche. 1995. Gonadoblastoma: molecular definition of the susceptibility region on the Y chromosome. *Am J Hum Genet* **57:** 1400-1407.

Valero, M.C., O. de Luis, J. Cruces, and L.A. Perez Jurado. 2000. Fine-scale comparative mapping of the human 7q11.23 region and the orthologous region on mouse chromosome 5G: the low-copy repeats that flank the Williams-Beuren syndrome deletion arose at breakpoint sites of an evolutionary inversion(s). *Genomics* **69:** 1-13.

Van Arsdell, S.W. and A.M. Weiner. 1984. Human genes for U2 small nuclear RNA are tandemly repeated. *Mol Cell Biol* **4:** 492-499.

Vollrath, D., S. Foote, A. Hilton, L.G. Brown, P. Beer-Romero, J.S. Bogan, and D.C. Page. 1992. The human Y chromosome: a 43-interval map based on naturally occurring deletions. *Science* **258:** 52-59.

Vollrath, D., J. Nathans, and R.W. Davis. 1988. Tandem array of human visual pigment genes at Xq28. *Science* **240:** 1669-1672.

Wang, P.J. 2004. X chromosomes, retrogenes and their role in male reproduction. *Trends Endocrinol Metab* **15:** 79-83.

Wang, P.J., J.R. McCarrey, F. Yang, and D.C. Page. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* **27:** 422-426.

Warburton, P.E., J. Giordano, F. Cheung, Y. Gelfand, and G. Benson. 2004. Inverted Repeat structure of the human genome: The X chromosome contains a preponderance of large highly homologous inverted repeats with contain testes genes. *Genome Res* **14:** 1861-1869.

Winderickx, J., L. Battisti, A.G. Motulsky, and S.S. Deeb. 1992. Selective expression of human X chromosome-linked green opsin genes. *Proc Natl Acad Sci U S A* **89:** 9710-9714.

Winge, O. 1927. The location of eighteen genes in Lebistes reticulatus. *J. Genet.* **18:** 1-43.

Wright, S. 1934. Physiological and evolutionary theories of dominance. *Am Nat* **68:** 24-53.

Yager, T.D., A.A. Dempsey, H. Tang, D. Stamatiou, S. Chao, K.W. Marshall, and C.C. Liew. 2004. First comprehensive mapping of cartilage transcripts to the human genome. *Genomics* **84:** 524-535.

Yamaguchi, T., A.G. Motulsky, and S.S. Deeb. 1997. Visual pigment gene structure and expression in human retinae. *Hum Mol Genet* **6:** 981-990.

Yamashita, T., M. Honda, H. Takatori, R. Nishino, N. Hoshino, and S. Kaneko. 2004. Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes. *Genomics* **84:** 867-875.

WEB SITE REFERENCES

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=omim. "OMIM, Online Mendelian Inheritance in Man"

CHAPTER 2

Gene conversion and testis transcription bias associated with palindromes on the primate X chromosome

(adapted from a manuscript in preparation for submission, with *Genome Research* formatting)

Author contributions:

The authors affiliated with the Washington University Genome Sequencing Center sequenced the chimpanzee BACs I identified as containing sequence homologous to human X chromosome palindromes. The authors affiliated with the National Human Genome Research Institute sequenced the orangutan and rhesus monkey BACs I identified as containing sequence homologous to human X chromosome palindromes. Steve Rozen wrote the custom Perl scripts for identifying palindromes from sequence data. Steve Rozen and Helen Skaletsky wrote other custom Perl scripts I utilized in my analysis of sequence data. Helen Skaletsky created the overgo probes I used when hybridizing the primate genomic BAC libraries for BACs containing sequence homologous to human X palindromes and also generated some of the primers I used for RT-PCR gene expression analysis of human X palindromes. I did all of the rest.

Gene conversion and testis transcription bias associated with palindromes on the primate X chromosome.

Jennifer R. Saionz[1], Helen Skaletsky[1], Steve Rozen[1], Robert W. Blakesley[2], Bradley I. Coleman[2], Tina Graves[3], Patrick Minx[3], Rick K. Wilson[3], Eric D. Green[2], David C. Page[1]

[1]Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA
[2]Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[3]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA

ABSTRACT

Recent genomic studies of the Y chromosome have revealed massive, testis-specific palindromes that span across 30% of the human Y chromosome and are subject to gene conversion. We conducted studies to determine whether similar palindromes exist on the human X chromosome and, if they exist, to what degree they share the features of the Y chromosomal palindromes. Electronic analysis revealed 24 palindromes on the X chromosome, composed of inverted repeats a few to hundreds of kilobases long and sharing greater than 99% nucleotide identity By combining RT-PCR with manual annotation, we showed experimentally that all of the genes associated with X palindromes are transcribed in the testis and a surprisingly large fraction are predominantly transcribed in the testis. We determined sequence from orthologous palindromes in chimpanzee, orangutan and rhesus monkey, revealing a common 25 million year-old origin for a portion of the palindromes. In all species studied, the palindrome arms underwent concerted evolution. Several examples of insertions and deletions greater than 100 bp between orthologous palindrome arms were subsequently homogenized to both palindrome arms. These homogenized indels (insertions or deletions) included a 14.6 kb deletion that is the largest example of an indel homogenization event due to gene conversion known in mammals. Like the human Y palindromes, the human X palindromes are a site of recurrent intra-chromatid recombination. Taken together with previous findings, a clear correlation exists between the enrichment for conserved palindromes on the sex chromosomes, a testis transcription bias associated with palindromes and recurrent gene conversion between palindrome arms.

INTRODUCTION

Current models of gene duplication predict divergence of duplicated gene copies (Lynch and Conery 2000; Ohno 1970). The origins of segmental duplications, low copy duplications of large sequences, are dated according to their nucleotide sequence identity under an assumption of constant divergence (Bailey et al. 2001; Lander et al. 2001). Despite the presuppositions of duplicated sequence divergence models, there are many well-studied examples of duplicated sequences that undergo concerted evolution, that is, they do not diverge, but instead remain highly similar. In mammals, examples of multiple copy sequence repeats include the X-linked color vision genes (Nathans et al. 1986), the rDNA repeats (Gonzalez and Sylvester 2001; Worton et al. 1988) and the U2 snRNA repeats (Pavelitz et al. 1995). In these cases, large tandem repeat arrays engage in concerted evolution, thereby preserving homogeneity of the repeat copies. The result renders the repeat units within a species more similar to each other than repeat units in different species. In some instances, the homogeneity of the repeat units is sufficiently extreme that the repeats within an array on a single chromosome are more similar to each other than repeats on allelic chromosomes in the same species (Liao et al. 1997). Examples of tandem repeat arrays undergoing concerted evolution, but unassociated with genes, such as the centromeric alpha satellite arrays (Warburton et al. 1993), further supports the idea that concerted evolution is as much a consequence of duplication as is divergence.

High copy duplications have more potential partners and presumably more opportunity for sequence homogenization than low copy duplications. Furthermore, many low copy segmental duplications can reach sizes of several tens of kilobases, if not

hundreds of kilobases, requiring homogenization over extremely long tract lengths. These ideas have led perhaps to the false supposition that low copy segmental duplications necessarily follow paths of divergence. A handful of segmental duplications are documented to be older than their high level of nucleotide identity implies. However, studies of their evolution in non-human primates have been limited to detecting presence of the duplication by hybridization (Aradhya et al. 2001; DeSilva et al. 1999; Park et al. 2002; Shaikh et al. 2000; Small et al. 1997) or PCR amplification of duplication boundary sequence (Warburton et al. 2004). Recent comparative sequence studies have yielded a more complete picture of sequence duplications. The remarkably large palindromes on the human Y chromosome have orthologous palindromes on the chimpanzee Y chromosome. Studies of the Y palindromes revealed both concerted evolution between paired palindrome arms and reduced sequence divergence between orthologous palindrome arms relative to nearby single copy sequence (Rozen et al. 2003).

Concerted evolution mechanisms include unequal chromatid exchange and gene conversion. Gene conversion is defined as the unidirectional transfer of information from one DNA strand to another (Szostak et al. 1983). Gene conversion models can incorporate formation of homologous recombination intermediates and allow for its occurrence in either mitotic or meiotic cells (Paques and Haber 1999). While tandemly repeated sequence can undergo either unequal chromatid exchange or gene conversion without gross chromosomal rearrangement, inverted repeats such as palindromes are restricted to gene conversion. Gene conversion in palindromes may be the target of selection for palindrome preservation or it may simply be a consequence of palindrome structure.

All of the genes within the human Y palindromes exhibit testis-predominant expression (Skaletsky et al. 2003). Concerted evolution of palindrome arms may provide the benefits of recombination to genes on this chromosome without a homolog (Rozen et al. 2003). An electronic survey of all palindromes in the human genome suggested that palindromes tend to contain testis expressed genes and speculated that palindromes might generate a unique chromatin structure for testis expression (Warburton et al. 2004).

To understand better the involvement of palindromes in the concerted evolution of duplicated sequence and their potential role in regulating gene activity in male germ cells, we undertook a study of the palindromes on primate X chromosomes. The human X chromosome contains a disproportionate number of palindromes relative to the rest of the genome and a suggested propensity for testis genes (Warburton et al. 2004). We sought to independently catalog all of the palindromes on the human X chromosome and to annotate manually and test empirically the transcription range of the genes within the palindromes. Preservation of palindromes on the chimpanzee and human Y chromosomes led us to search for X palindromes in primates further removed from humans and to characterize their sequence.

RESULTS

Palindromes comprise 1.8 percent of the human X chromosome

We searched the NCBI July 2003 Build of the human genome to identify palindromes. Our first round of electronic analysis uncovered inverted repeats over 5 kb separated by intervening spacers under 190 kb. We manually limited the analysis to palindromes with arms greater than 6 kb exhibiting arm-to-arm identity greater than 99%. The lower limit of arm lengths excluded recent LINE insertions. The lower limit of arm

identity ensured a sample of inverted duplications comparable to the palindromes on the human Y chromosome, all of which exhibit arm-to-arm identity greater than 99% identity (Skaletsky et al. 2003). Remarkably, of the 26 sets of inverted repeats uncovered by the first round electronic search and longer than 6 kb, only two exhibited less than 99% identity. Both displayed 97-98% identity and were excluded from further analysis.

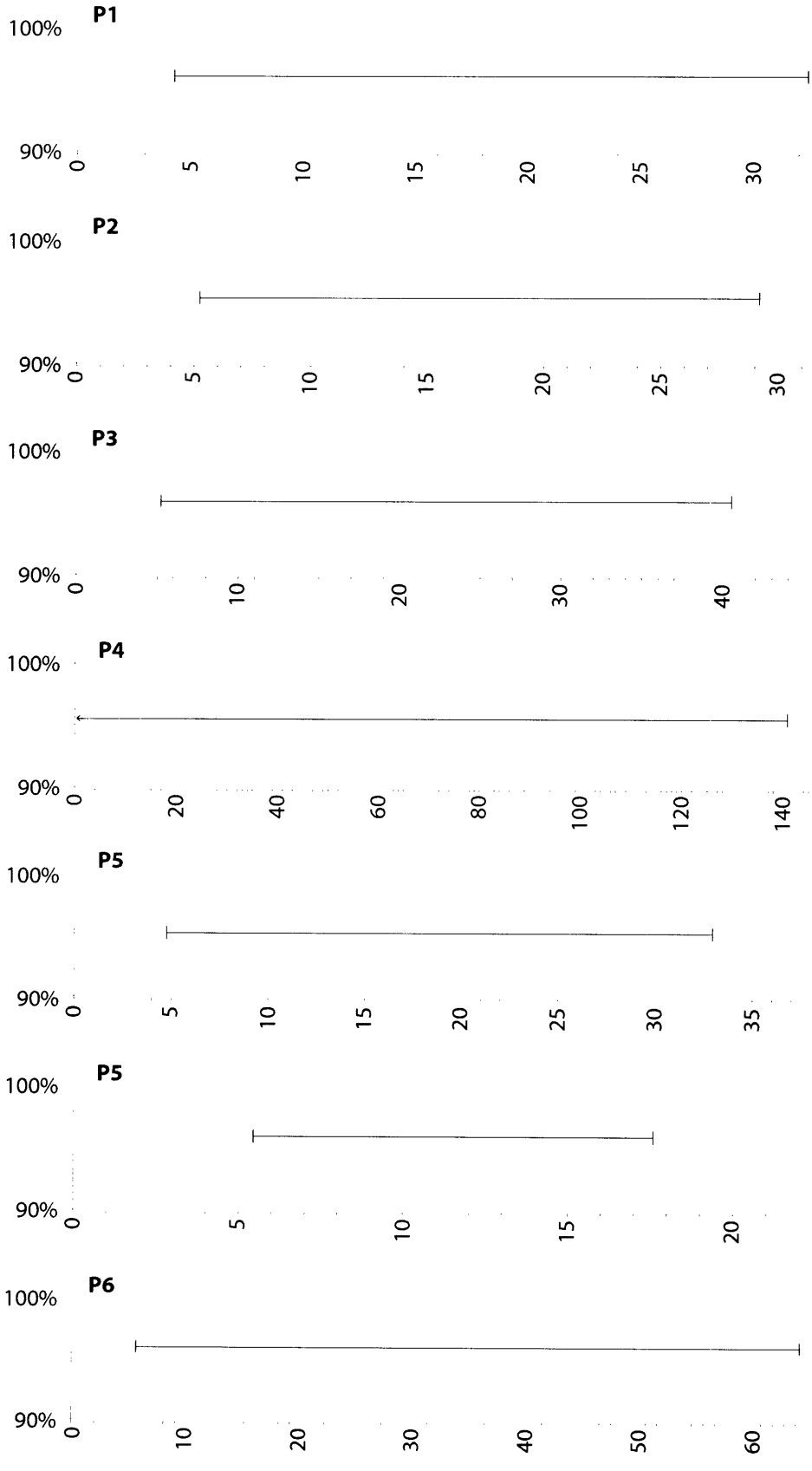
The 24 X palindromes range in size between 9.2 kb to over 140 kb arms with an average of 44 kb and together comprise 2.8 Mb or 1.8% of the 153 Mb X chromosome (Table 1). Nucleotide identity between palindrome arms varies between 99.42 to 99.98%. Several of the palindromes exhibit a trailing off of nucleotide identity at either the inner boundaries (between arm and spacer) or the outer boundaries (between arm and surrounding sequence) of the arm (Figure 1). Most of the palindromes are structured as two single inverted sequences; two, however, are more complex. P21 lies off center within the spacer of P20 (Figure 2). P4 and P5 are of related sequence composition (Figure 3), but a higher order genomic structure cannot be definitively ascertained at this point in time because of a gap in the map of the X chromosome between the two palindromes (Warburton et al. 2004). Half of the palindromes, 12 out of 24, are located within 15 Mb of the centromere. The remaining palindromes are spread out along the length of the long arm of the chromosome, but we observed no palindromes on the short arm further than 13 Mb away from the centromere (Tables 1 and 2).

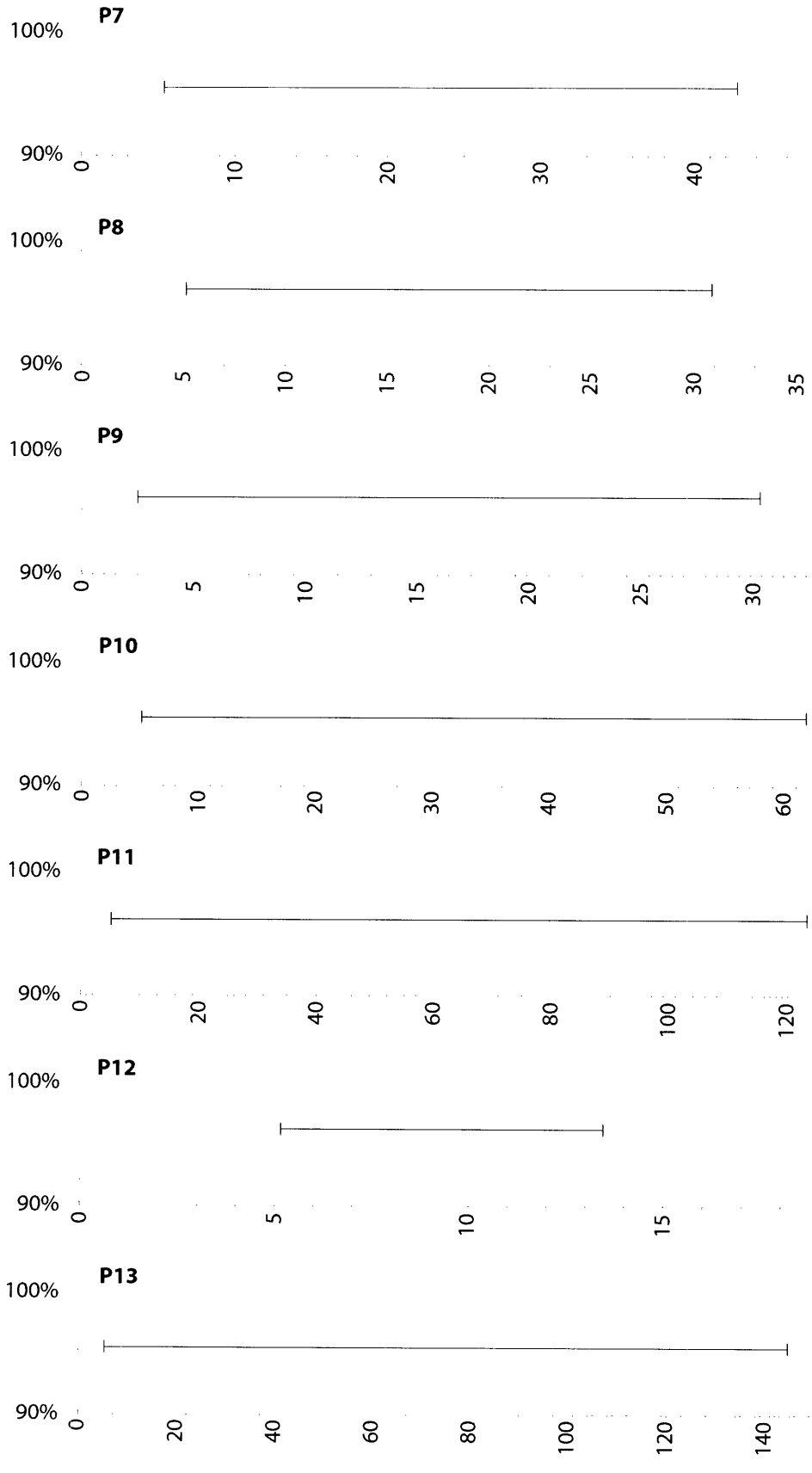
We sought to determine whether the dearth of palindromes on the short arm is an artifact of the long arm being more accurately and completely mapped than the short arm. We examined the number and size of X chromosome sequence contigs under the

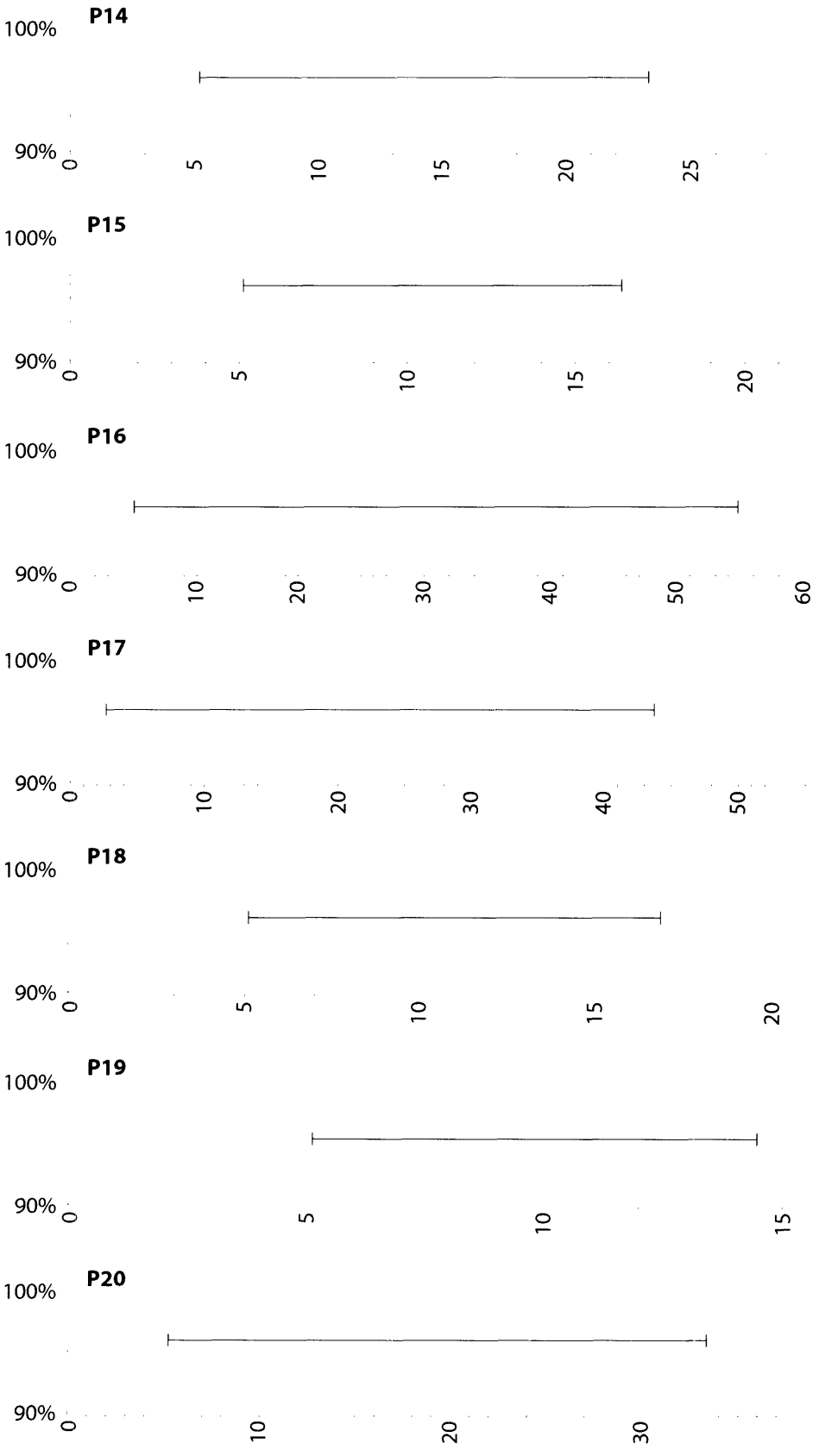
Table 1. Palindrome sizes, paired arm identity and gene content. Distance from centromere describes the distance in megabases from the centromere to the left-most edge of the palindrome according to the July 2003 UCSC Genome Browser. Negative numbers are on the short arms and positive numbers are on the long arm. Genes predominantly transcribed in the testis (see figure 6) are in bold. Genes located in the arms are duplicated identically in the paired arm. 1. There are three copies of GAGED2 in P5. 2. Transcription begins in P12 arms, but ends in P12 spacer. 3. P21 is located within the spacer of P20. 4. CSAG2 transcription pattern is documented in literature (Duan et al. 1999; Feller et al. 2000) and does not appear in figure 2.

	Distance from Centromere	Arm (kb)	Spacer (kb)	Identity (%)	Arm genes (duplicated)	Spacer genes (single copy)
P1	-9.7	29.0	2.8	99.94	***SSX4***	
P2	-6.7	25.0	7.4	99.94	***PRR6L1***	
P3	-6.3	36.4	99.9	99.98	*MAGED4*	
P4	-6.0	142.0	8.3	99.99	*GAGED3*, ***GAGED2***	
P5	-5.5	30.6	8.3	99.97	***GAGED2***[1]	
P6	-5.3	59.9	0.2	99.92	***SSX2***	
P7	-5.1	38.4	15.5	99.97	*TMEM29*	
P8	-2.6	26.6	12.9	99.98		*USP51*
P9	1.2	56.6	8.2	99.98		
P10	9.8	57.4	0.6	99.98	***CXorf49***	
P11	10.8	119.3	0.4	99.93	*DMRTC, CXorf50*	
P12	11.1	9.5	71.9	99.97		*RBM32A*[2], *RBM32B*[2]
P13	40.2	138.6	10.8	99.96	***NXF2***	
P14	42.0	20.3	60.0	99.93	*H2BFXP*	***H2BFM, H2BFWT***
P15	44.3	12.0	8.7	99.94		
P16	57.9	48.9	62.0	99.84	**AF317219**	**NM_139282**
P17	73.0	43.5	53.2	99.94	***ETD1, ETD2***	***CXorf48***
P18	79.4	12.7	3.9	99.42	***SPANXA***	
P19	84.6	10.5	0.2	99.96	***CXorf51***	
P20	87.4	29.1	164.8	99.87	***MAGEA9***	*FAM11A*[3]
P21	87.5	28.3	41.0	99.93		***MAGEA11***
P22	90.5	51.2	8.0	99.89	***MAGEA2, MAGEA6, CSAG2***[4]	***CSAG1, MAGEA12***
P23	92.0	11.4	37.6	99.85		*EMD, FLNA*
P24	92.3	35.5	21.8	99.95	***CTAG1***, *CXorf52*	

Figure 1. Percent nucleotide identity plots of paired palindrome arms. Percent nucleotide identity of a 1000 bp sliding window, 1 bp steps, constructed from alignment of palindrome paired arms and sequence contiguous to both sides of arms (for alignment of palindrome arms without surrounding sequence see accompanying CD or http://staffa.wi.mit.edu/page/saionz/Alignments/index.html). The Y-axis represents the percent nucleotide identity between the sequences. The X-axis measures the length of the paired arm alignment in kilobases. Horizontal bar delineates the arm of the palindrome. Outer boundaries are on the left, inner on the right. P5a percent identity plot generated from alignment of green and yellow sequences at outer boundaries of palindrome (see figure 3). P5b percent identity plot generated from alignment of yellow sequences surrounding spacer (see figure 3).


P1
100%
90%
0
5
10
15
20
25
30
P2
100%
90%
0
5
10
15
20
25
30
P3
100%
90%
0
10
20
30
40
P4
100%
90%
0
20
40
60
80
100
120
140
P5
100%
90%
0
5
10
15
20
25
30
35
P5
100%
90%
0
5
10
15
20
P6
100%
90%
0
10
20
30
40
50
60


P7
100%
90%
0
10
20
30
40
P8
100%
90%
0
5
10
15
20
25
30
35
P9
100%
90%
0
5
10
15
20
25
30
P10
100%
90%
0
10
20
30
40
50
60
P11
100%
90%
0
20
40
60
80
100
120
P12
100%
90%
0
5
10
15
P13
100%
90%
0
20
40
60
80
100
120
140


P14
100%
90%
0
5
10
15
20
25
P15
100%
90%
0
5
10
15
20
P16
100%
90%
0
10
20
30
40
50
60
P17
100%
90%
0
10
20
30
40
50
P18
100%
90%
0
5
10
15
20
P19
100%
90%
0
5
10
15
P20
100%
90%
0
10
20
30

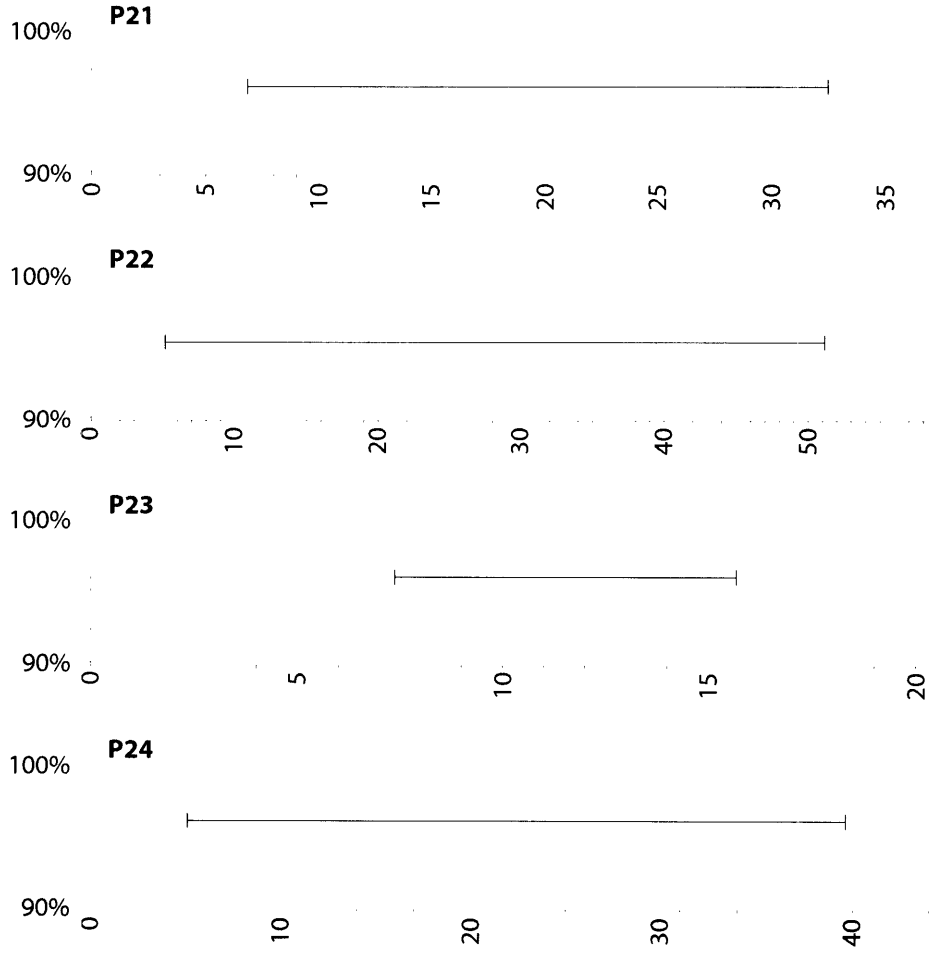

P21
100%
90%
0
5
10
15
20
25
30
35
P22
100%
90%
0
10
20
30
40
50
P23
100%
90%
0
5
10
15
20
P24
100%
90%
0
10
20
30
40

Figure 2. Palindrome P21 lies within the spacer of palindrome P20. Triangular dot plot in which 400000 bp from Xq28 is compared to itself. Within the plot each dot represents a match of 100% within a sliding window of 200 bp with 100 bp steps. Inverted repeats appear as vertical lines. Blue shaded arrows represent palindrome P21. Green shaded arrows represent palindrome P20. Red line arrows indicate genes.

w= 200 s=100

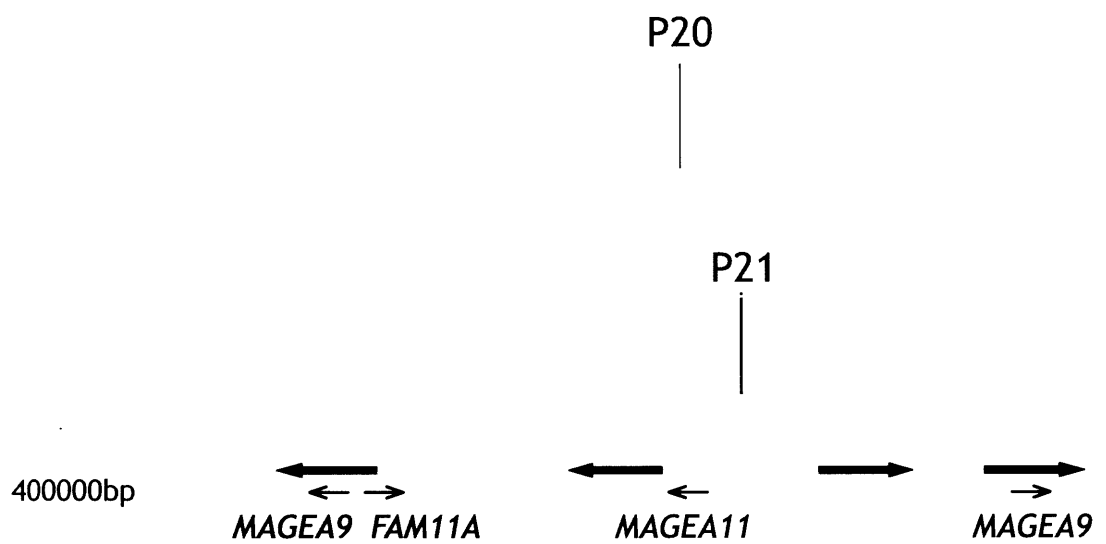


Figure 3. P4 and P5 form a complex higher order structure. Triangular dot plot in which 400000 bp from Xp11.22 is compared to itself. Within the plot each dot represents a match of 100% within a sliding window of 200 bp with 100 bp steps. Inverted repeats appear as vertical lines and direct repeats as horizontal lines. P4 and P5 are of related sequence composition, but a higher order genomic structure cannot be indisputably determined at this time due to a mapping gap in the X chromosome between the two palindromes. This gap sits at the end of P4 thereby preventing examination of its entirety. P5 is composed of two different repeat sequences. Repeat A (green) is present twice in inverted orientation. Repeat B (yellow) is present three times, once in one orientation, twice in tandem in the opposite orientation. The B repeats are immediately outside of the A repeats. *GAGED2* lies within Repeat B, and so appears three times within P5. The two repeats varieties and the unique spacer of P5 exist in P4 giving P4 two copies of *GAGED2*. Only a single base substitution in an intron of *GAGED2* differentiates the P4 *GAGED2* copies from the P5 *GAGED2* copies.

w= 200 s=100

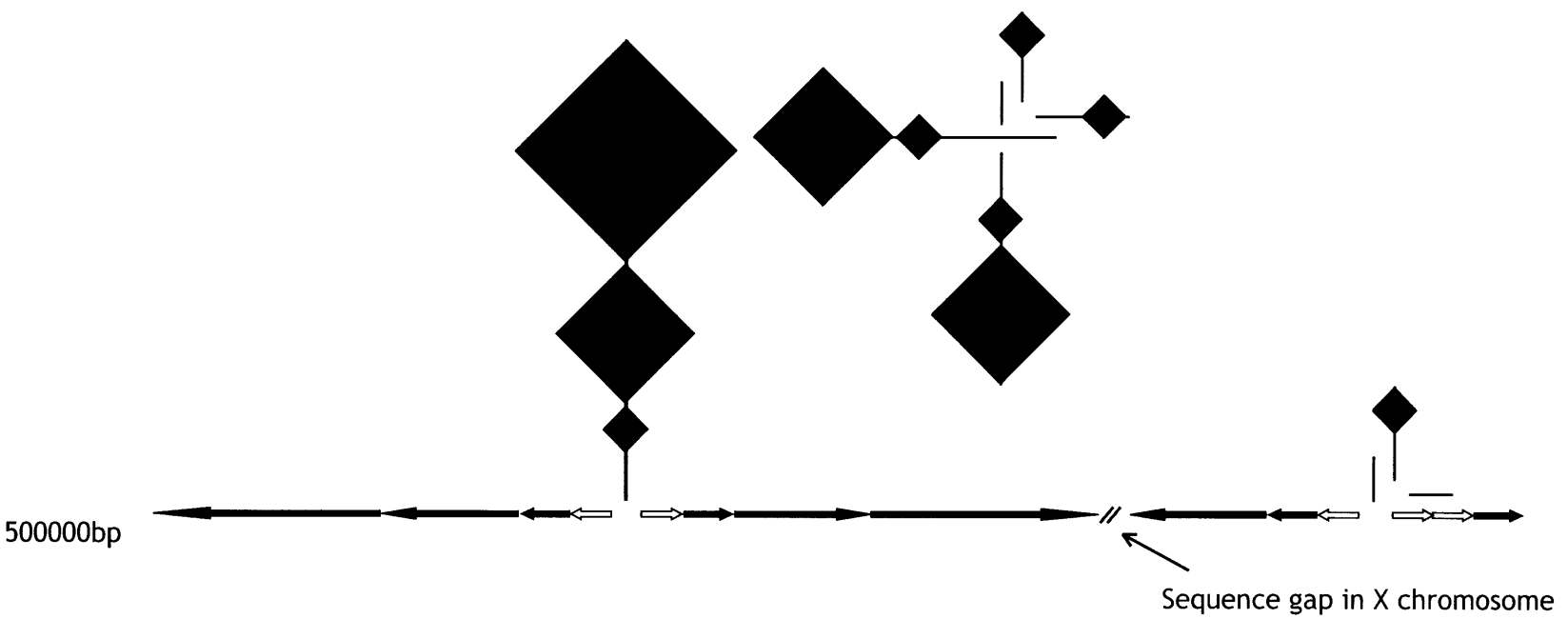


Table 2. Genomic position of human X palindromes. The palindrome boundary coordinates according to the UCSC genome browser July 2003 assembly.

	Arm1 begin	Spacer begin	Spacer end	Arm2 end
P1	47272742	47301377	47304896	47332477
P2	50327866	50352807	50360207	50385118
P3	50700617	50737009	50836890	50873313
P4	51020041	51162150	51170605	51312705
P5	51456330	51485101	51493548	51534876
P6	51665913	51724922	51725162	51784269
P7	51878073	51916123	51931676	51969542
P8	54446793	54473365	54486288	54512857
P9	61215266	61271836	61280043	61336611
P10	69769743	69827152	69827721	69885103
P11	70827693	70946844	70947262	71066382
P12	71082718	71091909	71164380	71173565
P13	100224493	100364950	100375715	100516231
P14	101995552	102014383	102077334	102096168
P15	104281370	104293439	104303068	104315137
P16	117926270	117974831	118037540	118086115
P17	132955658	132997644	133053970	133095968
P18	139354282	139366980	139370919	139383606
P19	144550981	144561423	144561502	144571944
P20	147360443	147389547	147554482	147583577
P21	147441866	147468462	147509478	147536074
P22	150464654	150511428	150529289	150576052
P23	152033502	152043518	152081144	152091159
P24	152251650	152287141	152308901	152344379

assumption that if the short arm had unobserved palindromes, they should be detected as unresolved gaps between map contigs. The number of contigs per megabase on the short arm is only slightly larger than the number of contigs per megabase on the long arm: 0.15 compared to 0.14. The average size of a short arm contig is 6.3 Mb, while the average size of a long arm contig is a slightly larger 7.1 Mb. Conceivably, undetected palindromes may exist in gaps in the present map of the X chromosome. However, there is no suggestion that the short arm harbors a larger density of undetected palindromes than the long arm.

Genes in palindromes display a testis-predominant transcription bias.

Evidence of a testis transcription bias in the Y palindromes is based on a systematic annotation of the Y chromosome (Skaletsky et al. 2003). A survey of palindromes in the entire human genome suggested a testis bias in all palindromes based on a cursory examination of publicly available EST data (Warburton et al. 2004). To test whether palindromes on the X chromosome display a transcriptional bias, we manually annotated the X palindromes and then evaluated the transcription pattern of all genes in both arms and spacers by RT-PCR.

We began our effort to describe all of the transcribed genes in the X palindromes by cataloging all genes from the NCBI annotation. To discover genes not yet annotated, we relied on the NCBI human EST databases in conjunction with experimental validation by RT-PCR. Only two palindromes, P9 and P15, contain no detectable transcription units. Our manual annotation of palindromes revealed over a dozen genes not uncovered in the Warburton *et al.* (2004) analysis. We discounted, as a chimeric EST and likely genomic contaminant, only a single gene that was included that analysis.

Unlike the Y palindromes, many of the X palindromes contain genes in the unique spacer sequence between the arms of the palindromes and a few contain genes that fall across the boundaries of the palindrome. *IKBKG* spans the outer boundary of P24 (Figure 4). *IKBKG* is a single copy gene with its transcriptional start 8kb away from the palindrome. Because the transcriptional start site and much of its coding sequence is outside of the palindrome, we did not include this gene in our analysis of transcriptional patterning of palindrome genes. A few of the genes span the inner boundaries of palindromes. In P12 there are two genes, *RBM32A* and *RBM32B*, that begin transcription in the arms and end transcription in the spacer. As can be predicted from their presumably identical promoter regions, their transcription patterns are the same (Figure 5 and Figure 6). The *CXorf48* gene and the *FAM11A* gene (Figure 2) also span the inner boundaries of the P17 and P20 palindromes, respectively. The transcriptional starts of both genes are within the spacers. Five copies of GAGED2 are embedded within the complex structure of duplications in P4 and P5: two copies in P4 and three copies in P5 (Figure 3).

To ascertain the transcription pattern of each of the palindrome genes we used RT-PCR. We included in our study genes located both in the arms and in the spacers. Forty-two out of 64 (66%) genes found in palindromes exhibit testis-predominant transcription (Figure 6). Thirty-five of 51 (69%) genes found in palindrome arms display testis-predominant transcription. (One gene, *CSAG2*, was not included in the RT-PCR analysis because we were unable to successfully amplify it by PCR. Because *CSAG2* is well characterized as being exclusively expressed in the testis among normal tissues (Duan et al. 1999; Feller et al. 2000) , we did include it in our analysis of palindrome

Figure 4. *IKBKG* spans the outer boundary of P24. Schematic drawing of P24 gene structures. Thick horizontal lines represent palindrome arms. Boxes represent exons. Arrows give direction of transcription. Gene names appear above the exons.

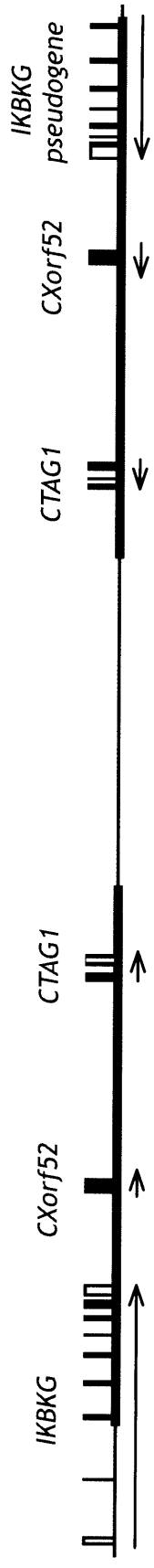
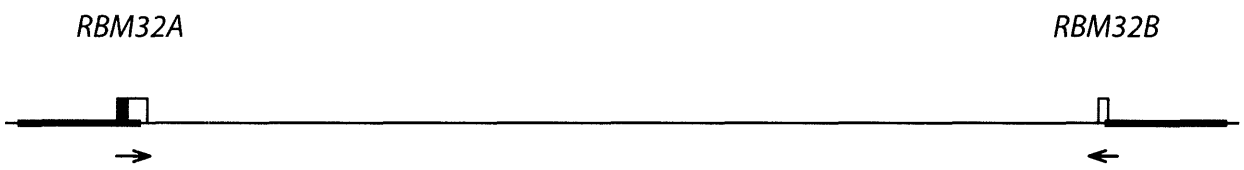

IKBKG
pseudogene
CXorf52
CTAG1
CTAG1
CXorf52
IKBKG

Figure 5. P12 genes start transcription in the arms of P12 and end in the spacer.

A. *RBM32A* and *RBM32B* transcripts begin in the arms and end in the spacer of P12. Thick lines represent palindrome arms. Open boxes are untranslated regions, shaded boxes are predicted translated regions. Red arrows give direction of transcription. Gene names appear above the genes. B. *RBM32A* predicted amino acid sequence aligned to *PABPC1*. *RBM32A* shares homology with poly(A) binding protein, cytoplasmic 1 in the RNA binding domains, but does not contain the consensus c-terminal poly(A) binding domain. RRM domains are boxed in red. C-terminal Poly(A) binding domain is boxed in blue. C. ClustalW nucleotide alignment of *RBM32I* and *RBM32B*. The two transcripts are highly similar at their 5' ends. The complete ORFs of both gene are unknown. Underlined bases mark inner boundaries of palindrome. Dashes represent gaps. See accompanying CD or http://staffa.wi.mit.edu/page/saionz/Figures/Figure_5C.txt

A.



B.

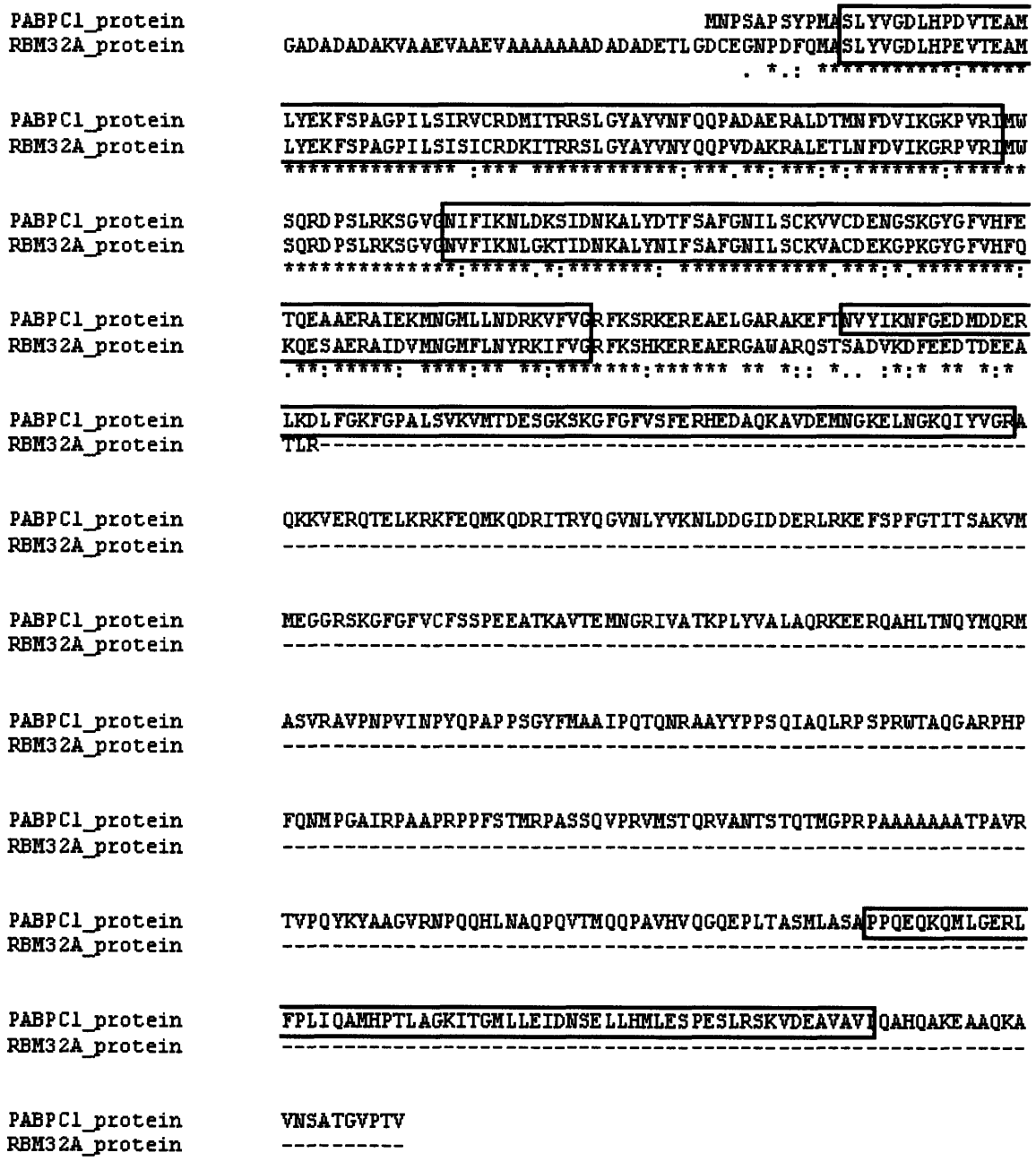
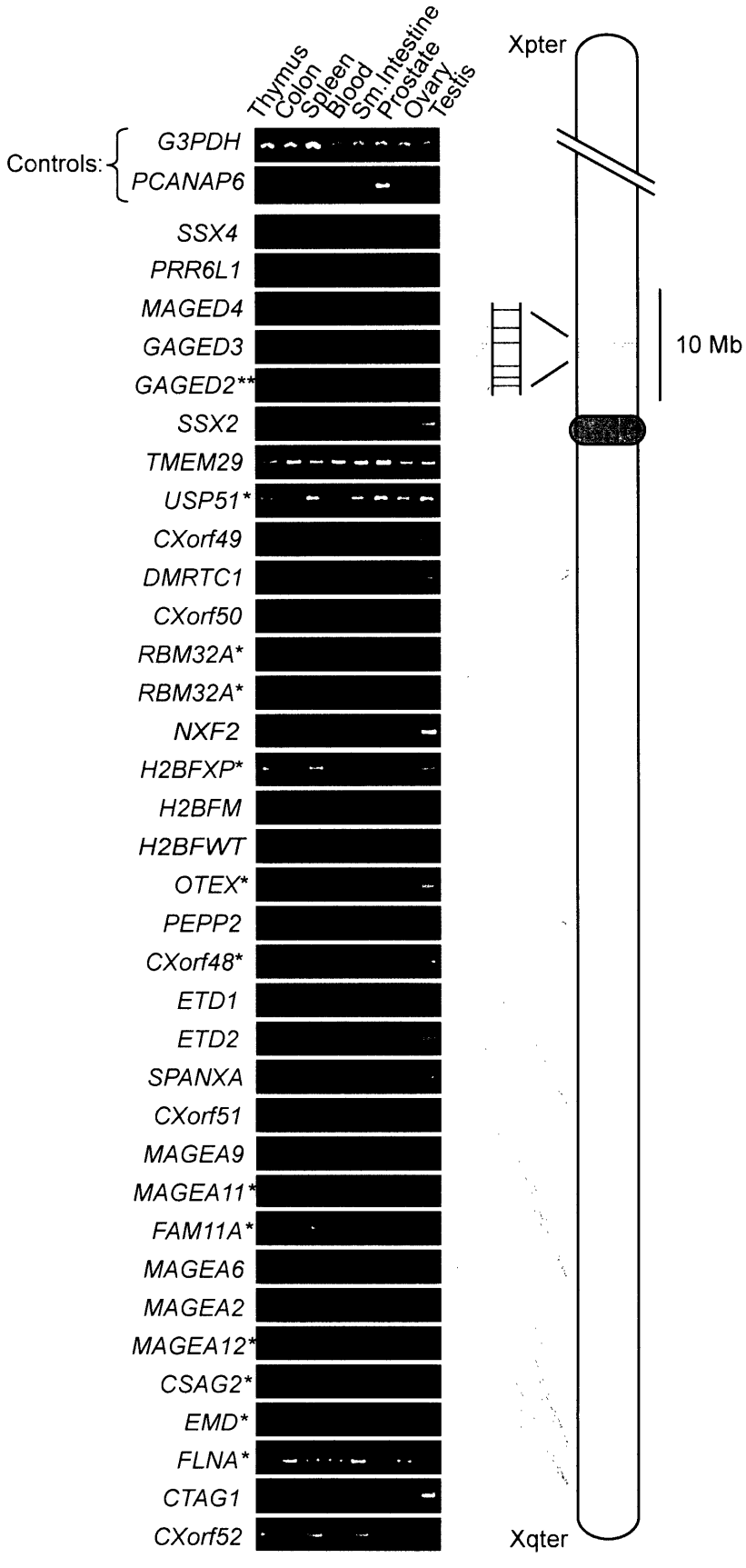

PABPC1_protein
RBM32A_protein
MNPSAPSYPMASLYVGDLHPDVTEAM
GADADADAKVAAEVAAEVAAAAAAADADADETLGDCEGNPDFQMASLYVGDLHPEVTEAM
LYEKFSPAGPILSIRVCRDMITRRSLGYAYVNFQQPADAERALDTMNFDVIKGKPVRIMW
LYEKFSPAGPILSISICRDKITRRSLGYAYVNYQQPVDAKRALETLNFDVIKGRPVRIMW
SQRDPSLRKSGVGNIFIKNLDKSIDNKALYDTFSAFGNILSCKVVCDENGSKGYGFVHFE
SQRDPSLRKSGVGNVFIKNLGKTIDNKALYNIFSAFGNILSCKVACDEKGPKGYGFVHFQ
TQEAAERAIEKMNGMLLNDRKVFVGRFKSRKEREAELGARAKEFTNVYIKNFGEDMDDER
KQESAERAIDVMNGMFLNYRKIFVGRFKSHKEREAERGAWARQSTSADVKDFEEDTDEEA
LKDLFGKFGPALSVKVMTDESGKSKGFGFVSFERHEDAQKAVDEMNGKELNGKQIYVGRA
TLR
QKKVERQTELKRKFEQMKQDRITRYQGVNLYVKNLDDGIDDERLRKEFSPFGTITSAKVM
MEGGRSKGFGFVCFSSPEEATKAVTEMNGRIVATKPLYVALAQRKEERQAHLTNQYMQRM
ASVRAVPNPVINPYQPAPPSGYFMAAIPQTQNRAAYYPPSQIAQLRPSPRWTAQGARPHP
FQNMPGAIRPAAPRPPFSTMRPASSQVPRVMSTQRVANTSTQTMGPRPAAAAAAATPAVR
TVPQYKYAAGVRNPQQHLNAQPQVTMQQPAVHVQGQEPLTASMLASAPPQEQKQMLGERL
FPLIQAMHPTLAGKITGMLLEIDNSELLHMLESPESLRSKVDEAVAVIQAHQAKEAAQKA
VNSATGVPTV

Figure 6. Palindrome genes are transcribed predominantly in the testis. On the right is a schematic of the X chromosome with palindrome approximate locations marked as a line on the chromosome. The centromere is represented by a filled oval. On the left are RT-PCR results for all palindrome genes across a panel of tissues. Genes are present in both arms of palindromes except when marked by an asterisk. Positive control for the RT-PCR is *G3PDH* and negative control is a prostate specific gene, *PCANAP6* (Xu et al. 2001). *Genes are present in single copy in the spacer of palindromes. **GAGED2 is in duplicate copies in P4, in triplicate copies in P5. PCR conditions were 94°C 30 seconds, 60°C 30 seconds, 72°C 1 minute for 30-40 cycles.


Thymus
Colon
Spleen
Blood
Sm.Intestine
Prostate
Ovary
Testis
Controls:
G3PDH
PCANAP6
SSX4
PRR6L1
MAGED4
GAGED3
GAGED2**
SSX2
TMEM29
USP51*
CXorf49
DMRTC1
CXorf50
RBM32A*
RBM32A*
NXF2
H2BFXP*
H2BFM
H2BFWT
OTEX*
PEPP2
CXorf48*
ETD1
ETD2
SPANXA
CXorf51
MAGEA9
MAGEA11*
FAM11A*
MAGEA6
MAGEA2
MAGEA12*
CSAG2*
EMD*
FLNA*
CTAG1
CXorf52
Xpter
10 Mb
Xqter

gene expression as a testis-predominant gene.) The 42 testis-predominant genes are spread out between 16 of the 24 palindromes in both arms and spacers. Only two palindromes, P14 and P21, contain their sole testis genes within their spacers.

The X chromosome has been suggested by some to contain more male benefit genes than expected by chance (Lercher et al. 2003; Rice 1984), especially male genes with premeiotic function (Khil et al. 2004; Wang et al. 2001). To determine if the large number of testis-predominant genes in palindromes is due simply to their linkage to a chromosome enriched for testis genes, we tested the number of testis-predominant genes on the X chromosome and autosomes with gene expression profiling data available from the GEO database (www.ncbi.nlm.nih/geo). We utilized the data set from a study by Su *et al.* (2004) that profiled gene expression with Affymetrix arrays to describe the tissue distribution of over 44,000 transcripts across 79 tissues and cell types. In analyzing the Su *et al.* data, we used only data obtained from the 67 normal tissues and cell types. We averaged the testis, germ cell testis and seminiferous tubule samples for each probe, then designated a gene testis-predominant when the expression level rank of the averaged testis was higher than any other tissue. For genes on the chip, the enrichment of testis genes on the non-palindrome X is significant when compared to the autosomes, but marginally so ($p< 1.7\times10^{-2}$, Fisher's exact test, two-sided, Table 3). When palindromes are included in the comparison of X to autosomes, the strength of the bias is increased. For X palindrome genes on the chip, 43.5% (10/23) are testis-predominant, with a significant enrichment of palindrome testis genes relative to the 3.4% (40/1177) non-palindrome X testis-genes ($p< 4.7\times10^{-9}$, Fisher's exact test, two-sided).

Table 3. X chromosome palindromes are enriched for testis genes.

	Autosomes	X chromosome (total)	X chromosome (no palindromes)	X palindromes
Testis genes	722	50	40	10
All genes	31794	1200	1177	23
% Testis genes	2.3%	4.2%	3.4%	43.5%

Two-sided exact Fisher's test p-value:

Autosomes vs. X chromosome (total): 8.8×10^{-5}

Autosomes vs. X chromosome (no palindromes): 1.7×10^{-2}

X chromosome (no palindromes) vs. X palindromes: 4.7×10^{-9}

Palindromes on the X chromosome have been conserved over 25 million years

Despite the recent duplication time implied by high levels of nucleotide identity between the arms of palindromes, palindromes on the Y chromosome have their origins before the diversification of the great apes (Rozen et al. 2003). We sought to determine if the X palindromes are also older than their paired-arm identities imply by pursuing a comparative sequencing strategy to trace the evolution of X palindromes. We searched for orthologous palindromes in the chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus*) and rhesus monkey (*Macaca mulatta*) by hybridization against BAC libraries from each of these species for clones homologous to sequence in the arms and spacers of eight human X palindromes (Table 4). We identified and then sequenced clones corresponding to seven of the palindromes (P6, P8, P13, P16, P17, P22 and P24), although we were not successful in obtaining clones for all seven palindromes in orangutan and rhesus. Discovery of six orthologous palindromes in chimpanzee, four in orangutan and three in rhesus monkey provides strong evidence that palindromes have been conserved on the X chromosome since, at least, the divergence of apes and old world monkeys 25 million years ago.

Nucleotide substitution rates of primate palindromes reveal concerted evolution

The paired arms in each of the primate palindromes in the present study exhibit arm-to-arm nucleotide identity greater than 99% (Table 5). The high paired arm identity in all orthologous palindromes indicates that concerted evolution has maintained these palindromes within each of the separate primate lineages.

Observation of a reduction in rate of nucleotide evolution of palindrome arms relative to nearby single-copy sequence in the Y palindromes of human and chimpanzee

Table 4. Orthologous sequence in primates generated for X chromosome palindromes. Accession numbers refer to the BACs sequenced in each species. Entries in parentheses do not contain definitive evidence of palindromes. P9 clones were not obtained in any library despite PCR amplification of P9 boundary sequence in chimpanzee. However, all libraries used contained about five-fold coverage of the X chromosome, leaving potential gaps in coverage. For P16 we obtained sequence in the orthologous region in chimpanzee, orangutan, and rhesus monkey, but could not detect a palindrome in the sequence. Sequence obtained from rhesus monkey BACs homologous to P17 aligned to only a single arm of the human sequence preventing ascertainment of palindrome existence. We did not obtain any homologous sequence in orangutan for P24. Previous results from southern blot hybridizations (Aradhya et al. 2001) did not provide evidence of sequence duplication in orangutan. Orthologous sequence in rhesus monkey contains duplicate copies of the P24 repeats, but orientation is uncertain. The duplication of P24 orthologous sequence in rhesus monkey suggests that P24 was secondarily lost in orangutan; however, it is unknown whether the ancestral duplication was in tandem or inverted orientation.

	Chimpanzee	Orangutan	Rhesus Monkey
P6:	AC146278	—	—
P8:	AC142343, AC145690	AC146951, AC146314	AC146528, AC146352
P9:	—	—	—
P13:	AC146276, AC144386, AC146277	AC146531	AC146491, AC148182
P16:	(AC145687, AC142344)	(AC146356)	(AC146353, AC148184)
P17:	AC146267, AC144383	AC146530, AC146843	(AC146489)
P22:	AC145689, AC144384	AC146919, AC148185	AC146490, AC146354, AC148183
P24:	AC144385, AC145688	—	(AC146529, AC146312)

Table 5. Palindromes undergo concerted evolution. All of the palindromes in the present study display greater than 99% identity between paired arms. Parentheses surrounding arm lengths and paired identity measurements indicate that the calculations are based on the available sequence, but that external boundaries were not included within the sequenced BACs.

	Species	Arm (kb)	Spacer (kb)	Identity (%)
P6	Human	59.9	0.2	99.92
	Chimpanzee	(10.3)	1.4	(100)
P8	Human	26.6	12.9	99.98
	Chimpanzee	28.6	10.7	99.98
	Orangutan	28.7	10.6	99.98
	Rhesus Monkey	(22.8)	11.7	(99.91)
P13	Human	138.6	10.8	99.96
	Chimpanzee	160.2	5.7	99.97
	Orangutan	(4.6)	10.8	(100)
	Rhesus Monkey	(33.7)	8.4	(99.99)
P17	Human	43.5	53.2	99.94
	Chimpanzee	41.9	56.2	99.96
	Orangutan	(32.3)	59.1	(99.30)
P22	Human	51.2	8.0	99.89
	Chimpanzee	44.3	19.4	99.86
	Orangutan	53.5	7.4	99.93
	Rhesus Monkey	21.96	72.5	99.95
P24	Human	35.5	21.8	99.95
	Chimpanzee	35.4	21.8	99.98

(Rozen et al. 2003) prompted us to examine the inter-specific nucleotide divergence in the X palindromes. The X palindromes in this study are as likely to exhibit higher nucleotide divergence between orthologous arms compared to nearby sequence as lower nucleotide divergence (Table 6). This result suggests that the palindromes in the present study are evolving, as a group, at approximately the same rate as the single-copy sequence nearby.

Inner boundaries of palindromes are frequently rearranged in different species

Some rearrangements of palindrome boundaries can be observed when comparing the different species. The rearrangements appear as small inversions, insertions or deletions across the boundaries. Most of the outer boundaries are conserved (Table 7). All five palindromes in the chimpanzee for which we have sequence covering both outer boundaries have the same outer boundaries as the orthologous human palindromes. In contrast to the conserved outer boundaries, most of the inner boundaries of the palindromes in this study are characterized by some degree of rearrangement of the sequence. Only two out of the six chimpanzee palindromes and none of the orangutan or rhesus monkey palindromes have inner boundaries conserved with orthologous human palindromes. In this data set, when a boundary sequence is rearranged in one species relative to the orthologous human sequence, that rearrangement is species-specific.

The rhesus P22 palindrome is extensively rearranged relative to its orthologs in human, chimpanzee and orangutan. The human P22 contains genes in both arms and spacer. The rhesus monkey P22 has no genes in the arms, but the spacer contains genes

Table 6. Nucleotide divergences of primate palindromes. Percent divergence in the same palindrome gives percent nucleotide divergence between paired arms of a palindrome within a species. Percent divergence from human gives percent nucleotide divergence between human and orthologous palindrome arms in another primate. Percent divergence from human outside palindrome gives the percent nucleotide divergence between human and another primate in nearby single copy sequence. Confidence intervals provided in parentheses. For sequence alignments see accompanying CD or http://staffa.wi.mit.edu/page/saionz/Alignments/index.html.

% divergence same palindrome arm-to-arm		% divergence from human arm to arm	% divergence from human outside palindrome	2-sided p-value
P8 human 0.02% (0.0-0.04)	chimp 0.03% (0.01-0.05)	0.63 (0.53-0.72)	1.07 (1.01-1.14)	6.22E-11
	orang 0.02% (0.00-0.04)	2.21 (2.01-2.40)	3.02 (2.92-3/12)	4.96E-11
	rhesus 0.09% (0.06-0.14)	4.79 (4.46-5.13)	5.90 (5.62-6.18)	7.64E-06
P13 human 0.05% (0.04-0.06)	chimp 0.03% (0.03-0.04)	0.79 (0.75-0.84)	1.11 (1.03-1.19)	8.75E-12
P17 human 0.06% (0.04-0.09)	chimp 0.04% (0.02-0.06)	0.98 (0.89-1.08)	0.73 (0.68-0.78)	1.33E-15
	orang 0.71% (0.62-0.80)	3.29 (3.10-3.49)	2.52 (2.40-2.64)	5.50E-12
P22 human 0.11% (0.08-0.14)	chimp 0.14% (0.10-0.17)	1.19 (1.09-1.30)	0.91 (0.86-0.95)	2.23E-08
	orang 0.07% (0.04-0.09)	3.03 (2.88-3.18)	2.82 (2.71-2.92)	2.24E-02
	rhesus 0.05% (0.02-0.09)	6.21 (5.85-6.57)	4.94 (4.73-5.14)	4.38E-10
P24 human 0.05% (0.03-0.07)	chimp 0.02% (0.00-0.03)	1.10 (0.99-1.21)	0.99 (0.94-1.04)	2.60E-02

Table 7. External palindrome boundaries are more conserved than internal palindrome boundaries. For each boundaries that there is sequence available the orthologous palindromes were compared. Conserved boundaries are marked with a plus sign (+) while non-conserved boundaries are marked with a minus sign (-). Comparisons were made for all available species for each boundary.

	Outer boundary		Inner boundary	
P6:			chimpanzee	-
P8:	chimpanzee	+	chimpanzee	-
	orangutan	+	orangutan	-
	rhesus monkey	+	rhesus monkey	-
P13:	chimpanzee	+	chimpanzee	-
			rhesus monkey	-
P17:	chimpanzee	+	chimpanzee	+
			orangutan	-
P22:	chimpanzee	+	chimpanzee	-
	orangutan	+	orangutan	-
	rhesus monkey	-	rhesus monkey	-
P24:	chimpanzee	+	chimpanzee	+

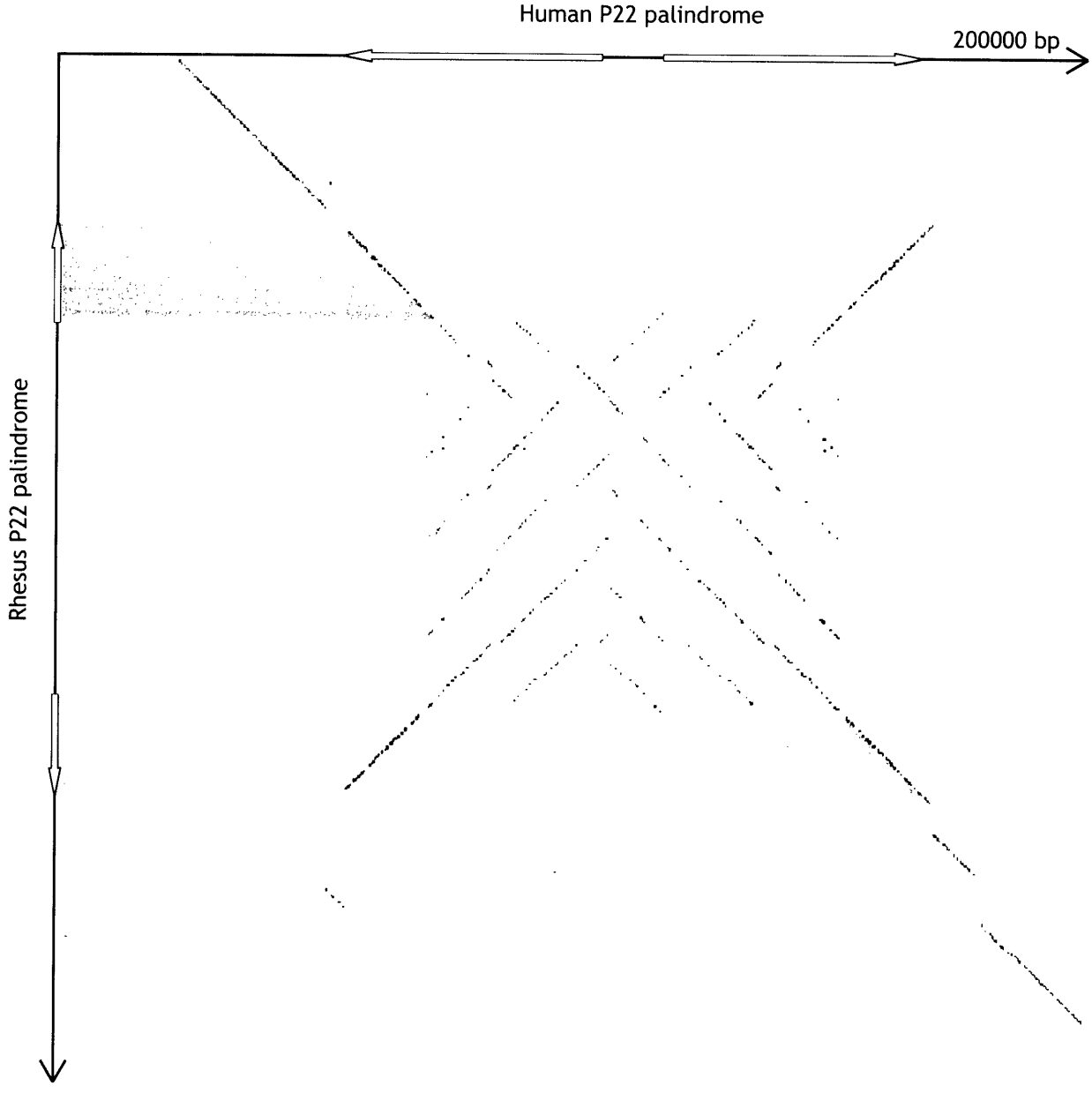
homologous to the human P22 arm and spacer genes (Figure 7). Interestingly, the orangutan and the chimpanzee P22s maintained the same gene order that as the human P22, yet the chimpanzee P22 underwent an internal gene conversion event homogenizing a spacer and arm gene (Figure 7c and 7d).

Insertions and deletions in palindrome arms are homogenized between arms

We characterized all of the insertions and deletions over 120 bp, aside from simple repeats, in the arms of orthologous palindromes. We uncovered several examples of species-specific identical insertions and deletions in both arms of a palindrome (Figure 8). We also observed instances of species-specific insertions in one arm, but not in the other arm of the same palindrome. We distinguished insertions from deletions by comparisons with the other species in the study. Only one indel was ambiguous as to whether it is an insertion or deletion. Analysis with RepeatMasker of inserted and deleted sequences in palindromes revealed that all likely insertions were generated by Alu or LINE element retroposition events, while likely deletions show no association with repetitive elements. The association of Alu and LINE elements with indels in palindromes is consistent with observations made on the chimpanzee chromosome 22 sequence (Watanabe et al. 2004). We observed insertions as large as 0.6 kb homogenized to both arms of a palindrome, while homogenized deletions were as large as 14.6 kb. The three examples of insertions in only a single palindrome arm were between 0.7 to 6.1 kb. In two of these instances, the palindrome arm-to-arm substitution rate remains above 99.9%. In the cases where the insertions and deletions are in both arms of the palindromes, the indel event must have occurred in one arm and through gene conversion was homogenized to the other palindrome arm. That there are more instances of

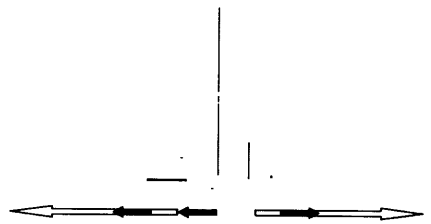
Figure 7. Comparisons between human and rhesus monkey P22 palindromes are characterized by extensive rearrangement. A. Dot plot of human and rhesus monkey P22 palindromes. Each dot represents a match of 100% within a sliding window of 40 bp with 20 bp steps. Shaded regions represent palindrome arms. Line arrows indicate 200 kb of genomic sequence including arms for each species. Open boxed arrows indicate palindrome arms. B. Alignment of MAGE predicted transcripts from rhesus monkey P22 ortholog. Start and stop sites are underlined. Dots denotes same base as the first sequence; dashes denote gaps. See accompanying CD or http://staffa.wi.mit.edu/page/saionz/Figures/Figure_7B.txt C. Triangular dot plot of chimpanzee P22 orthologous sequence. Each dot represents a match of 100% within a sliding window of 200 bp with 100 bp steps. Inverted repeats appear as vertical lines and direct repeats as horizontal lines. D. Homogenization of *CSAG* homologs in chimpanzee P22 palindrome. Percent identity plot of generated from alignment of sequence containing *CSAG2* homologs from chimpanzee P22 arm and spacer. Plot is based on a 1000 bp sliding window scale with 1 bp steps. The Y-axis represents the percent nucleotide identity between the sequences. The X-axis measures the length of the paired arm alignment in kilobases.

A.



C.

w= 200 s= 100



333232 bp

D.

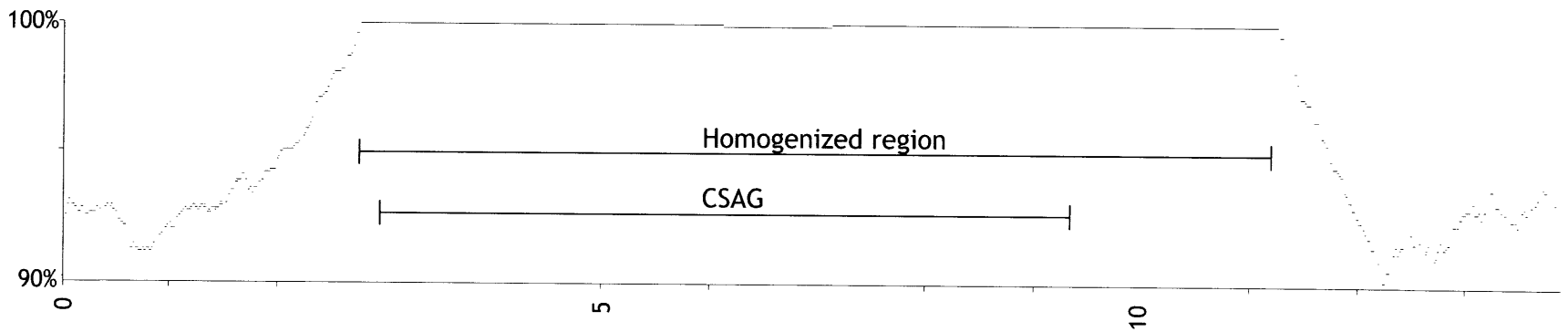
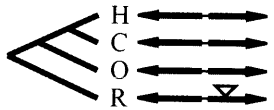
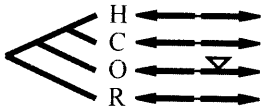


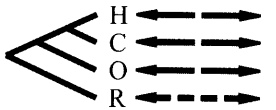
Figure 8. Insertion and deletions in palindromes. Insertions and deletions greater than 120 bp (not including simple repeats) from orthologous palindromes can be differentiated by scoring their presence or absence on phylogenetic trees. Completely sequenced palindromes represented as two inverted arrows. Orangutan and rhesus monkey P13s are incompletely sequenced and represented without the arrow heads. Rhesus monkey sequence orthologous to P17 includes a single arm. Rhesus monkey P22 palindrome is gray in instances where palindrome in rhesus does not include sequence with the indel in question. H, human; C, chimpanzee; O, orangutan; R, rhesus monkey



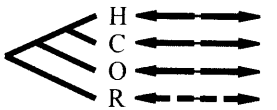
714 bp insertion in single arm of rhesus monkey P8.
Contains 595 bp of inverted L1PA5 LINE 3' end
7.4% diverged from consensus



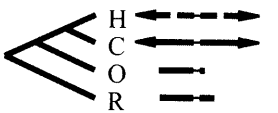
6138 bp insertion in single arm of orangutan P8.
Contains 6076 pb of L1PA3 LINE 3' end
3.4% diverged from consensus



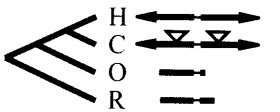
3162 bp deletion in both arms of
rhesus monkey P8



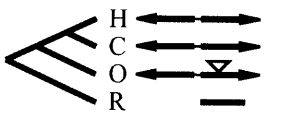
132 bp deletion in both arms of
rhesus monkey P8



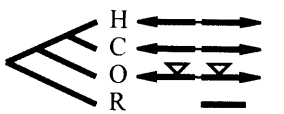
14555 bp deletion in both arms of chimpanzee P13
See supplementary figure 9



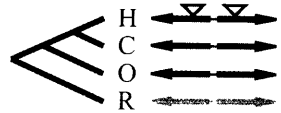
348 bp insertion in both arms of chimpanzee P13
Contains 300 bp of inverted L1PA2 LINE 3' end
1.6% diverged from consensus



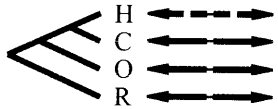
4181 bp insertion in single arm of orangutan P17
Contains 4105 bp of L1PA3 LINE 3' end
2.9% diverged from consensus



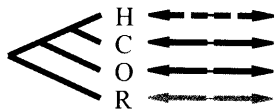
195 bp insertion in both arms of orangutan P17
Contains 140 bp of L1PA3 LINE 3' end
5.7% diverged from consensus



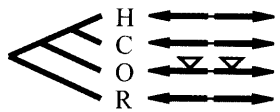
337 bp insertion in both arms of human P22
Contains complete AluYa5 (311bp)
0% diverged from consensus



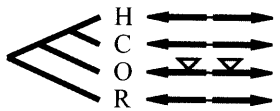
126 bp deletion in both arms of
human P22



235 bp deletion in both arms of human P22

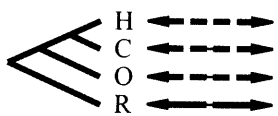


556 bp insertion in both arms of orangutan P22
Contains 432 pb of L1PA3 LINE 3' end
3.6% diverged from consensus



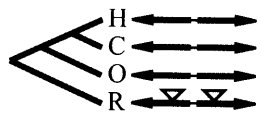
469 bp insertion in both arms of orangutan P22
Contains 454 bp of L1PA3 LINE 3' end
2.6% diverged from consensus

142 bp deletion
in hominid lineage



or

142 bp insertion
in rhesus P22



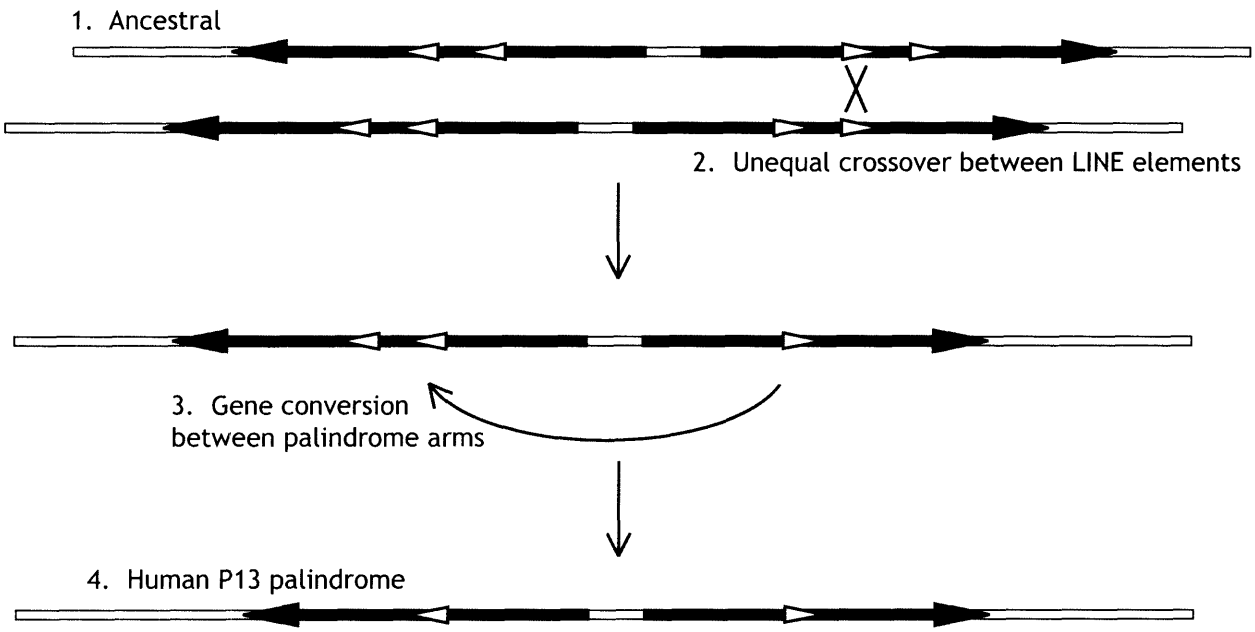
homogenized indels than single-arm indels suggests that the gene conversion rate is higher than the rate of insertion or deletion, assuming no bias in the rates of fixation.

Interestingly, we can infer the origin of the 14.6 kb deletion in the arm of human P13 (Figure 9). Analysis with RepeatMasker (Smit and Green) of the orthologous chimpanzee sequence and the human sequence surrounding the deletion reveals LINE sequences. The chimpanzee P13 arm contains two LINEs lying in tandem that are about 96% identical to each other. Recombination between the two LINEs after human diverged from chimpanzee would delete the intervening sequence to give the sequence in the human P13 arm. Alignment of the human LINE sequence with the chimpanzee LINEs reveals where the recombination event took place (Figure 9B). Presence of the deletion on both arms of human P13 suggests that gene conversion subsequent to the primary deletion homogenized the deletion to both arms.

DISCUSSION

At first glance the palindromes on the X and Y chromosomes appear similar, yet there are several significant differences between them. There are 24 palindromes on the X chromosome and only eight on the Y; however, close to 30% of the euchromatic Y is covered by palindromes (Skaletsky et al. 2003) and only 1.8% of the X. The arm lengths of the Y palindromes range between 9 kb to 1.5 Mb, with an average of 337 kb, while the X palindromes have an average arm length of 44 kb, ranging between 9 to 142 kb (P4 may be larger than calculated, see Supplementary figure S4). The enrichment for testis-predominant genes on the X is significant, but on the Y the bias is complete. The cause

Figure 9. Deletion within P13 arm in human ancestor by way of LINE-LINE recombination and gene conversion. A. A schematic drawing of the inferred deletion and homogenization events. Horizontal open triangles represent LINE elements. 1. Ancestral P13 palindrome. 2. Unequal crossing over between LINE elements. 3. Gene conversion between palindrome arms. 4. Human P13 palindrome. B. Alignment of chimpanzee LINE elements and human chimeric LINE element. See accompanying CD or http://staffa.wi.mit.edu/page/saionz/Figures/Figure_9B.txt


1. Ancestral
2. Unequal crossover between LINE elements
3. Gene conversion
between palindrome arms
4. Human P13 palindrome

of these differences may be the result of the differing selective pressures exerted on the two sex chromosomes.

Whereas the Y chromosome never recombines with a homolog during meiosis, the X chromosome does so during its passage through the female germline. Segmental duplications can provide substrates for non-allelic recombination that generates genomic disease (Ji et al. 2000; Lupski 1998). The high identity over generous stretches of sequence in palindromes should make these structures particularly prone to chromosomal mispairing and genomic disruption. The absence or reduction of recombination with a homologous chromosome that is unique to the sex chromosomes may have generated the enrichment of palindromes there relative to the autosomes (Warburton et al. 2004). However, recombination with a homolog in the female germline and the consequent risk of genomic disruption may have prevented the occurrence of palindromes on the X chromosome of the extraordinary size observed on the Y chromosome (Skaletsky et al. 2003).

The suppression of recombination between the ancestral X and Y chromosomes occurred in a stepwise fashion across the X chromosome. The modern human X is characterized by sequence strata ordered as they have most recently recombined with the Y chromosome (Lahn and Page 1999). We did not uncover any palindromes in the youngest strata on the short arm of the X. Suppression of recombination in these regions postdates the divergence of placental mammals from marsupials and monotremes. Perhaps the bias of palindromes to older parts of the X chromosome is due to the longer time the selective pressures that maintain palindromes on the non-recombining sex chromosomes have been in effect there.

We observed that palindromes are more frequently located near the centromere of the X chromosome on either arm. Both intrachromosomal and interchromosomal segmental duplications are more likely to be located near the centromere of chromosomes (Eichler 2001). Pericentromeric regions are characterized by reduced rates of recombination (Kong et al. 2002). Recombination repression was particularly well documented in a study of the human X chromosome (Mahtani and Willard 1998). Perhaps, reduction in allelic recombination rates near the centromere relaxes negative selection against palindromes and provides a haven for duplicated sequence. Centromere proximity may increase palindrome generation or preservation. While the Y chromosome does not recombine with an allelic chromosome, it may still be of note that the Y is sufficiently small that its entire euchromatic length is centromere proximal and all its palindromes are within 15 Mb of the centromere.

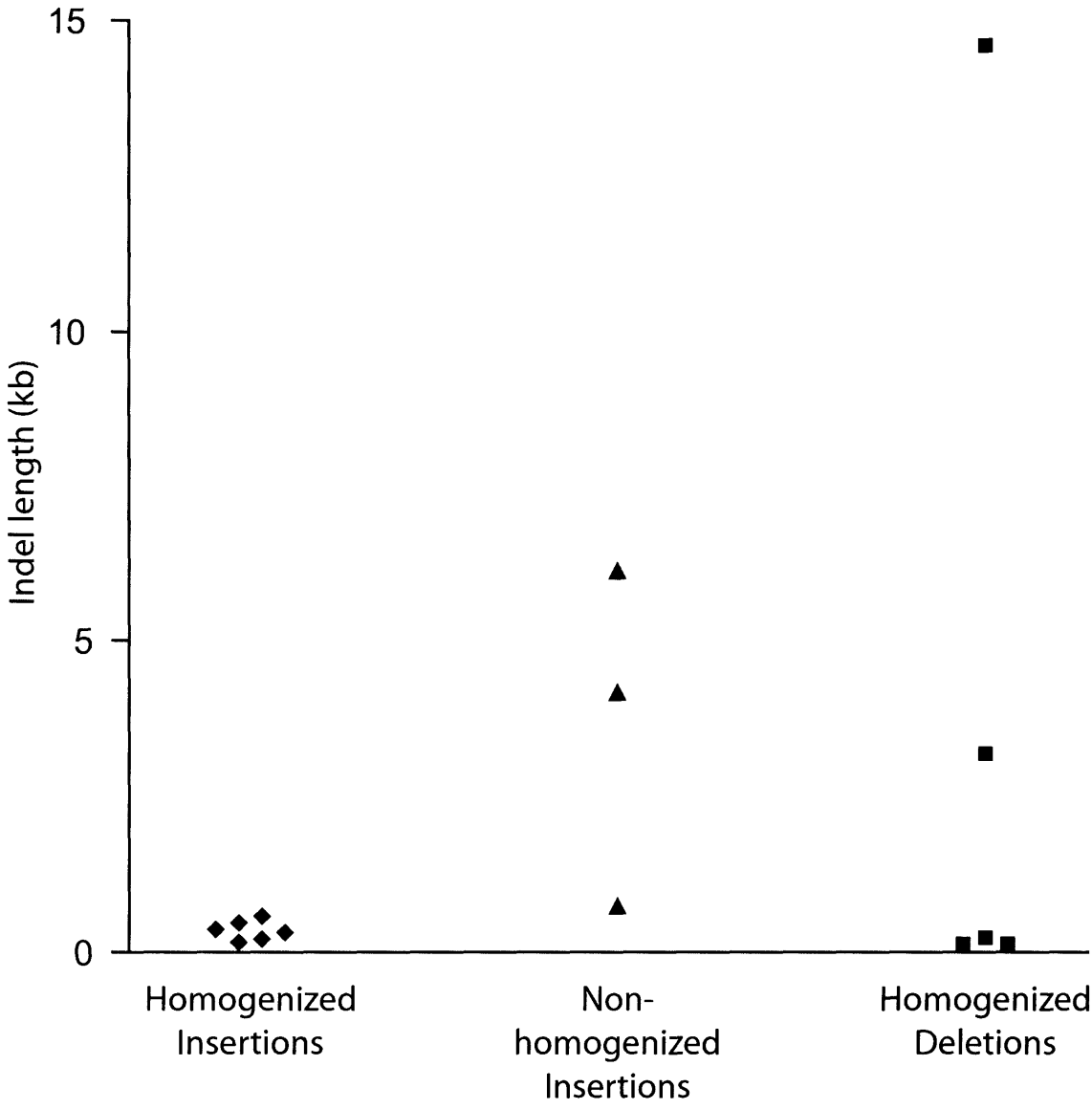
Our comparative study of orthologous palindromes provides evidence for concerted evolution of palindrome arms in all four primate lineages examined, including human. One of the most striking observations we made concerned the homogenization of insertions and deletions between palindrome arms. Presumably, the indels observed in both arms of a palindrome originated from an event occurring in a single arm later homogenized to the other by gene conversion. Transgenic studies in mammalian cell culture demonstrated that 1.5 kb sequence heterologies could be homogenized through gene conversion (Godwin and Liskay 1994). *Saccharomyces cerevisiae* has also been shown to be capable of gene conversion of a 5 kb insertion in meiotic cells (Kearney et al. 2001). We are not aware of any example in the published literature of gene conversion of an indel larger than the 14.6 kb deletion in P13.

Intriguingly, the insertions observed that were not homogenized to the opposing arm are all larger than the largest observed homogenized deletion (Figure 10). Furthermore, the homogenized deletions include examples that are as large or larger than the non-homogenized insertions. From our data, it would appear that homogenizing large deletions is more efficient than homogenizing large insertions. However, our sampling of indels is not large enough to draw conclusions from about the relative efficiencies of duplication versus deletion mechanisms, nor do we know the relative rates of insertion and deletion. However, our observations suggest that further study of insertions and deletions in palindrome arms might yield insight into the mechanisms of homogenization.

Most of the orthologous X palindromes have rearrangements at their inner boundaries and a few have rearrangements at their outer boundaries. The resolution of blocked recombination intermediates might generate rearrangement (Pavelitz et al. 1999). Heteroduplexes formed during intra-palindrome recombination should move along the DNA until heterologies are encountered at the boundary of the palindrome and recombination intermediates must be resolved. Perhaps inner boundary rearrangements are more common because the chromatin composition of the palindrome spacer influences heteroduplex resolution.

It has been suggested that the mammalian X chromosome is a haven for male-advantage genes (Hurst and Randerson 1999; Khil et al. 2004; Rice 1984; Wang et al. 2001). Our analysis of the Su *et al.* (2004) gene expression data provides additional evidence that the X chromosome is statistically enriched for testis genes relative to the autosomes. It must be remembered that despite the overrepresentation of testis genes on the X, less than 5% of the chromosome's genes are testis genes.

Figure 10. Arm-to-arm homogenization of insertions and deletions. All indels between orthologous palindrome arms over 100 bp composed of complex sequence were classified as insertions or deletions (see Figure 8) and as homogenized (present in both palindrome arms) or non-homogenized (present in only a single arm). No non-homogenized deletions were observed. The sizes of these indels are shown in base pairs.


15
10
5
0
Indel length (kb)
Homogenized Insertions
Non-homogenized Insertions
Homogenized Deletions

Approximately a quarter of X chromosome testis genes are found in palindromes and the vast majority of Y chromosome testis genes are. Perhaps palindromes serve to compartmentalize testis-genes. Palindromes may afford a more stable duplication structure than tandem repeats to facilitate increased gene dosage, while concerted evolution prevents divergence of gene copies. Palindromes may protect genes from mutagenesis due to incorrectly repaired double-strand breaks (Rozen et al. 2003). In yeast, regions of active transcription preferentially receive double strand breaks during meiosis due to their open chromatin conformation (Wu and Lichten 1994). Genes transcribed in the testis may be prone to double strand breaks due to the temporal proximity of their transcription to meiosis. However, mutations accrued in either meiotic or mitotic germ cells would necessarily be transmitted to the next generation. Homology directed double-strand break repair in palindrome arms could utilize the opposing arm as a pairing partner rather than engage in more mutagenic non-homologous end-joining.

The present study provides substantial gains in current understandings of duplicated sequence in the human genome. We have demonstrated that palindromes are part of a specialized class of segmental duplications characterized by concerted evolution and frequent gene conversion. Our efforts to carefully annotate all of the genes in human X palindromes allowed us to empirically describe the gene expression of palindrome genes and to draw a well-supported conclusion that a testis transcription bias exists within these palindromes. Most investigations into links between nucleotide sequence and its function are aimed at small sequence motifs. The propensity for testis predominant transcription within palindromes is an interesting observation of an association between a large scale genomic structure and its function.

METHODS

Palindrome discovery

X chromosome sequence from the July 2003 NCBI build 34 was downloaded. Palindrome discovery was performed using custom Perl code. This code used BLAST (http://blast.wustl.edu) to compare to itself 400 kb sequence segments, in 200 kb steps. High scoring results were subjected to dot plot analyses and sequence analysis as previously described (Kuroda-Kawaguchi et al. 2001).

Palindrome boundary determination

Palindrome arm boundaries were determined by prediction of regions most likely to have engaged in gene conversion recently and to be capable of gene conversion in the future. Percent nucleotide identity plots using a 1000 bp window and 1 bp steps were constructed from alignment of palindrome arms and contiguous sequence. The abrupt increase to ~0.1% nucleotide difference demarcates the palindrome boundaries.

Gene annotation

Known genes residing in palindromes were cataloged by searching the NCBI annotation. BLAST searches (Altschul et al. 1990) were performed against the NCBI human EST database with masked sequence (Smit and Green). Positive hits consisted of multiple ESTs or single ESTs that were spliced or contained poly(A) tails that did not align to genomic sequence. ESTs were aligned *in silico* against the palindrome sequence to elucidate gene structure, including its orientation and a putative open reading frame. Where multiple ESTs aligned to a putative gene, corresponding IMAGE clones were resequenced. PCR primers were designed manually and utilizing Primer3 (Rozen and

Skaletsky 2000) to amplify the putative genes and confirm transcription from cDNA pooled from a range of tissues.

Expression analysis

To determine the range of tissue expression for the palindrome genes we PCR amplified the palindrome genes from the Human Multiple Tissue cDNA panel (Clontech).

Identification and sequencing of chimpanzee, orangutan and rhesus monkey BACs

We screened high-density filters from the CHORI-251 male chimpanzee, CHORI-253 male orangutan and CHORI-250 male rhesus monkey BAC libraries (BACPAC resources) using hybridization probes designed to detect sequences in the arms and spacers of palindromes P6, P8, P9, P13, P16, P17, P22 and P24. PCR primers designed from orthologous human sequence confirmed that, among the candidate chimpanzee BACs identified by hybridization, 15 clones contained sequence orthologous to the human X palindromes. Further hybridizations with probes designed from genes within orthologous human palindromes against candidate orangutan and rhesus BACs initially identified, and subsequent BAC-end sequencing confirmed, that 23 contained sequence orthologous to human X palindromes CHORI-251 BAC clones were sequenced as previously described (Lander et al. 2001). CHORI-253 and CHORI-250 BAC clones were sequenced either as described by Blakesley and colleagues (2004) with some further refinement (e.g., gaps filled) through customized sequencing reactions.

Sequence analysis

Insertions and deletions between palindromes were identified by aligning orthologous palindrome arms using ClustalW under default parameters (Thompson et al.

1994). Alignment gaps greater than 120 bp not composed of simple repeats were ascertained manually. RepeatMasker (Smit and Green) was used to determine repetitive content of the indels and surrounding sequence. Nucleotide divergences were determined as previously described (Kuroda-Kawaguchi et al. 2001).

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Aradhya, S., T. Bardaro, P. Galgoczy, T. Yamagata, T. Esposito, H. Patlan, A. Ciccodicola, A. Munnich, S. Kenwrick, M. Platzer, M. D'Urso, and D.L. Nelson. 2001. Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. *Hum Mol Genet* **10:** 2557-2567.

Bailey, J.A., A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005-1017.

Blakesley, R.W., N.F. Hansen, J.C. Mullikin, P.J. Thomas, J.C. McDowell, B. Maskeri, A.C. Young, B. Benjamin, S.Y. Brooks, B.I. Coleman, J. Gupta, S.L. Ho, E.M. Karlins, Q.L. Maduro, S. Stantripop, C. Tsurgeon, J.L. Vogt, M.A. Walker, C.A. Masiello, X. Guan, G.G. Bouffard, and E.D. Green. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res* **14:** 2235-2244.

DeSilva, U., H. Massa, B.J. Trask, and E.D. Green. 1999. Comparative mapping of the region of human chromosome 7 deleted in williams syndrome. *Genome Res* **9:** 428-436.

Duan, Z., A.J. Feller, H.C. Toh, T. Makastorsis, and M.V. Seiden. 1999. TRAG-3, a novel gene, isolated from a taxol-resistant ovarian carcinoma cell line. *Gene* **229:** 75-81.

Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17:** 661-669.

Feller, A.J., Z. Duan, R. Penson, H.C. Toh, and M.V. Seiden. 2000. TRAG-3, a novel cancer/testis antigen, is overexpressed in the majority of melanoma cell lines and malignant melanoma. *Anticancer Res* **20:** 4147-4151.

Godwin, A.R. and R.M. Liskay. 1994. The effects of insertions on mammalian intrachromosomal recombination. *Genetics* **136:** 607-617.

Gonzalez, I.L. and J.E. Sylvester. 2001. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73:** 255-263.

Hurst, L.D. and J.P. Randerson. 1999. An eXceptional chromosome. *Trends Genet* **15:** 383-385.

Ji, Y., E.E. Eichler, S. Schwartz, and R.D. Nicholls. 2000. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* **10:** 597-610.

Kearney, H.M., D.T. Kirkpatrick, J.L. Gerton, and T.D. Petes. 2001. Meiotic recombination involving heterozygous large insertions in Saccharomyces cerevisiae: formation and repair of large, unpaired DNA loops. *Genetics* **158:** 1457-1476.

Khil, P.P., N.A. Smirnova, P.J. Romanienko, and R.D. Camerini-Otero. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat Genet* **36:** 642-646.

Kong, A., D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, and K. Stefansson. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31:** 241-247.

Kuroda-Kawaguchi, T., H. Skaletsky, L.G. Brown, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, S. Silber, R. Oates, S. Rozen, and D.C. Page. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat Genet* **29:** 279-286.

Lahn, B.T. and D.C. Page. 1999. Four evolutionary strata on the human X chromosome. *Science* **286:** 964-967.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg

L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan J. Szustakowki P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol Biol Evol* **20:** 1113-1116.

Liao, D., T. Pavelitz, J.R. Kidd, K.K. Kidd, and A.M. Weiner. 1997. Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. *Embo J* **16:** 588-598.

Lupski, J.R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14:** 417-422.

Lynch, M. and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151-1155.

Mahtani, M.M. and H.F. Willard. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res* **8:** 100-110.

Nathans, J., T.P. Piantanida, R.L. Eddy, T.B. Shows, and D.S. Hogness. 1986. Molecular genetics of inherited variation in human color vision. *Science* **232:** 203-210.

Ohno, S. 1970. *Evolution by gene duplication.* Springer-Verlag, New York.

Paques, F. and J.E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* **63:** 349-404.

Park, S.S., P. Stankiewicz, W. Bi, C. Shaw, J. Lehoczky, K. Dewar, B. Birren, and J.R. Lupski. 2002. Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs. *Genome Res* **12:** 729-738.

Pavelitz, T., D. Liao, and A.M. Weiner. 1999. Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. *Embo J* **18:** 3783-3792.

Pavelitz, T., L. Rusche, A.G. Matera, J.M. Scharf, and A.M. Weiner. 1995. Concerted evolution of the tandem array encoding primate U2 snRNA occurs in situ, without changing the cytological context of the RNU2 locus. *Embo J* **14:** 169-177.

Rice, W.R. 1984. Sex-chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735-742.

Rozen, S. and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365-386.

Rozen, S., H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, and D.C. Page. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423:** 873-876.

Shaikh, T.H., H. Kurahashi, S.C. Saitta, A.M. O'Hare, P. Hu, B.A. Roe, D.A. Driscoll, D.M. McDonald-McGinn, E.H. Zackai, M.L. Budarf, and B.S. Emanuel. 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* **9:** 489-501.

Skaletsky, H., T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K.

Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825-837.

Small, K., J. Iber, and S.T. Warren. 1997. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat Genet* **16:** 96-99.

Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.

Szostak, J.W., T.L. Orr-Weaver, R.J. Rothstein, and F.W. Stahl. 1983. The double-strand-break repair model for recombination. *Cell* **33:** 25-35.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Wang, P.J., J.R. McCarrey, F. Yang, and D.C. Page. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* **27:** 422-426.

Warburton, P.E., J. Giordano, F. Cheung, Y. Gelfand, and G. Benson. 2004. Inverted Repeat structure of the human genome: The X chromosome contains a preponderance of large highly homologous inverted repeats with contain testes genes. *Genome Res* **14:** 1861-1869.

Warburton, P.E., J.S. Waye, and H.F. Willard. 1993. Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin. *Mol Cell Biol* **13:** 6520-6529.

Watanabe, H., A. Fujiyama, M. Hattori, T.D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak, M. Kube, S. Taenzer, P. Galgoczy, M. Platzer, M. Scharfe, G. Nordsiek, H. Blocker, I. Hellmann, P. Khaitovich, S. Paabo, R. Reinhardt, H.J. Zheng, X.L. Zhang, G.F. Zhu, B.F. Wang, G. Fu, S.X. Ren, G.P. Zhao, Z. Chen, Y.S. Lee, J.E. Cheong, S.H. Choi, K.M. Wu, T.T. Liu, K.J. Hsiao, S.F. Tsai, C.G. Kim, O.O. S, T. Kitano, Y. Kohara, N. Saitou, H.S. Park, S.Y. Wang, M.L. Yaspo, and Y. Sakaki. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382-388.

Worton, R.G., J. Sutherland, J.E. Sylvester, H.F. Willard, S. Bodrug, I. Dube, C. Duff, V. Kean, P.N. Ray, and R.D. Schmickel. 1988. Human ribosomal RNA genes: orientation of the tandem array and conservation of the 5' end. *Science* **239:** 64-68.

Wu, T.C. and M. Lichten. 1994. Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* **263:** 515-518.

Xu, J., M. Kalos, J.A. Stolk, E.J. Zasloff, X. Zhang, R.L. Houghton, A.M. Filho, M. Nolasco, R. Badaro, and S.G. Reed. 2001. Identification and characterization of prostein, a novel prostate-specific protein. *Cancer Res* **61:** 1563-1568.

WEB SITE REFERENCES

Smit, A.F.A. and P. Green. http://repeatmasker.org "RepeatMasker".
http://blast.wustl.edu. "WU BLAST".
http://www.ncbi.nlm.nih.gov/geo/. "Gene Expression Omnibus (GEO) Homepage"

CHAPTER 3

Conclusions and further thoughts

Conclusions and further thoughts

The work presented in this thesis describes an association between a large-scale sequence structure and biological function. We characterized the palindrome content of the human X chromosome using bioinformatic approaches, experimentally determined the tissue expression pattern of the genes associated with those palindromes, and traced the evolution of several of the palindromes through comparative sequencing in primates. We found that 1.8% of the human X chromosome is organized in palindrome sequence structures. We found that all of the genes associated with palindromes are transcribed in the testis. Our studies of X palindromes in other primates demonstrated that palindromes do not represent recent duplications, but are conserved segmental duplications that display greater nucleotide identity than expected considering the age of duplication. Our comparative sequencing analysis demonstrated that gene conversion between palindrome arms occurs in all primate lineages examined. Previous work on the Y chromosome showed that gene conversion is ongoing between the massive palindrome arms in the human lineage and revealed a correlation between palindromes and testis specific genes (Skaletsky et al. 2003). A recent survey of the human genome indicated that palindromes are present throughout the genome, but are disproportionately located on the sex chromosomes (Warburton et al. 2004). All together, there is a clear association between palindrome sequence structures and testis transcription. This association suggests that palindromes themselves serve a functional role involving testis transcription. Because we have convincing evidence that gene conversion occurs in all palindromes examined, current thoughts on palindrome function have focused on the role gene conversion and

palindrome arm to arm recombination might play in testis biology. However, other potential functions may be provided by palindromes. Numerous observations we made in characterizing palindromes in humans and other primates suggest hypotheses that may be tested in the future.

Palindrome origins?

The particular spans of nucleotide sequences contained within orthologous palindrome arms differs between species. Inversions, deletions, or duplications of sequence at the inner or outer boundaries can change the composition of nucleotide sequence in a palindrome arm between species. For X-linked palindrome P22, the genes that are present in the arms of the human, chimpanzee and orangutan palindromes lie in the spacer of the rhesus monkey palindrome. The difference in gene organization in the rhesus monkey P22 may be the ancestral state. Alternatively, the difference might be a lineage specific event that could potentially be associated with palindrome death. That is, the rearrangement of palindrome genes from arm to spacer might indicate that rhesus monkey P22 no longer serves its purpose as a recombinational domain for testis genes. However, because the palindrome paired arm identity is greater than 99%, the imminent demise of rhesus monkey P22 is unlikely.

On the other hand, the absence of genes in the palindrome arms may tell us more about palindrome structural characteristics. Tracing the evolution of P22 in other primate lineages might inform us of the range of variation possible in the sequence composition in orthologous palindromes in different lineages. A complementary question is how much orthologous sequence is consistently maintained within the arms of orthologous

palindromes in different lineages. Characterization of palindromes in other species and mammalian lineages may tell us whether there are general characteristics (e.g., GC content) that are common to all palindromes and whether sequence composition changes after time when sequence resides in a palindrome conformation.

Why are there genes in the palindrome spacer?

The difference in gene organization between the rhesus monkey P22 and the human, chimpanzee and orangutan P22 genes, such that all of the rhesus monkey P22 genes lie in the spacer, argues against arm to arm recombination as the sole benefit provided to testis genes by their association with palindromes. Supporting evidence against recombination as the only function of palindromes is the finding that the human P8 palindrome is conserved in chimpanzee, orangutan and rhesus monkey. The sole gene associated with human P8 is located in the spacer, not in the arms of the palindrome. That gene is not testis specific, although, like all human palindrome genes, it is expressed within the testis. Thus, palindromes without genes in the arms to recombine with each other are also conserved.

Gene organization of other human X-linked palindromes, too, may be inconsistent with arm to arm recombination as the sole function of palindromes. There are 13 genes present in the spacers of human X palindromes. Three human X-linked palindromes have no arm genes at all, but do have spacer genes. Two human X-linked palindromes, one of the aforementioned and one other, have their only testis-predominant genes in their spacers. Of these, P8 is the only one for which we have sequence data confirming orthologous sequence in other species. Setting aside P8 as a potential exception, it is

possible that the other palindromes containing only spacer genes represent recent duplications that as such do not share the same function as the conserved palindromes. X-linked palindromes might fall into two categories, those that are conserved and provide a recombination function, and those that are recent and are only coincidental in their expression bias. In fact, Y palindromes do not contain genes within their spacers, thereby suggesting the possibility that no benefit is derived from a gene's association with Y palindromes unless the gene resides in the arms. It is also possible that the role played by palindromes on the Y chromosome differs from that played by palindromes on the X chromosome.

Recombination and palindromes

For the Y chromosome palindromes, there is a reduction in the rate of evolution between orthologous palindrome arms relative to surrounding single-copy sequences (Rozen et al. 2003). For the X-linked palindromes, where we looked at orthologous palindromes that included species further removed from humans than in the Y chromosome palindrome study, we did not see a consistent difference in nucleotide divergence between orthologous palindrome arms and surrounding sequence. On the Y chromosome as a whole, the rate of evolution is normally increased relative to the autosomes, while on the X chromosome, the rate of evolution is reduced relative to the autosomes (Miyata et al. 1987). We do observe evidence that recombination takes place between X chromosome palindrome arms. It may be that the rate of evolution is sufficiently low on the X already that the increased recombination in palindromes does not visibly affect it, whereas on the Y, the increase in recombination affects the rate of

nucleotide evolution significantly. The genes associated with the Y palindromes may derive the benefit of a reduced rate of nucleotide evolution due to the recombination between palindrome arms. On the X chromosome, other functions of palindromes may provide the benefits to palindrome associated genes to generate the selective advantage that maintains palindromes over evolutionary time.

Further experiments should be undertaken to obtain direct evidence of palindrome arm to arm recombination. Localization to palindromes of proteins with known involvement in recombination would be further evidence that palindrome arm to arm recombination is a normal event in the testis. Moreover, these types of experiments could resolve the issue of whether palindrome arm to arm recombination takes place in the mitotic or meiotic cells of the testis. A huge impediment to performing these experiments is obtaining sufficient tissue from human testis. A successful search in the mouse genome for palindromes associated with testis transcription or generation of mice transgenic for human palindromes should solve the problem of tissue availability.

A specialized palindrome specific chromatin?

The presence of genes in palindrome spacers suggests that association with a palindrome is sufficient to provide a benefit to these spacer genes that cannot potentially recombine. A palindrome specific chromatin structure might supply that benefit. Palindrome arm to arm pairing and the resulting stem-loop or cruciform structures might create a scaffold for a testis specific chromatin conformation. To test this hypothesis, experiments beyond bioinformatic analysis must be undertaken. Again, difficulties in

obtaining sufficient tissue from human testis to perform biochemistry and cell biology experiments could preclude such experiments unless mice can be utilized.

Palindromes and the sex body

Another potential functional role for palindromes in the testis might relate to the sex body. The X and Y chromosomes are transcriptionally silenced during male meiosis. The chromosomes condense to form the sex body, from which transcriptional machinery is excluded (Richler et al. 1994). Palindromes could form structures that protrude from the sex body thereby permitting access for transcription. An example of an X-linked gene that defies the transcriptional silencing of the sex chromosomes is *Ott*, a gene encoding a protein of unknown function (Kerr et al. 1996). Whether *Ott* is present in a palindrome is unknown. The hypothesis of palindrome mediated escape from male meiotic sex chromosome silencing could be explored in using *in situ* hybridization techniques to visualize transcripts and through immuno-histochemistry techniques to visualize transcriptional protein co-localization.

Assessing the limitations of current palindrome characterizations

The limitations of current characterizations of palindrome sequences dictated by the status of the human genome sequence must be noted. Palindromes are by their nature duplicated sequences that can share levels of nucleotide identity as high as that between alleles on homologous chromosomes (Skaletsky et al. 2003). Sequence variation between palindrome arms can be difficult to differentiate from polymorphism between allelic chromosomes. On the Y, which was known from the start of sequencing efforts to

harbor repetitive sequence, the decision to sequence a single individual's Y chromosome eased some of the difficulties of differentiating arm to arm variation from haplotypic variation. The X chromosome sequence was assembled from sequence obtained from bacterial artificial chromosomes (BAC) and cosmids from multiple libraries generated from different males and females, including individuals carrying more than two copies of the X chromosome. Except for P8, P10, P12, P18 and P24 where the palindrome sequence was derived entirely from a single clone, the sequence of the two palindrome arms may reflect variation between X haplotypes as well as variation between palindrome arms, thereby altering the apparent nucleotide differences between palindrome arms. Furthermore, if sufficient depth of coverage in sequencing individual clones was not achieved, all sequence variation between palindrome arms might not have been captured, thereby reducing the apparent nucleotide differences between palindrome arms.

To obtain results that accurately depict the nucleotide identity between palindrome arms would require measures that are expensive and would not significantly expand our understanding of palindrome characteristics or function. For those palindromes in the human X chromosome sequence that are not constructed from a single BAC, it would be necessary to identify and sequence BACs from a single male library covering those palindromes. It would be better to acknowledge the limitation of the current genomic sequence when drawing conclusions related to nucleotide identity and to remember these limitations when designing future sequencing efforts and sequencing analyses.

The genome projects currently in the pipeline will primarily use whole genome shotgun (WGS) sequence, the random sequencing of a large collection of clones of

various insert sizes. Large duplications that are highly similar, of which palindromes are an extreme subset, tend to be collapsed and are underrepresented in WGS assemblies (Eichler 1998; Eichler 2001; She et al. 2004). An analysis of the mouse genome predicted that over half of segmental duplications displaying greater than 95% identity were misassembled or collapsed (Bailey et al. 2004). In the rat genome, which was sequenced by a hybrid WGS and clone by clone approach, duplicated regions displaying greater than 98.5% nucleotide identity were more likely to be collapsed than less highly similar regions (Tuzun et al. 2004). Strategies that search for nonrandom distribution of WGS sequence reads can be utilized to look for regions that are likely to be duplicated (Bailey et al. 2002). Palindromes are unlikely to be easily uncovered from other mammalian genomes unless found during targeted sequencing of duplicated regions or as a result of a comparative sequencing strategy such as that employed in this thesis.

Conclusions

The comparative sequencing strategy undertaken in studying palindromes has firmly established that palindromes are not anomalies in the human genome sequence. The experimental approach to defining palindrome associated gene transcription ranges and the use of publicly available transcription data to confirm the significance of the transcription bias have provided strong evidence for a functional association between palindromes and the testis. The exact nature of the functional role of palindromes in the testis is a fascinating question whose answer is beyond the scope of my thesis project. In the future, biological experiments are needed to test the hypotheses of palindrome function posed by this thesis and to form new and better hypotheses for testing.

REFERENCES

Bailey, J.A., D.M. Church, M. Ventura, M. Rocchi, and E.E. Eichler. 2004. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res* **14:** 789-801.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Eichler, E.E. 1998. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res* **8:** 758-762.

Eichler, E.E. 2001. Segmental duplications: what's missing, misassigned, and misassembled--and should we care? *Genome Res* **11:** 653-656.

Kerr, S.M., M.H. Taggart, M. Lee, and H.J. Cooke. 1996. Ott, a mouse X-linked multigene family expressed specifically during meiosis. *Hum Mol Genet* **5:** 1139-1148.

Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu, and T. Yasunaga. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* **52:** 863-867.

Richler, C., G. Ast, R. Goitein, J. Wahrman, R. Sperling, and J. Sperling. 1994. Splicing components are excluded from the transcriptionally inactive XY body in male meiotic nuclei. *Mol Biol Cell* **5:** 1341-1352.

Rozen, S., H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, and D.C. Page. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423:** 873-876.

She, X., Z. Jiang, R.A. Clark, G. Liu, Z. Cheng, E. Tuzun, D.M. Church, G. Sutton, A.L. Halpern, and E.E. Eichler. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431:** 927-930.

Skaletsky, H., T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825-837.

Tuzun, E., J.A. Bailey, and E.E. Eichler. 2004. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res* **14:** 493-506.

Warburton, P.E., J. Giordano, F. Cheung, Y. Gelfand, and G. Benson. 2004. Inverted Repeat structure of the human genome: The X chromosome contains a preponderance of large highly homologous inverted repeats with contain testes genes. *Genome Res* **14:** 1861-1869.

APPENDIX A

Support Files

A CD-ROM containing the following files in plain text format accompanies this manuscript. These files can also be found on the web at http://staffa.wi.mit.edu/page/saionz/

Figures:

Figure 5C
Figure 7B
Figure 9B

Alignments:

Human palindrome paired arms
Species comparisons between orthologous palindrome arms and between orthologous sequence near to palindromes