The Pennsylvania State University

The Graduate School

Department of Biology

STUDIES OF GENE EXPRESSION EVOLUTION:

GENES ON THE INACTIVE X CHROMOSOME AND DUPLICATE GENES

A Dissertation in

Biology

by

Chungoo Park

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2010

The dissertation of Chungoo Park was reviewed and approved* by the following:

Kateryna D. Makova
Associate Professor of Biology
Dissertation Co-Advisor
Chair of Committee

Laura Carrel
Associate Professor of Biochemistry and Molecular Biology
Dissertation Co-advisor

Francesca Chiaromonte
Professor of Statistics

Webb Miller
Professor of Biology and Computer Science and Engineering

Claude dePamphilis
Professor of Biology

Douglas R. Cavener
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School

ABSTRACT

Understanding the determinants of the rate of protein evolution is one of the major goals in molecular evolution. Among the potential variables, expression abundance is one of the most important factors for determining protein evolutionary rates; the variation in gene expression appears to contribute to the evolutionary divergence and phenotypic diversity among species and individuals. Here we perform studies to characterize variation in gene expression patterns on the inactive X chromosome, and across duplicate genes in mammals. Specifically, several questions are addressed in greater detail in this dissertation. First, what genomic signals determine the expression status of genes on the inactive X chromosome? Second, does selection operate differently on genes that escape inactivation vs. genes that are inactivated? Third, do genomic features and motifs predict candidate X-linked mental retardation (XLMR) genes? Fourth, what drives the rapid expression divergence observed between human paralogs? To investigate these issues, we use genome-scale gene expression data and bioinformatic analyses. We find that (1) the majority of the sequences enriched in the vicinity of inactivated genes are found within L1 repeats (indicating an involvement of L1 repeats in X chromosome inactivation), and these sequences capture most of the genomic signal determining inactivation; some unique or overrepresented motifs in boundary regions (indicating that they are candidates for the boundary elements separating genes with different X inactivation profiles) are also found; (2) escape genes experience stronger purifying selection than inactivated genes at both the protein-coding and gene expression levels, and this effect largely results from the importance of function and dosage of escape genes; (3) sequence motifs that are mutually exclusively overrepresented in either XLMR or non-XLMR genes effectively capture genomic signals to distinguish between them; and (4) turnover of transcription start sites, structural heterogeneity of coding sequences, and divergence of *cis*-regulatory regions between duplicate gene copies play a pivotal role in determining the

expression divergence of duplicate genes. Results from these studies provide valuable insights into the regulation of inactive X expression and understanding the X chromosome inactivation mechanism, and will further aid in our understanding of long-range control of gene expression on the X chromosome. Moreover, they provide important information for understanding human transcriptome heterogeneity, complexity, and evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have been privileged over the past six years to work with an outstanding group of colleagues who have taught me more about biology, genomics, molecular evolution, statistics, computer science etc than they will ever know. In this space, I would like to record my gratitude to a number of individual who have given me help and encouragement over the period in which this dissertation was written. First and most of all, I wish to express my sincere appreciation to my Ph.D. advisor Dr. Kateryna Makova. With enthusiasm, inspiration, and great efforts, she provided lots of encouragement and sound advice, and helped me to enjoy the research. I would also like to acknowledge my indebtedness to my thesis co-advisor, Dr. Laura Carrel for her continued support, insight and invaluable advice. I would also like to express my appreciation to my committee members, Drs. Francesca Chiaromonte, Webb Miller, and Claude dePamphilis for their encouragement and advice. I would also like to thank the members of the Makova lab, extant and extinct, especially Melissa Wilson Sayres, Yogeshwar Kelkar, Dr. Hiroki Goto, and Dr. Erika Kvikstad. I would also like to acknowledge all the members of the Center for Comparative Genomics and Bioinformatics (CCGB) and the Institute for Molecular Evolutionary Genetics (IMEG) at Penn State who have provided welcoming academic communities. I am also indebted to many Korean students in Penn State, especially Dr. HyunGwan Lee with his family, and Dr. YoungBum Park. I cannot picture my life in State College without them. I also thank Dr. Doheon Lee at KAIST and Dr. Hong-Gil Nam at POSTECH, for their continuous encouragement and support. Without the financial contributions from the following sources, my research and graduate school experience would have been severely lacking in monies: a Braddock fellowship, the J. Ben and Helen D. Hill

memorial fund award and teaching assistantships from the Department of Biology; a University graduate fellowship from The Pennsylvania State University; travel grants from the Department of Biology and IMEG; research assistantships and research support from a NIH (to KD and LC). Finally, I would like to express my deepest appreciation to my family. As always, my family has been there, providing all sorts of tangible and intangible support, especially my parents and parents-in-law. Thank you for your support and prayers. To my wife Sunghee Kim who is a companion for my life, who has supported me in innumerable ways, and loved through many demanding times. Thank you for encouraging me and for sacrificing much these past years. To my son, Brenden Sungmin Park, who always make me smile. Thank you for being my son.

Chapter 1

Introduction

Understanding the determinants of the rate of protein evolution is one of the major goals in molecular evolution (Pal et al. 2006). Among the potential variables, expression abundance is one of the most important factors for determining protein evolutionary rates (Xia et al. 2009). Generally, highly expressed genes evolve slowly because of their functional importance and/or selection for their translational robustness (Rocha and Danchin 2004; Drummond et al. 2005), and consequently tend to have low rates of protein evolution. Gene expression also provides important information for understanding molecular and cellular phenotypes. Indeed it has revealed that gene expression differs among individuals (reviewed in (Cheung and Spielman 2009; Skelly et al. 2009)). Thus, variation in gene expression appears to contribute to the evolutionary divergence and phenotypic diversity among species and individuals, respectively. Here we perform studies to characterize variation in gene expression patterns (1) on the inactive X chromosome, and (2) across duplicate genes in mammals. Specifically, two questions are addressed in greater detail in this dissertation. First, what genomic factors determine whether genes are or are not expressed on the inactive X chromosome? Second, what drives the rapid expression divergence observed between human paralogs? To investigate these issues, we use genome-scale gene expression data and bioinformatic analysis.

X chromosome inactivation

Background

Sex chromosome evolution

Genomes of therian mammals possess X and Y sex chromosomes: XX determines femaleness and XY determines maleness. The X chromosome is large and gene-rich, while the Y is small and has very few genes. Recently, comparative mapping between the entire human X chromosome and the chicken genome showed that while most of the human Xp (especially between the telomere and Xp11.3) aligned with chicken chromosome 1q, human Xq matched contiguous blocks from chicken chromosome 4p (Ross et al. 2005). Such intensive comparisons supported Ohno's proposition that mammalian sex chromosomes originated from a pair of autosomes (Ohno 1967). The finding that the platypus X chromosomes share homology with the chicken sex chromosome and have no homology with therian X additionally provided evidence that mammalian sex chromosomes began differentiation after the divergence of the monotremes (166 million years (My) ago) (Marshall Graves 2008; Veyrunes et al. 2008).

The human X chromosome represents the putative ancestral eutherian X chromosome (Bourque et al. 2005). Genes mapping to the short arm of the human X, called the X-added region (XAR), are autosomal in marsupialia (Graves 1995). The long arm of human X, called the X-conserved region or XCR, however, is shared with the marsupial X chromosomes, supporting that the XAR was added to the eutherian X chromosome between the marsupial-eutherian divergence (148 My ago) and the eutherian radiation (105 My ago) (Veyrunes et al. 2008). Lahn and Page (Lahn and Page 1999) estimated evolutionary age of the human X chromosome using the nucleotide divergence (synonymous rate, K_S) between 19 X-Y homologous gene pairs and

determined four evolutionary strata on the X. Strata 1 and 2, corresponding to the XCR, exhibit the largest pairwise divergence scores. Additionally, strata 3 and 4 map to the XAR. A newly defined fifth stratum was defined by exploring inversion events in stratum 4 (Ross et al. 2005), and divergence times and evolutionary origins for XAR and XCR were reevaluated (Veyrunes et al. 2008; Delbridge et al. 2009; Wilson and Makova 2009).

Dosage compensation between males and females

Human X and Y sex chromosomes are highly dimorphic. The human X chromosome is ~155 Mb long and contains ~1098 genes (Ross et al. 2005). In contrast, only ~156 genes have been identified on the human Y chromosome, which is ~60 Mb in length (Skaletsky et al. 2003). Interestingly, only 54 genes on the human X chromosome have functional homologs on the human Y chromosome. Such a small number of shared genes between the X and Y chromosomes cause a dosage imbalance between males and females. To equalize X-linked gene expression in mammals, one of female's X chromosomes is silenced, a process called X chromosome inactivation (XCI) (Avner and Heard 2001; Park and Kuroda 2001; Chow et al. 2005). In marsupials, the paternal X chromosome is preferentially inactivated in all tissues (Richardson et al. 1971), but some loci on the inactive X chromosome are frequently reactivated in somatic cells (Graves 1996; Johnston et al. 2002), indicating XCI is less stable in marsupials than in eutherians. Interestingly, a non-coding XIST gene (which plays critical roles in XCI in eutherians; see below) does not exist in marsupials X chromosomes, and thus the marsupial XCI is essentially different from inactivation in eutherians (reviewed in (Deakin et al. 2009)). The non-random paternal X inactivation is also observed in the extraembryonic tissues of mouse (Takagi and Sasaki 1975; West et al. 1977); however, humans lose the paternal X inactivation (Zeng and Yankowitz 2003).Intriguingly, later in development, at fetal and adult stages in animal development of

mouse and human, either the maternal or the paternal X chromosome is active (Payer and Lee 2008). Thus, human and mouse females are mosaics in this respect. In contrast, monotreme mammals do not use XCI to achieve dosage compensation. Two possible mechanisms have been theoretically considered: downregulation of X chromosome expression in females (50%) or upregulation of X chromosome expression in males (200%) (Reik and Lewis 2005). Recently, it has been discovered that only partial and locus-specific inactivation occurs in monotremes (Deakin et al. 2008).

XCI is controlled by the XIST gene, found in the X Inactivation Center (XIC) mapping to Xq13 in human. The XIST gene encodes an untranslated RNA and is expressed only from the inactive X chromosome. After calculating how many X chromosomes there are in a cell (the 'counting' step), if more than one X is present, one X chromosome is determined to remain active (the 'choice' step), then the XIST RNA gene on the soon-to-be inactive X chromosome becomes upregulated and coats the X chromosome from which it is expressed (Prothero et al. 2009). According to the recent chromatin immunoprecipitation (ChIP) experiments, the XIST RNA may be associated with various histone modifications to repress expression of genes on the inactive X chromosome. For example, hypoacetylatation and methylation of histone H3 and H4, and macroH2A recruitment are highly associated with heterochromatin and silencing for inactive X chromosome (Gilbert and Sharp 1999; Csankovszki et al. 2001; Heard et al. 2001; Goto et al. 2002; Plath et al. 2003; Silva et al. 2003; Kohlmaier et al. 2004; Okamoto et al. 2004; Marks et al. 2009; Mietton et al. 2009). CTCF binding sites also function as epigenetic regulators for X inactivation to silence one of two female X chromosomes (Chao et al. 2002). Additionally, aberrant CpG-island methylation on the inactive X chromosome (Monk 1986) and late replication of the inactive X chromosome (Taylor and Miner 1968; Schmidt and Migeon 1990) show obvious differences between the inactive and active X chromosomes.

Genes that escape XCI

Recently, Carrel and Willard (Carrel and Willard 2005) generated a comprehensive X inactivation profile of genes on the human X chromosome, and showed that about 15% of genes assayed escape X inactivation, *i.e*. are expressed from both the active and inactive X chromosomes ("escape genes"). Most of the genes escaping X inactivation reside on the short arm of the X chromosome, comprised of the evolutionarily young strata. The short arm of the X chromosome has a relatively high density of X/Y gene pairs, as compared to the long arm (Ross et al. 2005). The genes with Y-linked homologues are likely to escape X inactivation, because they do not require dosage compensation. Thus, it is not surprising that the short arm of the X chromosome has many genes escaping X inactivation, and most of the genes which are subject to X inactivation reside on the long arm of the X chromosome (Ross et al. 2005). However, many genes on the X chromosome's short arm without Y-linked homologues also escape X inactivation (Carrel and Willard 2005). This means that even with XCI, such X-linked genes continue to cause dosage imbalances between males and females. On the inactive X chromosome, interestingly, 10% of genes escape inactivation in only some females ("heterogeneous genes"); the other 75% of genes are subject to inactivation in all females ("inactivated genes") (Carrel and Willard 2005). The mouse X chromosome has fewer escape genes than the human X chromosome (13 out of 393 assayed; 3.3% (Yang et al. 2010) versus 94 out of 612 assayed; 15% (Carrel and Willard 2005)); these genes are scattered individually along the mouse X chromosome, whereas they cluster together on the human X (Coleman et al. 1996; Tsuchiya and Willard 2000; Yang et al. 2010).

Several factors have been proposed to account for the regulation of escape genes on inactive X chromosome. *The interspersed repeat elements*. Many studies found that escape genes are enriched for Alu transposable elements and (GATA)n simple repeats, and are depleted of LINE1, LTRs and MIRs transposable elements (Tsuchiya et al. 2004; Carrel et al. 2006; McNeil

et al. 2006; Wang et al. 2006). Indeed, full-length L1 elements expressed from the inactive X chromosome regulate the expression of an inactivated gene which has an escape genes in close neighborhood (Chow et al. 2010). *CTCF insulators*. CCCTC-binding factor (CTCF), a widely expressed 11-zinc finger transcription factor, may function as a DNA insulator to prevent the propagation of X inactivation signals (Ciavatta et al. 2006). Indeed, Filippova et al. (Filippova et al. 2005) reported that human ElF2S3 and mouse Eif2s3x and Jarid1c (they are escape genes, and are adjacent to an inactivated gene) have CTCF binding sites at their 5' ends. In addition to the genetic factors listed above, epigenetic regulators have been also key players in the regulation of escape genes. *Histone modification*. It has been known that histone modification plays an important role in gene expression (Rice and Allis 2001; Turner 2002). For example, lysines (K) in the histone N-terminal tails are frequently modified with acetyl or methyl groups, and serines (S) are likely to be modified by phosphates. Such modifications of histone tails create binding sites for chromatin-modifying enzymes and alter chromatin accessibility. Some studies indeed found that escape genes are marked by the absence of H3K27me3 histone modification associated with the inactive X chromosome in a chromosome-wide fashion (Marks et al. 2009; Yang et al. 2010). *CpG island methylation*. Ke and Collins (Ke and Collins 2003) observed a lower density of CpG islands in the vicinity of genes escaping XCI and suggested that CpG islands may provide a mechanism for genes to escape XCI; however, this pattern was not observed in chromosome-wide analysis (Carrel and Willard 2005).

In spite of the several evidences mentioned above, it is not clear whether these factors themselves are necessary to establish escape gene expression (reviewed in (Prothero et al. 2009)). It has been shown that some histone modifications are highly associated with heterochromatin and silencing for XCI, and to control these types of heterochromatin, two not mutually exclusive models have been postulated (Prothero et al. 2009). First, when inactive heterochromatin is propagated, "way stations" (which function as booster elements to propagate the inactivation

signal (Gartler and Riggs 1983; Lyon 1998)) to be positioned throughout the inactive X chromosome are required. If a gene resides far from the way stations, the gene is likely to be an escape gene. Second, a genomic element on the boundary regions between escape and inactivated genes may prevent the spread of the inactive X heterochromatin. This model ("boundary elements model") assumes the presence of insulators. A combination of the two is also possible. Way stations might boost the propagation of inactive heterochromatin, and boundary elements might prevent the inactivation signals from approaching close to escape genes.

Dissertation outline for X chromosome studies

In X chromosome inactivation studies, we have explained the following three issues to gain further insights into factors regulating XCI and evolution of escape genes. (1) What genomic landmarks render most genes silent while leaving others expressed on the inactive X chromosome in mammalian females? (2) Do escape genes tend to be subject to a unique selective regime? (3) What genomic elements in the boundary regions between escape and inactivated genes may prevent the spread of the inactive X heterochromatin?

Chapter 2 is concerned with the question of what genomic signals determine the expression status of genes on the inactive X chromosome. We utilized an experimentally derived comprehensive inactivation profile from human X chromosome (Carrel and Willard 2005) and developed bioinformatic approaches to identify candidate sequences that potentially predict the inactivation status of X-linked genes in human. We found that the majority of the sequences enriched in the vicinity of inactivated genes lying evolutionary young strata were found within L1 repeats, indicating an involvement of L1 repeats in X chromosome inactivation. Using linear discriminant analysis (LDA) trained by occurrences of all sequences we discovered, expression status was correctly predicted for 84% and 91% of escape and inactivated genes, respectively. It

suggests that these sequences capture most of the genomic signal determining inactivation and may play a role in controlling expression status of genes on inactive X chromosome.

Chapter 3 is about the evolution of escape genes. Do selection operate differently on escape than inactivated or heterogeneous genes? There are two possibilities for escape genes to be retained as expressed genes on the inactivated X chromosome. First, insufficient evolutionary time may have passed to acquire XCI by all genes on the X; this hypothesis, not invoking selective forces, is supported by an excess of escape genes in the evolutionary young strata (Jegalian and Page 1998; Lahn and Page 1999; Carrel and Willard 2005; Ross et al. 2005). Second, selection may have played a role in maintaining escape genes. For example, compared with 46,XX individuals, Turner syndrome individuals (45, X), who do not have any biallelically expressed X-linked genes have a distinct phenotype including premature ovarian failure and short status (Good et al. 2003; Bondy 2006). This observation suggests the importance of function and of precise gene dosage of escape genes. We hypothesized that, to be retained as expressed genes on the inactivated X chromosome, escape genes might evolve non-neutrally. Indeed, we observed that escape genes experience stronger purifying selection than inactivated genes at both the protein-coding and gene expression levels. This effect largely results from the importance of function and dosage of escape genes.

Chapter 4 is devoted to studies of the boundary elements model (namely, a genomic element at the boundary region between escape and inactivated genes may prevent the spread of the inactive X heterochromatin). We utilized an experimentally derived inactivation profile from several eutherian mammals and developed bioinformatic approaches to identify candidate sequences that potentially determine the inactivation status of X-linked genes. Three approaches were explicitly addressed for this study: (1) identify overrepresented motifs in the boundary regions using chromosome-wide human XCI data; (2) identify motifs uniquely present in a boundary region (as compared with non-boundary regions) studied in detail in a number of

eutherian mammals (see below), and (3) identify motifs present in a boundary region studied in detail in a number of eutherian mammals (not requiring uniqueness of these motifs in this boundary). We found some unique or overrepresented motifs in boundary regions, indicating that they are the candidates for the boundary elements separating genes with different XCI profiles. The resulting data set needs to be evaluated in future experimental studies and may provide valuable insights into the regulation of escape gene expression.

Chapter 5 is about a bioinformatic analysis of candidate gene prioritization for X-linked mental retardation (XLMR). The same bioinformatic approaches (chapter 2) were applied to this XLMR study. We detected sequence motifs that were mutually exclusively overrepresented in either XLMR or non-XLMR genes, and built LDA classifiers using them. The classification accuracies for both XLMR and non-XLMR genes were high (> 80%), implying that the sequence motifs effectively capture genomic signals to distinguish between XLMR and non-XLMR genes.

Gene duplication

Nowadays, genome-wide expression experiments have produced accurate gene expression data and have revealed expression divergence patterns between duplicate genes (Wray et al. 2003; Khaitovich et al. 2006; de Hoon and Hayashizaki 2008; Mortazavi et al. 2008). Studies of expression divergence are necessary for understanding the emergence of new gene functions after duplication events (Li et al. 2005).

Background

Gene duplication mechanisms

Gene duplication traditionally occurs through chromosomal (or genome) duplications, retropositions, or unequal crossing over (Zhang 2003; Hurles 2004). Chromosomal duplications can result from misalignment between homologous chromosomes during cell division and generate a duplicate for every gene on the chromosome (Zhang 2003; Hurles 2004). Substantial evidence of these large-scale duplications has been observed in the flowering plant lineage but not in the animal lineage (Wendel 2000; Blanc and Wolfe 2004). Mismatching of homologous chromosomes can also cause unequal crossing over that commonly generates tandem duplication. Based on the position of unequal crossing over, the tandem duplication can involve part of a gene, a complete gene, or multiple genes. Sometimes, deletions and insertions can occur in intervening sequences by these recombination events (Zhang 2003; Hurles 2004). Retroposition can also contribute to the generation of duplicate genes. Retroposition occurs through reverse transcription of RNA with subsequent insertion of the cDNA back into the genome. Duplicated genes generated by retroposition were commonly considered to be pseudogenes due to lack of regulatory elements (promoters) required for their expression. However, multiple studies have shown that retroposition generates a significant number of active genes (Betran and Long 2002; Long et al. 2003; Emerson et al. 2004).

Gene duplication models

Because of the importance of gene duplication in evolution (Ohno 1970), it is crucial to know how duplicate genes evolve and what determines their destiny. According to the classical models, gene duplication may result in one of the following: (1) creation of a pseudogene because

of degenerative mutations ("nonfunctionalization") (Harrison et al. 2002), (2) gain of a new function by one duplicate gene ("neofunctionalization") (Ohno 1970), (3) division of the parental gene's function between the two duplicate copies after the duplication event ("subfunctionalization") (Force et al. 1999; Lynch and Force 2000), or (4) a combination of both neofunctionalization and subfunctionalization ("subneofunctionalization") (He and Zhang 2005). The classical models basically assume that duplications do not affect fitness (namely, fixation of a duplicated copy is a neutral process) (Innan and Kondrashov 2010); however, the fixation also plays an important role in determining the maintenance of gene duplication. Thus, several population genetic models for gene duplication evolution have been also proposed (Conant and Wolfe 2008; Innan 2009; Innan and Kondrashov 2010).

Rapid expression divergence between duplicate genes

Many studies found that duplicate genes diverged rapidly in their expression (reviewed in (Li et al. 2005)). Although it has been debated whether adaptive mutations are more likely to occur in the protein-coding region than in the regulatory regions of genes or vice versa (Wray et al. 2003; Hoekstra and Coyne 2007; Carroll 2008), both coding and regulatory region divergence is well correlated with expression divergence of duplicate genes. For example, Gu and colleagues (Gu et al. 2002), Conant and Wagner (Conant and Wagner 2004), and Makova and Li (Makova and Li 2003) found a positive correlation between coding-sequence divergence and expression divergence between duplicate genes in yeast, *C. elegans*, and human, respectively; (Zhang et al. 2004) and (Castillo-Davis et al. 2004) found a significant correlation between expression divergence of duplicate genes and extent of their shared cis-regulatory motifs in yeast and *C elegans*, respectively. Moreover, some further factors have been shown to influence the expression divergence of duplicate genes: *trans*-acting factors (Zhang et al. 2004), post-

transcriptional regulation (e.g., microRNA (Li et al. 2008)), and transcriptional reprogramming (Kafri et al. 2005).

Dissertation outline for duplicate genes study

Chapter 6 is concerned with the question of what drives the expression divergence of human paralogs on a genome-wide scale. Three subquestions are explicitly addressed for this main question: (1) how frequently does the turnover of transcription start sites (TSSs) occur between duplicate genes? (2) how often are duplicate genes structurally identical at birth and does this influence expression divergence? (3) Does divergence of cis-regulatory regions influence the rate of expression divergence in duplicate genes? We observed a frequent turnover of TSSs between duplicate genes and a high proportion of young duplicate genes with incompletely similar structures. These two factors significantly influence expression divergence between duplicate genes. The proportion of aligned sequences in *cis*-regulatory regions and expression similarity between the two copies are also strongly correlated.

References

Avner P, Heard E. 2001. X-chromosome inactivation: counting, choice and initiation. *Nature reviews* **2**(1): 59-67.

Bell AC, Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**(6785): 482-485.

Betran E, Long M. 2002. Expansion of genome coding regions by acquisition of new genes. *Genetica* **115**(1): 65-80.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* **16**(7): 1667-1678.

Bondy CA. 2006. Turner's syndrome and X chromosome-based differences in disease susceptibility. *Gend Med* **3**(1): 18-30.

Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome research* **15**(1): 98-110.

Carrel L. 2006. Molecular biology. "X"-rated chromosomal rendezvous. *Science (New York, NY* **311**(5764): 1107-1109.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Carrel L, Park C, Tyekucheva S, Dunn J, Chiaromonte F, Makova KD. 2006. Genomic environment predicts expression patterns on the human inactive X chromosome. *PLoS genetics* **2**(9): e151.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**(1): 25-36.

Castillo-Davis CI, Hartl DL, Achaz G. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome research* **14**(8): 1530-1536.

Chao W, Huynh KD, Spencer RJ, Davidow LS, Lee JT. 2002. CTCF, a candidate trans-acting factor for X-inactivation choice. *Science (New York, NY* **295**(5553): 345-347.

Cheung VG, Spielman RS. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature reviews* **10**(9): 595-604.

Chow JC, Yen Z, Ziesche SM, Brown CJ. 2005. Silencing of the mammalian X chromosome. *Annual review of genomics and human genetics* **6**: 69-92.

Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E et al. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**(6): 956-969.

Ciavatta D, Kalantry S, Magnuson T, Smithies O. 2006. A DNA insulator prevents repression of a targeted X-linked transgene but not its random or imprinted X inactivation. *Proceedings of the National Academy of Sciences of the United States of America* **103**(26): 9958-9963.

Coleman MP, Ambrose HJ, Carrel L, Nemeth AH, Willard HF, Davies KE. 1996. A novel gene, DXS8237E, lies within 20 kb upstream of UBE1 in Xp11.23 and has a different X inactivation status. *Genomics* **31**(1): 135-138.

Conant GC, Wagner A. 2004. Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. *Proceedings* **271**(1534): 89-96.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature reviews* **9**(12): 938-950.

Csankovszki G, Nagy A, Jaenisch R. 2001. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *The Journal of cell biology* **153**(4): 773-784.

Deakin JE, Chaumeil J, Hore TA, Marshall Graves JA. 2009. Unravelling the evolutionary origins of X chromosome inactivation in mammals: insights from marsupials and monotremes. *Chromosome Res* **17**(5): 671-685.

Deakin JE, Hore TA, Koina E, Marshall Graves JA. 2008. The status of dosage compensation in the multiple X chromosomes of the platypus. *PLoS genetics* **4**(7): e1000140.

de Hoon M, Hayashizaki Y. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *BioTechniques* **44**(5): 627-628, 630, 632.

Delbridge ML, Patel HR, Waters PD, McMillan DA, Marshall Graves JA. 2009. Does the human X contain a third evolutionary block? Origin of genes on human Xp11 and Xq28. *Genome research* **19**(8): 1350-1360.

Diaz-Perez S, Ouyang Y, Perez V, Cisneros R, Regelson M, Marahrens Y. 2005. The element(s) at the nontranscribed Xist locus of the active X chromosome controls chromosomal replication timing in the mouse. *Genetics* **171**(2): 663-672.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **102**(40): 14338-14343.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science (New York, NY* **303**(5657): 537-540.

Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ, Nguyen DK, Tsuchiya KD, Disteche CM. 2005. Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Developmental cell* **8**(1): 31-42.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-1545.

Gartler SM, Riggs AD. 1983. Mammalian X-chromosome inactivation. *Annual review of genetics* **17**: 155-190.

Gilbert DM. 2002. Replication timing and transcriptional control: beyond cause and effect. *Current opinion in cell biology* **14**(3): 377-383.

Gilbert SL, Sharp PA. 1999. Promoter-specific hypoacetylation of X-inactivated genes. *Proceedings of the National Academy of Sciences of the United States of America* **96**(24): 13825-13830.

Good CD, Lawrence K, Thomas NS, Price CJ, Ashburner J, Friston KJ, Frackowiak RS, Oreland L, Skuse DH. 2003. Dosage-sensitive X-linked locus influences the development of amygdala and orbitofrontal cortex, and fear recognition in humans. *Brain* **126**(Pt 11): 2431-2446.

Goto Y, Gomez M, Brockdorff N, Feil R. 2002. Differential patterns of histone methylation and acetylation distinguish active and repressed alleles at X-linked genes. *Cytogenetic and genome research* **99**(1-4): 66-74.

Graves JA. 1995. The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding. *Bioessays* **17**(4): 311-320.

-. 1996. Mammals that break the rules: genetics of marsupials and monotremes. *Annual review of genetics* **30**: 233-260.

Gu Z, Nicolae D, Lu HH, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **18**(12): 609-613.

Hansen RS, Wijmenga C, Luo P, Stanek AM, Canfield TK, Weemaes CM, Gartler SM. 1999. The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **96**(25): 14412-14417.

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**(6785): 486-489.

Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M. 2002. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome research* **12**(2): 272-280.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2): 1157-1164.

Heard E. 2004. Recent advances in X-chromosome inactivation. *Current opinion in cell biology* **16**(3): 247-255.

Heard E, Rougeulle C, Arnaud D, Avner P, Allis CD, Spector DL. 2001. Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell* **107**(6): 727-738.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution; international journal of organic evolution* **61**(5): 995-1016.
Hurles M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS biology* **2**(7): E206.
Illingworth RS, Bird AP. 2009. CpG islands--'a rough guide'. *FEBS letters* **583**(11): 1713-1720.
Innan H. 2009. Population genetic models of duplicated genes. *Genetica* **137**(1): 19-37.
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews* **11**(2): 97-108.
Jegalian K, Page DC. 1998. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**(6695): 776-780.
Johnston PG, Watson CM, Adams M, Paull DJ. 2002. Sex chromosome elimination, X chromosome inactivation and reactivation in the southern brown bandicoot Isoodon obesulus (Marsupialia: Peramelidae). *Cytogenetic and genome research* **99**(1-4): 119-124.
Kafri R, Bar-Even A, Pilpel Y. 2005. Transcription control reprogramming in genetic backup circuits. *Nature genetics* **37**(3): 295-299.
Ke X, Collins A. 2003. CpG islands in human X-inactivation. *Annals of human genetics* **67**(Pt 3): 242-249.
Khaitovich P, Enard W, Lachmann M, Paabo S. 2006. Evolution of primate gene expression. *Nature reviews* **7**(9): 693-702.
Kohlmaier A, Savarese F, Lachner M, Martens J, Jenuwein T, Wutz A. 2004. A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS biology* **2**(7): E171.
Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science (New York, NY* **286**(5441): 964-967.
Lee JT. 2005. Regulation of X-chromosome counting by Tsix and Xite sequences. *Science (New York, NY* **309**(5735): 768-771.
Li J, Musso G, Zhang Z. 2008. Preferential regulation of duplicated genes by microRNAs in mammals. *Genome biology* **9**(8): R132.
Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet* **21**(11): 602-607.
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nature reviews* **4**(11): 865-875.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**(1): 459-473.
Lyon MF. 1998. X-chromosome inactivation: a repeat hypothesis. *Cytogenetics and cell genetics* **80**(1-4): 133-137.
Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research* **13**(7): 1638-1645.
Marks H, Chow JC, Denissov S, Francoijs KJ, Brockdorff N, Heard E, Stunnenberg HG. 2009. High-resolution analysis of epigenetic changes associated with X inactivation. *Genome research* **19**(8): 1361-1373.
Marshall Graves JA. 2008. Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annual review of genetics* **42**: 565-586.
McNeil JA, Smith KP, Hall LL, Lawrence JB. 2006. Word frequency analysis reveals enrichment of dinucleotide repeats on the human X chromosome and [GATA]n in the X escape region. *Genome research* **16**(4): 477-484.
Mietton F, Sengupta AK, Molla A, Picchi G, Barral S, Heliot L, Grange T, Wutz A, Dimitrov S. 2009. Weak but uniform enrichment of the histone variant macroH2A1 along the inactive X chromosome. *Molecular and cellular biology* **29**(1): 150-156.
Monk M. 1986. Methylation and the X chromosome. *Bioessays* **4**(5): 204-208.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**(7): 621-628.

Ohno S. 1967. Sex chromosomes and sex-linked genes. *Berlin:Springer-Verlag*.

-. 1970. Evolution by gene duplication. *New York, New York: Springer*.

Okamoto I, Otte AP, Allis CD, Reinberg D, Heard E. 2004. Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science (New York, NY* **303**(5658): 644-649.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nature reviews* **7**(5): 337-348.

Park Y, Kuroda MI. 2001. Epigenetic aspects of X-chromosome dosage compensation. *Science (New York, NY* **293**(5532): 1083-1085.

Payer B, Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. *Annual review of genetics* **42**: 733-772.

Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. 2003. Role of histone H3 lysine 27 methylation in X inactivation. *Science (New York, NY* **300**(5616): 131-135.

Prothero KE, Stahl JM, Carrel L. 2009. Dosage compensation and gene expression on the mammalian X chromosome: one plus one does not always equal two. *Chromosome Res* **17**(5): 637-648.

Reik W, Lewis A. 2005. Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nature reviews* **6**(5): 403-410.

Rice JC, Allis CD. 2001. Histone methylation versus histone acetylation: new insights into epigenetic regulation. *Current opinion in cell biology* **13**(3): 263-273.

Richardson BJ, Czuppon AB, Sharman GB. 1971. Inheritance of glucose-6-phosphate dehydrogenase variation in kangaroos. *Nature: New biology* **230**(13): 154-155.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular biology and evolution* **21**(1): 108-116.

Ross MT Grafham DV Coffey AJ Scherer S McLay K Muzny D Platzer M Howell GR Burrows C Bird CP et al. 2005. The DNA sequence of the human X chromosome. *Nature* **434**(7031): 325-337.

Schmidt M, Migeon BR. 1990. Asynchronous replication of homologous loci on human active and inactive X chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* **87**(10): 3685-3689.

Silva J, Mak W, Zvetkova I, Appanah R, Nesterova TB, Webster Z, Peters AH, Jenuwein T, Otte AP, Brockdorff N. 2003. Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Developmental cell* **4**(4): 481-495.

Simon I, Tenzen T, Reubinoff BE, Hillman D, McCarrey JR, Cedar H. 1999. Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature* **401**(6756): 929-932.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**(6942): 825-837.

Skelly DA, Ronald J, Akey JM. 2009. Inherited variation in gene expression. *Annual review of genomics and human genetics* **10**: 313-332.

Takagi N, Sasaki M. 1975. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature* **256**(5519): 640-642.

Taylor JH, Miner P. 1968. Units of DNA replication in mammalian chromosomes. *Cancer research* **28**(9): 1810-1814.

Tsuchiya KD, Greally JM, Yi Y, Noel KP, Truong JP, Disteche CM. 2004. Comparative sequence and x-inactivation analyses of a domain of escape in human xp11.2 and the conserved segment in mouse. *Genome research* **14**(7): 1275-1284.

Tsuchiya KD, Willard HF. 2000. Chromosomal domains and escape from X inactivation: comparative X inactivation analysis in mouse and human. *Mamm Genome* **11**(10): 849-854.

Turner BM. 2002. Cellular memory and the histone code. *Cell* **111**(3): 285-291.

Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Grutzner F, Deakin JE, Whittington CM, Schatzkamer K et al. 2008. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome research* **18**(6): 965-973.

Wendel JF. 2000. Genome evolution in polyploids. *Plant molecular biology* **42**(1): 225-249.

West JD, Frels WI, Chapman VM, Papaioannou VE. 1977. Preferential expression of the maternally derived X chromosome in the mouse yolk sac. *Cell* **12**(4): 873-882.

Wilson MA, Makova KD. 2009. Evolution and survival on eutherian sex chromosomes. *PLoS genetics* **5**(7): e1000568.

Wang Z, Willard HF, Mukherjee S, Furey TS. 2006. Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS computational biology* **2**(9): e113.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Molecular biology and evolution* **20**(9): 1377-1419.

Xia Y, Franzosa EA, Gerstein MB. 2009. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS computational biology* **5**(6): e1000413.

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Pequignot E. 1999. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**(6758): 187-191.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome research* **20**(5): 614-622.

Zeng SM, Yankowitz J. 2003. X-inactivation patterns in human embryonic and extra-embryonic tissues. *Placenta* **24**(2-3): 270-275.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**(6): 292-298.

Zhang Z, Gu J, Gu X. 2004. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet* **20**(9): 403-407.

Chapter 2

Genomic Environment Predicts Expression Patterns on the Human Inactive X Chromosome

This chapter has been published in Public Library of Sciences Genetics 2006, 2(9): e151, and was formatted for that journal. Authors for this original manuscript are Laura Carrel, Chungoo Park, Svitlana Tyekucheva, John Dunn, Francesca Chiaromonte, and Kateryna D. Makova. CP performed the experiments, analyzed the data, and contributed analysis tools. LC, FC, and KDM conceived and designed the experiments. ST performed the experiments. LC, ST, FC, and KDM analyzed the data. ST, and JD contributed analysis tools. LC, FC, and KDM wrote the paper.

Abstract

What genomic landmarks render most genes silent while leaving others expressed on the inactive X chromosome in mammalian females? To date, signals determining expression status of genes on the inactive X remain enigmatic despite the availability of complete genomic sequences. Long interspersed repeats (L1s), particularly abundant on the X, are hypothesized to spread the inactivation signal and are enriched in the vicinity of inactive genes. However, both L1s and inactive genes are also more prevalent in ancient evolutionary strata. Did L1s accumulate there because of their role in inactivation or simply because they spent more time on the rarely recombining X? Here we utilize an experimentally derived inactivation profile of the entire human X chromosome to uncover sequences important for its inactivation, and to predict expression status of individual genes. Focusing on Xp22, where both inactive and active genes

reside within evolutionarily young strata, we compare neighborhoods of genes with different inactivation states to identify enriched oligomers. Occurrences of such oligomers are then used as features to train a linear discriminant analysis classifier. Remarkably, expression status is correctly predicted for 84% and 91% of active and inactive genes, respectively, on the entire X, suggesting that oligomers enriched in Xp22 capture most of the genomic signal determining inactivation. To our surprise, the majority of oligomers associated with inactivated genes fall within L1 elements, even though L1 frequency in Xp22 is low. Moreover, these oligomers are enriched in parts of L1 sequences that are usually underrepresented in the genome. Thus, our results strongly support the role of L1s in X inactivation, yet indicate that a chromatin microenvironment composed of multiple genomic sequence elements determines expression status of X chromosome genes.

Synopsis

To match the amount of gene product produced in males (XY), most genes in mammalian females (XX) are active on one X chromosome and inactivated on the other. However, some genes "escape" inactivation and are expressed from both X chromosomes. This study investigates sequences that may control whether a gene undergoes or escapes X chromosome inactivation, including DNA sequences previously thought of as non-functional or "junk." Earlier work suggested that one such sequence, L1 interspersed repeats, may be associated with inactivation, but the extent of such association, and whether it represented a consequence of the evolutionary history of X, remained unclear. This study utilized recently generated chromosome-wide data on sequence and gene expression for human X, with a particular focus on the Xp22 region, which is evolutionarily young and has had no time to accumulate many L1 elements. A rigorous statistical analysis identified with high accuracy a set of short sequences that discriminate between genes

undergoing and those escaping X chromosome inactivation. Interestingly, the majority of such sequences enriched in the vicinity of inactivated genes were found within L1s. These results strengthen the case for an involvement of L1s in X chromosome inactivation and suggest other DNA elements that might also play a role.

Introduction

X chromosome inactivation (XCI) is an extraordinary example of long-range gene regulation, extending over 150 Mb (megabases) and transcriptionally silencing genes on one X chromosome in females in order to equalize X-linked gene dosage with XY males (reviewed in [1,2]). XCI initiates during early embryogenesis and requires the presence of the *XIST* gene (in *cis*), whose RNA transcript closely associates with and coats the inactive X chromosome [2]. Upon inactivation, the X chromosome is heavily epigenetically modified in many ways typical of other silenced loci, including the incorporation of methylated DNA and modified histones [3].

Notwithstanding the chromosome-wide nature of XCI, not all genes on the X are silenced [4–6]. These genes that "escape" XCI lack at least some epigenetic alterations characterizing the rest of the chromosome [5]. Recently, in conjunction with completion of the sequence of the human X chromosome [7], a comprehensive human X inactivation profile was established [6]. A total of 15% of assayed genes escape XCI; their distribution and organization is highly non-random and mirrors the evolutionary history of the X. The X-specific portion of the X is partitioned into five strata that show increasing levels of sequence divergence with increasing distance from the distal tip of Xp [7,8]. Genes that escape inactivation are primarily found within the youngest strata that map to Xp22. Furthermore, such genes are clustered, suggesting that they are controlled at the level of chromosome domains [6,9].

Consideration of escape genes is important for understanding how XCI spreads and is

maintained in *cis* along the chromosome. Specific *cis*-acting sequences on the X may direct chromatin modifications or *XIST* RNA to specific sites along the chromosome, or might be involved in other aspects of regulating XCI. Studies of X;autosome translocations in human and mouse, and analysis of ectopic X inactivation of mouse *Xist* transgenes lend support for the involvement of *cis* regulatory sequences in the spreading of XCI. Although autosomal sequences on these chromosomes can be inactivated, autosome gene inactivation and spreading of *XIST* RNA as well as of epigenetic markers of inactivation are incomplete and in some cases discontinuous [10–15]. These studies suggest that the X may be organized in a manner distinct from that of autosomes and may be more receptive to transcriptional inactivation.

Such observations led Gartler and Riggs to hypothesize that specific sequences, "booster elements" or "way stations," could propagate an inactivation signal [16]. Such sequences need not be unique to the X, but should be more highly represented on the X than on autosomes. Subsequently, Lyon proposed that the repetitive element LINE-1 (L1) may function as such a booster [17], based on cytological studies showing L1 enrichment on the X in human and mouse [18,19]. Complete sequencing of the X confirmed this enrichment; L1 elements are approximately 2-fold enriched on the X compared to autosomes [7]. However, the distribution of L1 elements fluctuates along the X, with the highest proportion in the evolutionarily oldest strata. Notably, preliminary analysis suggested that sequences adjacent to genes escaping inactivation are depleted in L1s [20], although this study did not consider differences in L1 density along the X chromosome and escape gene organization. Moreover, although the study by Bailey et al. [20] lent support to the L1 hypothesis, it did not consider an alternative (but not mutually exclusive) model: escape genes may be associated with different *cis* regulatory sequences that prevent these genes from either initiating or stably maintaining inactivation.

The recently established X inactivation profile and completed X chromosome sequence [6,7] prompted us to reinvestigate the role of genomic sequences in XCI. To find sequences that

may influence X inactivation state, we computationally identified overrepresented motifs in the neighborhoods of both inactivated genes and genes that escape XCI in Xp22. These enriched sequences correctly predict the inactivation state of most genes along the entire X chromosome.

Results

Description of the Escape and Inactivated Subgenomes Analyzed in Xp22

We focused our analysis on the Xp22 region for the following reasons. First, Xp22 contains about equal numbers of genes that are transcriptionally silent on inactive X and of genes that escape inactivation. In fact, among 103 genes assayed in Xp22 [6], 30% (31 genes) are subject to inactivation and 39% (40 genes) escape inactivation. (The other genes exhibit heterogeneous expression patterns between different inactive Xs tested.) This is in contrast with the rest of the X chromosome, where the overwhelming majority of genes are inactivated (66%, or 339 out of 515 genes assayed), and only a small percentage of genes escape inactivation (6%, or 31 genes) [6]. Second, within Xp22, inactivated genes and genes escaping XCI are located in the same, relatively young, evolutionary strata (part of stratum 3 and strata 4–5). Thus, comparison of silenced and escape genes within Xp22 is expected to highlight XCI signals and not the evolutionary differences between strata. Third, we hypothesized that, if L1 interspersed repetitive elements were involved in XCI, analysis of a region in which their overall density is low could reveal either local L1 organizational differences or additional XCI regulatory elements. Indeed, only 15% of the Xp22 sequence is covered by L1 elements, as compared with 29% for the whole X chromosome [7]. The pseudoautosomal region (also located in Xp22) was excluded from our analysis.

To delineate sequences determining the inactivation status of genes in Xp22, we divided

this region into two subgenomes, *I* (for inactivated) and *E* (for escaping inactivation). The Xp22 genomic sequences were compiled on the basis of the X inactivation profile [6], including regions upstream and downstream from the transcription start site (TSS) of each gene. We considered three distances surrounding the TSSs of genes: ±50 kilobases (kb), ±100 kb, and ±250 kb. Thus, based on these distances, three pairs of *I* and *E* subgenomes were investigated: I_{50} and E_{50}, I_{100} and E_{100}, and I_{250} and E_{250} (Table 2-1). Each subgenome consisted of several "contigs": uninterrupted genomic sequences upstream and downstream of the TSS of a gene with a particular expression pattern. Frequently, the region surrounding a specific gene overlapped the region surrounding an adjacent gene with the same expression profile. In this case, the genomic sequences around TSSs of both (or sometimes several) genes were merged into the same contig. Overlapping surrounding sequences of adjacent genes with *different* inactivation patterns were excluded. The subgenome pairs were constructed to keep the frequency of repetitive elements and genomic length approximately equal between the two subgenomes (Table S2-1). Notably, the frequency of only one type of repetitive element (ERV class I) differed by more than 2-fold between any two subgenome pairs. L1 repeats were at low frequency in both subgenomes, but were slightly more abundant in the *I* subgenomes compared to the *E* subgenomes (1.4- to 1.6-fold difference), e.g., the L1 difference in the ±50-kb subgenomes is 13.3% versus 9.7%, both notably lower than the 29% X chromosome average [7].

Table **2-1**: Gene Number and Length of Contigs for the *E* and *I* Subgenomes within Xp22 (Used to Discover Overrepresented Oligomers).

Distance Surrounding TSS	Number of Genes (Length) in *E*	Number of Genes (Length) in *I*
±50 kb	31 (2,051 kb)	25 (2,072 kb)
±100 kb	17 (1,880 kb)	13 (1,821 kb)
±250 kb	9 (1,864 kb)	6 (1,753 kb)

Gene lists are given in Table S1.

Analysis of Oligomers Enriched in Either *E* or *I* Subgenomes

We next developed an XCI profile–driven computational approach to contrast genomic sequences adjacent to genes that are inactivated or escape inactivation. We compared the frequency of all possible oligomers of specified length between the *I* and *E* subgenomes. Initially, 8-, 12-, 16-, 20-, and 24-mers were examined separately for each of the three subgenome pairs. An oligomer was considered to be overrepresented in a subgenome if (1) it was present at least ten times in that subgenome; and (2) its frequency was at least 5-fold higher in that subgenome compared to the other subgenome. The oligomers that were identified using these initial criteria were further evaluated with a permutation test (see Methods) that assessed statistical significance of the overrepresentation. We focused our further analysis on 12-mers because they had the highest total number of different oligomers overrepresented for the *E* or *I* subgenomes.

Two additional operations were performed on the significantly ($p < 0.01$) overrepresented 12-mers (Figure 2-1). First, overlapping 12-mers were merged into longer oligomers. Second, such oligomers identified at different distances surrounding genes (±50 kb, ±100 kb, or ±250 kb) of the *E* subgenome were pooled (and merged) into a single set. This allowed the oligomers that were found to be overrepresented only at one or two distances to be considered in the further analysis of all three distances from TSSs. We followed the same procedure for 12-mers identified in the *I* subgenome. The resulting set consisted of 110 and 138 different oligomers overrepresented for the *E* and *I* subgenomes, respectively (Figure 2-1, Table S2-2). These are called "overrepresented oligomers" in the remainder of the manuscript. Remarkably, the majority of overrepresented oligomers (74% for *E* and 60% for *I*) were also significantly enriched on the X chromosome compared with autosomes ($p < 0.05$, permutation test). Focusing only on the oligomers enriched on chromosome X compared to autosomes had little effect on our quantitative

results and did not alter our conclusions (unpublished data); therefore, all 248 (110 + 138) overrepresented oligomers were used in the analyses described below.

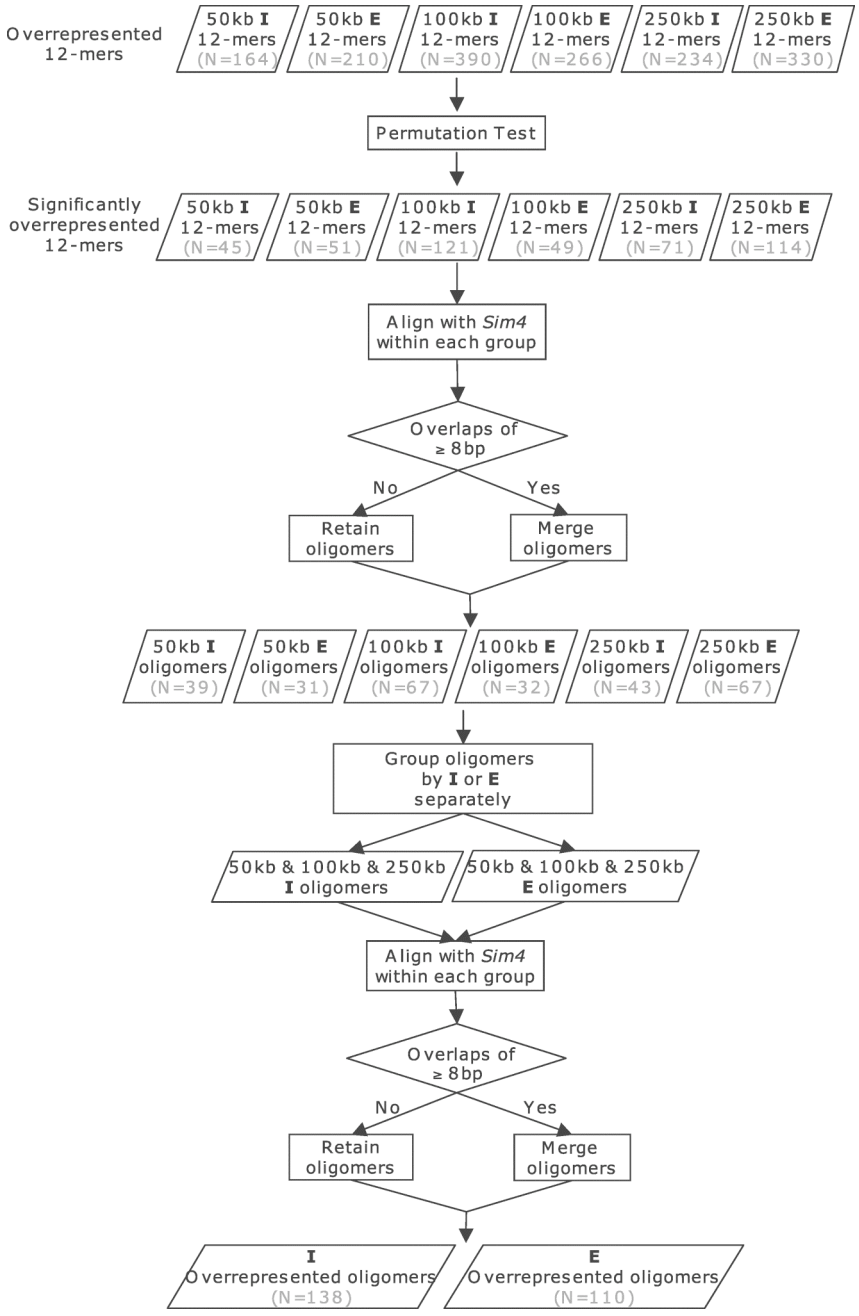


Figure **2-1**: Procedure Used to Obtain Overrepresented Oligomers Starting from the Overrepresented 12-Mers. The overrepresented 12-mers were defined with the initial criteria: at least ten occurrences and at least 5-fold enrichment. See Results and Methods for a detailed description.

Interestingly, oligomers overrepresented in the *E* and *I* subgenomes mapped within different sequence classes (Table 2-2). Indeed, 38% of oligomers overrepresented in the *E* subgenome were located within *Alu* repeats (the corresponding value for the *I* subgenome is only 9%). In contrast, 64% of the oligomers overrepresented in the *I* subgenome were within L1 repeats (compared with only 4% for *E*). Intriguingly, although the majority of L1 sequences in Xp22 (as well as on the X chromosome and in the whole human genome) are truncated at the 5′ end and frequently include only 3′UTR sequences [21], the oligomers enriched in the *I* subgenome had a substantially different distribution; they were enriched in ORF1 and ORF2, but depleted from the 3′UTR (Figure 2-2).

Table **2-2**: Assignment of Overrepresented Oligomers to Interspersed Repetitive Elements (Repeats).

RepeatType/Subgenome	*I*nactivated	*E*scape
DNA/MER1	4	1
LINE (L1)	88 (or 64%)	4 (or 4%)
LINE (except L1)	1	0
LTR/MaLR	3	15
LTR/ERV	0	7
SINE/Alu	12 (or 9%)	42 (or 38%)
SINE/MIR	3	0
Simple-repeat/low complexity	0	3
Occasionally present in repeats[a]	25	35
Unique	2	3
Total	138	110

An overrepresented oligomer was considered to be part of a repeat if its genomic coordinates were annotated as part of a repeat in at least 50% of its genomic occurrences in a studied subgenome. See Table S2 for a list of individual overrepresented oligomers.
[a]This category describes oligomers that are located in repeats infrequently (<50% of occurrences in a studied subgenome).

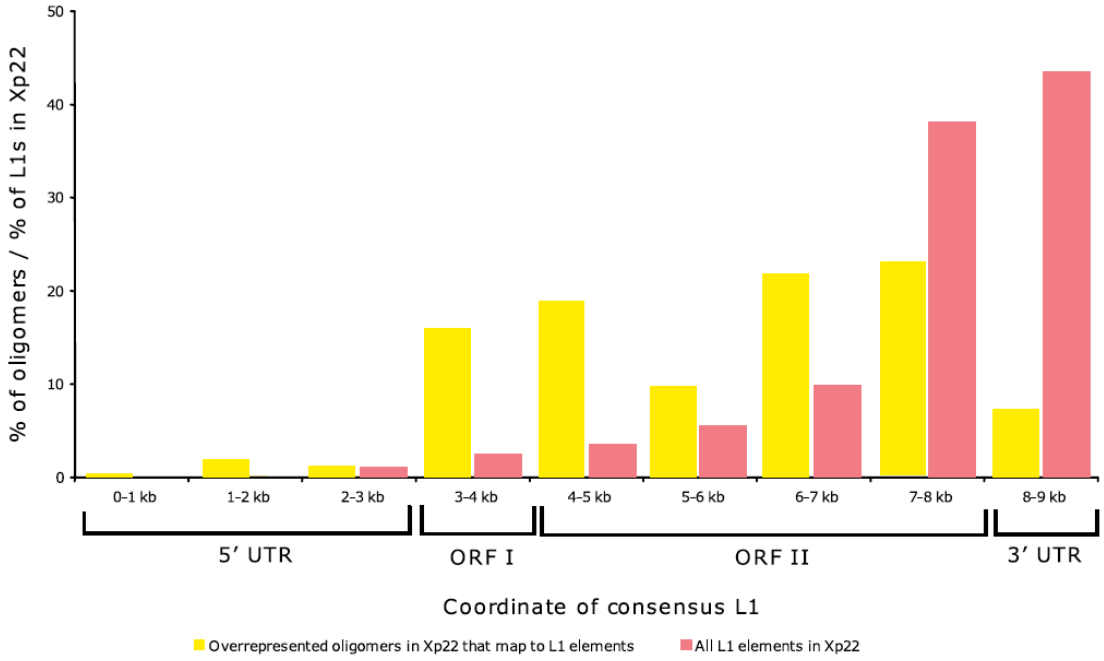


Figure **2-2**: The Distributions over the Length of L1 Element of Overrepresented Oligomers Found in *I* Subgenome (Yellow Bars) and of All L1 Sequences within Xp22 (Red Bars). Only overrepresented oligomers mapping frequently to L1s (>50% of their genomic occurrences in the *I* subgenome) are shown. Although the full-length L1 is approximately 7 kb long, the alignment of L1 subfamilies was approximately 9 kb long. ORF, open reading.

Classification of Genes as Either Inactivated or Escaping Inactivation Based on Surrounding Oligomers

To predict the inactivation status of genes (either *E* or *I*), we used linear discriminant analysis (LDA) [22]. In the application of LDA to the present study, genes were units to be classified, and counts of overrepresented oligomers surrounding TSSs of genes were classification features. Our training data consisted of the Xp22 genes that comprised our original *E* and *I* subgenomes, together with additional X chromosome genes for which expression status had been confirmed by a second assay in primary fibroblasts [6] (Tables 2-3 and S2-3). Although overrepresented oligomers were derived from Xp22, extending the training set with additional X

chromosome genes allowed us to "learn" the role of these sequence elements in XCI, not just within Xp22, but more generally on X.

Table **2-3**: The Numbers of Genes Analyzed for Training and Test Datasets.

Set/Distance from TSS	±50 kb *(E/I)*	±100 kb *(E/I)*	±250 kb *(E/I)*
Training set (largely Xp22)	34/59	21/42	13/29
Test set of Xp22 genes	5/7	16/13	18/13
Test set of X genes (outside of Xp22)	14/283	10/264	6/236
Training set of all X genes[a]	53/349	47/319	37/278

Pseudoautosomal regions were excluded. See Table S3 for a complete gene list.
[a]Includes Xp22 genes.

LDA based on counts of overrepresented oligomers had excellent performance on the training set; correct classification rates assessed by leave-one-out cross-validation were ≥85% for both *E* and *I* classes at each of the three distances surrounding TSSs (Figure 2-3A, Table 2-4). Next, we investigated whether the LDA classifier trained this way could predict expression status in two non-overlapping test sets, namely: (1) Xp22 genes not used in training, and (2) X chromosome genes outside of Xp22 and excluding pseudoautosomal regions; here also, training set genes were not included (Table 2-3 and S2-3). For these two sets, the counts of the overrepresented *E* and *I* oligomers (originally discovered from Xp22) surrounding each gene were calculated, and the XCI state was predicted.

We achieved high correct classification rates (≥81%) for Xp22 test genes at all three distances examined (Figure 2-3B, Table 2-4). Thus, our classifier effectively captures crucial genomic differences between *E* and *I* genes in the Xp22 region. Classification performance for X chromosome test genes increased with the distance surrounding TSSs (Figure 2-3C, Table 2-4). At ±250 kb, we were able to reach correct classification rates of 83% and 72% for *E* and *I* genes,

respectively, whereas performance at ±50 kb and ±100 kb was somewhat lower. Higher performance on Xp22 than on other X chromosome test genes could be due to the fact that the training data largely included Xp22 genes. As a consequence, the classifier may be capturing genomic features prevalent in Xp22 in addition to the XCI signals we are seeking.

Table **2-4**: Success Rates of LDA.

Set Analyzed	Parameter	±50 kb	±100 kb	±250 kb
Training set (largely Xp22)	τ	0.51	0.39	0.26
	Success in *E*	85%	90%	85%
	Success in *I*	93%	93%	93%
Test set of Xp22 genes	τ	0.3	0.4	0.34
	Success in *E*	100%	81%	83%
	Success in *I*	86%	100%	92%
Test set of X genes (Xp22 excluded)	τ	0.07	0.17	0.15
	Success in *E*	86%	80%	83%
	Success in *I*	38%	56%	72%
Training set of all X genes (Xp22 included)	τ	0.3	0.4	0.29
	Success in *E*	72%	77%	84%
	Success in *I*	80%	91%	91%

The tuning parameter τ was selected to maximize the sum of correct classification rates of *E* and *I* classes.

To overcome this problem, we used overrepresented oligomers derived from Xp22 to train LDA on all X chromosome genes (excluding pseudoautosomal regions). In other words, we replaced our original, mostly Xp22-based training set with a new one comprising genes from the entire X chromosome. The new set included all genes from the initial training set as well as from the two test sets discussed above. Leave-one-out cross-validation on this new, chromosome-wide training set yielded correct classification rates of 84% and 91% for *E* and *I* genes, respectively, for ±250 kb from TSS (Figure 2-3D, Table 2-4). This represents a substantial improvement in performance relative to previous success rates for the test set of X chromosome genes (see above). Moreover, the chromosome-wide training may in fact be less influenced by Xp22

"landscape" features and thus captures XCI signals more effectively. Subsequent results concerning correctly and erroneously classified genes were based on the outcomes of this analysis.

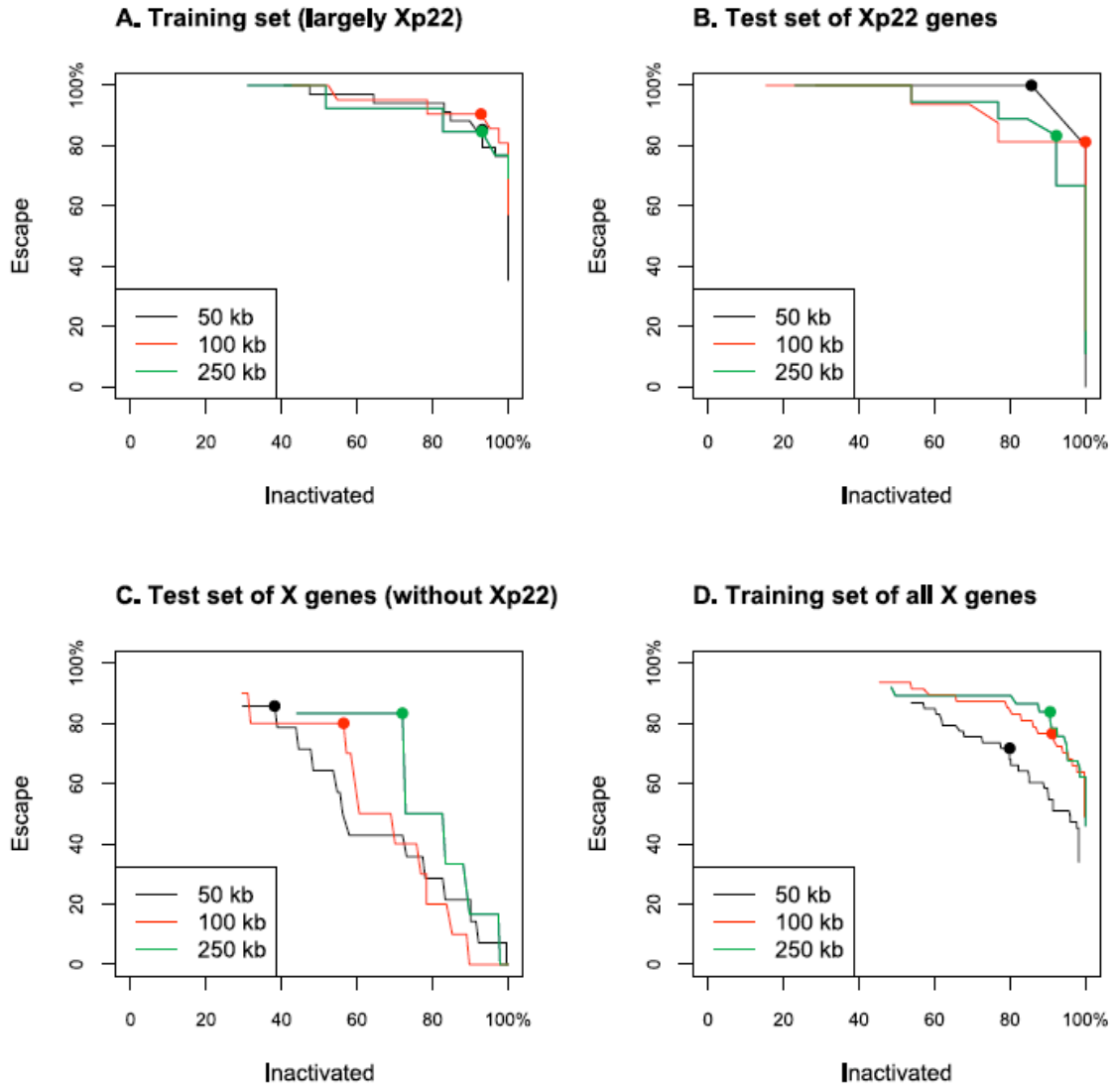


Figure **2-3**: LDA Classification Success Rates for Different Values of the Tuning Parameter τ.

(A) Training set derived largely, but not exclusively, from Xp22 (See Table S2-3).
(B) Test set of Xp22 genes, with training performed on genes in (A)
(C) Test set of X genes outside of Xp22, with training performed on genes in (A).
(D) Training set of all X genes, including genes in Xp22. Dots indicate optimal values of τ (see Table 2-4 and Methods).

Genes Classified Correctly and Misclassified Genes

Classification performance for LDA trained on all X chromosome genes is visualized with respect to chromosome location in Figure 2-4. Large *E* and *I* domains are, for the most part, well predicted. Our overall high rate of correct classification, despite evolutionary differences that have influenced the X sequence composition, argues strongly that these enriched oligomers successfully capture differences in XCI and not simply genomic differences in X chromosome sequence.

Notwithstanding the good overall classification performance, chromosome location does seem to influence our ability to predict inactivation status, as specific regions have higher assignment errors for both *E* and *I* genes. This is most apparent for genes within Xp11.3–Xp11.4 (~40–48 Mb). The reasons for incorrect classifications in this particular region are puzzling. Gene density and repeat content fluctuate greatly within Xp11.3–Xp11.4; however, genes at other X chromosome locations with even more dramatic fluctuations in these parameters are classified correctly. The region also contains an evolutionary breakpoint between strata 2 and 3, although it is unclear what role this could play in misclassification of both *E* and *I* genes.

Escape genes are particularly well classified within Xp22 domains. A plausible explanation is that the overrepresented oligomers were derived from within this region, although from a small subset of the correctly classified genes. Nonetheless, very large escape domains are not present elsewhere on the X, and smaller escape regions may not show adequate enrichment for classification purposes. Supporting this idea, we failed to correctly classify the only two non-domain escape transcripts, Hs.458197 and *SH3BGRL,* included in this study (at all three distances from TSSs); other non-domain escape genes were omitted because of the proximity of adjacent inactivated genes. In another instance, both *E* and *I* transcripts in and surrounding a <250-kb escape domain in Xp11.1 (including KIAA0522) are assigned incorrectly at the only scorable

distance, ±50 kb. This could suggest that both *E* and *I* signatures were detected, but that classifications were confounded by nearby genes of differing inactivation status. Although chromosome-wide classifications were most successful at ±250 kb from the TSS, domains of a different size may have different signatures, and analysis of smaller distances may be necessary to correctly assess a larger number of escape genes outside of Xp22. Classification performance on the whole X also likely reflects repeat element landscape differences for both *E* and *I* genes. At ±250 kb, misclassified *I* genes have strikingly lower L1 concentration than correctly classified genes (17.4%, $n = 26$ genes, versus 24.3%, $n = 252$ genes), whereas L1 concentration of misclassified *E* genes is much higher than at their correctly assigned counterparts (27.6%, $n = 6$ genes, versus 11.1%, $n = 31$ genes).

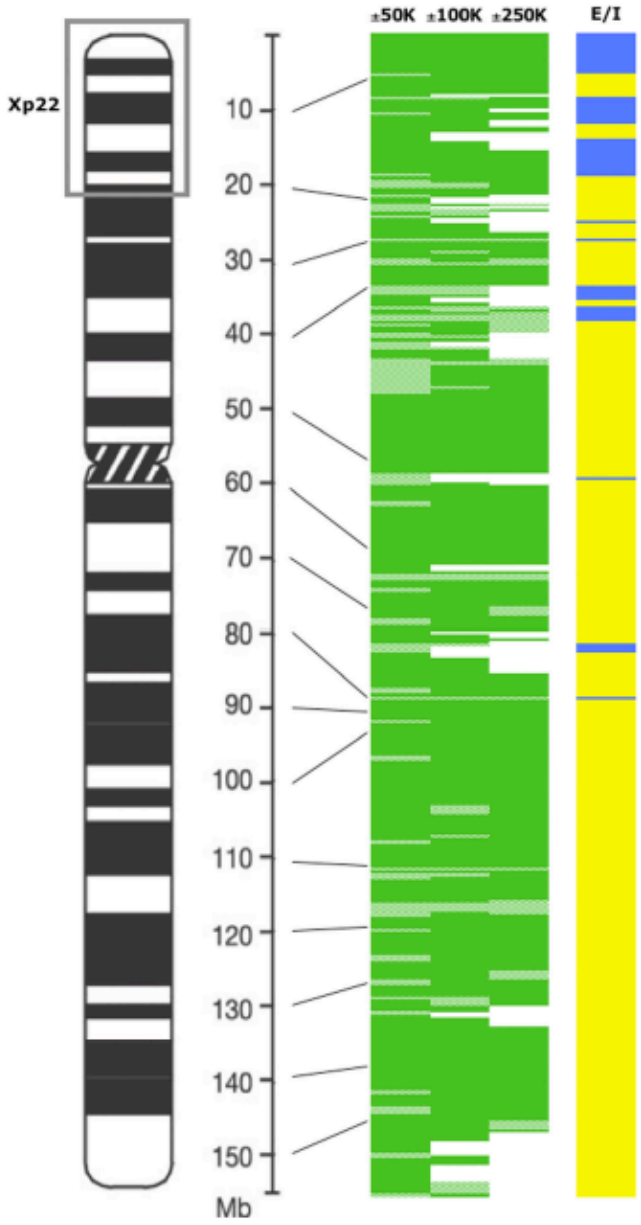


Figure **2-4**: The Distribution of Correctly and Incorrectly Classified Genes along the X Chromosome. Dark green indicates correctly classified genes; light green indicates misclassified genes. X inactivation expression patterns [6] for genes included in this study: yellow indicates inactivated genes, and blue indicates escape genes. Not all genes were analyzed at all distances because sequences that included adjacent genes with *different* inactivation patterns were excluded from analysis (see Methods). These gene distances remain uncolored.

Discussion

Unlike elsewhere on the X, an extraordinarily high proportion of genes in Xp22 escape X inactivation [6]. For this reason, and because these sequences have similar evolutionary origin, we rooted our computational approach within Xp22 in an effort to identify regulatory elements involved in XCI. This reduced the risk of uncovering genomic and evolutionary X chromosome features unrelated to XCI. Notably, using only oligomers identified as enriched in Xp22, we were able to successfully predict XCI status for the vast majority of genes along the entire X chromosome. The approach presented here is completely dependent on an experimentally derived XCI profile that was obtained by assaying human inactive X chromosomes in somatic cell hybrids [6]. A subset of these genes were validated in primary cell lines and confirmed that incorrect assignment of XCI status in the hybrids is rare [6]. Nonetheless, such genes would contribute to misclassification in the present study.

The hypotheses that interspersed regulatory elements (booster elements [16]) control spreading of XCI, and that other elements regulate escape genes, predict that the intrinsic sequence composition of regions surrounding genes with different inactivation statuses should differ. Our strategy made no assumptions about the identity of any overrepresented sequences; however, employing stringent enrichment criteria, we indeed found most overrepresented oligomers in both the *I* and *E* subgenomes within known classes of interspersed repeats. The potential regulatory role of repetitive elements is an emerging theme in epigenetics; other computational studies have reported increased density of interspersed repeats near imprinted loci that were utilized to predict similarly regulated genes [23,24]. The roles of repeats in both spreading and in escape from XCI are discussed separately below.

A high proportion of the enriched *I* oligomers map to L1 sequences. These results substantially strengthen the L1 hypothesis [17], as we started with a region of relatively low L1

frequency and were able to identify L1 sequences that are highly enriched near inactivated genes. The location of overrepresented oligomers within the L1 consensus sequence differs significantly from the L1 sequences within Xp22 or the whole X (Figure 2-2), suggesting that their enrichment may be functional and is not simply due to evolutionary mechanisms that have led to higher repetitive element levels on the X than on autosomes [7,25]. Nonetheless, the L1 oligomers found here predominantly map to primate-specific L1s (unpublished data), and presumably reflect the recent evolutionary origin of Xp22 [7,8] from which they were identified. Young L1s are attractive candidates for spreading XCI within Xp22 because it was hypothesized that formerly autosomal genes must acquire certain sequence characteristics to be inactivated [26]. A sequence necessary for spreading or maintaining XCI could function as a binding site for *XIST* RNA or heterochromatin proteins. Such a recognition motif would likely include non-conserved nucleotides and therefore be represented in older L1s as well. We are currently in the process of adapting our computational approach to score imperfect matches, which may effectively identify such sequences and improve classification success even further. Moreover, our study identified additional oligomers mapping to other repeats that were important for classification on the human X. The involvement of these sequences in XCI must be considered in future studies.

Are repetitive elements also involved in regulating escape from XCI? Previous X sequence analysis concluded that reduced density of L1s may be necessary but not sufficient to establish domains that escape XCI [20]. Our analysis has identified overrepresented oligomers in *E* subgenomes that may instead (or additionally) regulate expression of genes escaping XCI. The predictions were more accurate for escape genes in distal Xp than elsewhere on the chromosome (Figure 2-4). Perhaps this is because escape domains in this region are larger and, therefore, signatures are easier to capture. Alternatively, smaller escape domains or isolated escape genes may be regulated in a different manner and would explain why many were poorly classified. Sequences such as insulators, boundary elements, or barriers flank other coordinately regulated

genes [27]. Functional analysis of the junctions between domains will be necessary to establish the role of these sequences or to identify other elements. Indeed, CTCF-bound insulators flank several escape genes [28], although their role in X inactivation has not yet been definitively established. Such sequences would not have been identified with the present approach, but may be further distilled from chromosome landscape features through a comparative study. On the human X, domains that escape inactivation largely include at least one gene with Y homology that would be required to escape inactivation for appropriate male:female gene dosage [6]. Repetitive element control of XCI could predict that boundaries of escape domains will shift in different species, but likely that the genes with Y homologs remain expressed on the inactive X. Notably, this prediction appears true for one domain studied in human and mouse [29]. Whether this prediction will hold elsewhere on the X and in different species is unknown because although X chromosome sequences are currently available for several mammals, a comprehensive XCI profile exists only for the human X.

Successful classification for most genes within large escape domains, particularly within Xp22, does support a role for enriched oligomers in regulating the expression of these genes. Many of these overrepresented oligomers map to *Alu* repeats. A plausible function for such escape motifs would be to prevent methylation at CpG islands. Interestingly, a previous study found that sequences within *Alu*s were good genome-wide predictors for CpG islands that were resistant to de novo methylation [30].

Another interesting oligomer overrepresented in the *E* subgenome maps to a simple repeat $(GATA)_n$ that was recently proposed by others to determine escape from XCI [31]. We investigated this particular repeat, asking whether occurrences of $(GATA)_n$ alone could predict the inactivation status of genes using a naive Bayes classifier (the same that was used with the LDA for oligomer-based classification). Training and test sets were defined and used as before. $(GATA)_n$ performed well in predicting the status of the Xp22 genes used in training, yet was not

as effective as our oligomer-based LDA predictions on the test sets. Indeed, even when cross-validating genes on the whole X chromosome, using $(GATA)_n$ affords correct classification rates of at most 70% (Table S2-4). As $(GATA)_n$ enrichment was identified by comparison of the most distal 7.5 Mb of Xp to the rest of the X chromosome [31], in contrast to the gene-directed approach that we have employed for our study, it is possible that overrepresentation of this simple repeat could reflect landscape differences unrelated to X inactivation or, in any event, could be only one of several discriminating factors.

Five oligomers did not initially appear to map within repetitive sequences. These oligomers were examined in more detail, and none is composed of a truly unique sequence. Two of the oligomers enriched in the *E* subgenome map to localized repeats within Xp22. The oligomer CAGTGGTTCTTCC is found within a 30-bp repeat motif within the VCX gene family that has multiple members mapping to Xp22.31. Similarly, AAAGCCAGTTAC is part of a tandem repeat that encompasses 650 bp, also within Xp22.31. It is not surprising that our enrichment strategy identified such repeated sequences, but their focused location makes them unlikely candidates to play a role in XCI. Three other oligomers, AAACCATATCAC, identified as enriched in *E* sequences, and *I*-enriched GGGCCGGGCGCA and AAAAATGTTTAA, were not found in repeats within the Xp22 subgenomes, according to our conservative definition requiring both start and end coordinates of an oligomer to be within the repeat. However, closer examination established that each of these frequently was directly adjacent to or occasionally overlapped with known repeat elements. Although, unfortunately, not identifying new candidates controlling X inactivation, these oligomers do give further support to the role of repeat sequences in predicting expression patterns on the chromosome.

Future efforts will focus on identifying motifs that may further improve prediction of XCI status. In this study we only considered inactivated and escape genes that were adjacent to genes with similar inactivation status. Further analyses will need to incorporate more complex

patterns of inactive X expression, including genes within domains that show the opposite inactivation pattern, and heterogenous genes that escape inactivation only in a subset of inactive Xs tested [6]. Even for the genes considered in this study, it is very likely that additional parameters may provide substantial predictive contributions. Features to investigate include CpG islands, gene density, location within an escape domain particularly with respect to domain boundaries, and distance from the *XIST* locus. This idea is supported by a recent computational study that suggested L1 and *Alu* repetitive elements as important predictors for inactivated and escape genes respectively, and identified additional parameters that may also influence inactive X expression [32]. Genomic features that control XCI will further aid in our understanding of long-range control of gene expression and the impact of repetitive elements throughout the genome.

Methods

Transcripts

We utilized a comprehensive inactivation profile of X chromosome genes assayed in fibroblast-derived somatic cell hybrids containing one inactivated X chromosome [6]. Genes were considered to be X inactivated if silenced in all nine somatic hybrids tested or if expressed in only a single hybrid (0/9 or 1/9). Genes were scored as escaping XCI if expressed in eight or nine out of nine somatic hybrids tested (8/9 or 9/9). The TSSs for X chromosome genes were from Supplementary Table S2-3 in [6]. We assumed positive strand to be the coding strand for genes represented by ESTs (expressed sequence tags) with unknown strand orientation. This assignment is not expected to influence our results because the majority of genes represented by ESTs with unknown strand orientation were shorter than 1 kb.

Oligomer enrichment analysis

A series of Perl programs (available upon request) were developed to analyze the genomic sequences located in the subgenomes. Each possible oligomer of a specified size (8-, 12-, 16-, 20-, and 24-mers) was sequentially counted within each subgenome. Exact matches were required. Counts of oligomers with reverse complementary sequence were combined.

To evaluate the significance of overrepresented 12-mers, we implemented a random permutation test for each of the three subgenome pairs (E_{50} and I_{50}, E_{100} and I_{100}, and E_{250} and I_{250}) separately. Contigs were broken into nonoverlapping 2-kb fragments. *E* and *I* labels were removed, and the 2-kb fragments were randomly distributed to either a mock *I* or a mock *E* subgenome. The two mock subgenomes were equal in size. This process was repeated 1,000 times. To determine the empirical *p*-value for each 12-mer, we calculated the number of permutations in which this 12-mer was present at least ten times and overrepresented at least 5-fold in one mock subgenome compared to the other mock subgenome. The 12-mers that satisfied these criteria in fewer than ten out of 1,000 randomizations ($p < 0.01$) were considered significantly overrepresented.

Since we determined significance of overrepresentation for hundreds of 12-mers simultaneously, we needed to adjust for multiple testing. Using a false discovery rate approach [33], we verified that all 12-mers significantly overrepresented according to the permutation test had extremely low false discovery rates ($q < 0.01$). This can be explained by the high stringency of the overrepresentation criteria we set even before applying the permutation test. Thus, our dataset has few false positives after applying initial overrepresentation criteria (at least ten occurrences and at least 5-fold enrichment) and likely very few (if any) false positives after the permutation test.

After identifying significantly overrepresented 12-mers within each subgenome, we merged overlapping 12-mers to avoid scoring them twice. Using *sim4* with default parameters [34], we aligned all significantly overrepresented 12-mers identified for a subgenome against each other. The 12-mers with aligned regions of ≥8 bp (exact match) were merged to generate oligomers. This resulted in six groups of metamers, one for each subgenome (E_{50}, I_{50}, E_{100}, I_{100}, E_{250}, and I_{250}; Figure 2-1).

We next grouped all oligomers identified in each of the *E* subgenomes (E_{50}, E_{100}, and E_{250}) and aligned them against each other using *sim4* with default parameters [34]. Again, oligomers with aligned regions of ≥8 bp (exact match) were merged. Oligomers identified in the three *I* subgenomes underwent similar treatment. This resulted in two groups of oligomers: *I*- and *E*-overrepresented oligomers (Figure 2-1).

Overrepresented oligomers were assigned to interspersed repetitive elements if both start and end genomic coordinates of oligomers were within interspersed repeats as annotated by Repeatmasker (RepBase Update 10.04, version 20050523). For overrepresented oligomers mapping to L1s, we also calculated their coordinates within L1 sequences. The 25 full-length consensus sequences of L1 families [35] were aligned using CLUSTALW [36] with default parameters to derive the L1 consensus sequence. The overrepresented oligomers were aligned to this consensus sequence using BLAST [37] with the following parameters: −F F, −W 7, −r 4, and −q −5.

LDA

To calculate the number of occurrences of a particular overrepresented oligomer in a subgenome, we counted the number of times at least one of the initial 12-mers used in "assembling" this oligomer was present in a subgenome. Several hits within an oligomer at a

particular genomic location were counted only once. For instance, an overrepresented oligomer AAAAACAAGCAATG was created by merging two 12-mers, AAAAACAAGCAA and AAACAAGCAATG. If a subgenome had sequence AAAAACAAGCAATG at a particular genomic coordinate, it was counted only once, even though it had matches to two different initial 12-mers (AAAAACAAGCAA and AAACAAGCAATG). If a subgenome had sequence AAAAACAAGCAACC at some other genomic coordinate, it was also counted once because one 12-bp match (AAAAACAAGCAA) to the overrepresented oligomer could be found. If AAAAACAAGCAATG and AAAAACAAGCAACC were the only two occurrences of this overrepresented oligomer in a subgenome, its total count was 2 (this is just to illustrate how we counted overrepresented oligomers; in reality we required at least ten occurrences in a subgenome).

The counts of overrepresented oligomers in the ±50-kb, ±100-kb, and ±250-kb windows surrounding the TSSs were used to predict gene inactivation status. These counts formed a *p*-dimensional predictor vector $X = (X_1,...X_p)$, where p was equal to 110 + 138 = 248, the number of overrepresented oligomers for both the *E* and the *I* subgenome. Since the dimension exceeded the number of genes in the training set (Table 2-3), we first reduced the dimension by principal components analysis on the normalized predictor vector. Normalization consisted of subtracting the mean and dividing by the standard deviation for each predictor (vector coordinate). We used the first five principal components because they captured a substantial amount of the variability in the original data and were optimal in the subsequent classification analysis. Thus, features used in training and testing the LDA classifier formed a five-dimensional vector $Z = (Z_1,\ldots Z_5)$.

Following [38], the LDA direction L was computed using singular value decomposition of the matrix $W^{-1/2}BW^{-1/2}$, where W and B are the within and between variance-covariance matrices of Z, respectively. The LDA score of a gene with features $Z(g)$ is thus given by $\lambda(g)=L'Z(g)$, and the gene is classified depending on the value of this score relative to a threshold

c. The threshold is expressed by a convex combination of the average LDA scores for the two classes *(I* and *E)* in the training data. The tuning parameter was selected to maximize the sum of correct classification rates for *E* genes and for *I* genes.

Correct classification rates on the training datasets were computed by leave-one-out cross-validation: at each round, one gene was withheld and the classifier was trained on the remaining genes, and then the withheld gene was classified. Correct classification rates for test sets were obtained by applying the trained classifier to the test sets.

References

1. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B (2002) Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 36: 233–278.
2. Chow JC, Yen Z, Ziesche SM, Brown CJ (2005) Silencing of the mammalian X chromosome. Annu Rev Genomics Hum Genet 6: 69–92.
3. Heard E (2005) Delving into the diversity of facultative heterochromatin: The epigenetics of the inactive X chromosome. Curr Opin Genet Dev 15: 482–489.
4. Shapiro LJ, Mohandas T, Weiss R, Romeo G (1979) Non-inactivation of an X-chromosome locus in man. Science 204: 1224–1226.
5. Brown CJ, Greally JM (2003) A stain upon the silence: Genes escaping X inactivation. Trends Genet 19: 432–438.
6. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434: 400–404.
7. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The DNA sequence of the human X chromosome. Nature 434: 325–337.
8. Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. Science 286: 964–967.
9. Miller AP, Willard HF (1998) Chromosomal basis of X chromosome inactivation: Identification of a multigene domain in Xp11.21-p11.22 that escapes X inactivation. Proc Natl Acad Sci U S A 95: 8709–8714.
10. Russell LB (1963) Mammalian X-chromosome action: Inactivation limited in spread and region of origin. Science 140: 976–978.
11. White WM, Willard HF, Van Dyke DL, Wolff DJ (1998) The spreading of X inactivation into autosomal material of an X;autosome translocation: Evidence for a difference between autosomal and X-chromosomal DNA. Am J Hum Genet 63: 20–28.
12. Duthie SM, Nesterova TB, Formstone EJ, Keohane AM, Turner BM, et al. (1999) Xist RNA exhibits a banded localization on the inactive X chromosome and is excluded from autosomal material in cis. Hum Mol Genet 8: 195–204.

13. Sharp AJ, Spotswood HT, Robinson DO, Turner BM, Jacobs PA (2002) Molecular and cytogenetic analysis of the spreading of X inactivation in X;autosome translocations. Hum Mol Genet 11: 3145–3156.
14. Popova BC, Tada T, Takagi N, Brockdorff N, Nesterova TB (2006) Attenuated spread of X-inactivation in an X;autosome translocation. Proc Natl Acad Sci U S A 103: 7706–7711.
15. Lee JT, Jaenisch R (1997) Long-range cis effects of ectopic X-inactivation centres on a mouse autosome. Nature 386: 275–279.
16. Gartler SM, Riggs AD (1983) Mammalian X-chromosome inactivation. Annu Rev Genet 17: 155–190.
17. Lyon MF (1998) X-chromosome inactivation: A repeat hypothesis. Cytogenet Cell Genet 80: 133–7.
18. Korenberg JR, Rykowski MC (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell 53: 391–400.
19. Boyle AL, Ballard SG, Ward DC (1990) Differential distribution of long and short interspersed element sequences in the mouse genome: Chromosome karyotyping by fluorescence in situ hybridization. Proc Natl Acad Sci U S A 87: 7757–7761.
20. Bailey JA, Carrel L, Chakravarti A, Eichler EE (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis. Proc Natl Acad Sci U S A 97: 6634–6639.
21. Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246: 401–417.
22. Seber GAF (1984) Multivariate observations. New York: John Wiley & Sons. 686 p.
23. Allen E, Horvath S, Tong F, Kraft P, Spiteri E, et al. (2003) High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. Proc Natl Acad Sci U S A 100: 9940–9945.
24. Luedi PP, Hartemink AJ, Jirtle RL (2005) Genome-wide prediction of imprinted murine genes. Genome Res 15: 875–884.
25. Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9: 657–663.
26. Jegalian K, Page DC (1998) A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. Nature 394: 776–780.
27. West AG, Fraser P (2005) Remote control of gene transcription. Hum Mol Genet 1: R101–R111.
28. Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ, et al. (2005) Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. Dev Cell 8: 31–42.
29. Tsuchiya KD, Greally JM, Yi Y, Noel KP, Truong JP, et al. (2004) Comparative sequence and x-inactivation analyses of a domain of escape in human Xp11.2 and the conserved segment in mouse. Genome Res 14: 1275–1284.
30. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM (2006) DNA motifs associated with aberrant CpG island methylation. Genomics 87: 572–579.
31. McNeil JA, Smith KP, Hall LL, Lawrence JB (2006) Word frequency analysis reveals enrichment of dinucleotide repeats on the human X chromosome and [GATA]n in the X escape region. Genome Res 16: 477–484.
32. Wang Z, Willard HF, Muhkerjee S, Furey T (2006) Evidence of influence of genomic DNA sequence on human X chromosome inactivation. PLoS Comput Biol. In press.
33. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.

34. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8: 967–974.
35. Khan H, Smit A, Boissinot S (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16: 78–87.
36. Thompson JD, Higgins DG, Gibson TJ (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.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
38. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: Data mining, inference, and prediction. New York: Springer-Verlag. 533 p.

Chapter 3

Strong Purifying Selection at Genes Escaping X Chromosome Inactivation

This chapter has been published in Molecular Biology and Evolution 2010, Jun 9. (Epub ahead of print), and was formatted for that journal. Authors for this original manuscript are Chungoo Park, Laura Carrel, and Kateryna D. Makova. CP, LC, and KDM conceived and designed the experiments. CP performed the experiments, analyzed the data, and contributed analysis tools. CP and KDM wrote the paper.

Abstract

To achieve dosage balance of X-linked genes between mammalian males and females, one female X chromosome becomes inactivated. However, approximately 15% of genes on this inactivated chromosome escape X chromosome inactivation (XCI). Here, using a chromosome-wide analysis of primate X-linked orthologs, we test a hypothesis that such genes evolve under a unique selective pressure. We find that escape genes are subject to stronger purifying selection than inactivated genes and that positive selection does not significantly affect the evolution of these genes. The strength of selection does not differ between escape genes with similar vs. different expression levels in males vs. females. Intriguingly, escape genes possessing Y homologs evolve under the strongest purifying selection. We also found evidence of stronger conservation in gene expression levels in escape than inactivated genes. We hypothesize that divergence in function and expression between X and Y gametologs is driving such strong purifying selection for escape genes.

Introduction

The genomes of therian mammals possess two sex chromosomes – X and Y (with XX females and XY males). To achieve dosage balance between males and females, one female X chromosome is inactivated (Payer and Lee 2008). However, not all genes on the inactive X become silenced. A comprehensive profile of the human inactive X (Carrel and Willard 2005) revealed that 75% of genes are inactivated in all females ("inactivated genes"); 15% of genes escape inactivation in all females ("escape genes"); and 10% of genes escape inactivation in some females ("heterogeneous genes").

Selection may operate differently on escape than inactivated or heterogeneous genes. For instance, compared with 46,XX individuals, Turner syndrome individuals (45,X) lack biallelically expressed X-linked genes and exhibit distinct phenotypes including premature ovarian failure and short stature (Good et al. 2003; Bondy 2006). This observation suggests importance of function and of precise gene dosage of escape genes. Most X-linked genes with Y homologs escape XCI (Ross et al. 2005), which may serve to achieve dosage compensation between males and females. Relatedly, many mammalian X and Y homologs differ substantially in function and expression (Wilson and Makova 2009). Selection may act differently on escape genes with vs. without Y homologs, as most candidate Turner syndrome X-linked genes have Y homologs (Burgoyne 1989; Fisher et al. 1990; Ellison et al. 1997; Rao et al. 1997).

Distinct expression levels of escape genes between male and female brain (Skuse 2005; Xu and Disteche 2006) indicate that adaptation of sexually dimorphic features may also contribute to their evolutionary trajectories. Indeed, genes expressed primarily in one sex (e.g., some escape genes in females) tend to evolve under adaptive pressure (reviewed in (Ellegren and Parsch 2007). Thus, we hypothesize that escape genes might be subject to a unique selective regime.

Here we compared selective pressure between genes with different XCI states using a nonsynonymous-to-synonymous substitution rate ratios (K_A/K_S) and gene expression data. X-linked genes were classified into *inactivated*, *escape*, or *heterogeneous* groups based on the rodent/human somatic cell hybrids assay and the primary human cell line assay (Carrel and Willard 2005) Methods as presented in the Supplemental Information; Fig. S3-1; Table S3-1).

Results and Discussion

First, to contrast selective pressure between escape, heterogeneous and inactivated genes, the K_A/K_S ratio in the human-chimpanzee-macaque phylogenetic tree was computed for each orthologous gene group (Methods). Only genes with one-to-one orthology were utilized; the final data set included 346 genes (32 escape, 41 heterogeneous, and 273 inactivated genes; Methods; Fig. S1). We assumed conservation of the XCI profile among primates (consistent with unpublished results from the Carrel laboratory). We observed (Fig. 3-1A) that the median K_A/K_S ratio was significantly lower for escape genes than for either inactivated (0.087 vs. 0.147, $P < 0.0001$) or heterogeneous genes (0.087 vs. 0.162; $P = 0.0034$). All P values were computed with the permutation test (Methods). The same comparison was repeated separately for newly X-specific genes located in the X-added region (XAR) and anciently X-specific genes in the X-conserved region (XCR; Fig. 3-1B-C; (Ross et al. 2005). In both cases, the median K_A/K_S ratio was significantly lower for escape than inactivated genes (0.108 vs. 0.158 for XAR genes, $P = 0.0228$; 0.041 vs. 0.143 for XCR genes, $P = 0.0214$). This ratio was also lower for escape than heterogeneous genes for both XAR and XCR; the difference was significant for XCR genes (0.041 vs. 0.198; $P = 0.0061$) and non-significant for XAR genes (0.108 vs. 0.146; $P = 0.2779$), likely due to a small number of heterogeneous genes in the XAR. These findings suggest stronger purifying selection operating on escape than either inactivated or heterogeneous genes.

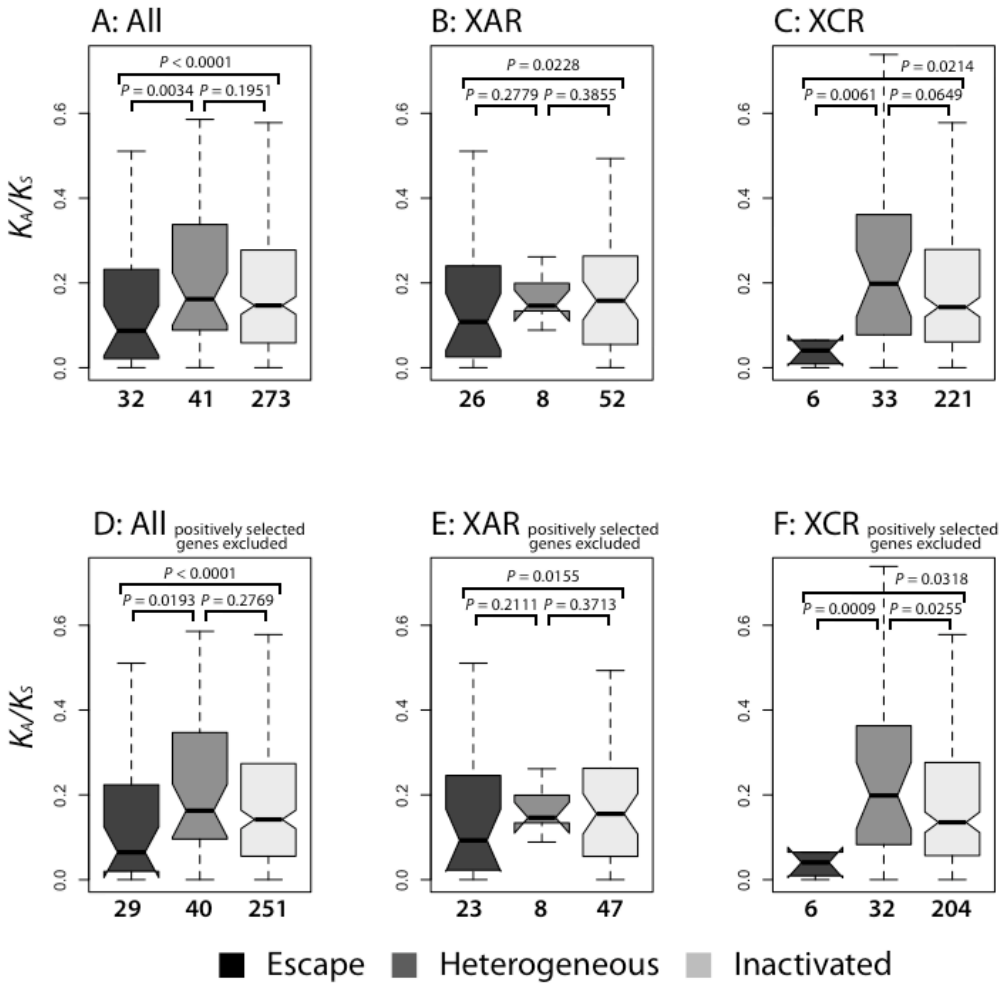


Figure **3-1**: A comparison of K_A/K_S ratios among escape, heterogeneous, and inactivated human-chimpanzee-macaque orthologous genes. (A) All orthologs. (B) Orthologs located in the X-added region (XAR). (C) Orthologs located in the X-conserved region (XCR). (D) All orthologs excluding genes with positively selected codons. (E) Orthologs located in XAR excluding genes with positively selected codons. (F) Orthologs located in XCR excluding genes with positively selected codons. The number of genes considered is given below each plot. In the box plots, edges and vertical dashed lines represent quartiles and range, respectively. Notches indicate standard deviations of the median. Outliers are not shown.

Heterogeneous genes had K_A/K_S ratios not significantly different from those of inactivated genes (Fig. 3-1A-C). When orthologous genes from orangutan and marmoset were added (Methods), escape genes again had significantly lower median K_A/K_S ratios than inactivated genes (0.084 vs.

0.128; $P = 0.0345$; Fig. S3-2); human-chimpanzee-macaque gene trios were used in the further analysis due to the high quality of these genome sequences

Second, we addressed whether the above pattern could be due to positive selection preferentially acting on inactivated genes. Among the 346 genes tested, seven genes had K_A/K_S ratio greater than one, however, this was not significant for any gene (Table S3-2). We also identified and excluded 26 genes with positively selected codons (Methods; Table S3-3). The remaining escape genes still had significantly lower K_A/K_S ratios than the remaining inactivated (0.065 vs. 0.142; $P < 0.0001$) or heterogeneous genes (0.065 vs. 0.163; $P = 0.0193$; Fig. 3-1D) independent of gene location in XAR vs. XCR (Fig. 3-1E-F). Thus, positive selection does not drive the observed lower K_A/K_S ratio for escape vs. inactivated genes. Admittedly, the tests utilized here to test for positive selection are weak. However, our conclusions of its minimal role in our results are supported by generally low proportion of adaptive amino acid substitutions in hominoids (Zhang and Li 2005; Eyre-Walker and Keightley 2009).

Third, as many escape genes have low inactive X relative to active X expression levels and are therefore essentially dosage-compensated, we evaluated whether selection pressure varied based on differences in expression levels between males and females (Nguyen and Disteche 2006; Johnston et al. 2008). Lymphoblast expression microarray data, available for 23 out of 32 escape genes, was utilized to divide them into non-dosage-compensated and dosage-compensated (as in Table 2 of (Johnston et al. 2008). We assumed the XCI pattern was the same between fibroblast and lymphoblastoid cells (although see (Talebizadeh, Simon, and Butler 2006). 15 non-dosage-compensated escape genes (with expression levels significantly different between males and females; HDHD1A, STS, PNPLA4, CA5B, EIF1AX, EIF2S3, ZFX, USP9X, DDX3X, FUNDC1, UTX, UBE1, JARID1C, SMC1L1, and RPS4X) and 8 dosage-compensated ones (with expression levels not significantly different between males and females; RAB9A, SEDL, FAM51A1, AP1S2, CTPS2, RBBP7, MGC39350, and ARHGAP4) had similar K_A/K_S ratios

(0.065 and 0.070, respectively). Thus, purifying selection in such genes does not depend on male-female expression differences.

Figure **3-2**: A comparison of K_A/K_S ratios among escape genes with Y homologs, escape genes without Y homologs, heterogeneous genes, and inactivated genes (as computed for the human-chimpanzee-macaque orthologous trios). The number of genes considered is listed below each plot.

Fourth, to evaluate whether the possession of a Y homolog – another way to achieve dosage balance – might be related to the tendency to evolve under strong purifying selection, we compared K_A/K_S ratios between escape genes with vs. without functional Y homologs (Fig. 3-2), as classified in (Ross et al. 2005). Nine escape genes with functional Y copies (NLGN4X, CXORF15, EIF1AX, ZFX, USP9X, DDX3X, UTX, SMCX, and RPS4X) had lower K_A/K_S ratios

than the remaining 23 escape genes (0.029 vs. 0.132; $P = 0.0745$). Moreover, the median K_A/K_S ratio was significantly lower for escape genes with a Y homolog than for either inactivated genes (0.029 vs. 0.147; $P = 0.0025$) or heterogeneous genes (0.029 vs. 0.162; $P = 0.0135$), while the median K_A/K_S ratio was not significantly different for escape genes without Y homologs vs. either inactivated genes or heterogeneous genes (data not shown). Thus, strong purifying selection in escape genes is primarily determined by existence of functional Y homologs.

Table **3-1**: Multiple regression models for K_A/K_S ratio in X-linked genes.

Predictors	*P*-value	RCVE*
Y[†]	0.047	0.934
Dosage[§]	0.604 (NS[‡])	0.058
R^2		0.193

*RCVE: relative contribution to the variability explained. [†]Y: existing versus not existing Y homolog. [§]Dosage: dosage-compensated versus non-dosage-compensated. [‡]NS: not significant

Because some factors mentioned above (i.e., XAR/XCR evolutionary strata, possession of a Y homolog, XCI status, and dosage balance) might correlate with each other, we performed a multiple regression analysis to evaluate the relative contribution of each factor to explaining the total variability in the *KA/KS* ratio. Local recombination rate was also added to the model, because recombination affects the efficacy of natural selection. Using all X-linked genes with three predictors (XAR/XCR strata, XCI status, and recombination rate; Table S3-4), we were able to predict only 1.5% of variation in K_A/K_S, however the XCI status was the only marginally significant predictor. This is not surprising, because our data set is overloaded by inactivated genes, thus diluting any signal attributed to escape genes. In a separate regression model, we considered escape genes separately and explained variation in their *KA/KS* with two predictors (possession of a Y homolog and dosage balance; Table 3-1). This model explained 19.3% of variability in their *KA/KS* and indicated the possession of a Y homolog to be a significant

predictor. Thus, the XCI status (Escape vs. non-Escape) and possession of a Y homolog are in part the determining the *KA/KS* ratio of genes on the X.

Fifth, we questioned whether escape genes with Y-linked homologs influence selection acting on neighboring escape genes. This is of particular interest because escape genes are clustered, with each cluster containing an escape gene with a functional Y homolog (Carrel and Willard 2005). We investigated the K_A/K_S ratios in six clusters, each including at least one escape gene with a functional Y homolog and at least one additional escape gene (Table S3-4). For any cluster the median difference in K_A/K_S ratios was not significantly different between an escape gene with a Y homolog and other escape genes belonging to the same cluster vs. between an escape genes with a Y homolog and escape genes from other clusters (Table S3-4). Thus, escape genes with Y homologs do not determine the selective pressure of neighboring escape genes.

Finally, we examined whether distinct selective pressure acts on the expression of escape genes. Utilizing microarray expression data from brain, heart, kidney, and liver in six humans and five chimpanzees (Khaitovich et al. 2005), we compared the median ratio of human-chimpanzee expression divergence between escape and inactivated gene groups. In all four tissues escape genes had lower expression divergence between species than inactivated genes (Fig. S3-3). This difference was significant for heart (0.227 vs. 0.298; $P < 0.0001$) and kidney (0.118 vs. 0.231; $P < 0.0001$), but not significant for brain (0.092 vs. 0.096; $P = 0.4306$) and liver (0.274 vs. 0.289; $P = 0.4123$). These findings suggest stronger selective constraints for not only the coding region but also the expression of escape over inactivated genes. Here we could not study heterogeneous genes and escape genes with Y homologs separately, due to the paucity of tissue-specific data.

In summary, we observed that escape genes experience stronger purifying selection than inactivated genes at both the protein-coding and gene expression levels. This effect largely results from the importance of function and dosage of escape genes, as observed when their gene dosage is altered in individuals with Turner syndrome (Good et al. 2003; Bondy 2006). Strong purifying

selection acting at escape genes is primarily driven by genes possessing Y-linked homologs. Although some X and Y gametologous pairs retained similar functions (Watanabe et al. 1993; Sekiguchi et al. 2004), the majority of such pairs diverged in both function and expression (Wilson and Makova 2009). Y gametologs usually acquired male-specific functions (Skaletsky et al. 2003) and evolved more rapidly than the corresponding X homologs (Wyckoff, Li, and Wu 2002; Wilson and Makova 2009). Thus, escape genes possessing Y-homologs might evolve under strong purifying selection because they fulfill important functions different from their counterparts on the Y (see Table S3-5 here and Table S4 in (Wilson and Makova 2009). This conclusion is further supported by the fact that most candidate genes for Turner syndrome have Y chromosome homologs (Burgoyne 1989; Fisher et al. 1990; Ellison et al. 1997; Rao et al. 1997).

Alternatively, or additionally, strong purifying selection acting on escape genes with Y homologs might be due to the dominance effect. Indeed, if we assume that such genes evolve largely like autosomal genes (i.e., Y copy provides a potent copy of its X chromosomal counterpart), then their dominance is distinct from this for inactivated genes, or escape genes without a Y homolog, whose expression is always hemizygous. Differences in dominance might lead to differences in the K_A/K_S ratio, because mutations with small selective effects – such mutations are more likely to contribute to the nonsynonymous rate – tend to be more dominant (Kondrashov and Koonin 2004).

Methods

Classification of X-linked genes

The X-linked genes were classified into three groups: inactivated genes, i.e. genes that are subject to XCI in all females tested; escape genes, i.e. genes that escape XCI in all females

tested; or heterogeneous genes, i.e. genes that exhibit XCI in some, but not all, females assayed. The XCI profile was established by one of the two assays. A gene whose XCI profile was derived from the *rodent/human somatic cell hybrids assay* (Carrel and Willard 2005) was considered to escape XCI or be X inactivated if it was expressed in at least 8 hybrids (8 out of 9, or 9 out of 9) or at most 1 hybrid (0 out of 9, or 1 out of 9), respectively. Otherwise, a gene was assigned to the heterogeneous group (except for genes expressed in 2 out of 9, or in 7 out of 9 hybrids – such genes, representing borderline cases, were not classified to any of the groups).

In the *primary human fibroblast cell line assay* (Carrel and Willard 2005), a gene was considered to be expressed on the inactive X in a female if in a cell line derived from this female it was expressed from the inactive X with at least 5% of its expression level on the active X. If a gene was expressed in more than 80% of individual cell lines tested, then it was classified as an escape gene. If a gene was expressed in less than 20% of individual cell lines tested, then it was classified as an inactivated gene. Other genes were assigned to the heterogeneous group.

Identification of orthologous genes

Alignments of primate species, obtained from the 44-way multiz alignments at the UCSC Genome Browser (Karolchik et al. 2008), were utilized to identify orthologous gene trios (human-chimpanzee-macaque) and quintets (human-chimpanzee-orangutan-macaque-marmoset). Protein-coding sequence (CDS) regions for the four non-human primate species were determined by the genomic coordinates of human RefSeq mRNAs available at the UCSC Genome Browser (Karolchik et al. 2008).

For functional Y homologs, we considered genes expressed from the Y chromosome at both the RNA and protein level.

Permutation tests

To evaluate the statistical significance of differences in median K_A/K_S ratios between the three gene groups, we implemented a permutation test. The XCI status labels were removed from each gene, and genes were randomly reassigned to one of the three groups (i.e., escape, inactivated, or heterogeneous), each group having the same number of genes as in the corresponding original group. This procedure was repeated 10,000 times. The empirical *P* values were determined by counting the number of permutations in which the differences in median K_A/K_S ratios between permuted groups were greater than those between original groups. A similar permutation test was also carried out to determine the statistical significance in median difference in the K_A/K_S ratios between an escape gene with a Y homolog and other escape genes belonging to the same vs. other clusters. In each permutation for a particular cluster, its escape gene with a Y homolog was kept constant, while the cluster labels (i.e. belonging to this or other clusters) for other genes were removed and reassigned randomly, and median differences in the K_A/K_S ratios were calculated as mentioned above. This procedure was repeated 10,000 times. The empirical *P* values were determined by counting the number of permutations in which the differences in median K_A/K_S ratios between permuted groups were greater than those between original groups.

Statistical tests of selection

We utilized the *codeml* module in the PAML software package (Yang 2007) to calculate the K_A/K_S ratio and test various evolutionary models. The K_A/K_S ratio for each orthologous gene group was computed using the M0 (one-ratio) model. To evaluate whether the K_A/K_S ratio was significantly different from 1 for a particular orthologous gene group, twice the log-likelihood difference between models with the estimated K_A/K_S ratio vs. the fixed K_A/K_S ratio ($K_A/K_S = 1$)

was compared with the χ^2-distribution with one degree of freedom (Yang and Bielawski 2000). To test whether selection acted on individual codon sites, M1a (nearly neutral model) was compared with M2a (selection model), and M7 (β-distribution neutral model) was compared with M8 (β-distribution selection model), again in each case using the likelihood ratio test. Additionally, to detect codon sites evolving under positive selection in particular lineages, we evaluated the improved branch-site likelihood model (Yang and Nielsen 2002; Zhang, Nielsen, and Yang 2005). Each specific branch of the phylogeny was selected as foreground and the others were designated as background. Codon sites within the foreground branch were tested for positive selection similarly via a comparison between MA (selection) and MA0 (neutral) models. Absolute K_A and K_S values of all genes tested were listed in Table S3-7.

Local recombination rate

We obtained recombination rates from the UCSC Genome Browser decode data track for build hg18 (Kong et al. 2002).

Multiple regression analysis

RCVE (Kvikstad et al. 2007; Kelkar et al. 2008) was used to assess the contribution of each predictor to explaining the total variability:

$$RCVE = \frac{R^2_{full} - R^2_{reduced}}{R^2_{full}}$$

where R^2_{full} and $R^2_{reduced}$ are the R^2 for the full model and the model except for the predictor of interest, respectively. Linear multiple regression analysis was conducted using R statistical package.

References

Bondy, C. A. 2006. Turner's syndrome and X chromosome-based differences in disease susceptibility. Gend Med **3**:18-30.

Burgoyne, P. S. 1989. Mammalian sex determination: thumbs down for zinc finger? Nature **342**:860-862.

Carrel, L., and H. F. Willard. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature **434**:400-404.

Ellegren, H., and J. Parsch. 2007. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet **8**:689-698.

Ellison, J. W., Z. Wardak, M. F. Young, P. Gehron Robey, M. Laig-Webster, and W. Chiong. 1997. PHOG, a candidate gene for involvement in the short stature of Turner syndrome. Hum Mol Genet **6**:1341-1347.

Eyre-Walker, A., and P. D. Keightley. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol **26**:2097-2108.

Fisher, E. M., P. Beer-Romero, L. G. Brown, A. Ridley, J. A. McNeil, J. B. Lawrence, H. F. Willard, F. R. Bieber, and D. C. Page. 1990. Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. Cell **63**:1205-1218.

Good, C. D., K. Lawrence, N. S. Thomas, C. J. Price, J. Ashburner, K. J. Friston, R. S. Frackowiak, L. Oreland, and D. H. Skuse. 2003. Dosage-sensitive X-linked locus influences the development of amygdala and orbitofrontal cortex, and fear recognition in humans. Brain **126**:2431-2446.

Johnston, C. M., F. L. Lovell, D. A. Leongamornlert, B. E. Stranger, E. T. Dermitzakis, and M. T. Ross. 2008. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. PLoS Genet **4**:e9.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850-1854.

Nguyen, D. K., and C. M. Disteche. 2006. Dosage compensation of the active X chromosome in mammals. Nat Genet **38**:47-53.

Payer, B., and J. T. Lee. 2008. X Chromosome Dosage Compensation: How Mammals Keep the Balance. Annu Rev Genet.

Rao, E., B. Weiss, M. Fukami, et al., (17 co-authors). 1997. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. Nat Genet **16**:54-63.

Ross, M. T., D. V. Grafham, A. J. Coffey, et al., (282 co-authors). 2005. The DNA sequence of the human X chromosome. Nature **434**:325-337.

Sekiguchi, T., H. Iida, J. Fukumura, and T. Nishimoto. 2004. Human DDX3Y, the Y-encoded isoform of RNA helicase DDX3, rescues a hamster temperature-sensitive ET24 mutant cell line with a DDX3X mutation. Exp Cell Res **300**:213-222.

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, et al., (40 co-authors). 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423**:825-837.

Skuse, D. H. 2005. X-linked genes and mental functioning. Hum Mol Genet **14 Spec No 1**:R27-32.

Talebizadeh, Z., S. D. Simon, and M. G. Butler. 2006. X chromosome gene expression in human tissues: male and female comparisons. Genomics **88**:675-681.

Watanabe, M., A. R. Zinn, D. C. Page, and T. Nishimoto. 1993. Functional equivalence of human X- and Y-encoded isoforms of ribosomal protein S4 consistent with a role in Turner syndrome. Nat Genet **4**:268-271.

Wilson, M. A., and K. D. Makova. 2009. Evolution and survival on eutherian sex chromosomes. PLoS Genet **5**:e1000568.

Wyckoff, G. J., J. Li, and C. I. Wu. 2002. Molecular evolution of functional genes on the mammalian Y chromosome. Mol Biol Evol **19**:1633-1636.

Xu, J., and C. M. Disteche. 2006. Sex differences in brain expression of X- and Y-linked genes. Brain Res **1126**:50-55.

Zhang, L., and W. H. Li. 2005. Human SNPs reveal no evidence of frequent positive selection. Mol Biol Evol **22**:2504-2507.

Bergen, A. W., M. Pratt, P. T. Mehlman, and D. Goldman. 1998. Evolution of RPS4Y. Mol Biol Evol **15**:1412-1419.

Carrel, L., and H. F. Willard. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature **434**:400-404.

Christensen, J., K. Agger, P. A. Cloos, D. Pasini, S. Rose, L. Sennels, J. Rappsilber, K. H. Hansen, A. E. Salcini, and K. Helin. 2007. RBP2 belongs to a family of demethylases, specific for tri-and dimethylated lysine 4 on histone 3. Cell **128**:1063-1076.

Clemson, C. M., J. C. Chow, C. J. Brown, and J. B. Lawrence. 1998. Stabilization and localization of Xist RNA are controlled by separate mechanisms and are not sufficient for X inactivation. J Cell Biol **142**:13-23.

Hansen, R. S., T. K. Canfield, A. M. Stanek, E. A. Keitges, and S. M. Gartler. 1998. Reactivation of XIST in normal fibroblasts and a somatic cell hybrid: abnormal localization of XIST RNA in hybrid cells. Proc Natl Acad Sci U S A **95**:5133-5138.

Kahan, B., and R. DeMars. 1975. Localized Derepression on the Human Inactive X Chromosone in Mouse-Human Cell Hybrids. Proc Natl Acad Sci U S A **72**:1510-1514.

Karolchik, D., R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. 2008. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res **36**:D773-779.

Kelkar, Y. D., S. Tyekucheva, F. Chiaromonte, and K. D. Makova. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res **18**:30-38.

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.

Kvikstad, E. M., S. Tyekucheva, F. Chiaromonte, and K. D. Makova. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS Comput Biol **3**:1772-1782.

Takeuchi, T., Y. Yamazaki, Y. Katoh-Fukui, R. Tsuchiya, S. Kondo, J. Motoyama, and T. Higashinakagawa. 1995. Gene trap capture of a novel mouse gene, jumonji, required for neural tube formation. Genes Dev **9**:1211-1222.

Wang, W., L. R. Meadows, J. M. den Haan, N. E. Sherman, Y. Chen, E. Blokland, J. Shabanowitz, A. I. Agulnik, R. C. Hendrickson, C. E. Bishop, and et al. 1995. Human H-Y: a male-specific histocompatibility antigen derived from the SMCY protein. Science **269**:1588-1590.

Wilson, M. A., and K. D. Makova. 2009. Evolution and survival on eutherian sex chromosomes. PLoS Genet **5**:e1000568.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**:1586-1591.

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol **15**:496-503.

Yang, Z., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol **19**:908-917.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22**:2472-2479.

Zinn, A. R., R. K. Alagappan, L. G. Brown, I. Wool, and D. C. Page. 1994. Structure and function of ribosomal protein S4 genes on the human and mouse sex chromosomes. Mol Cell Biol **14**:2485-2492.

Chapter 4

Studies of boundary elements model at the boundary region between escape and inactivated genes

Introduction

According to the recent work by Carrel and Willard (Carrel and Willard 2005), approximately 15% (94/612) of genes assayed on the X chromosome escape X inactivation. This indicates that they are expressed from both the active and inactive X chromosomes, and frequently cause dosage imbalances between males and females. Note that some escape genes achieve dosage compensation because of possessing Y homologs (Johnston et al. 2008). Escape genes also tend to experience stronger purifying selection than inactivated genes at both protein-coding and gene expression levels (Park et al. 2010).

What molecular mechanism(s) make(s) genes escape X inactivation? It has been shown that some histone modifications are highly associated with heterochromatin and silencing for XCI (Chadwick and Willard 2004; Marks et al. 2009), and to control these types of heterochromatin, two not mutually exclusive models have been proposed (Prothero et al. 2009). First, when inactive heterochromatin is propagated, "way stations" positioned throughout the inactive X chromosome might be required ("the way stations model"). If a gene resides far from the way stations, such gene is likely to be an escape gene. Second, a genomic element such as insulator or barrier (for example, CTCF binding site (Filippova et al. 2005)) at the boundary region between escape and inactivated genes may prevent the spread of the inactive X heterochromatin ("the boundary elements model"). A combination of these two models is also possible. Way stations

might boost the propagation of inactive heterochromatin, while boundary elements might prevent the inactivation signals from reaching escape genes.

Although to date many studies have been carried out to test the way stations model and have partially supported this model (reviewed in (Prothero et al. 2009) and see Chapter 2), it is insufficient in cases when escape and inactivated genes are in close proximity; and many escape and inactivated genes are indeed juxtaposed (Tsuchiya et al. 2004; Carrel and Willard 2005). To validate sequences identified in our previous work (Chapter 2) and to enable detection of sequences important for regulation at XCI boundaries, we undertook a comparative genomics approach under the hypothesis that conserved XCI patterns may be regulated by conserved sequences. Here, to find sequences (located in the boundary regions between escape and inactivated genes) that may influence X inactivation status, we utilized the following three approaches: (1) identified overrepresented motifs in the boundary regions using chromosome-wide human XCI data (first approach); (2) identified motifs uniquely present in a boundary region (as compared with non-boundary regions) studied in detail in a number of eutherian mammals (see below)(second approach), and (3) identified motifs present in a boundary region studied in detail in a number of eutherian mammals (not requiring uniqueness of these motifs in this boundary)(third approach). We also asked whether any underlying genomic factors (i.e., CTCF insulators, histone modifications, CpG islands, and transposable elements; see Chapter 1) are likely to function as boundary elements separating genes with different X inactivation profiles.

Results

A comprehensive computational analysis of boundary regions using chromosome-wide human XCI data (first approach)

Defining boundary regions separating genes with different X inactivation patterns

Using information about the XCI profile in human (based on the rodent/human somatic cell hybrids assay and the primary human cell line assay (Carrel and Willard 2005)), we focused on nine escape gene clusters and 18 boundary regions surrounding them (Table 4-1). In this study, a set of escape genes consisting of at least one escape gene (no inactivated or heterogeneous genes were included; see Methods for detail) was defined as an escape gene cluster, and intergenic sequences residing between transcription start and transcription end sites of genes with different XCI profiles (i.e., one from escape gene cluster and the other one from its closest neighboring inactivated gene) were referred to as a boundary region. Of the nine escape gene clusters, five and four were located in the X-added region (XAR) and in the X-conserved region (XCR), respectively. Using a working assumption (based on experimental results from the Carrel laboratory) that XCI profile is conserved among eutherian mammals excluding rodents, we extracted human-chimpanzee-macaque-dog alignments (from the 28-way vertebrate alignments (Miller et al. 2007)) corresponding to the human coordinates of the boundary regions. Nucleotide substitution rates at the boundary regions were compared with these at the control regions (here, we used intron regions that were located within 250 kb on either side of the middle point of each boundary region as the control sites, which were considered to evolve neutrally; see Methods for details). As a result, approximately 51.7 kb of boundary regions had significantly lower substitution rates than the neighboring introns ($P < 0.05$, permutation test, see Methods for details; Table S4-1). We searched for overrepresented motifs in these conserved sequences.

Table **4-1**: Boundary regions surrounding escape gene clusters for the first approach.

Escape gene cluster			Length of boundary regions (upstream/downstream)	Evolutionary stratum
5' border gene	Gene lists in cluster*	3' border gene		
PRPS2	FAM9C RAB9A TRAPPC2 (SEDL)** OFD1 GPM6B GEMIN8 (FAM51A1)	GLRA2	659 kb (169 / 490)	Stratum 3
PIGA	TMEM27 CA5BP (CA5BL) CA5B AP1S2 GRPR CTPS2 S100G (CALB3) SYAP1 CXorf15 RBBP7	SCML1	341 kb (41 / 300)	Stratum 3
MAP7D2 (FLJ14503)	EIF1AX	RPS6KA3	18 kb (10 / 8)	Stratum 3
KLHL15	EIF2S3 ZFX	PDK3	277 kb (28 / 249)	Stratum 3
ATP6AP2	CXorf38 MED14 (CRSP2) USP9X DDX3X	CASK	163 kb (22 / 141)	Stratum 3
RBM10	UBA1 (UBE1) INE1 PCTK1	ZNF157	129 kb (4 / 125)	Stratum 2
TSPYL2	JARID1C IQSEC2 (KIAA0522) SMC1A (SMC1L1)	HSD17B10 (HADH2)	103 kb (103 / 0.4)	Stratum 2
PDZD4 (PDZK4)	L1CAM AVPR2 ARHGAP4 ARD1A RENBP HCFC1	IRAK1	60 kb (31 / 29)	Strata 1
G6PD	GAB3	DKC1	105 kb (100 / 5)	Stratum 1

*- gene name from RefSeq; ** - gene name from (CARREL and WILLARD 2005)

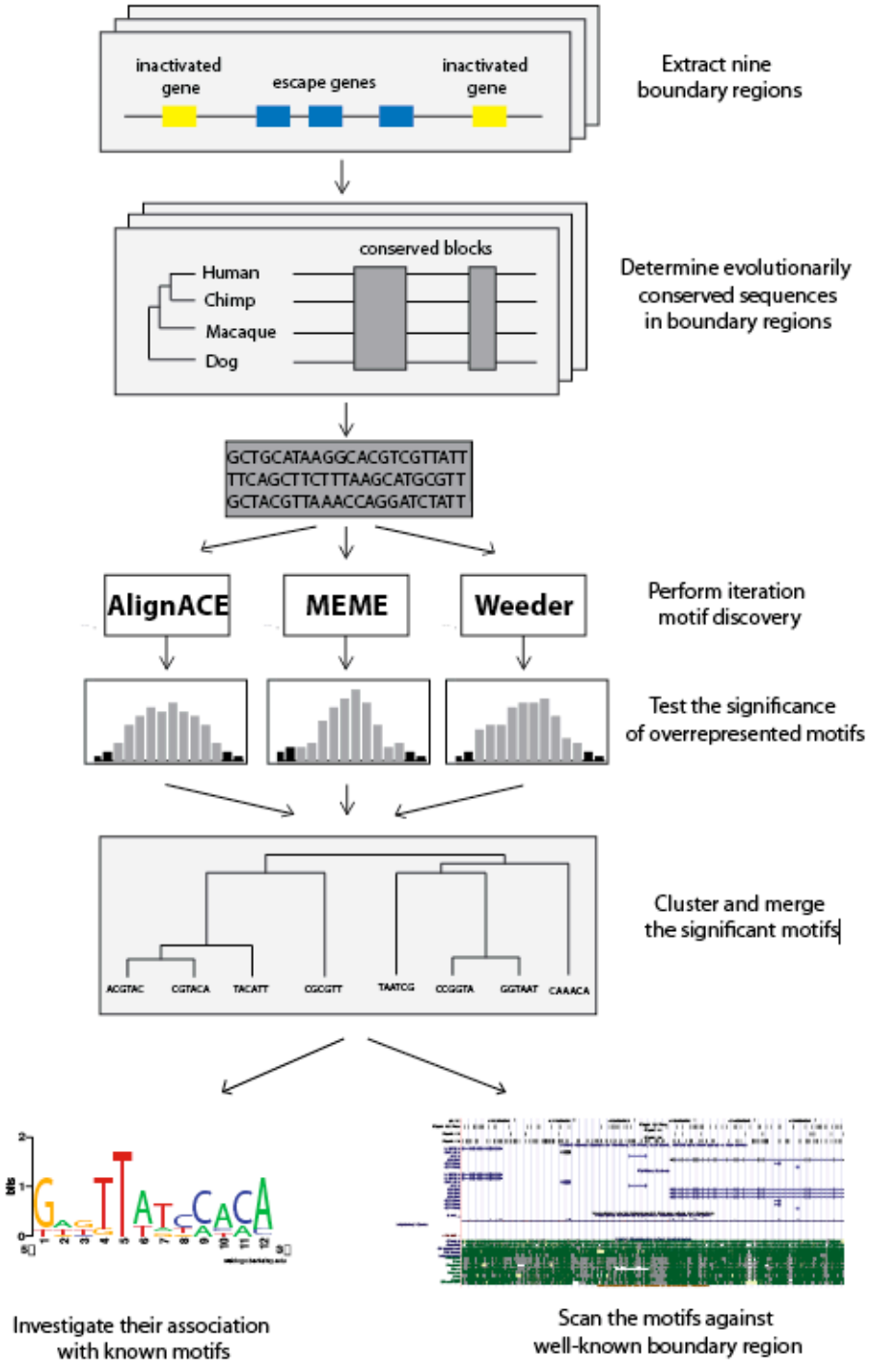


Figure **4-1**: Computational pipeline for identification of candidate motifs in boundary regions for the first approach.

Finding overrepresented motifs

We established a computational pipeline to identify candidate motifs in the boundary regions between escape and inactivated genes on inactive X chromosome (Fig. 4-1). Three different motif discovery algorithms (AlignACE (Hughes et al. 2000), MEME (Bailey and Elkan 1994), and Weeder (Pavesi et al. 2004)) were applied to identify motifs, and the hypergeometric enrichment score (Harbison et al. 2004) was calculated to measure the statistical overrepresentation of each motif.

After clustering overlapping motifs identified by these three algorithms, 81 non-redundant overrepresented motifs were identified (Table S4-2). We compared the sequences of these 81 motifs with the sequences of known transcription factor binding sites in JASPAR CORE database (version 3.0) (Vlieghe et al. 2006). Ten out of 81 motifs significantly overlapped with known transcription factor binding sites (Table 4-2). Interestingly, seven out of ten motifs represented binding sites of transcription factors containing zinc finger domains. This finding echoes other results suggesting an association of zinc finger containing proteins (e.g., CTCF) with XCI (Filippova et al. 2005; Ciavatta et al. 2006).

Table **4-2**: Known transcription factor binding sites overlapping with discovered motifs from the first approach.

Class	Number of discovered motifs
ZN-FINGER, C2H2	7
REL	1
HMG	1
STAT	1
Total	**10**

Reinvestigation of the motifs found in well-studied boundary regions

We explored in detail the distribution of 81 non-redundant motifs in the boundary region between ATP6AP2 (inactivated gene) and CXorf38 (escape gene), because, in an experimental analysis conducted in the Carrel laboratory, this boundary was found experimentally to be conserved in a number of eutherian mammals, except for rodents (Fig. 4-2). Interestingly, we identified three motifs (TTTAAAGCCAAG, GGAAATATCCAA, and CAGAAAAGGCAGA) that were (1) overrepresented in this boundary region as compared with intergenic regions that were surrounded by genes having the same XCI profile (defined as non-boundary regions); and (2) conserved among human, chimpanzee, macaque, and dog, but not mouse and rat. Although these motifs might be candidate motifs for XCI that can be tested in future wet-lab experiments, further studies are necessary to confirm this because the same motifs were also observed in both upstream and downstream of ATP6AP2-CXorf38 boundary region (data not shown). This suggests that, according to the boundary elements model, they are unlikely to discriminate escape genes from inactivated genes on the inactive X chromosome. For our further analysis, we focused on one escape gene cluster (consisting of four escape genes, namely, CXorf38, MED14, USP9X, and DDX3X; herein called USP9X gene cluster) including two boundaries (ATP6AP2-CXorf38 and DDX3X-CASK; Fig. 4-2) because, as mentioned above, conservation of these boundaries was confirmed by experimental analyses in several eutherian species.

A comprehensive computational analysis for boundary regions in USP9X gene cluster

To identify candidate boundary elements in two regions (ATP6AP2-CXorf38 and DDX3X-CASK boundaries), we considered two approaches based on whether uniqueness of the elements within the boundary regions should be required: (1) we searched for candidate sequence

elements uniquely present at the boundary region between escape and inactivated genes (this corresponds to the second approach above); (2) we searched for candidate sequence elements overrepresented in the boundary region (but they were not required to be unique) (this corresponds to the third approach above).

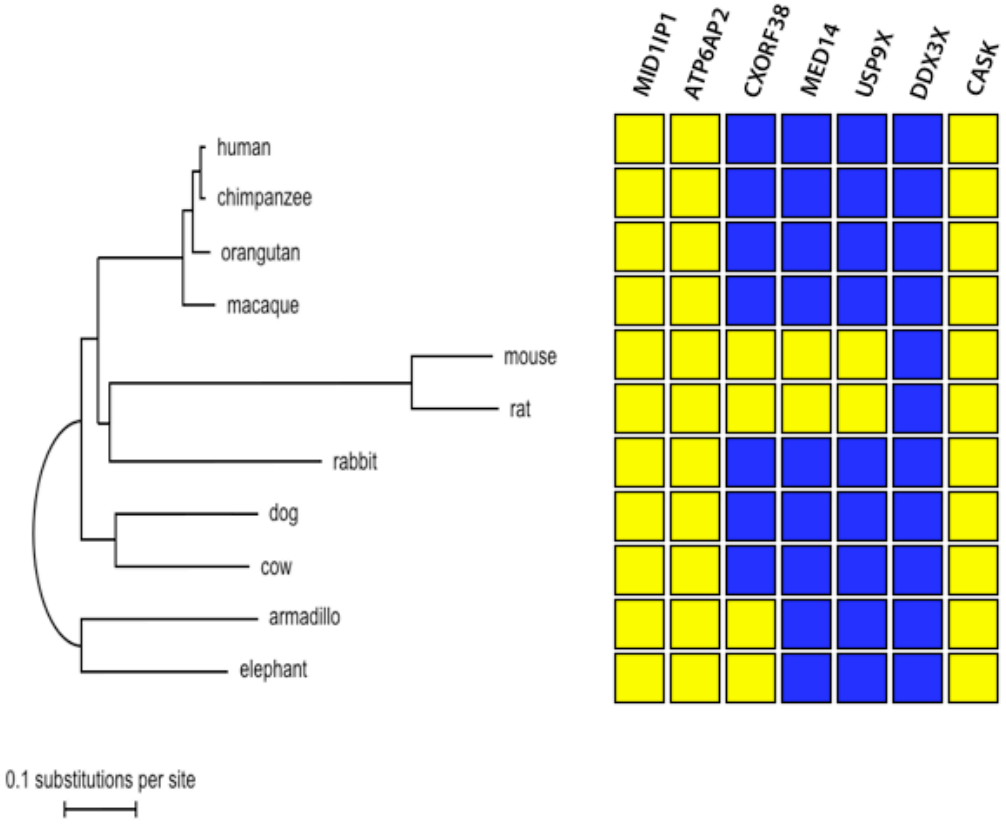


Figure **4-2**: XCI in mammals. Blue and yellow depict escape and inactivated genes, respectively. A Phylogenetic tree was built by (Murphy et al. 2007) and (Prasad et al. 2008).

Preliminary comparative sequence analysis of boundary regions

We first examined sequence similarity of the boundary region in several species to test the possibility that boundary regions possess specific DNA sequences controlling inactive X expression. If particular motifs functioning as insulators to prevent the spread of inactivation exist

in the boundary region, we may observe matching genomic segments between two boundary sequences from species having the same XCI profile. Several diagonal matches were detected in comparison between each pair of non-rodent species, and between each pair of rodent species especially in the ATP6AP2-CXorf38 boundary (Fig. 4-3). However, such diagonal matches were disrupted between non-rodent and rodent species. These observations are consistent with the XCI similarity and difference between species and suggests that candidate boundary elements might exist in the ATP6AP2-CXorf38 boundary, and should be conserved among species having the same XCI profile.

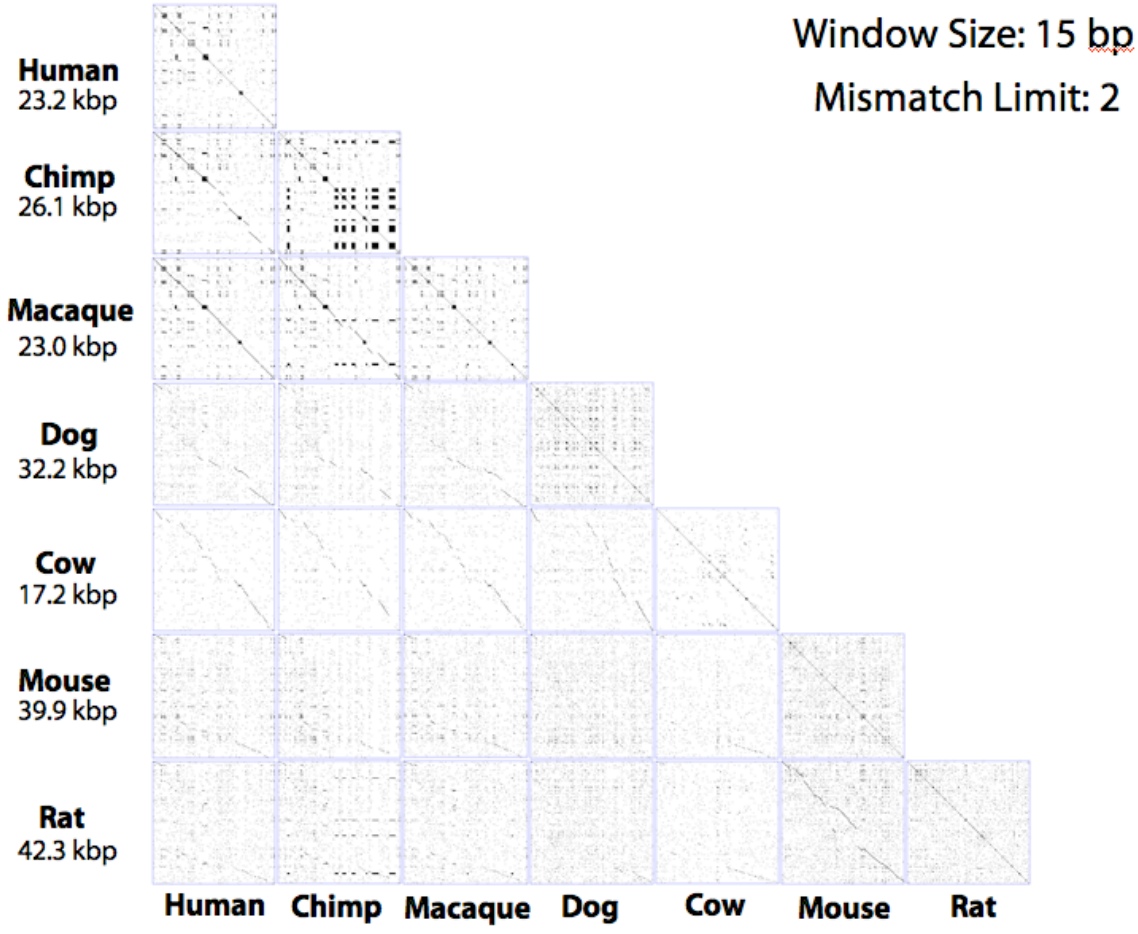


Figure **4-3**: Dot pots of ATP6AP2-Cxorf38 boundary sequences.

Identification of motifs uniquely present in the boundary region (second approach)

We assume that the XCI profile conserved among several species was inherited from their last common ancestor, and thus, should be under control of conserved candidate elements. Accordingly, we identified motifs from conserved sequence blocks in boundary regions. Based on the experimentally validated XCI profile for genes in the USP9X gene cluster of several eutherian species (namely, XCI profiles are conserved among all eutherian mammals assayed except rodents; armadillo and elephant were not taken into account in this study because additional analyses are required to confirm their XCI profiles; Fig. 4-2), we limited our analysis to four species (i.e., human, chimpanzee, macaque, and dog; for which high coverage and quality annotation are now available) and extracted their multiple sequence alignments from the UCSC 44-way vertebrate alignments (hg18) (Karolchik et al. 2008) using the human coordinates of the boundary regions. In this analysis, we also added the intragenic regions of genes composing boundaries (i.e., intron sequences from ATP6AP2, CXorf38, DDX3X, or CASK) into the boundary regions to consider the possibility that boundary elements might be present in introns. After comparing nucleotide substitution rates at the boundary region with those at their neighboring intron sequences, which were used as neutrally evolving controls (see Methods for details), we identified that approximately 1.7 kb of ATP6AP2-CXorf38 and 26.3 kb of DDX3X-CASK boundaries had significantly lower substitution rates than their control regions respectively (False Discovery rate (FDR) - q-value < 0.5 corresponding to $p < 0.038$ and $p < 0.026$ for each boundary, respectively). When ancestral repeats (AR; (Hardison et al. 2003)) and four-fold degeneration sites (4D) were used as control regions, we observed similar substitution rates (i.e., 0.362 for intron, 0.349 for AR, and 0.355 for 4D in ATP6AP2-CXorf38 boundary; 0.339 for intron, 0.340 for AR, and 0.340 for 4D in DDX3X-CASK boundary).

Table **4-3**: Number of unique motifs in boundary regions from the second approach.

	ATP6AP2-CXorf38 boundary			DDX3X-CASK boundary		
	Motif size					
Matching type	12 bp	14 bp	16 bp	12 bp	14 bp	16 bp
Perfect	74	107	145	1655	3655	3202
One mismatch	126	1987	2527	3443	43738	54946
Two mismatch	16	6406	19417	997	129026	419878

Note: there are no unique motifs with 8 and 10 bp

To identify motifs that were unique to boundary regions, we used the following procedure. (Note that (1) these analyses were performed separately for the ATP6AP2-CXorf38 and DDX3X-CASK boundaries (see Fig. S4-1), (2) initially 8-, 10-, 12-, 14-, and 16-mers were examined separately for each boundary, and (3) motifs with up to two mismatches were allowed.) First, from conserved sequence blocks in a boundary region, all possible motifs found were enumerated. They were required be observed in all four species. Second, motifs present also in non-boundary regions (see Fig. S4-1) were excluded. Third, a number of unique motifs were found in the boundary regions (Table 4-3). After merging overlapping motifs into longer motifs (see Methods for details) in order to avoid scoring them multiple times, we found 145 and 1,536 different motifs for the ATP6AP2-CXorf38 and DDX3X-CASK boundaries, respectively. Interestingly, among these motifs, 24 motifs were observed in both boundaries (herein called "shared motifs"), indicating that they are potential candidates for the boundary elements. To test whether the shared motifs may act as boundary elements controlling inactive X expression, rodent XCI profiles were utilized additionally because there is a shift in the XCI boundary between rodents and four other eutherin species (Fig. 4-2). If the motifs are absent from rodent APT6AP2-

CXorf38 regions, but are located in rodent DDX3X-CASK boundary regions (or possibly located in rodent USP9X-DDX3X regions; Fig. 4-2), then such motifs are the strongest candidates for boundary elements. We scanned shared motifs against rodent USP9X gene cluster region as well as rodent APT6AP2-CXorf38 and DDX3X-CASK boundary regions. 14 shared motifs (out of 24) were absent from the rodents. Seven shared motifs existed in both boundary and non-boundary regions of rodents. Rodent ATP6AP2-CXorf38 region had two shared motifs, even though this is not a boundary region in rodent (i.e., rodent ATP6AP2 and CXorf38 are inactivated genes), indicating they are just conserved motifs in this region unrelated to the XCI control. Notably, one shared motif (TAGTTTGGTGGGGCT) was observed in rodent DDX3X-CASK boundary region (Fig. 4-4). This finding can be explained by assuming that rodents might have lost the boundary element in the ATP6AP2-CXorf38 region, and the boundary element might be conserved in the DDX3X-CASK region. Although this one shared motif might be the candidate motif for XCI that can be tested in future wet-lab experiments, however, further studies are necessary to confirm this. If the conserved motif acts as the boundary element, it is likely to be located in the orthologous regions of non-rodent species. However, despite that the mouse DDX3X-CASK boundary contains the shared motif, it was incongruently observed in the intragenic region of CASK and of ATP6AP2 in human (Fig. 4-4).

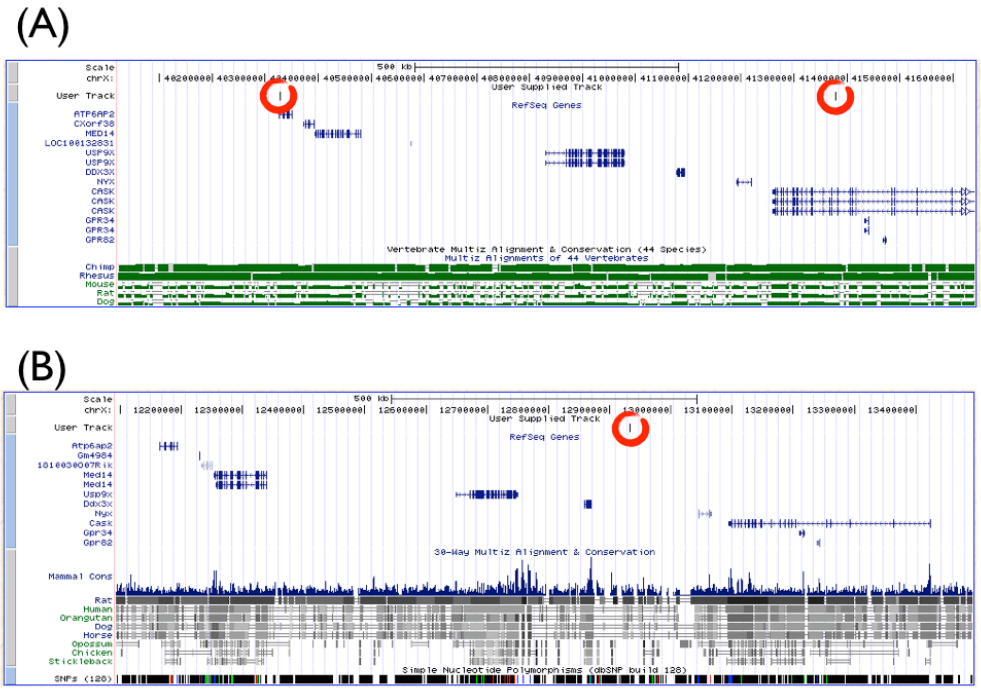


Figure **4-4**: Genomic locations of shared motif (NTAGTTTGGTGGGGCTN) identified by using the second approach in human (A) and mouse (B). Red circle indicates the shared motif.

Identification of motifs that are enriched at the boundary regions (third approach)

Alternatively, we attempted to identify motifs present in the boundary regions, but without requiring them to be uniquely present in such regions. We employed the same computational approaches that we described above for finding motifs uniquely present in the boundary regions except for the following modifications. First, motifs were identified directly from boundary regions and not from their conserved sequence blocks. This was done because we might have failed to detect conserved motifs if they resided in non-aligning regions among species (Dermitzakis et al. 2003; Siggia 2005). Second, we only considered 8-mers with up to four mismatches (i.e., the shortest motifs allowing more degenerate sites). Finally, we did not consider the possibility that boundary elements might be located at the intragenic regions of genes

composing boundaries (i.e., ATP6AP2, CXorf38, DDX3X, or CASK), and thus we only considered intergenic regions (demarcated by transcription end and start sites).

Table **4-4**: Number of motifs present in the boundary regions, regardless of their uniqueness from the third approach.

	Motif size
Matching type	8 bp
Perfect	179
One mismatch	363
Two mismatch	669
Three mismatch	11
Four mismatch	0

Using boundary regions separating genes with different X inactivation patterns in rodents (mouse and rat) as well as four other eutherian species (human, chimpanzee, macaque, and dog) at once, we identified motifs that were present in the ATP6AP2-CXorf38 boundaries in four species excluding two rodents (i.e., the motifs were absent from rodent APT6AP2-CXorf38 regions) and simultaneously in the DDX3X-CASK boundaries in all six species (Table 4-4 and Fig. S4-2). Overlapping motifs were merged and, as a result, 290 different motifs were found. When the frequencies of these motifs were visualized with respect to the USP9X gene cluster in human, a relatively large number of motifs was observed at both the ATP6AP2-CXorf38 and DDX3X-CASK boundaries (Fig. 4-5), suggesting that these motifs might act as boundary elements and successfully capture differences in XCI. Interestingly, many of these motifs were also observed around the BCOR gene region (Fig. 4-5). Although the BCOR's XCI profile has not been assayed yet, the possibility of involvement of this gene in determining the XCI profile of

surrounding genes (especially the USP9X escape gene cluster) needs to be evaluated in future studies.

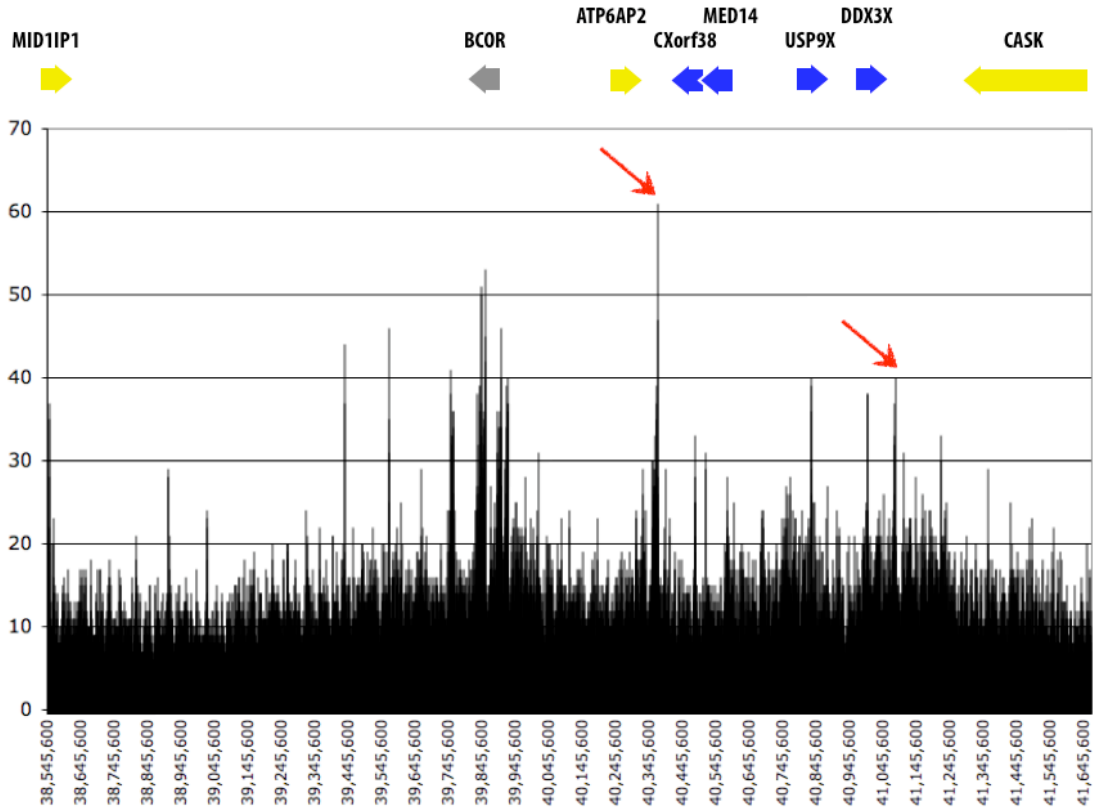


Figure **4-5**: Frequency of motifs overrepresented in the boundary regions, regardless of their uniqueness according to the third approach. X-axis indicates human genomic coordinates. Y-axis indicates number of motifs. Blue and yellow arrows depict escape and inactivated genes, respectively. Gray arrow depicts a gene whose XCI profile has not been assayed yet. Red sharp arrows indicate the highest peak in each boundary region.

Genomic factors examined are not sufficient to explain how genes escape XCI according to the boundary elements model

We asked whether any underlying genomic factors (i.e., CTCF insulators, histone modifications, CpG islands, and transposable elements; see Chapter 1) are likely to function as boundary elements separating genes with contrasting XCI patterns. We investigated the

distribution of these features at the USP9X gene cluster and its two boundaries vs. the neighboring genomic regions including closest inactivated genes (i.e., ATP6AP2, MID1IP1 and CASK) in several eutherian mammals.

Transposable elements. Long terminal repeat element (LTR) and long interspersed repetitive element 1 (L1) were tested in human, dog, mouse, and rat USP9X cluster, because escape genes regions are depleted of such elements (Tsuchiya et al. 2004; Ross et al. 2005; Carrel et al. 2006). These transposable elements were correlated with neither escape nor inactivated genes (Fig. S4-3). The ATP6AP2-CXorf38 and DDX3X-CASK boundaries also had no distinct LTR (Fig. S4-3A) and L1 (Fig. S4-3B) distribution patterns, as compared with neighboring genomic regions including ATP6AP2, MID1IP1 and CASK genes.

CpG islands. The genes within the USP9X gene cluster in human and mouse had conserved CpG islands in their promoter regions (Fig. S4-4) in spite of the fact that only DDX3X is an escape gene in mouse (Fig. 4-2). Although, interestingly, the ATP6AP2-CXorf38 boundary had one CpG island in human, but none in mouse; no difference in the number of CpG islands was observed in the DDX3X-CASK boundary between human and mouse (Fig. S4-4).

CTCF insulators. We observed that the genomic occupancy of CTCF binding sites appears to be almost randomly distributed in the USP9X gene cluster of human, dog, mouse, and rat (Fig. S4-5). There were no CTCF binding sites in their ATP6AP2-CXorf38 boundary, and only human and dog had CTCF binding sites in the DDX3X-CASK boundary.

Histone modifications. We tested whether trimethylated H3K27 in particular could reveal XCI profiles on USP9X gene cluster, because escape genes tend to be depleted of H3K27me3 (Marks et al. 2009; Yang et al. 2010). As found in previous studies (Marks et al. 2009; Yang et al. 2010), some escape genes (e.g., USP9X and DDX3X) were clearly marked by the absence of H3K27me3 (Fig. S4-6). However, there was no manifest signal difference between the other escape genes (i.e., CXorf38 and MED14) and inactivated genes. The H3K27me3 profiles also did

not display differentiation between each boundary region and its adjacent up/downstream regions (Fig. S4-6).

Overrepresented oligomers from (Carrel et al. 2006). Finally, we explored whether primary DNA sequence elements found in our previous study (Carrel et al. 2006) could determine expression status of the genes in the USP9X gene cluster and of the neighboring inactivated genes (e.g., including ATP6AP2, MID1IP1 and CASK). In human, as expected, the USP9X gene cluster was enriched in escape oligomers and depleted of inactivated oligomers (Fig. S4-7). However, these patterns were not consistent in dog, mouse and rat. Our oligomers also did not discriminate between each boundary region and its adjacent up/downstream regions.

In summary, although several genomic hallmarks described above have been proposed to explain the XCI mechanism (see Chapter 1), we found no evidence that they account for escape gene expression, and how escape genes regulation especially in USP9X gene cluster is controlled should be subject to future research.

Caveats to our approaches

Our analyses may have several caveats. First, the difficulty of the computational identification of regulatory elements is well recognized (Harbison et al. 2004; Hu et al. 2005; MacIsaac and Fraenkel 2006). Although we adopted the computational approach to score imperfect matches and thus to identify functionally significant non-conserved motifs, such moderate stringency decreased the power of distinguishing escape from inactivated genes. Second, recently, Chow et al (Chow et al. 2010) suggested that inactivated genes in a boundary region can be repressed by endogenous interfering (si)RNAs. This finding may propose that the possibility that potential boundary elements may be located outside of the ATP6AP2-CXorf38 and DDX3X-CASK intergenic regions. Third, although there was no clear evidence that several genomic

hallmarks (i.e., CTCF insulators, histone modifications, CpG islands, and transposable elements) play an important role in XCI mechanism escape gene expression, we can not rule out the possibility of the involvement of some epigenetic factors in the escape gene expression. XCI profile was assayed in the fibroblast cell lines by measuring the level of DNA methylation; however, CTCF binding and histone modification data were derived from many different non-fibroblast cells. Indeed, inactive X expression levels and functional activity of regulatory elements (i.e., enhancers and insulators) tend to vary between different tissues (Talebizadeh et al. 2006; Xi et al. 2007; Heintzman et al. 2009).

Conclusions

We utilized an experimentally derived inactivation profile from several eutherian mammals and developed bioinformatic approaches to identify candidate sequences that potentially determine inactivation status of X-linked genes. We found some unique or overrepresented motifs in boundary regions, indicating that they are the candidates for the boundary elements separating genes with different XCI profiles. The resulting data set needs to be evaluated in future experimental studies and may provide valuable insights into regulation of escape gene expression.

Methods

Analysis using genome-wide human XCI data (first approach)

Boundary regions

Using X chromosome inactivation (XCI) profile from rodent/human somatic cell hybrids (Carrel and Willard 2005), genes were classified into escape or inactivated genes if they were expressed in at least 8 hybrids (8 or 9 out of 9) or at most one hybrid (0 or 1 out of 9), respectively. Each escape gene cluster was required to consist of at least one escape gene (we excluded genes on pseudoautosomal regions and genes adjacent XIST RNA) and to be surrounded by only inactivated genes. (namely, particular gene clusters were excluded if their adjacent genes (1) did not have XCI profiles or (2) had heterogeneous expression patterns (i.e., neither escape nor inactivated genes)). To define two boundary regions for each escape gene cluster (i.e., the upstream and downstream intergenic regions of the escape gene cluster), the following two criteria were considered: (1) genes that had XCI profile represented by ESTs were excluded; (2) to alleviate the effects altering the XCI of genes due to the lack of epigenetic features in somatic cell hybrids (Kahan and DeMars 1975; Clemson et al. 1998) we additionally utilized XCI profile from (Kahan and DeMars 1975) primary human cell assay (Carrel and Willard 2005), and gave preference to its assignments if two assay systems led to incongruent classification for particular genes. In the primary human fibroblast cell line assay (Carrel and Willard 2005), a gene was considered to be expressed on the inactive X in a female if in a cell line derived from this female it was expressed from the inactive X with at least 5% of its expression level on the active X. If a gene was expressed in more than 80% of individual cell lines tested, then it was classified as an escape gene. If a gene was expressed in less than 20% of

individual cell lines tested, then it was classified as an inactivated gene. Other genes were assigned to the heterogeneous group.

Evolutionarily conserved sequences in boundary regions

From the 28-way vertebrate genomic alignments (Miller et al. 2007) available from the UCSC Genome Browser, we extracted a set of human genome segments where all four species (i.e., human, chimpanzee, macaque, and dog) were aligned on X chromosome. To determine sequences under selective constraints in boundary regions, a multiple alignment corresponding to each boundary region was broken into 100 bp windows (because this size is likely to provide enough resolution to identify functional regulatory elements (Stone et al. 2005)), and nucleotide substitution rate for each window was estimated using the REV model (Yang 1993) implemented in the BASEML module of the PAML package (Yang 2007). The level of selective constraint was evaluated by comparison between substitution rate in each window of the boundary regions and that in neutrally evolving control regions. We used intron regions that were located within 250 kb on either side of the middle point of each boundary region as the control sites (Gaffney et al. 2008). First intron and splice site (first and last 100bp of each intron) regions were eliminated in our control sites because they tend to evolve under functional constraints (Keightley and Gaffney 2003; Gaffney and Keightley 2006). The nucleotide substitution rate for each control region was estimated as described above and compared with that for each window of the corresponding boundary region.

To test whether the substitution rate in each window from a boundary region is significantly lower than that in corresponding control site, we implemented a statistical test. Similar to the boundary regions, each control region was broken into 100-bp windows, and the number of windows with substitution rates lower than or equal to the observed substitution rate

from window in boundary region were used to compute an empirical *P* value. If at most 50 control windows had lower substitution rate than the boundary region window compared ($P < 0.05$), such window was defined as having sequences under strong selective constraints. We concatenated all these windows from both upstream and downstream regions of each active gene cluster and utilized them (total 9 boundary blocks) for the further study.

Motif discovery

To find overrepresented motifs from evolutionarily conserved sequences in boundary regions, we applied three different motif discovery programs: AlignACE (Hughes et al. 2000), MEME (Bailey and Elkan 1994), Weeder (Pavesi et al. 2004). The AlignACE which is based on Gibbs sampling algorithm was run 10 times with default settings and with different random number seeds because this approach can increase the motif space and thus suffers from the local maximum problem only minimally. From each running of AlignACE, top 50 motifs that were sorted in descending order of the MAP score (maximum a priori log likelihood) were requested for further analysis (total 500 motifs with redundancy). MEME involving expectation-maximization algorithm was run using "zoops" (zero or one occurrence per sequence) model and option with reverse complement ("-revcomp"). The program MEME was carried out twice: once under motif width range of 7 to 12 and the next time under their range of 13 to 18. As a result, 100 (top 50 motifs from each run) motifs were retained for further analysis. To run Weeder which uses enumerative method, we utilized "large" option (i.e., it reports motifs of the specified sizes with mismatch: 6, 8, 10, and 12 motif length with 1, 2, 3, and 4 mismatch, respectively) and obtained top 50 motifs from each run for different motif length (a total of 200 motifs).

Statistical significance of motifs

To examine the statistical overrepresentation of the output motifs, an enrichment score for each motif was computed (Harbison et al. 2004). The score indicates the degree to which the frequency of the motif in the boundary regions would be equal or greater than that in non-boundary regions. A *P*-value for the enrichment score was calculated by the following formula (Harbison et al. 2004):

$$p = \sum_{i=b}^{\min(B,t)} \frac{\binom{B}{i}\binom{T-B}{t-i}}{\binom{T}{t}}$$

where B = number of boundary regions (i.e., 9); T = total number of intergenic regions (i.e., sum of B and number of non-boundary regions); b = number of boundary regions that match a target motif; and t = number of intergenic regions that match a target motif.

To calculate the *P*-value for each motif, we needed to define (1) what is meant by the non-boundary regions and (2) the optimal cutoffs for the position weight matrix (PWM) of a target motif (e.g., a sequence may be considered to contain the target motif if at least one site in the sequence aligns to the target motif with a certain cutoff thresholds of the PWM of its motif). For the first case, 147 intergenic regions that were surrounded by genes having same XCI profile were extracted and used as non-boundary regions. For the second case, we implemented a permutation test to determine the optimal cutoff value for each motif. The sequences of each boundary block were randomized 100 times and a set of 900 randomized sequences was generated. These randomized sequences were used here as a control set because they are no longer able to correspond to binding sites, but their nucleotide frequencies are identical to original frequencies. We scanned these control sequences for the PWM of each motif and obtained candidate motifs having a minimal PWM score of 0.5. Then, a PWM score was retrieved at the

top 1% percentile ($P < 0.01$), and was used as a cutoff of the corresponding motif. All motifs from AlignACE and Weeder except for MEME (because it already provides probability matrices) were converted to PWMs using pseudocount value of 1 (Nishida et al. 2008).

Clustering motifs

Although the use of multiple different motif discovery methods improves performance of finding transcription factors (Harbison et al. 2004; Hu et al. 2005), combining these outcomes might typically produce redundancy. We used MATLIGN (Kankainen and Loytynoja 2007) to cluster similar motifs and generate a set of non-redundant motifs. Briefly, Euclidean distance (because it was evaluated as having significantly better performance than the other methods (Gupta et al. 2007)) and agglomerative hierarchical clustering were used as scoring function and clustering method, respectively. The statistical significance of similarity between motifs was assessed by False Discovery rate (FDR; (Benjamini and Hochberg 1995)) of < 1%.

Interpretation of motifs

To compare between discovered non-redundant motifs and known transcription factor binding sites, we aligned the motifs against all binding sites in JASPAR CORE database (version 3.0) (Vlieghe et al. 2006) using MATLIGN (Kankainen and Loytynoja 2007) with the Euclidean distance scoring function. For each motif one binding site having the highest score was reported if this mapping was significantly supported by FDR of < 10%. We used Patser (Hertz and Stormo 1999) with default options to scan the genomic sequences of interest against target motifs.

Scanning the USP9X gene cluster for candidate elements controlling XCI

Transposable elements and CpG islands

RepeatMasker (Smit and Green 1999) tables at the UCSC Genome Browser (hg18) (Karolchik et al. 2008) were utilized to map the coordinates of transposable elements into the genomic region of the USP9X gene cluster. The data were apportioned into 50-kb windows with 5 kb slides. To classify L1s into young vs. old in human, dog, mouse, and rat, were considered Repbase DB (http://www.girinst.org/repbase/update/browse.php) as well as some published studies (Bailey et al. 2000; Goodier et al. 2000). Species-specific L1s were considered to be young L1s. The genome averages of L1s were obtained from several genome project papers (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005). The coordinates of human and mouse CpG islands were obtained from the UCSC Genome Browser (Karolchik et al. 2008).

CTCF insulators

To predict candidate CTCF binding sites, we used program STORM ((Schones et al. 2007); http://rulai.cshl.edu/storm/). A weight matrix for CTCF motif was obtained from ((Kim et al. 2007); http://bioinformatics-renlab.ucsd.edu/rentrac/wiki/CTCF_Project). In addition to the CTCF motif matrix, all 13804 CTCF binding sites from IMR90 cells were also obtained, and their genomic coordinates (hg17) were converted to hg18 using the Lift-Over utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver) in Galaxy ((Blankenberg et al. 2010); http://g2.bx.psu.edu/). We set a STORM score > 14 as a criterion to define candidate CTCF binding sites. This score was derived from the average STORM score of predicted CTCF binding sites that overlapped with experimentally identified sites from the IMR90 cells. Genomic

sequences for the upstream boundary regions in four species (Human-hg18, Dog-canFam2, Mouse-mm9, and Rat-rn4) were obtained using Galaxy ((Blankenberg et al. 2010); http://g2.bx.psu.edu/).

Histone modifications

To delineate epigenetic modifications determining the XCI status, the data on ENCODE histone modifications were retrieved from the UCSC Genome Browser (Karolchik et al. 2008), using the hg18 version of the human genome. We utilized Chip-seq data generated by the Broad/MGH group (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=161293465&c=chrX&g=wgEncodeBroadChipSeq). Five cell lines (HMEC, Human Mammary Epithelial Cells; HSMM, Normal Human Skeletal Muscle Myoblasts; HUVEC, Human Umbilical Vein Endothelial Cell; NHEK, Normal Human Epidermal Keratinocytes; NHLF, Normal Human Lung Fibroblasts) were considered for this study because they are mature cells and do not have abnormal karyotypes. Signal values were normalized by subtracting the corresponding control signals (they also can be obtained from the UCSC Genome Browser).

Motif analysis

We obtained 138 and 110 overrepresented oligomers labeled as inactivated and escape, respectively, from (Carrel et al. 2006). Using in-house Perl scripts, the coordinates of oligomers were mapped into genomic regions. The frequencies of oligomers were apportioned into 50-kb windows with 5-kb slides.

Analysis using comparative XCI data in the USP9X gene cluster (second and third approaches)

Defining evolutionarily conserved sequences in boundary regions

Using experimentally validated XCI profile for genes in the USP9X gene cluster of several species (Fig. S4-1), intergenic regions between ATP6AP2 and CXorf38, and between DDX3X and CASK were defined as "ATP6AP2-CXorf38" and "DDX3X-CASK" boundaries, respectively. Alignments of species, obtained from the 44-way multiz genomic alignments at the UCSC Genome Browser (Karolchik et al. 2008), were utilized to identify human-chimpanzee-macaque-dog and mouse-rat orthologous sequences. Genic regions for the non-human species were determined by the genomic coordinates of human RefSeq mRNAs available at the UCSC Genome Browser (Karolchik et al. 2008). To determine sequences under selective constraints in boundary regions, each boundary region was broken into 50-bp windows with 10-bp slides, and nucleotide substitution rate for each window was estimated using the REV model (Yang 1993) implemented in the BASEML module of the PAML package (Yang 2007). The degree of selective constraint was evaluated by comparison between substitution rate in each window from boundary regions and that in neutrally evolving control regions. Introns, ancestral repeats (AR; (Hardison et al. 2003)), and four-fold degeneration sites (4D) were used as the neutrally evolving control regions in this study. To reduce the effects of regional variation in control sites, genomic regions that were located within 500 kb on either side of the middle point of each boundary region were chosen. To improve the reliability of the control regions, first intron and splice sites (non-rigorously first and last 100 bp of each intron) were eliminated because they are likely to evolve under functional constraint (Keightley and Gaffney 2003; Gaffney and Keightley 2006), and repeats including MER121 family were excluded because they evolve under strong selection (Kamal et al. 2006).

To test whether substitution rate in each window from a boundary region is significantly lower than that in the corresponding control site, we implemented a statistical test. Similarly to the boundary regions, each control region was broken into 50-bp windows with 10-bp slides, and the number of windows with substitution rates lower than or equal to the observed substitution rate from window in boundary region were used to compute an empirical *P* value. To correct for multiple testing, we used the false discovery rate approach (Storey and Tibshirani 2003).

Finding motifs

A series of Perl programs were designed to identify motifs, with one or more mismatches allowed. From sequences in a boundary region we exhaustively enumerated motifs with 1-bp sliding. They were defined as motifs with perfect match. Reverse complement motifs were considered together. For motifs with one mismatch allowed, all possible motifs with one mismatch were generated from motifs with perfect match. For example, given a particular 8-bp motif (e.g., ATGCCGTA), a total of 8 possible motifs with one mismatch (e.g., NTGCCGTA, ANGCCGTA, ATNCCGTA, ATGNCGTA, ATGCNGTA, ATGCCNTA, ATGCCGNA, and ATGCCGTN) were generated and defined as motifs with one mismatch. Similar procedures were applied for two, three, and four mismatches.

Clustering motifs

Motifs were merged if they overlapped according to their genomic coordinate by at least 75% (for example, 6 bp for 8-mers).

References

Bailey JA, Carrel L, Chakravarti A, Eichler EE. 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* **97**(12): 6634-6639.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**(1): 289-300.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al* **Chapter 19**: Unit 19 10 11-21.

Carrel L, Park C, Tyekucheva S, Dunn J, Chiaromonte F, Makova KD. 2006. Genomic environment predicts expression patterns on the human inactive X chromosome. *PLoS genetics* **2**(9): e151.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Chadwick BP, Willard HF. 2004. Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **101**(50): 17450-17455.

Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E et al. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**(6): 956-969.

Ciavatta D, Kalantry S, Magnuson T, Smithies O. 2006. A DNA insulator prevents repression of a targeted X-linked transgene but not its random or imprinted X inactivation. *Proceedings of the National Academy of Sciences of the United States of America* **103**(26): 9958-9963.

Clemson CM, Chow JC, Brown CJ, Lawrence JB. 1998. Stabilization and localization of Xist RNA are controlled by separate mechanisms and are not sufficient for X inactivation. *The Journal of cell biology* **142**(1): 13-23.

Dermitzakis ET, Bergman CM, Clark AG. 2003. Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. *Molecular biology and evolution* **20**(5): 703-714.

Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ, Nguyen DK, Tsuchiya KD, Disteche CM. 2005. Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell* **8**(1): 31-42.

Gaffney DJ, Blekhman R, Majewski J. 2008. Selective constraints in experimentally defined primate regulatory regions. *PLoS genetics* **4**(8): e1000157.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. *PLoS genetics* **2**(11): e204.

Gibbs RA Weinstock GM Metzker ML Muzny DM Sodergren EJ Scherer S Scott G Steffen D Worley KC Burch PE et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982): 493-521.

Goodier JL, Ostertag EM, Kazazian HH, Jr. 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Human molecular genetics* **9**(4): 653-657.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome biology* **8**(2): R24.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99-104.

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome research* **13**(1): 13-26.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**(7243): 108-112.

Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)* **15**(7-8): 563-577.

Hu J, Li B, Kihara D. 2005. Limitations and potentials of current motif discovery algorithms. *Nucleic acids research* **33**(15): 4899-4913.

Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**(5): 1205-1214.

Johnston CM, Lovell FL, Leongamornlert DA, Stranger BE, Dermitzakis ET, Ross MT. 2008. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS genetics* **4**(1): e9.

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**(9): 418-420.

Kahan B, DeMars R. 1975. Localized Derepression on the Human Inactive X Chromosone in Mouse-Human Cell Hybrids. *Proceedings of the National Academy of Sciences of the United States of America* **72**(4): 1510-1514.

Kamal M, Xie X, Lander ES. 2006. A large family of ancient repeat elements in the human genome is under strong selection. *Proceedings of the National Academy of Sciences of the United States of America* **103**(8): 2740-2745.

Kankainen M, Loytynoja A. 2007. MATLIGN: a motif clustering, comparison and matching tool. *BMC bioinformatics* **8**: 189.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic acids research* **36**(Database issue): D773-779.

Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proceedings of the National Academy of Sciences of the United States of America* **100**(23): 13402-13406.

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**(6): 1231-1245.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**(7069): 803-819.

MacIsaac KD, Fraenkel E. 2006. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS computational biology* **2**(4): e36.

Marks H, Chow JC, Denissov S, Francoijs KJ, Brockdorff N, Heard E, Stunnenberg HG. 2009. High-resolution analysis of epigenetic changes associated with X inactivation. *Genome research* **19**(8): 1361-1373.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome research* **17**(12): 1797-1808.

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome research* **17**(4): 413-421.

Nishida K, Frith MC, Nakai K. 2008. Pseudocounts for transcription factor binding sites. *Nucleic acids research*.

Park C, Carrel L, Makova K. 2010. Strong Purifying Selection at Genes Escaping X Chromosome Inactivation. *Mol Biol Evol (in press)*.

Pavesi G, Mereghetti P, Mauri G, Pesole G. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research* **32**(Web Server issue): W199-203.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Molecular biology and evolution* **25**(9): 1795-1808.

Prothero KE, Stahl JM, Carrel L. 2009. Dosage compensation and gene expression on the mammalian X chromosome: one plus one does not always equal two. *Chromosome Res* **17**(5): 637-648.

Ross MT Grafham DV Coffey AJ Scherer S McLay K Muzny D Platzer M Howell GR Burrows C Bird CP et al. 2005. The DNA sequence of the human X chromosome. *Nature* **434**(7031): 325-337.

Schones DE, Smith AD, Zhang MQ. 2007. Statistical significance of cis-regulatory modules. *BMC bioinformatics* **8**: 19.

Siggia ED. 2005. Computational methods for transcriptional regulation. *Current opinion in genetics & development* **15**(2): 214-221.

Smit A, Green P. 1999. RepeatMasker at, http://www.repeatmasker.org.

Stone EA, Cooper GM, Sidow A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annual review of genomics and human genetics* **6**: 143-164.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16): 9440-9445.

Talebizadeh Z, Simon SD, Butler MG. 2006. X chromosome gene expression in human tissues: male and female comparisons. *Genomics* **88**(6): 675-681.

Tsuchiya KD, Greally JM, Yi Y, Noel KP, Truong JP, Disteche CM. 2004. Comparative sequence and x-inactivation analyses of a domain of escape in human xp11.2 and the conserved segment in mouse. *Genome research* **14**(7): 1275-1284.

Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic acids research* **34**(Database issue): D95-97.

Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M An P et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RD, Chenoweth JG, Tesar PJ, Furey TS et al. 2007. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS genetics* **3**(8): e136.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome research* **20**(5): 614-622.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution* **10**(6): 1396-1401.

-. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8): 1586-1591.

Chapter 5

A Computational Approach to Candidate Gene Prioritization for X-Linked Mental Retardation using Clinically-Informed Binary Filtering and Motif-Based Linear Discriminatory Analysis

This chapter was submitted to BMC genomics, and was formatted for that journal. Authors for this original manuscript are Zane Lombard, Chungoo Park, Kateryna D. Makova, and Michele Ramsay. CP and ZL performed the bioinformatics analyses, interpreted the results, and wrote the first draft of the manuscript. MR proposed the initial idea, ZL did preliminary investigation, KDM stimulated the addition of the sequence motif-based analysis. MR, KDM interpreted the results. All authors contributed to writing the paper. CP and ZL contributed equally to this work and are considered co-first authors.

Abstract

Background: recently, many computational candidate gene selection and prioritization methods have been developed. There are often two central approaches employed for these in silico selection and prioritization techniques – the examination of similarities to known disease genes and/or the evaluation of functional annotation of genes. Each of these approaches has its own caveats. Here we employ a previously described method of candidate gene prioritization based mainly on gene annotation, in accompaniment with a technique based on the evaluation of pertinent sequence motifs or signatures, in an attempt to refine the gene prioritization approach. We apply this approach to X-linked mental retardation (XLMR), a group heterogeneous disorders for which some of the underlying genetics is known.

Results: using the gene annotation-based binary filtering method yielded a ranked list of putative XLMR candidate genes with good plausibility of being associated with XLMR development. In parallel, a motif finding approach based on linear discriminatory analysis (LDA) was employed to identify short sequence patterns that may discriminate XLMR from non-XLMR genes. High rates (>80%) of correct classification was achieved, suggesting that the identification of these motifs effectively capture genomic signals found near XLMR vs. non-XLMR genes. The intercept between these approaches were also evaluated and discussed. The computational tools developed for the motif-based LDA is integrated into the freely available genomic analysis portal Galaxy (http://g2.bx.psu.edu/).

Conclusions: combining gene annotation information and sequence motif-orientated computational candidate gene prediction methods, highlighted a compelling list of plausible candidate genes, as has been demonstrated for XLMR.

Background

The identification and characterization of genes and genetic variants that result in disease, or contribute to disease susceptibility, is a critical objective in medical research. Such findings have contributed to improvements in diagnosis, prognosis and therapy [1]. The typical approach taken for disease gene discovery for monogenic traits involves the identification of affected families, genotyping and linkage analysis. Subsequent fine mapping of the identified linked region is performed to focus the candidate region and reduce the number of putative candidate genes, mutation detection in which uncovers the genetic cause of the disorder [2-3]. This approach has been successfully employed in pinpointing the genetic contributors of more than 1000 disorders, including Huntington disease [4], Duchenne muscular dystrophy [5] and cystic fibrosis [6]. With the successful identification of disease genes for many single-gene disorders,

the focus has shifted to diseases with a complex, multifactorial etiology [7-8]. The candidate gene approach has often been used in the search for complex disease genes, but with the advent of massively parallel sequencing and genotyping, genome-wide approaches (such as genome-wide association studies [GWAS]) are starting to take precedent [9]. However, these approaches can result in large sets of potentially implicated genes, with the challenge then being to identify the actual genes involved with disease pathogenesis – a potentially laborious and costly exercise.

Recently, many computational candidate gene selection and prioritization methods have been developed in part to provide new avenues to pinpoint disease genes or to prioritize genes from a large list of candidates [10-17]. Most often one of the two approaches is taken to identify and prioritize putative disease genes – either investigating similarities (including sequence resemblance) to know disease genes or evaluating functional annotation of genes. The first approach is based on the premise that differences exist between genes identified to be involved in disease, and other human genes, including differences in gene sequence and structure. Such approach has been implemented in the selection of candidate disease genes, as well as for other objectives such as the discovery of sequences important in X-chromosome inactivation, and the subsequent prediction of expression status of individual genes [18]. The second, annotation-based approach centers on the hypothesis that similar diseases may be influenced by genes with comparable features (such as function). This method relies heavily on the terms used to describe genes and their related products, and standardized ontologies utilized across databases (such as Gene Ontology (GO) and eVOC) are imperative to the efficiency of this approach.

In an attempt to increase the probability of correctly identifying disease related genes by producing a short list of highly probably candidate genes, a combined approach could reduce uncertainty. This could be done by combining multiple independent lines of evidence, each by itself lacking sufficient power. In this paper, a previously described method of candidate gene selection based primarily on gene annotation [19] is complemented by the evaluation of pertinent

sequence motifs or signatures to refine the gene selection approach [18]. The intercept between these approaches is evaluated, and the usefulness of combining the two is discussed. We have applied this approach to X-linked mental retardation (XLMR), a group of related but heterogeneous disorders for which some of the underlying genetics is known.

Results

Annotation-based gene prioritization using clinically-informed binary filtering method

The complete set of X chromosome genes (a total of 814 X-linked genes; Ensembl v49) was subjected to candidate gene selection for XLMR using a previously described method [19]. Briefly, gene annotation terms found to be pertinent to XLMR were identified through literature and data-mining (a total of 40 terms; summarized in *Table 5-1*). Each term was used as a selection criterion to populate a gene list (containing all genes within the Ensembl database annotated with that term). Candidate genes were then prioritized using a binary evaluation grid that assessed the term gene lists against the X-linked gene list, and genes were scored (see Methods for details) accordingly**.** X-linked genes with the most matches to the annotation-derived gene lists (27/40 being the most matches for any gene) were ranked as strong candidate XLMR genes. *Figure 1* shows the relative enrichment for known XLMR genes as compared with non-XLMR genes. *Table S5-1* summarizes the prioritization of X-linked genes as XLMR candidates using this approach, serving as a test for the sensitivity of this approach. As expected, the categories ranked highly by this method (i.e. more terms matched, therefore stronger support of being a likely candidate) has a higher enrichment for known XLMR genes than for non-XLMR genes. Conversely, the enrichment becomes moderated as one progress down the ranked list. Out of the 814 X-linked genes, 255 (31.3%) were present in at least ten of the 40 criteria lists

Table **5-1**: Annotation terms identified to be pertinent to XLMR using a literature- and data-mining approach

ANNOTATION TERM CATEGORIES[1]			
Anatomical site	**Biological Process**	**Phenotype**	**Animal model homology**
Developmental	Development	Seizures	*Phenotype*
Liver	Transcription	Epilepsy	Behaviour/Neurological
Central nervous system	Metabolism	Acidosis	Nervous system related
Respiratory	Phosphorylation	Microcephaly	Embryogenesis
Cerebellum	Brain development	Tremor	
Kidney			*Timing*
Hippocampus			Pre-Embryonic
Spinal cord			Embryonic
Cerebral cortex			Fetal
Testis			
Brain stem			*Anatomy*
Peripheral nerve			TS[2]8-9 Ectoderm
Cerebrum			TS10-13 Neural Ectoderm
Substantia nigra			TS14-26 CNS
Cardiovascular			
Adrenal gland			
Thyroid			
Ovary			
Amygdala			
Musculoskeletal			
Ganglion			
Hypothalamus			

[1]These terms were used to extract gene lists that were compared to all X chromosome genes in a binary filtering process. Annotation terms are divided into four categories based on their ontological classification.

[2]TS – Theiller stage: A term used to denote the stage of development of a mouse as described by Theiler in "The House Mouse: Atlas of Mouse Development" (Springer-Verlag, New York, 1989)

indicating their match to the annotation terms. This is a 6.5-fold enrichment for the XLMR genes since 10% (82/814) of the genes on the X chromosome are known XLMR genes and 66% (54/82) of the XLMR genes match at least ten annotation categories in the criteria lists. Among the genes matching at least ten annotation terms, there were an additional 201 genes that have not previously been associated with XLMR, but and a set of such genes these may involve part of become interesting candidates XLMR genes of interest. They can, in turn, be ranked according to the number of categories matched. Genes that matched very few of the identified annotation

terms (two or less) are considered to be less likely to be associated with XLMR, and based on this premise, 22.5% (183/814) of the X-linked genes are unlikely to be involved in XLMR. Only two known XLMR gene (*BRWD3* and *SLC9A6*) were among these low-ranked genes.

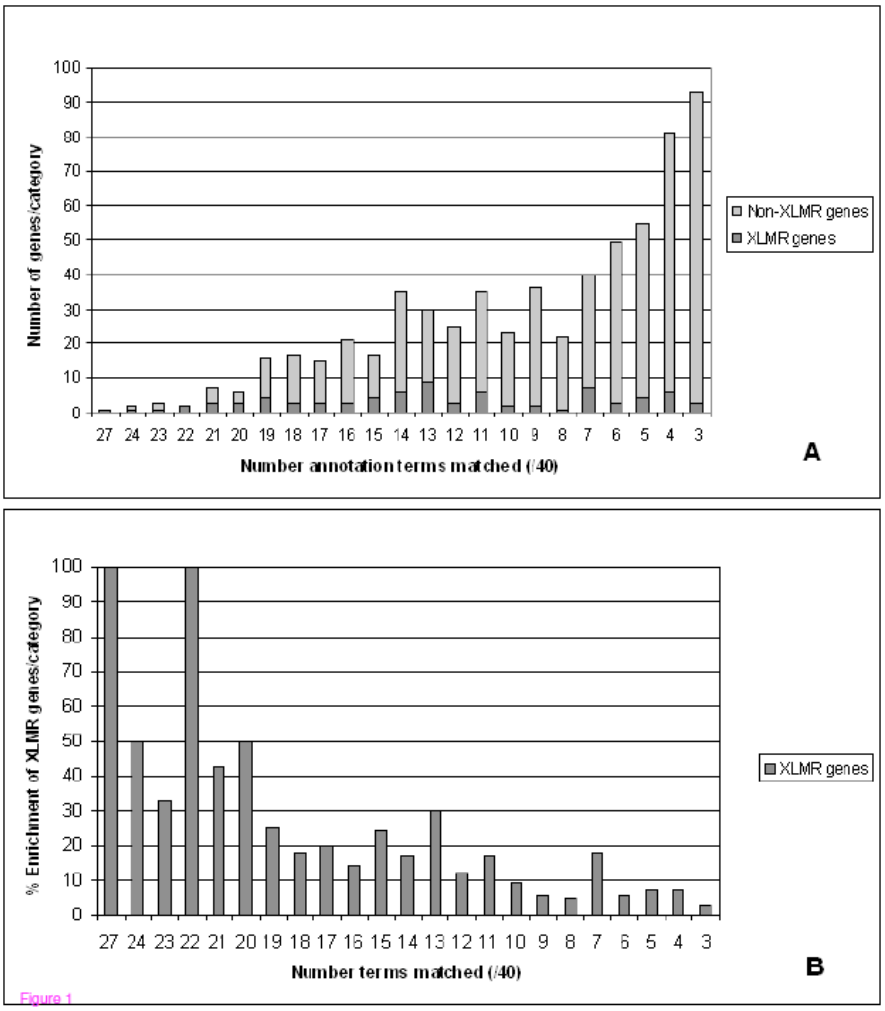


Figure **5-1**: Ranking of genes on the X chromosome as XLMR candidates using a binary filtering process. (A) depicts the number of known XLMR and non-XLMR genes in each category. A noted enrichment of XLMR genes is seen in the categories depicting a higher number of matches to annotation terms. (B) shows this enrichment as a percentage.

MECP2, a known XLMR causative gene (OMIM *300005), had the most annotation term matches (27/40) and matched at least one term in all four nomenclature categories (i.e., anatomical site terms, biological process, animal homology, and phenotype; see Methods for

details) used in prioritization. *MECP2* is a widely expressed transcriptional repressor and has two conserved functional domains, the methyl-CpG binding domain and the transcription repression domain. *MECP2* displays extreme allelic heterogeneity, with more than 100 different mutations in the *MECP2* gene being described in patients with Rett Syndrome [20]. Further to this, *MECP2* mutations have also been shown to produce non-syndromic male fatal neonatal encephalopathy, progressive spasticity and non-syndromic Angelman and Prader–Willi-like phenotypes [20-22]. Rett syndrome is a prime example of the locus heterogeneity associated with XLMR as *MECP2* mutations account for only approximately 70–80% of cases, whereas locus heterogeneity is hypothesized to explain the occurrence of the syndrome among *MECP2* negative cases [20]. It is therefore likely that other genes that have been prioritized by the binary ranking method may be genetic candidates for Rett syndrome (among other forms of mental retardation).

The dystrophin gene (*DMD*) was ranked second highest (24/40 matches) and mutations in this gene have been associated with X-linked mental retardation in some, but not all cases. *DMD* is one of the largest known genes, measuring 2.4 Mb, and was identified as the gene responsible for Duchenne (DMD) and Becker (BMD) muscular dystrophies. Dystrophin mRNA is present in brain tissue and is therefore thought to explain mental retardation, as observed in some DMD patients [23]. Dystrophin in brain is transcribed from a different promoter from that used in muscle. Chelly et al. (1990) [24] demonstrated that the brain-type promoter of the dystrophin gene is highly specific to neurons.

The top 50 ranked genes not previously linked to XLMR were compared to a list of genes from chromosomal regions that have previously been linked to XLMR syndromes, but for which the causative gene has not been pinpointed to date (summarized in *Table S5-2*) Four genes, in particular, were present in highly overlapping critical regions for XLMR and were also highly ranked becoming compelling XLMR candidates for further investigation: *EFNB1* (18/40 matched annotations), *NONO* (19/40), *PDZD11* (11/40) and *TAF9B* (18/40).

Prioritization based on sequence motifs

Gene sets were selected to identify distinguishing sequence-based features in the immediately upstream regions of the transcriptions start sites (TSSs) between the sets. The X-linked genes were divided into two groups: genes demonstrated to be involved in XLMR and genes that have not been identified as being involved in XLMR (called non-XLMR). In this study, 81 and 486 genes were labeled as XLMR and non-XLMR genes, respectively. The remaining X-linked genes were excluded from the analysis because of the lack of annotated putative transcription start sites. To consider the effects of the observed nonrandom distribution of XLMR genes along the X chromosome [25-26] and the possible influence of distinct origins and evolutionary pressures on XLMR genes [27], genes were labeled as belonging to the X-added region (XAR; representing evolutionary young strata) or the X-conserved region (XCR; representing ancestral mammalian X), determined by their genomic location [28-29] and were analyzed separately (*Table 5-2*).

Table **5-2**: Number of genes considered for sequence-based prioritization.

	XAR	**XCR**	**TOTAL**
XLMR	25	56	**81**
Non-XLMR	110	376	**486**
TOTAL	**135**	**432**	

To detect sequence motifs (hereafter called "oligomers") that may discriminate XLMR from non-XLMR genes (separately for XAR and XCR genes), genomic sequences (hereafter called "subgenomes") were compiled to form four subgenomes. They included regions upstream from the transcription start sites (TSS) of each gene (hereafter called "contigs") and four variable distances were considered for the contigs: 5 kilobases (kb), 10 kb, 50 kb, and 100 kb upstream from the TSS of each gene. If a contig overlapped with a neighboring contig and both shared the

same profile (XLMR or non-XLMR), the two were merged into a single, larger contig. If a contig overlapped with a contig from the opposite profile (e.g., XLMR and non-XLMR), then both contigs were discarded. Finally, if a contig from one profile overlapped with any part of a gene that was classified in the opposite profile, the contig was discarded (*Table S5-3*).

To identify the overrepresented oligomers in each subgenome, the following criteria were considered: first, each oligomer was required to occur at least ten times in the subgenome and second, to be retained for a particular subgenome (e.g. XLMR; XAR), the oligomer was required to occur five times more frequently in that subgenome compared to the alternative (e.g. non-XLMR, XAR). For the analysis described below, only 12-mers within 10 kb, 50 kb, and 100 kb of each gene were used as they had the highest total number of overrepresented oligomers (*Table S4*; too few overrepresented oligomers were found in the 5 kb subgenomes, so this scale was omitted from the further analysis; see Methods for details). Permutation tests were performed to evaluate the significance of the overrepresentation of oligomers (*Table S5-5*). Subsequently, overlapping (among different subgenomes) oligomers were merged into longer ones resulting in 268 (two from XLMR and 266 from non-XLMR) and 584 (11 from XLMR and 573 from non-XLMR) overrepresented oligomers found for XAR and XCR, respectively.

Table **5-3**: Number of genes for training and success rates of LDA.

Set Analyzed	**Parameter**	**10 kb (Genes)**	**50 kb (Genes)**	**100 kb (Genes)**
Training and test set of genes in XAR	τ	0.96	0.43	0.4
	Success in XLMR	100% (19)	100% (15)	100% (9)
	Success in non-XLMR	45% (84)	91% (40)	96% (26)
Training and test set of genes in XCR	τ	0.87	0.75	0.62
	Success in XLMR	87% (38)	100% (16)	100% (7)
	Success in non-XLMR	52% (257)	82% (141)	96% (74)

τ is a tuning parameter, which was selected to maximize the sum of correct classification rates for XLMR and non-XLMR sets

Statistical analyses were performed to predict whether, using the set of overrepresented oligomers, genes could be classified as putative XLMR genes or non-XLMR genes. Linear discriminant analysis (LDA [30]) was used as a classifier to distinguish between XLMR and non-XLMR genes, considering XAR and XCR genes separately. Genes and counts of overrepresented oligomers surrounding TSS of genes, identified above were used as the units and features for classification. Utilizing two distinct sets of training data (either genes from the XAR or XCR), both LDA classifiers achieved classification accuracy of greater than 80% for both XLMR and non-XLMR classes except for the 10 kb range (*Table 5-3* and *Table S5-6*). The high rates of correct classification imply that classifiers based on counts of overrepresented oligomers effectively capture genomic signals found near XLMR vs. non-XLMR genes. In the case of the 100-kb range, correct classification rates for both classifiers were ≥ 96 % for XLMR and non-XLMR classes (*Figure 5-2* and *Table 5-3*). All genes tested in XAR with the 100-kb distance were perfectly classified (*Table S5-7*). For genes in XCR with the 100-kb distance, two genes (out of seven) in XLMR class and three genes (out of 75) in non-XLMR class were incorrectly classified. Interestingly, among five misclassified genes using the 100-kb distance, one gene (*STARD8*) was also incorrectly classified as an XLMR gene at the 10-kb and 50-kb distances (*Table S5-7*), indicating that this gene may be a strong candidate XLMR gene. *STARD8* did not receive a particularly high ranking using the annotation-based approach (6/40) and might have been missed as a putative XLMR gene by this method.

Table **5-4**: Number of genes with ten or more matched categories using the annotation approach that were classified correctly by sequence-based method.

Length of contigs	Number of genes tested	Genes classified successfully		
		10 kb (Genes)	50 kb (Genes)	100 kb (Genes)
> 10 kb	101	54.5% (55)	52.5% (53)	52.5% (53)
> 50 kb	42	59.5% (25)	78.6% (33)	88.1% (37)
> 100 kb	22	59.1% (13)	81.8% (18)	86.4% (19)

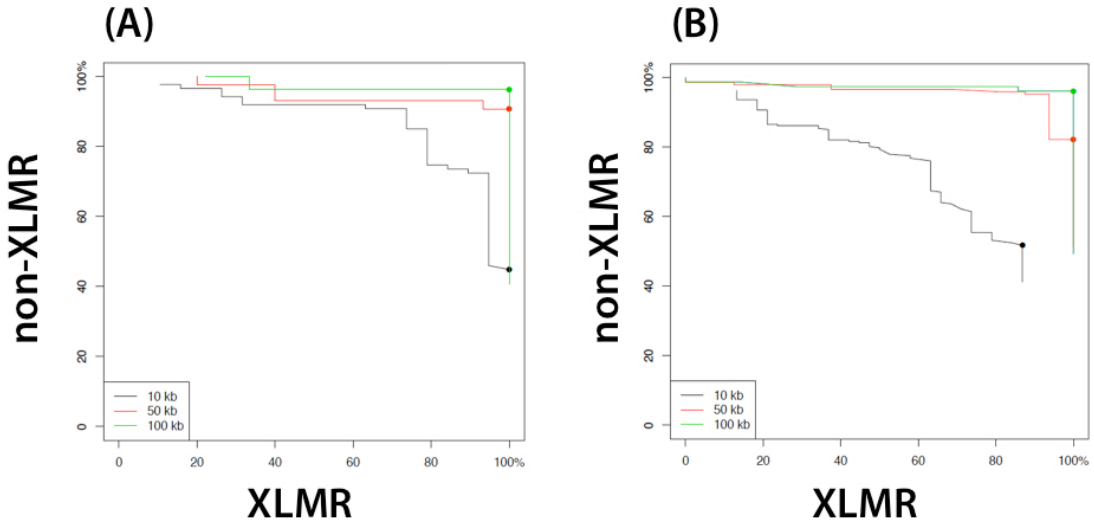


Figure **5-2**: LDA classification success rates for different values of the turning parameter τ. (A) All XAR genes were used for training and test sets. (B) All XCR genes were used for training and test sets. Leave-one-out cross-validation was utilized to calculate correct classification rates. Dots indicate optimal values of τ (see Table 5-3).

Combined analysis of annotation-based and sequence motif assessment in classifying X-linked genes as putative XLMR genes

To evaluate which genes are classified as high-probability XLMR genes by both of the assessment approaches, a step-wise approach was used. The 255 top-ranked genes from the annotation-based approach (matching at least 10/40 criteria) were analyzed using the XLMR vs. non-XLMR discriminatory sequence-based classifiers. For this analysis, 154 genes (out of 255) were excluded; either because of the lack of annotated TSSs (18 genes) or because of the short length of contigs (136 genes; namely, the distance from the TSS of gene to its closest neighbor gene was less than 10 kb). The remaining 101 genes with at least 10-kb contigs were classified as XLMR [31] or non-XLMR genes and assessed using the sequence-based method to determine whether they had a XLMR or non-XLMR signature. In the case of genes whose contig length was greater than 50 kb (42 genes), 78.6% (33/42) had signatures as expected ("correctly classified"), and for contig lengths greater that 100 kb (22 genes), 86.4% (19/22) were correctly classified (*Table 5-4* and *Table 5-5*). Of particular interest would be the high-ranked candidates that were "incorrectly classified" as XLMR when there is no prior information that they had previously been linked to XLMR. Further analysis of genes with >50-kb contigs, using the sequence-based classifiers with the 50-kb and 100-kb distances, revealed that 14 genes were classified as putative XLMR genes. Five of them are known XLMR genes (*AP1S2*, *ARHGEF9*, *BCOR*, *HUWE1*, and *ZDHHC9*) and the remaining nine (*APLN*, *ZC4H2*, *MAGED4*, *MAGED4B*, *RAP2C*, *FAM156A*, *FAM156B*, *TBL1X*, and *UXT*) are strong candidates for XLMR and merit further analysis (*Table 5-6*). Interestingly, only one of these nine genes is located in the XAR.

Table **5-5**: Classification of genes with > 50 kb contigs that were in the top 16 matched categories.

Nr of annotation terms matched (/40	HGNC ID	XLMR gene[1]	50 kb	100 kb	Dist (kb)[2]	Strata
24	DMD	1	X	X	71.7	XAR
23	GPM6B	0	X	X	68	XAR
19	AP1S2	1	O	O	61.4	XAR
19	FGF13	0	X	X	157.8	XCR
19	ZDHHC9	1	O	X	60.7	XCR
18	TAF9B	0	X	X	133	XCR
18	*FAM156A*	0	O	X	90.8	XCR
18	*FAM156B*	0	O	X	90.8	XCR
17	ARHGEF9	1	X	O	259.5	XCR
17	TMEM47	0	X	X	285.5	XAR
16	NAP1L2	0	X	X	182.7	XCR
16	NSBP1	0	X	X	80.3	XCR
16	PCSK1N	0	X	X	56.7	XCR
16	UXT	0	O	X	64.4	XCR
15	BCOR	1	O	O	253.8	XAR
15	ENOX2	0	X	X	155	XCR
15	HUWE1	1	O	X	224.7	XCR
15	LAS1L	0	X	X	132.8	XCR
15	THOC2	0	X	X	124	XCR
14	CXorf61	0	X	X	323.4	XCR
14	TBL1X	0	O	X	190.3	XAR
14	TMLHE	0	X	X	154.8	XCR
13	C1GALT1C1	0	X	X	102.9	XCR
13	KAL1	0	X	X	58.6	XAR
13	PPP2R3B	0	X	X	237.4	XAR
13	RNF12	0	X	X	65.4	XCR
13	ZMAT1	0	X	X	77	XCR
12	GABRE	0	X	X	103	XCR
12	*ZC4H2*	0	O	X	431.8	XCR
12	MAGED4	0	X	O	115.6	XCR
12	MAGED4B	0	X	O	115.6	XCR
12	MORC4	0	X	X	64.2	XCR
12	RAI2	0	X	X	302.1	XAR
11	APLN	0	O	X	84.5	XCR
11	KLHL13	0	X	X	83.1	XCR
11	MUM1L1	0	X	X	93.4	XCR
11	RAP2C	0	O	X	160.1	XCR
11	SLITRK4	0	X	X	72.2	XCR
11	SYTL4	0	X	X	88.2	XCR
10	CHRDL1	0	X	X	300.5	XCR
10	PHF16	0	X	X	72.4	XAR
10	STAG2	0	X	X	60.9	XCR

"O" and "X" represent that a gene was classified as XLMR and non-XLMR genes, respectively. Light gray indicates genes that are known XLMR genes (Chiurazzi et al., 2008) as well as candidate XLMR genes by both binary filtering and sequence-based approaches. Dark gray indicates genes that are classified as putative XLMR genes by two approaches, but not explicitly annotated as XLMR genes.

[1]: indicate if gene is a known XLMR (=1) or not (=0).

[2]: distance from the TSS of gene to its closest neighbor gene.

Table **5-6**: Nine genes highlighted as XLMR candidates by both the annotation and sequence motif method.

RANK (/40)*	HGNC ID	DESCRIPTION	LOCA -TION	FUNCTION
18	FAM156A	Family with sequence similarity 156, member A	Xp11.23	Function Unknown
18	FAM156B	Family with sequence similarity 156, member B	Xp11.22	Function Unknown
16	UXT	Ubiquitously-expressed transcript	Xp11.23-p11.22	Plays a role in facilitating receptor-induced transcriptional activation [74]
14	TBL1X	Transducin (beta)-like 1X-linked	Xp22.3	Plays an essential role in transcription activation mediated by nuclear receptors [75]
12	MAGED4	Melanoma antigen family D, 4	Xp11	Mainly tumour cell proliferation [76]
12	MAGED4B	Melanoma antigen family D, 4B	Xp11	Mainly tumour cell proliferation [76]
12	ZC4H2	Zinc finger, C4H2 domain containing	Xq11.1	Hepatocellular carcinoma-associated antigen [28]
11	APLN	Apelin	Xq25	Neuropeptide involved in the regulation of body fluid homeostasis and cardiovascular functions [77]
11	RAP2C	Member of RAS oncogene family	Xq25	Involved in serum response element mediated gene transcription [56]

*: The ranking obtained using the annotation method for these genes are indicated

Discussion

Mental retardation is defined as a disability characterized by significant limitations both in intellectual functioning and adaptive behaviour [32]. The importance of genes on the X-chromosome in the cause of mental retardation has been recognized for decades, largely due to

the fact that males outnumber females in nearly all surveys of mental retardation by approximately a third [33]. For the more than 200 forms of XLMR described, 82 causative genes have been officially catalogued [31] and ongoing research means that causative genes are being continually identified [34-37].

Mammalian X and Y chromosomes diverged independently after evolving from a pair of autosomes [38-40]. The X chromosome is particularly gene-poor, has an overall low GC-content compared with the genome average [28] and is highly enriched for interspersed repeats. It can be divided into the following evolutionary domains: the X-added region (XAR) and the X-conserved region (XCR). In this study we use XLMR as an example to test a combined gene annotation and sequence-based assessment model for candidate disease gene prediction.

The annotation-based gene prioritization method using binary filtering identified 201 X-linked genes that were annotated in at least ten out of the 40 categories chosen as representing biological categories expected to have a link to XLMR, but which had not previously been associated with XLMR. The annotation categories included gene expression site, protein function, related phenotype and animal homology. Of the top ranked genes that had not previously been associated with a mental retardation phenotype, several are compelling candidates for XLMR based on their function.

The highest ranked novel candidate XLMR gene was *TIMP1* (24/40 matches) that encodes a natural inhibitor of the matrix metalloproteinases (MMPs), a group of peptidases involved in degradation of the extracellular matrix. *TIMP1* was shown to reside within a genetic hotspot for neurodegenerative disorders on the X chromosome [41]. It has recently been shown that MMP9-mediated TIMP1-regulated extracellular proteolysis is a novel mechanism contributing to synaptic plasticity. It is thought that the MMP9/TIMP1 system could be involved in a broad range of physiological and pathological phenomena in various brain regions, including developmental reorganization of the cerebellum, hippocampus-dependent learning and long-term

potentiation [42]. Four other high ranking genes that have not yet been identified as XLMR genes in the literature are particularly highlighted as compelling candidates to include in an experimental investigation of XLMR association – namely *EFNB1 (18/40 matches), NONO (19/40), PDZD11 (11/40) and TAF9B (18/40).* An investigation of these four genes showed some links between their protein products and brain function. The *EFNB1* gene product is involved in central nervous system development, angiogenesis, and neural synapses formation and maturation, as well as axon guidance. It has also been shown to be involved in normal morphogenesis of skeletal elements, and mutations within the gene have been correlated to craniofrontonasal syndrome [43-44]. This observation can possibly be extrapolated to the association of craniofacial anomalies associated with the XLMR syndromes that are linked to the regions overlapping with the location of *EFNB1*. *PDZD11* encodes a novel protein that regulates brain copper homeostasis [45]. Aberrant copper homeostasis is implicated in neurodegenerative disorders such as Alzheimer disease and this suggests that impairment of copper efflux could negatively influence neuronal function and in this way contribute to rapid neuronal degeneration [46-47]. Although there is no direct prior evidence that *NONO* and *TAF9B* are involved in XLMR, These two genes are transcriptional regulators, and thus could potential have a regulatory effect on another gene product that exerts an influence on XLMR development [48-49].

The major caveat of the clinically-informed binary filtering method is that it relies on current gene annotation data in the public domain. These data are incomplete and therefore it is feasible that the method has a biased probability of prioritizing better annotated genes over those that are not as well annotated, regardless of whether these genes are relevant to the disease of interest.

Here we also used an evaluation method that depends not on gene annotation, but on DNA sequence that is readily available in the public domain, because it may be less biased. Such a method would require some information that would allow one to build training sets of genes to

identify motifs that will discriminate between high probability candidate genes and weak candidates. We used a prioritization method based on sequence motifs to explore the feasibility of differentiating XLMR genes from non-XLMR genes. In order to construct the training data sets we need to take into account the evolutionary origin of segments of the X chromosome, as explained above. It is known that the X chromosome can be divided into two evolutionary distinct domains, XAR and XCR. Therefore, for this investigation LDA classifiers were constructed for the XAR and XCR genes separately to consider the possible influence of distinct origins and evolutionary pressures on XLMR genes [27]. This showed high rates of correct classification for known XLMR genes and genes on the X chromosome not previously associated with XLMR. We tested whether a classifier trained on XAR genes could predict the XLMR status of XCR genes, and vice versa. A low success rate was achieved when each LDA classifier was used to predict the XLMR status in another test set (*Figure S5-1C, D* and *Table S5-6*). It suggests that each set of overrepresented oligomers captures distinct genomic signals for XLMR genes in either the XAR or XCR, but not both.

STARD8 (also known as *DLC3*), which is not a known XLMR gene, was identified as a strong candidate XLMR gene by the sequence-based classification approach. A partial deletion of *STARD8* is associated with the craniofrontonasal syndrome especially in females [50]. *STARD8* is also involved in the growth and metastasis of tumor cells [51] as are *DLC1* and *DLC2*. Structurally, this protein is composed of three protein domains: a sterile α-motif, a RhoGAP (Rho GTPase-activating protein) domain, and a START domain (StAR, steroidogenic acute regulatory protein, related lipid transfer). King and colleagues (2002) [52] revealed that the StAR is expressed in glia and neurons of the mouse brain and has a role in the production of neurosteroids supporting *STARD8* as a tentative biological role player in genetic disorders related to brain functioning.

One of the limitations of the sequence-based motif approach is the stipulation that a gene must have an annotated TSS and must have a region of at least 10 kb upstream of the TSS that does not overlap with another gene, in order to be assessed. This meant that out of the top 255 genes identified using the annotation-based binary filtering method, only less that half (only 101) could be assessed.

Interestingly, *STARD8* was not ranked within the top 255 genes using the annotation criteria as it only matched six of the 40 annotation criteria used for assessment. Of these criteria matches four were related to the anatomical site of expression (pinpointing the brain as a significant expression location of this gene) and the other two were related to phenotype simile in the mouse. *STARD8* is not a very extensively studied gene, which could explain its relatively low level of annotation and subsequent poor ranking obtained with the annotation-based evaluation.

Conversely of the four non-XLMR genes, *EFNB1*, *NONO*, *PDZD11* and *TAF9B*, singled out using the annotation-based selection, three could not be assessed by the sequence motif method as they failed the inclusion criteria. Only *TAF9B* was included in this analysis, and was classified as a non-XLMR gene by this method, perhaps indicating that this gene is not a strong XLMR candidate. Unfortunately, one cannot make the conclusion that the other three genes are therefore better candidates as they were not evaluated by both methods.

Despite these limitations, a set of nine genes not previously associated with XLMR have emerged as high likelihood candidates by both prioritization approaches. These genes represent the overlap of the highest ranking genes among those that could be assessed by both methods. A summary of these genes and their main functions is found in *Table 6*. Evaluation of these genes' function and expression site among other features reveal that these genes do have some putative link to mental retardation. First most, all of the genes are expressed in the central nervous system or brain in normal tissues and have all been linked to a neurological-related phenotype or function in the mouse model [53]. *APLN* has been shown to play a critical role in fluid homeostasis and

pressure/volume homeostasis in the brain [54]. This particular function is of utmost importance to the developing brain and research has shown that *APLN* is involved in modifications of the microvasculature in the immature brain, affecting cerebral blood flow during a hypoxic insult [55]. *RAP2C* has a more general function as it forms part of the Ras family, which regulates a wide variety of cellular functions that include cell growth, differentiation, and apoptosis [56]. *TBLX1* has previously been associated with lissencephaly, a rare brain formation disorder caused by defective neuronal migration resulting in a lack of brain fold and grooves development [57]. Similarly, *UXT* has been shown to interact with *RCAN1*, which codes for the Down syndrome candidate region 1 (DSCR1) protein, and it thought to play a role in the mental disability features of this syndrome [58]. It is interesting that all nine of the prioritized genes have a link to cancer. This could indicate an important role of expression regulation mechanisms such as epigenetic modifications, RNA interference and nonsense-mediated mRNA decay in XLMR development. These mechanisms are often disregulated in cancers, and show redundancy related to their role in human development and pathology [59-61].

As information in public databases becomes more complete, gene annotation will improve, increasing the likelihood of detecting appropriate candidates and decreasing the bias due to uneven annotation of genes. When using a sequence motif-based approach examining the TSSs of genes, one would be limited in compiling training sets in regions that are gene rich, resulting in genome overlaps and leading to the exclusion of genes for analysis. More extensive sequenced-based analysis could investigate other gene regions, for example the 3'UTR, for motifs that may be related to RNAi mechanisms [62]. Combining gene annotation information and sequence motif-orientated computational candidate gene prediction methods, it is possible to identify a set of plausible candidate genes for disease, as has been demonstrated for XLMR.

Methods

Annotation-based gene prioritization using binary filtering

A list of HGNC IDs for all genes on the X chromosome (n=814) were obtained from Ensembl (v49) and were prioritized for XLMR candidature using a previously described computational method [19]. This is a gene prioritization method based on annotation and employs the assessment of various data sources to establish whether a candidate gene and the relevant protein product exhibit the biological characteristics expected to presume a link to a particular disease. Annotation terms were divided into four categories based on their ontological classification – anatomical site, function, phenotype and animal homology (no category takes precedent over another). Each term was then used as a selection criterion to populate a gene list (containing all genes within the Ensembl database annotated with that term) where after all X-linked genes were prioritized for XLMR using a binary evaluation grid. The most recent list of cloned XLMR genes (obtained from Chiurazzi et al, 2008 [31]) was compared to the ranked list to assess the ability of this approach to identify known XLMR genes. Gene annotation terms found to be pertinent to XLMR were identified through literature and data-mining as follows:

Anatomical site terms: all scientific abstracts related to XLMR were obtained from PubMed. The online literature mining tool Dragon Disease Explorer (DDE; http://research.i2r.a-star.edu.sg/DRAGON/DE/) was used to extract eVOC ontology terms [63] from this body of literature. Anatomical site ontology terms found to be associated with XLMR were then used to populate the annotation term lists, as described above.

Biological process: the online literature mining tool Dragon TF Association Miner (DTFAM) [64] was used to extract all GO terms from the abstracts of disease-related literature. Of the terms extracted, terms falling in the molecular function (binding) and cellular component

ontologies were not included in the analysis, as these terms were considered non-specific with regard to XLMR and non-specific in general.

Animal homology: human orthologues to the following categories of mouse genes were obtained by data-mining the Jackson Laboratory Mouse Genome database [53]: Genes associated with phenotypes associated with XLMR; genes expressed at different developmental stages and genes expressed in the developing brain.

Phenotype: all scientific abstracts related to XLMR were obtained from PubMed. Dragon TF Association Miner (DTFAM) [64] was used to extract phenotype ontology terms from this body of literature. Phenotype terms found to be associated with XLMR were then used to populate the annotation term lists, as described above.

The binary evaluation was performed as follows: A gene in the X-linked list was assigned a 1 when that gene was also present in an annotation term list. If the gene was absent from that list it was assigned a 0. For each of the X chromosome genes a final binary score was calculated, simply by summing all binary scores for each of the terms used. Then all genes were ranked based on this score, with those having higher scores being higher in the rank list. Genes in the list that matched most annotation terms (i.e. those genes that obtaining the most 1-scores in the binary matrix) received the highest rank as XLMR candidates. Similarly, genes that matched very few or none of the terms have a lower rank and are considered to be weak candidates.

Prioritization based on sequence motifs

To perform the analysis for sequence-based prioritization a total 87 non-redundant candidate XLMR genes were collected from four main references: Euro-MRX consortium (http://www.euromrx.com/en/database.html; 18 genes), 83 genes from [65], 62 genes from [66] and 69 genes collected from additional experimental literature. Because it can be difficult to

define the candidate non-XLMR genes, as many genes as possible that are likely involved in XLMR were considered. A majority of genes (80 out of 87) gathered here were already utilized in the annotation-based approach and seven candidate XLMR genes were additionally obtained. In this analysis, it was assumed that X-linked genes, excluding the 87 genes listed above, are non-XLMR genes, resulting in 647 X-linked genes being classified as non-XLMR genes. The reference sets of X-linked genes were obtained from the UCSC Refseq track (hg18) and the Ensembl HUGO track (release 48 of NCBI build 36).

The motif discovery method used in [18] was adopted to detect significantly overrepresented oligomers from each subgenome (i.e., the XLMR and non-XLMR subgenome), for the XAR and XCR genes separately. Briefly, first all possible oligomers found within each subgenome were enumerated and sequentially counted. Five specified sizes (8-, 12-, 16-, 20-, and 24-mers) of oligomers were considered. Counts of oligomers were combined with counts of their reverse complement sequences. Exact matches were required. To define an oligomer as an overrepresented oligomer, two criteria were used: the oligomer should (1) occur at least 10 times in the relevant subgenome and (2) should be enriched at least five-fold in the relevant subgenome as compared to the other subgenome (e.g. to be defined as an oligomer for XLMR using the 10 kb distance as an example, an oligomer must occur at least five times more often in the 10 kb region upstream of an XLMR gene vs. the frequency of the same oligomer in the 10 kb region upstream of a non-XLMR gene). Permutation tests were performed to evaluate whether the overrepresented oligomers identified were significantly overrepresented in one subgenome compared to the other. For this permutation test, the subgenomes (XLMR and non-XLMR subgenomes) were pooled together and then divided into nonoverlapping 2-kb fragments. Each 2-kb fragment was randomly assigned to either a pseudo-XLMR or a pseudo-non-XLMR subgenome until the two pseudo subgenomes were equal to the two actual subgenomes in size. Within each pseudo subgenome the oligomers that satisfied the above two criteria for overrepresented oligomers were identified. This

process was repeated 1,000 times, and those oligomers that were identified in fewer than 50 out of the 1000 pseudo subgenome trials ($p < 0.05$) were considered significantly overrepresented.

The putative transcription start sites (TSSs) of genes were identified using the methods described in [67]. Briefly, using data from two sets of 5'-end-tag-capture technologies (i.e., CAGE [68] and PET[69]), TSS-tag clusters for genes were identified. To map the TSS-tag clusters to their corresponding genes, the following two criteria were considered. First, the strand of a TSS-tag cluster must be identical to the strand of a gene. Second, a TSS-tag cluster must be located in the 5' upstream region from the coding start site of a gene. If multiple TSS-tag clusters were identified for a single gene, then only the TSS-tag site supported by the highest number of tags was selected to be used as the representative TSS. If multiple TSS-tag clusters had the same highest tag score, the TSS cluster closest to the coding start site was chosen to serve as the representative TSS. To ensure the reliability of the TSS-tag data, TSS-tag clusters with a single tag were excluded. Using RefSeq [70], H-Invitational [71] and human ESTs [72] from the UCSC genome browser server (hg18), TSS-tag clusters were also discarded if the genomic coordinate of the 5' end of the putatively corresponding cDNAs or ESTs did not overlapped with the TSS-tag clusters. As a result, putative transcription start sites for 81 XLMR and 486 non-XLMR genes were obtained from experimental data.

LDA analysis was performed as reported in [18] with the only difference being that the *P* value for the *p*-dimensional predictor vector was 268 for XAR and 584 for XCR classifiers (*Table 2*). Leave-one-out cross-validation was utilized to calculate correct classification rates. The computational tools for Principal Component Analysis (PCA), LDA and ROC visualization were developed and integrated into a freely available genomic analysis portal Galaxy (http://g2.bx.psu.edu/; [73]).

References

1. O'Connor TP, Crystal RG: **Genetic medicines: treatment strategies for hereditary disorders.** *Nature reviews* 2006, **7:**261-276.
2. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.** *Nature genetics* 2003, **33 Suppl:**228-237.
3. Haines JLaP-V, M.A.: **Designing a study for identifying genes in complex traits.** In *Genetic analysis of complex disease.* Second edition. Edited by Haines JLaP-V, M.A. New Jersey: Wiley-Liss; 2006: 455-467
4. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, et al.: **A polymorphic DNA marker genetically linked to Huntington's disease.** *Nature* 1983, **306:**234-238.
5. Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM: **Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals.** *Cell* 1987, **50:**509-517.
6. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al.: **Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA.** *Science (New York, NY* 1989, **245:**1066-1073.
7. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease.** *Science (New York, NY* 2008, **322:**881-888.
8. Orr N, Chanock S: **Common genetic variation and human disease.** *Advances in genetics* 2008, **62:**1-32.
9. Fan JB, Chee MS, Gunderson KL: **Highly parallel genomic assays.** *Nature reviews* 2006, **7:**632-644.
10. Keerthikumar S, Bhadra S, Kandasamy K, Raju R, Ramachandra YL, Bhattacharyya C, Imai K, Ohara O, Mohan S, Pandey A: **Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach.** *DNA Res* 2009, **16:**345-351.
11. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y: **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** *Nucleic acids research* 2008, **36:**W377-384.
12. Calvo B, Lopez-Bigas N, Furney SJ, Larranaga P, Lozano JA: **A partially supervised classification approach to dominant and recessive human disease gene prediction.** *Computer methods and programs in biomedicine* 2007, **85:**229-237.
13. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *American journal of human genetics* 2006, **78:**1011-1025.
14. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics (Oxford, England)* 2002, **18 Suppl 2:**S110-115.
15. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA: **G2D: a tool for mining genes associated with disease.** *BMC genetics* 2005, **6:**45.

16. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates.** *Nucleic acids research* 2005, **33:**1544-1552.
17. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G: **GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases.** *Nucleic acids research* 2005, **33:**W758-761.
18. Carrel L, Park C, Tyekucheva S, Dunn J, Chiaromonte F, Makova KD: **Genomic environment predicts expression patterns on the human inactive X chromosome.** *PLoS genetics* 2006, **2:**e151.
19. Lombard Z, Tiffin N, Hofmann O, Bajic VB, Hide W, Ramsay M: **Computational selection and prioritization of candidate genes for fetal alcohol syndrome.** *BMC genomics* 2007, **8:**389.
20. Renieri A, Meloni I, Longo I, Ariani F, Mari F, Pescucci C, Cambi F: **Rett syndrome: the complex nature of a monogenic disease.** *J Mol Med* 2003, **81:**346-354.
21. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY: **Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2.** *Nature genetics* 1999, **23:**185-188.
22. Meloni I, Bruttini M, Longo I, Mari F, Rizzolio F, D'Adamo P, Denvriendt K, Fryns JP, Toniolo D, Renieri A: **A mutation in the rett syndrome gene, MECP2, causes X-linked mental retardation and progressive spasticity in males.** *American journal of human genetics* 2000, **67:**982-985.
23. Chamberlain JS, Pearlman JA, Muzny DM, Gibbs RA, Ranier JE, Caskey CT, Reeves AA: **Expression of the murine Duchenne muscular dystrophy gene in muscle and brain.** *Science (New York, NY* 1988, **239:**1416-1418.
24. Chelly J, Gilgenkrantz H, Lambert M, Hamard G, Chafey P, Recan D, Katz P, de la Chapelle A, Koenig M, Ginjaar IB, et al.: **Effect of dystrophin gene deletions on mRNA levels and processing in Duchenne and Becker muscular dystrophies.** *Cell* 1990, **63:**1239-1248.
25. Ropers HH, Hamel BC: **X-linked mental retardation.** *Nature reviews* 2005, **6:**46-57.
26. Ropers HH, Hoeltzenbein M, Kalscheuer V, Yntema H, Hamel B, Fryns JP, Chelly J, Partington M, Gecz J, Moraine C: **Nonsyndromic X-linked mental retardation: where are the missing mutations?** *Trends Genet* 2003, **19:**316-320.
27. Delbridge ML, McMillan DA, Doherty RJ, Deakin JE, Graves JA: **Origin and evolution of candidate mental retardation genes on the human X chromosome (MRX).** *BMC genomics* 2008, **9:**65.
28. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al: **The DNA sequence of the human X chromosome.** *Nature* 2005, **434:**325-337.
29. Graves JA: **The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding.** *Bioessays* 1995, **17:**311-320.
30. Hastie T, Tibshirani R, Friedman JH: *The elements of statistical learning : data mining, inference, and prediction.* New York: Springer; 2001.
31. Chiurazzi P, Schwartz CE, Gecz J, Neri G: **XLMR genes: update 2007.** *Eur J Hum Genet* 2008, **16:**422-434.
32. Katz G, Lazcano-Ponce E: **Intellectual disability: definition, etiological factors, classification, diagnosis, treatment and prognosis.** *Salud Publica Mex* 2008, **50 Suppl 2:**s132-141.
33. Lehrke R: **Theory of X-linkage of major intellectual traits.** *Am J Ment Defic* 1972, **76:**611-619.

34. Giannandrea M, Bianchi V, Mignogna ML, Sirri A, Carrabino S, D'Elia E, Vecellio M, Russo S, Cogliati F, Larizza L, et al: **Mutations in the small GTPase gene RAB39B are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly.** *American journal of human genetics* 2010, **86:**185-195.
35. Dibbens LM, Tarpey PS, Hynes K, Bayly MA, Scheffer IE, Smith R, Bomar J, Sutton E, Vandeleur L, Shoubridge C, et al: **X-linked protocadherin 19 mutations cause female-limited epilepsy and cognitive impairment.** *Nature genetics* 2008, **40:**776-781.
36. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'Meara S, Latimer C, Dicks E, Menzies A, et al: **A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation.** *Nature genetics* 2009, **41:**535-543.
37. Depienne C, Trouillard O, Saint-Martin C, Gourfinkel-An I, Bouteiller D, Carpentier W, Keren B, Abert B, Gautier A, Baulac S, et al: **Spectrum of SCN1A gene mutations associated with Dravet syndrome: analysis of 333 patients.** *J Med Genet* 2009, **46:**183-191.
38. Lahn BT, Page DC: **Four evolutionary strata on the human X chromosome.** *Science (New York, NY* 1999, **286:**964-967.
39. Graves JA, Koina E, Sankovic N: **How the gene content of human sex chromosomes evolved.** *Current opinion in genetics & development* 2006, **16:**219-224.
40. Waters PD, Wallis MC, Marshall Graves JA: **Mammalian sex--Origin and evolution of the Y chromosome and SRY.** *Semin Cell Dev Biol* 2007, **18:**389-400.
41. Thiselton DL, McDowall J, Brandau O, Ramser J, d'Esposito F, Bhattacharya SS, Ross MT, Hardcastle AJ, Meindl A: **An integrated, functionally annotated gene map of the DXS8026-ELK1 interval on human Xp11.3-Xp11.23: potential hotspot for neurogenetic disorders.** *Genomics* 2002, **79:**560-572.
42. Wilczynski GM, Konopacki FA, Wilczek E, Lasiecka Z, Gorlewicz A, Michaluk P, Wawrzyniak M, Malinowska M, Okulski P, Kolodziej LR, et al: **Important role of matrix metalloproteinase 9 in epileptogenesis.** *J Cell Biol* 2008, **180:**1021-1035.
43. Compagni A, Logan M, Klein R, Adams RH: **Control of skeletal patterning by ephrinB1-EphB interactions.** *Dev Cell* 2003, **5:**217-230.
44. Wieland I, Jakubiczka S, Muschke P, Cohen M, Thiele H, Gerlach KL, Adams RH, Wieacker P: **Mutations of the ephrin-B1 gene cause craniofrontonasal syndrome.** *American journal of human genetics* 2004, **74:**1209-1215.
45. Stephenson SE, Dubach D, Lim CM, Mercer JF, La Fontaine S: **A single PDZ domain protein interacts with the Menkes copper ATPase, ATP7A. A new protein implicated in copper homeostasis.** *J Biol Chem* 2005, **280:**33270-33279.
46. Madsen E, Gitlin JD: **Copper and iron disorders of the brain.** *Annu Rev Neurosci* 2007, **30:**317-337.
47. Schlief ML, Gitlin JD: **Copper homeostasis in the CNS: a novel link between the NMDA receptor and copper homeostasis in the hippocampus.** *Mol Neurobiol* 2006, **33:**81-90.
48. Salton M, Lerenthal Y, Wang SY, Chen DJ, Shiloh Y: **Involvement of matrin 3 and SFPQ/NONO in the DNA damage response.** *Cell Cycle* 2010, **9**.
49. Frontini M, Soutoglou E, Argentini M, Bole-Feysot C, Jost B, Scheer E, Tora L: **TAF9b (formerly TAF9L) is a bona fide TAF that has unique and overlapping roles with TAF9.** *Mol Cell Biol* 2005, **25:**4638-4649.
50. Twigg SR, Matsumoto K, Kidd AM, Goriely A, Taylor IB, Fisher RB, Hoogeboom AJ, Mathijssen IM, Lourenco MT, Morton JE, et al: **The origin of EFNB1 mutations in craniofrontonasal syndrome: frequent somatic mosaicism and explanation of the paucity of carrier males.** *American journal of human genetics* 2006, **78:**999-1010.

51. Durkin ME, Ullmannova V, Guan M, Popescu NC: **Deleted in liver cancer 3 (DLC-3), a novel Rho GTPase-activating protein, is downregulated in cancer and inhibits tumor cell growth.** *Oncogene* 2007, **26:**4580-4589.
52. King SR, Manna PR, Ishii T, Syapin PJ, Ginsberg SD, Wilson K, Walsh LP, Parker KL, Stocco DM, Smith RG, Lamb DJ: **An essential component in steroid synthesis, the steroidogenic acute regulatory protein, is expressed in discrete regions of the brain.** *J Neurosci* 2002, **22:**10613-10620.
53. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM, et al: **The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology.** *Nucleic acids research* 2005, **33:**D471-475.
54. Ladeiras-Lopes R, Ferreira-Martins J, Leite-Moreira AF: **The apelinergic system: the role played in human physiology and pathology and potential therapeutic applications.** *Arq Bras Cardiol* 2008, **90:**343-349.
55. Gustavsson M, Mallard C, Vannucci SJ, Wilson MA, Johnston MV, Hagberg H: **Vascular response to hypoxic preconditioning in the immature brain.** *J Cereb Blood Flow Metab* 2007, **27:**928-938.
56. Guo Z, Yuan J, Tang W, Chen X, Gu X, Luo K, Wang Y, Wan B, Yu L: **Cloning and characterization of the human gene RAP2C, a novel member of Ras family, which activates transcriptional activities of SRE.** *Mol Biol Rep* 2007, **34:**137-144.
57. Emes RD, Ponting CP: **A new sequence motif linking lissencephaly, Treacher Collins and oral-facial-digital type 1 syndromes, microtubule dynamics and cell migration.** *Hum Mol Genet* 2001, **10:**2813-2820.
58. Silveira HC, Sommer CA, Soares-Costa A, Henrique-Silva F: **A calcineurin inhibitory protein overexpressed in Down's syndrome interacts with the product of a ubiquitously expressed transcript.** *Braz J Med Biol Res* 2004, **37:**785-789.
59. Froyen G, Bauters M, Voet T, Marynen P: **X-linked mental retardation and epigenetics.** *J Cell Mol Med* 2006, **10:**808-825.
60. Lukong KE, Chang KW, Khandjian EW, Richard S: **RNA-binding proteins in human genetic disease.** *Trends Genet* 2008, **24:**416-425.
61. Chang TC, Mendell JT: **microRNAs in vertebrate physiology and human disease.** *Annu Rev Genomics Hum Genet* 2007, **8:**215-239.
62. Wang X, Wang G, Shen C, Li L, Mooney SD, Edenberg HJ, Sanford JR, Liu Y: **Using RNase sequence specificity to refine the identification of RNA-protein binding regions.** *BMC genomics* 2008, **9 Suppl 1:**S17.
63. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, et al: **eVOC: a controlled vocabulary for unifying gene expression data.** *Genome Res* 2003, **13:**1222-1230.
64. Pan H, Zuo L, Choudhary V, Zhang Z, Leow SH, Chong FT, Huang Y, Ong VW, Mohanty B, Tan SL, et al: **Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining.** *Nucleic acids research* 2004, **32:**W230-234.
65. de Brouwer AP, Yntema HG, Kleefstra T, Lugtenberg D, Oudakker AR, de Vries BB, van Bokhoven H, Van Esch H, Frints SG, Froyen G, et al: **Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium.** *Human mutation* 2007, **28:**207-208.
66. Ropers HH: **X-linked mental retardation: many genes for a complex disorder.** *Current opinion in genetics & development* 2006, **16:**260-269.

67. Park C, Makova KD: **Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes.** *Genome biology* 2009, **10:**R10.
68. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nature genetics* 2006, **38:**626-635.
69. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, et al: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.** *Nature methods* 2005, **2:**105-111.
70. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic acids research* 2007, **35:**D61-65.
71. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, et al: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones.** *PLoS biology* 2004, **2:**e162.
72. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update.** *Nucleic acids research* 2004, **32:**D23-26.
73. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists.** *Curr Protoc Mol Biol* 2010, **Chapter 19:**Unit 19 10 11-21.
74. Schroer A, Schneider S, Ropers H, Nothwang H: **Cloning and characterization of UXT, a novel gene in human Xp11, which is widely and abundantly expressed in tumor tissue.** *Genomics* 1999, **56:**340-343.
75. Yoon HG, Chan DW, Huang ZQ, Li J, Fondell JD, Qin J, Wong J: **Purification and functional characterization of the human N-CoR complex: the roles of HDAC3, TBL1 and TBLR1.** *EMBO J* 2003, **22:**1336-1346.
76. Ito S, Kawano Y, Katakura H, Takenaka K, Adachi M, Sasaki M, Shimizu K, Ikenaka K, Wada H, Tanaka F: **Expression of MAGE-D4, a novel MAGE family antigen, is correlated with tumor-cell proliferation of non-small cell lung cancer.** *Lung Cancer* 2006, **51:**79-88.
77. Tatemoto K, Hosoya M, Habata Y, Fujii R, Kakegawa T, Zou MX, Kawamata Y, Fukusumi S, Hinuma S, Kitada C, et al: **Isolation and characterization of a novel endogenous peptide ligand for the human APJ receptor.** *Biochem Biophys Res Commun* 1998, **251:**471-476.

Chapter 6

Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes

This chapter has been published in Genome Biology 2009, 10: R10, and was formatted for that journal. Authors for this original manuscript are Chungoo Park and Kateryna D. Makova. CP and KDM designed the experiments and wrote the manuscript. CP performed data analyses.

Abstract

Background: Gene expression divergence is one manifestation of functional differences between duplicate genes. Although rapid accumulation of expression divergence between duplicate gene copies has been observed, the driving mechanisms behind this phenomenon have not been explored in detail.

Results: We examine which factors influence expression divergence between human duplicate genes, utilizing the latest genome-wide data sets. We conclude that the turnover of transcription start sites between duplicate genes occurs rapidly after gene duplication and that gene pairs with shared transcription start sites have significantly higher expression similarity than those without shared transcription start sites. Moreover, we find that most (55%) duplicate gene pairs do not retain the same coding sequence structure between the two duplicate copies and this also contributes to divergence in their expression. Furthermore, the proportion of aligned sequences in *cis*-regulatory regions between the two copies is positively correlated with expression similarity. Surprisingly, we find no effect of copy-specific transposable element insertions on the divergence of duplicate gene expression.

Conclusions: Our results suggest that turnover of transcription start sites, structural heterogeneity of coding sequences, and divergence of *cis*-regulatory regions between copies play a pivotal role in determining the expression divergence of duplicate genes.

Background

Because of the importance of gene duplication in evolution [1-5], it is crucial to know how duplicate genes diverge and which factors determine their destiny. Recently, genome-wide analyses of microarray data [6] have revealed patterns of expression divergence in duplicate genes, which are necessary for understanding the emergence of new functions after gene duplication. Numerous studies indicated that genes diverge rapidly in their expression after duplication [7-12]. Population genetic models proposed directional selection and relaxation of selective constraints as possible forces driving the evolution of expression in duplicate genes, although the relative frequency of these two scenarios in the evolution of paralogs is still being debated [4,5,13]. These population genetic models have been implemented under the assumption that two duplicated gene copies are structurally and functionally identical immediately after duplication. However, this assumption is sometimes violated. First, genes duplicated via retrotransposition lose regulatory sequences and include additional sequences at each side (for example, poly(A) tails at 3′ terminus and short direct repeats at both termini), so that retrotransposed copies differ from the corresponding parental genes [4,13,14]. Second, tandem duplication by unequal crossing over might not include the entire coding sequence and/or regulatory elements specifying expression of a parental gene. Indeed, Katju and Lynch [15] demonstrated that more than half of newborn duplicate genes in *Caenorhabditis elegans* represent not complete, but rather partial or chimeric duplications. Such structural heterogeneity may play

an important role in rapid expression divergence between human duplicate genes as well; however, it has not been considered in detail in previous studies.

Transposable elements (TEs) represent another factor that might account for the expression divergence of duplicate genes, since several studies provided evidence of TEs altering gene expression. Jordan and colleagues [16] showed that almost 25% of human promoter regions as well as many other *cis*-regulatory elements contain, or at least overlap with, TE-derived sequences. This result was later confirmed by another study [17]. A specific example of the importance of TEs in the regulation of gene expression comes from the *CYP19* gene, which encodes the aromatase enzyme, important for estrogen biosynthesis [18]. Because of the recent insertion of a long terminal repeat into the first exon of one of the isoforms of human *CYP19*, the gene gained expression in placenta, while its mouse ortholog has no long terminal repeat and is not expressed there [19].

Finally, alternative promoter usage by duplicate genes should be considered as a mechanism for rapid expression divergence. Recent comprehensive studies concluded that many known genes in the human genome are expressed from alternative promoters [20-23]. Similarly, approximately 22% of genes in the ENCODE regions have functional alternative promoters [24]. The alternative promoters provide a heterogeneity in tissue-specific expression patterns and levels, developmental activity, and translational efficiency [25-27]. As a result, the use of alternative promoters might be one of the major sources for achieving transcriptome diversity and one of the routes by which duplicate genes acquire divergence in their expression.

To investigate what drives expression divergence of human paralogs on a genome-wide scale, we addressed the following three questions in the present study: how frequently the turnover of transcription start sites (TSSs) occurs between duplicate genes; how often duplicate gene copies (their coding sequences) differ from each other structurally; and whether the density of copy-specific TEs within *cis*-regulatory regions influences expression divergence in duplicated

genes. We utilized the gene expression profile available for 61 non-redundant and non-pathogenic human tissues [28], the largest comprehensive expression profile of human genes available to date, and assessed the contributions of TSS turnover, coding sequence structural heterogeneity, and TE integration to divergence in duplicate gene expression.

Results

Identification of duplicate genes

Utilizing two different methods, FASTA and TRIBE-MCL, we identified 6,536 and 7,027 non-redundant human duplicate gene pairs, respectively (see Materials and methods for details). These pairs represented 3,313 and 3,555 gene families, respectively. After filtering out duplicate gene pairs with synonymous rate (K_S) >2 and/or lacking a start codon, we obtained 2,790 and 2,750 duplicate gene pairs using the former and the latter methods, respectively. A total of 1,600 duplicate gene pairs overlapped between these two data sets (Additional data file 2). All subsequent analyses were carried out for duplicate genes identified with each of the two methods. Because the results were similar, we present the results only for duplicate genes identified with the FASTA method (2,790 gene pairs in group A), as this method is stricter for clustering proteins into families compared with the TRIBE-MCL method [29,30].

From human U133A and GNF1H oligonucleotide arrays [28], we defined 14,505 genes that mapped to probes with a one-to-one correspondence (see Materials and methods), thus minimizing cross-hybridization. Among these genes, we were able to detect 2,924 non-redundant duplicate gene pairs belonging to 1,792 multiple gene families. After filtering out duplicate gene pairs with K_S >2 and/or lacking a start codon, we obtained 1,015 duplicate gene pairs (group B,

representing a subset of group A). In the remainder of the manuscript, we consider duplicate genes of group B when gene expression is investigated and duplicate genes of group A otherwise.

Turnover of TSSs between duplicate genes

Initially, we analyzed the divergence in the position of TSSs between copies in each duplicate gene pair. Using tag clusters, which were built by grouping overlapping tags (namely, 5′-end-sequences) with the same strand, from large-scale tag clustering of the cap analysis of gene expression (CAGE) [20] and the paired-end ditags (PETs) [31], putative TSSs of each gene were identified (see Materials and methods). From 2,790 duplicate gene pairs in group A, we excluded duplicate gene pairs that were duplicated by retrotransposition or for which at least one copy lacked a TSS(s) identified by either CAGE or PETs. As a result, 1,124 duplicate gene pairs were retained. To evaluate sharing of TSSs between duplicate genes, we compared the sequences of genomic regions surrounding putative TSSs (as identified by CAGE or PETs) between the two copies for each of these 1,124 duplicate gene pairs. We considered 110 bp (-20 bp to +90 bp) surrounding each TSS (later called the 'TSS region'), because there was a clear peak in the average sequence similarity between TSSs of duplicate genes in this region (Additional data file 3) and because several studies indicated that a region of this size surrounding TSSs was well conserved between human and mouse orthologs [32,33]. Sequence similarity between all possible combinations of TSS regions from each duplicate gene pair was considered. If at least one pair of TSS regions had an identity greater than 60%, it was defined as a TSS(s) shared between the two duplicate copies. As a result, 13.6% (153 out of 1,124) of duplicate gene pairs had shared TSSs.

We observed that the relative frequency of gene pairs with shared TSSs decreases with increasing K_S, a proxy of time since duplication (Figure 6-1). The L-shaped distribution observed in Figure 6-1 implies a rapid turnover of TSSs after gene duplication. Already at $K_S = 0.1$,

corresponding to only about 33 million years ago since duplication [34], a mere 64% of duplicate genes share TSSs. Considering an instantaneous K_S rate according to [35] did not alter our results (Additional data file 4).

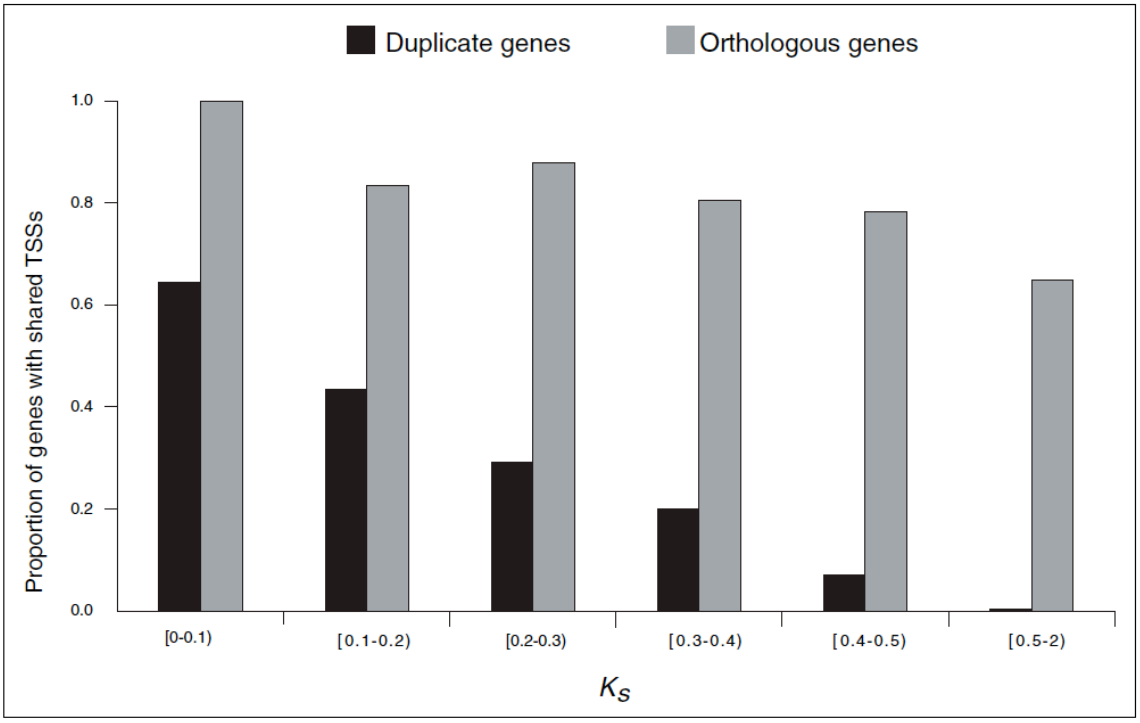


Figure **6-1**: The decline in the proportion of group A duplicate gene pairs with shared TSSs (shown in black) depending on the time since duplication (approximated by K_S). The proportion of human-mouse orthologous genes with conserved TSSs is shown for comparison (in gray); in this case variation in K_S is due to regional variation in substitution rates.

Interestingly, the turnover of TSSs between human duplicate genes was much more rapid than between human-mouse orthologs. Indeed, for 1,610 human-mouse orthologs considered (see Materials and methods), the mean K_S was 0.61 (with a 95% confidence interval of 0.60-0.63), while the proportion of orthologs with shared TSSs was 0.71, several fold higher than the proportion of human duplicate genes with similar K_S (Figure 6-1).

To estimate the relationship between TSS usage patterns (for example, shared TSSs versus non-shared TSSs) and gene duplication mechanisms, the duplicate genes were divided into

three classes: retrotransposed duplicate genes, tandem, and nontandem duplications (see Materials and methods for details). The relative frequencies of gene pairs with shared TSSs in each class were calculated (thus, we analyzed 1,124 non-retransposed genes as above plus 220 retrotransposed genes). Duplicate gene copies in which one of the pair has one exon and the duplicate copy has multiple exons were called retrotransposed duplicate gene copies. We found that among paralogs with shared TSSs, the majority of pairs represented tandem duplicates (Additional data file 1).

Interestingly, about 30% (67 out of 220) of retrotransposed duplicate gene pairs retained the same TSSs (Additional data file 1). To evaluate whether the retrotransposed gene pairs with shared TSSs tend to undergo stronger purifying selection than those without shared TSSs, the median nonsynonymous-to-synonymous rate ratios (K_A/K_S) were compared between these two groups of genes; however, no significant difference was detected (0.475 versus 0.499; $P > 0.1$, Mann-Whitney U test).

Next, to test whether the turnover of TSSs may contribute to the expression divergence in duplicate genes, the Pearson correlation coefficient of expression values ($R_{expression}$; calculated for 61 non-redundant tissues) between the two copies in each pair was computed and compared among group B duplicate gene pairs with shared TSSs versus those without shared TSSs (a total of 581 group B pairs with available TSS data were included in the analysis). Duplicate genes with shared TSSs had significantly higher $R_{expression}$ values than those without shared TSSs (0.437 versus 0.080; $P < 0.01$, Mann-Whitney U test). It is conceivable that the significant difference in $R_{expression}$ values is due to different synonymous rates in genes with shared TSSs versus those without shared TSSs. Indeed, we observed that all duplicate genes (belonging to group B) with shared TSSs had $K_S < 0.4$, while more than 97% of gene pairs without shared TSSs had $K_S \geq 0.4$. However, if only genes with $K_S < 0.4$ were considered, the gene pairs with shared TSSs still had

higher (but not significantly so) $R_{expression}$ values than those without shared TSSs (0.437 versus 0.140; $P > 0.05$, Mann-Whitney U test).

The 60% identity threshold among the TSS regions that was tentatively inferred from substitution rates between human and mouse ortholog core promoters [36] may be inadequate for estimating the sharing of TSSs among human paralogous genes. Thus, we reclassified the sharing of TSSs between copies of duplicate genes using several identity thresholds (40%, 50%, 70%, and 80%). Although the numbers of duplicate genes with shared TSSs in each bin varied with the threshold, the frequency of gene pairs with shared TSSs decreased over divergent time independent of the threshold used (Additional data file 5), consistent with the pattern observed with the 60% identity threshold (Figure 6-1). Moreover, regardless of the identity threshold, the $R_{expression}$ values were significantly higher in duplicate genes with shared TSSs versus those without shared TSSs (data not shown).

Structural heterogeneity in coding regions of human duplicate genes

By reconstructing the full-length coding sequences via concatenating exons from multiple splicing variants for each gene separately, each pair of duplicate genes was classified into one of two structural categories: completely similar and incompletely similar. If the proportion of aligned sequences was greater than 0.9, duplicate gene pairs were categorized as completely similar and as incompletely similar otherwise. For some analyses, incompletely similar duplicate gene copies were classified in one of the three non-overlapping groups: 5′ similar, 3′ similar, and neither 5′ nor 3′ similar. If alignments between the two copies started at the start codons of both copies, then such duplicates were classified as 5′ similar. Alternatively, if the alignments ended at the stop codons of both copies, we classified the duplicate genes as 3′ similar.The remaining duplicate gene pairs were labeled as neither 5′ nor 3′ similar.

After excluding genes that lacked start/stop codons or consensus splice sites, 2,591 duplicate gene pairs were retained (from 2,790 pairs of group A; for group B, 889 duplicate gene pairs were retained). We found that 55% (1,429 out of 2,591) of duplicate gene pairs had incompletely similar structures. As expected from the divergence of the coding sequence over time, the proportion of duplicate gene pairs with completely similar structures decreased gradually with divergence between the two duplicate copies, approximated by K_S (Figure 6-2). Considering an instantaneous K_S rate according to [35] did not alter our results (Additional data file 6). Interestingly, even at the smallest duplicate gene divergence (K_S <0.1), the proportion of genes with completely similar structures was only 80% (Figure 6-2). Although this finding might be affected by misannotations, our results suggest that some duplicate genes might have acquired structural differences during duplication.

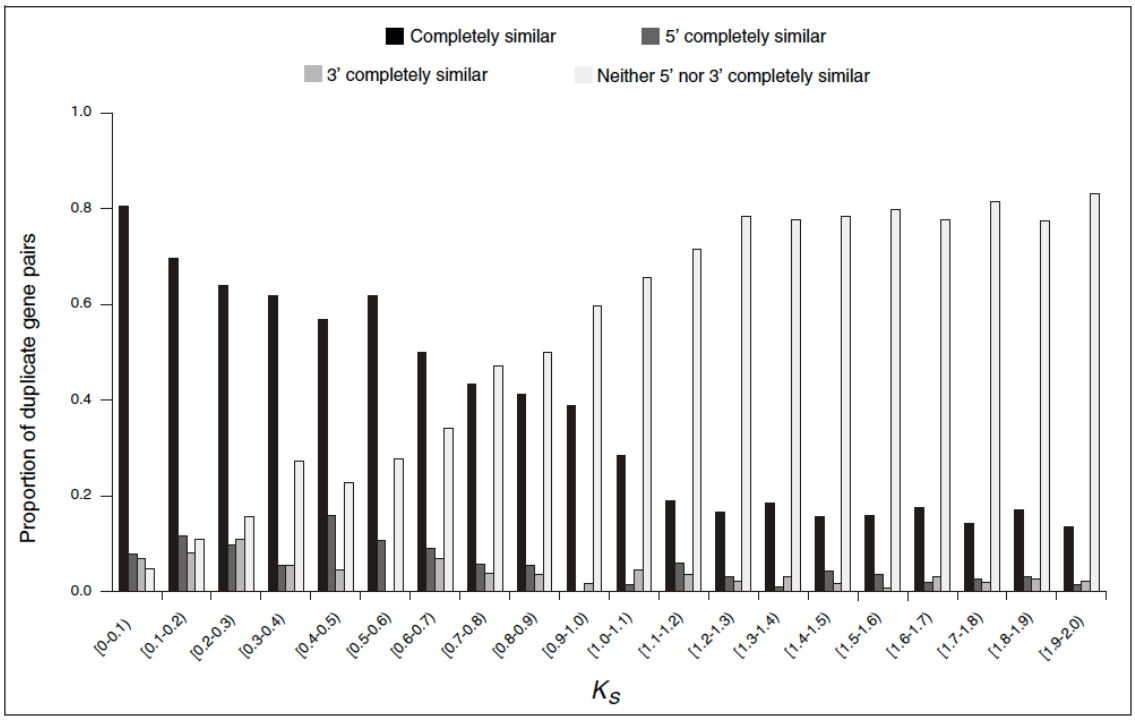


Figure **6-2**: Proportion of group A duplicate gene pairs classified by coding sequence structural heterogeneity.

To analyze whether the incompletely similar structures of duplicate genes can lead to expression divergence, we compared the relationship between $R_{expression}$ and K_S for duplicate genes with completely versus incompletely similar structures. Before addressing this issue, retrotransposed duplicate genes (a total of 108 out of 889 genes retained in group B) were excluded because, as retrotransposition does not include a promoter, it can lead to expression divergence regardless of structural heterogeneity in coding sequence between duplicates. We found that: the correlation coefficient between $R_{expression}$ and K_S for duplicate gene pairs with completely similar structures was significantly lower than that for pairs with incompletely similar structures (R = -0.315 versus R = -0.001; Fisher's z test, z = -4.028, $P < 0.001$; Kolmogorov-Smirnov test for normality, $P < 0.010$; Figure 6-3 and Table 6-1); and duplicate genes with completely similar structures had significantly higher y-intercepts of regression lines than duplicate genes with incompletely similar structures (0.407 versus 0.134; z = 2.672, $P < 0.01$). These observations suggest that, immediately after duplication, the expression pattern is more similar for duplicate gene pairs retaining the same versus acquiring different coding sequence structures, and that divergence of gene expression is more dependent on evolutionary time for duplicate gene pairs with completely versus incompletely similar structures. To estimate the importance of sharing of 5′ regions of coding sequences between duplicate gene copies, which can be an indirect indicator of common transcription regulation mechanisms, we separately considered duplicate gene pairs completely similar at the 5′ end only (a total of 24 gene pairs from group B that were otherwise genes with incompletely similar structures and calculated the correlation coefficient between their $R_{expression}$ and K_S. The correlation was negative, but not significant (Table 6-1). When duplicate gene pairs having completely similar and 5′ similar structures were considered together, the correlation coefficient between $R_{expression}$ and K_S was somewhat lower than that for duplicate gene pairs with completely similar structures (Table 6-1), although the difference was not significant (z = -0.093, $P > 0.1$). We observed that there was no

correlation between $R_{expression}$ and K_S for duplicate genes with 3′ similar structure and with neither 5′ nor 3′ similar structure (Table 6-1). These results suggest that maintenance of the entire coding region (and not just of its 5′ or 3′ portion) is important for determining gene expression profile after duplication.

Table **6-1**: The relationship between K_S and $R_{expression}$ in each structural category using group B duplicate gene pairs.

Structural categories	Number of gene pairs	K_A/K_S*	K_S*	$R_{expression}$*	Pearson correlation coefficient of K_S versus $R_{expression}$ (*P*-value)
Completely similar	214	0.296 (0.237)	1.153 (1.225)	0.213 (0.162)	-0.315 (<0.001)
5' similar	24	0.391 (0.311)	1.292 (1.501)	0.053 (0.026)	-0.157 (NS)
3' similar	23	0.302 (0.311)	1.365 (1.610)	0.346 (0.249)	0.019 (NS)
Neither 5' nor 3' similar	520	0.551 (0.456)	1.565 (1.658)	0.126 (0.063)	0.017 (NS)
Incompletely similar (the sum of the above three categories)	567	0.534 (0.444)	1.545 (1.646)	0.132 (0.068)	-0.001 (NS)
Completely and 5' similar	238	0.307 (0.246)	1.167 (1.263)	0.197 (0.151)	-0.307 (<0.001)

*Values are mean (median). NS, not significant.

To estimate differences in selective pressure among duplicate genes in different structural categories, their K_A/K_S ratios were compared (Table 6-1). We observed that K_A/K_S was significantly lower for duplicate genes with completely similar structures than for those with incompletely similar structures ($P < 0.001$, Mann-Whitney U test; Table 6-1), suggesting that the former genes are subject to stronger purifying selection than the latter genes.

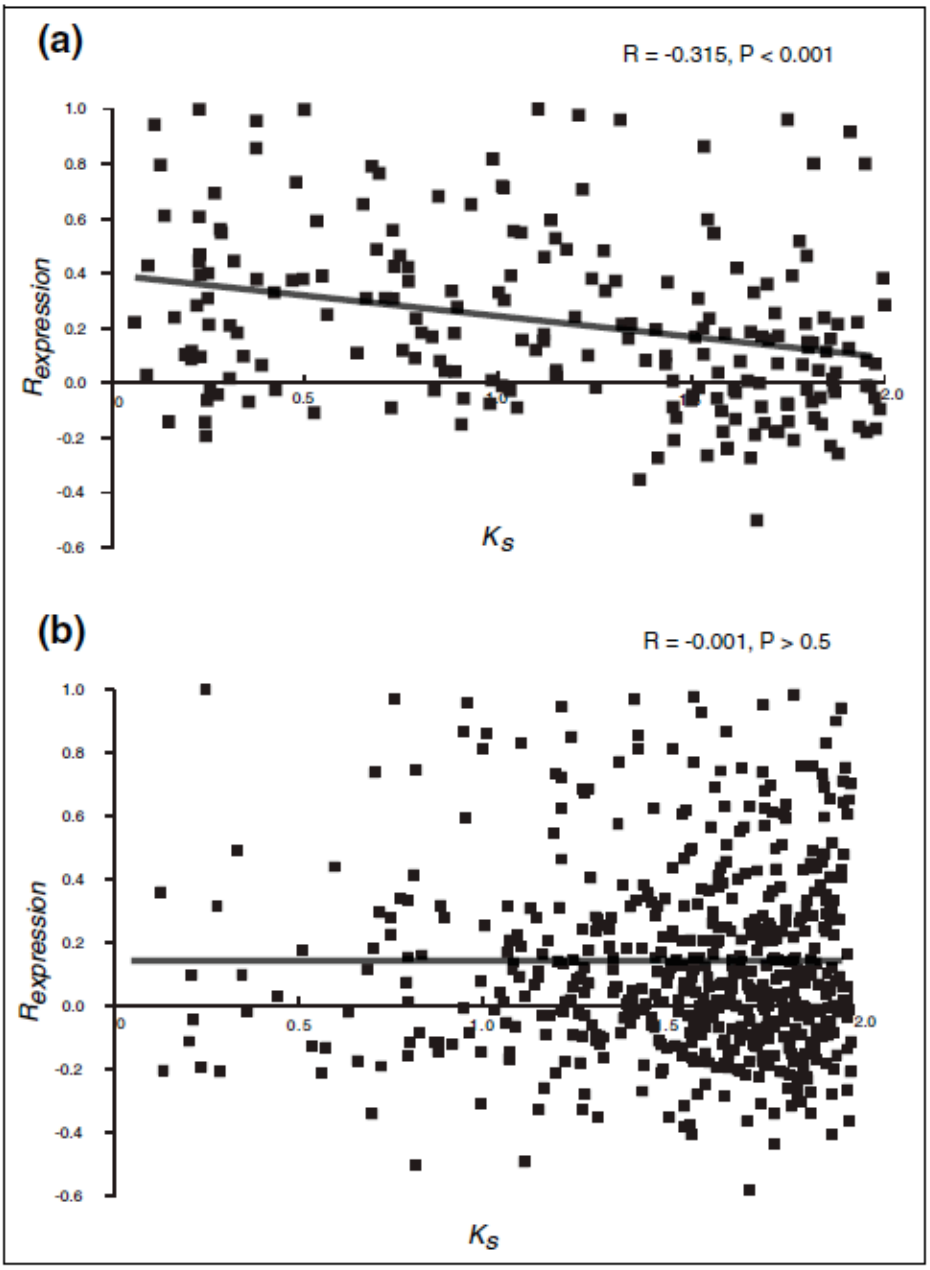


Figure **6-3**: The relationship between K_S and $R_{expression}$ for group B duplicate genes with **(a)** completely similar structures and **(b)** incompletely similar structures.

Divergence of *cis*-regulatory sequences between duplicate genes

Next, we evaluated the relative contribution of *cis*-regulatory divergence to differences in expression between copies of duplicate genes in each pair. The 2-kb (from -1.5 kb to +0.5 kb) genomic regions surrounding TSSs were used as putative *cis*-regulatory sequences and their divergence was estimated with REALIGNER [37]. For genes with multiple TSSs, a TSS supported by the highest number of CAGE/PET tags was selected. This analysis was limited to group B duplicate genes with completely similar structures (a total of 158 duplicate gene pairs). We found a significant positive correlation (R = 0.242, $P < 0.01$) between the proportion of aligned sequences in the *cis*-regulatory region (P_{cis}) and $R_{expression}$. This implies that the divergence of *cis*-regulatory regions leads to expression divergence in duplicate genes. After duplicate genes created by retrotransposition (a total of 23 gene pairs) were excluded, the correlation coefficient was even higher (R = 0.252, $P < 0.01$). Through comparison between K_S (which may serve as a neutral proxy, although see [38]) on the one hand and the proportion (corrected for multiple hits using HKY85 model) of aligned sequences in the *cis*-regulatory region on the other hand in each non-retrotransposed duplicate gene pair, we estimated whether the *cis*-regulatory regions evolved neutrally. We found that for 107 out of 135 duplicate gene pairs compared, K_S was significantly higher ($P < 0.001$, Wilcoxon signed-rank test) than the proportion of aligned sequences in the *cis*-regulatory region, suggesting that purifying selection acts at *cis*-regulatory regions.

To investigate whether copy-specific TEs influence divergence in duplicate gene expression, we identified such TEs (TEs that integrated in the *cis*-regulatory region of only one duplicate gene copy of a pair after duplication) in the same 2-kb regions surrounding TSSs of the above 158 duplicate genes pairs (excluding 23 retrotransposed duplicate pairs; see Materials and methods). However, no significant correlation was found between the proportion of copy-specific TEs and either P_{cis} for duplicate genes or $R_{expression}$ (data not shown). This suggests that the effect

of copy-specific TEs on divergence in duplicate gene expression may be at best minor, although this issue requires additional studies.

Table **6-2**: Multiple regression models for expression divergence in duplicate genes.

Predictors	*P*-value	RCVE*
Cis†	4.2×10^{-2} (NS‡)	0.075
TSS§	9.9×10^{-5}	0.277
Tandem¶	2.7×10^{-6}	0.405
K_A × Cis	1.1×10^{-2} (NS)	0.118
K_S × Cis	2.7×10^{-2} (NS)	0.088
Structure¥ × Tandem	1.7×10^{-3}	0.180
TSS × Tandem	1.1×10^{-5}	0.354
ω# × Cis	3.1×10^{-3}	0.159
R^2		0.093

*RCVE: relative contribution to the variability explained (see Materials and methods for more details). †Cis: divergence of *cis*-regulatory sequences in 2 kb surrounding TSS (see Materials and methods for more details). ‡NS: not significant after Bonferroni correction for multiple tests. §TSS: shared versus not shared TSSs. ¶Tandem: tandem versus nontandem organization of duplicate genes. ¥Structure: structural heterogeneity in coding sequences. #ω:K_A/K_S ratio.

Interplay of multiple predictors in explaining divergence of paralogous gene expression

Because several factors studied above might be interrelated, we conducted multiple regression analysis to estimate the relative contribution of each factor to explaining the total variability in $R_{expression}$. A total of four continuous predictors (K_A, K_S, the K_A/K_S ratio, and divergence of *cis*-regulatory sequences (labeled 'Cis') and three categorical predictors (shared versus not shared TSSs (labeled 'TSS'); completely versus incompletely similar gene structure (labeled 'Structure'); and tandem versus non-tandem gene organization (labeled 'Tandem')) as

well as all possible pairwise interaction terms were used to build a regression model. After pruning nonsignificant terms, the final multiple regression model explained approximately 10% of the variation in $R_{expression}$ and consisted of eight predictors (Table 6-2). Five of these predictors remained significant after applying Bonferroni correction for multiple tests (Table 6-2). These predictors included: Tandem, TSS, and interaction terms between Structure and Tandem, between TSS and Tandem, and between K_A/K_S ratio and Cis (Table 6-2). Our computation of the relative contribution of the variability explained (RCVE) for significant predictors (see Materials and methods for details) indicated that each of them makes a sizeable input into the model.

Discussion

Although it has been shown that duplicate genes diverge rapidly in their expression [10,39-41], little is known about which factors influence their expression divergence at the genomic level [42]. In this study, we investigated three such factors: structural heterogeneity of coding sequences, turnover of TSSs, and divergence of *cis*-regulatory regions (including insertions of copy-specific TEs).

Our results indicate that structural differences in coding sequences are common among human duplicate genes. We observed a high proportion of duplicate genes with structural differences even among young duplicates (K_S <0.1), which is consistent with the findings for *C. elegans* duplicate genes [15]. Thus, genes might already be structurally different at the point of duplication. In general, duplication by unequal crossing over might not contain the entire coding sequence of a parental gene, and indeed, for the majority of individual young duplicate gene pairs with incompletely similar structures in our data set (for approximately 90% of duplicate pairs of group A), both copies reside on the same chromosome. Over time, duplicate genes accumulate mutations leading to amino acid changes, premature stop codons, and atypical splicing [4,14,43].

These mutations might lead to decreasing numbers of duplicate genes retaining their ancestral structure and lead to more rapid divergence in expression and function.

Alteration of TSSs between duplicate gene copies is likely to have a direct impact on expression divergence. Using sequence similarity analysis, we examined whether duplicate genes share their TSSs. A large number of duplicate genes with distinct TSSs between the two copies were observed and these duplicate gene copies usually had different expression patterns. Although we did not directly estimate the fitness effects of turnover of TSSs on retention of duplicate genes, alteration of TSSs provides a means for the realization of several models of gene duplication evolution (for example, subfunctionalization and neofunctionalization [44,45]).

Additionally, we observed that *cis*-regulatory regions of duplicate genes diverge with time since duplication. This is consistent with several previous reports [46-48]. We investigated a potential impact of the density of copy-specific TEs on the divergence of duplicate gene expression and, surprisingly, found no major effect. This result corroborates recent findings regarding orthologous mammalian promoters; in human core promoters, the density of most observed repeat classes was significantly below the genomic average, suggesting that insertion of TEs in *cis*-regulatory regions is prevented by purifying selection [36].

Using multiple regression analysis, we observed that shared versus not shared TSS ('TSS'), completely versus incompletely similar structure ('Structure'), divergence of *cis*-regulatory sequences ('Cis'), the K_A/K_S ratio, and tandem versus non-tandem duplicate gene organization played an important role in determining divergence in duplicate gene expression. It is worth noting that all three novel predictors introduced in this manuscript (TSS, Structure, and Cis) significantly influence divergence in duplicate gene expression alone and/or through interaction with other predictors. Interestingly, K_S, a proxy of evolutionary time, was not a significant predictor in our model. However, as noted above, evolutionary time influences alterations in other predictors and, therefore, the influence of K_S on $R_{expression}$ might be observed

through significance of predictors dependent on K_S. While interaction terms are not straightforward to interpret, the finding that several of them significantly contributed to the model suggests that considering multiple correlated factors might be essential for understanding patterns of duplicate gene expression divergence.

In this study, expression pattern was used as an indicator of evolution of biological functions after gene duplication. Several studies have suggested that gene expression density and breadth (for example, in housekeeping versus tissue-specific genes) has significantly influenced the evolution of proteins [49-52]. In addition to gene expression, which is likely a strong predictor [53,54], several additional factors have been implicated in protein evolution. Such factors include gene dispensability [55,56], protein stability and interaction network [57,58] as well as codon usage [54,59]. Although these variables individually explain only a small fraction of variation in the rate of protein evolution, studying them might provide important insights into divergence between duplicate genes.

Most gene evolution models have assumed that two duplicate gene copies are expressed equally immediately after duplication. However, similarly to coding sequences, promoter regions might also be incompletely duplicated between copies; this possibility needs to be evaluated in future studies. Frequently, because of the complex evolutionary dynamics of promoter sequences [47,60,61], it is difficult to distinguish incomplete promoter duplication from rapid promoter evolution after duplication.

Reconstruction of ancestral gene expression state can be performed using a parsimony-based procedure in multi-gene families [62], instead of using the pairwise analysis employed here. However, rigorous filtering for potential cross-hybridization of transcripts of genes from the same multi-gene family in our study makes such ancestral reconstruction difficult. Thus, additional studies using different types of expression data may allow us to decompose the expression

divergence of genes in multi-gene families and thus provide us with additional methodological insights for understanding gene expression divergence.

In the present study, as expected, we observed a significant negative correlation between the synonymous rate and Pearson correlation coefficient of expression values between duplicate gene copies; however, the resulting correlation was weaker than in our previous study [10]. There might be several potential reasons explaining this difference (for example, different K_S thresholds used in the two studies and a greater number of tissues used in the present study). However, the major advance of the present study compared with the previous one [10] is a more rigorous filtering for potential cross-hybridization of transcripts of two duplicate gene copies to the same probe, and thus we consider the present results more robust.

Conclusion

The present study represents the first report of the effects of structural differences in coding region and of unique TSSs on the divergence of duplicate gene expression. Our observations of frequent turnover of TSSs between duplicate genes and a high proportion of young duplicate genes with incompletely similar structures contradict the assumptions of classic gene duplication models, according to which duplicate genes are considered to be equal both structurally and functionally at the point of duplication [4,13,14]. Although potential incomplete duplication of promoters will be the subject of future studies, our investigation of factors contributing to expression divergence of duplicate genes provides important information for understanding human transcriptome heterogeneity, complexity, and evolution.

Materials and methods

Identification of duplicate gene pairs

To cluster genes into families, we downloaded 48,218 protein sequences of consensus coding sequences, known and novel genes from Ensembl (release 38 of NCBI build 36) and independently used the FASTA [63] and TRIBE-MCL [64] methods to define duplicate gene families. Briefly, for the FASTA method, each protein sequence was used as a query to search against all other protein sequences using FASTA [65] with E < 10. Two protein sequences formed a link if: the aligned region was >80% of the longer protein; and the identity between two proteins was ≥30% for alignments longer than 150 amino acids or $\geq(0.01n + 4.8L^{-0.32[1+\exp(-L/1000)]})$ otherwise, where L is the alignable length between two proteins and $n = 6$. The formula above was derived from empirical data, which suggested that a higher sequence identity was required for shorter proteins [66]. These gene pairs were grouped into gene families according to the single linkage clustering algorithm. For gene families derived by TRIBE-MCL, we downloaded the gene annotations through BioMart in the Ensembl database, and considered gene families with at least two members.

To identify independent pairs of duplicate genes within each gene family, we sorted gene pairs in ascending order of K_S and selected the pair with the lowest K_S. After excluding genes that had been picked, we chose the next gene pair with the lowest K_S. These steps were repeated for each gene family. All genes encoding proteins were realigned using CLUSTALW [67], and the yn00 module [68] of PAML [69] was used to calculate K_S. We counted duplicate gene pairs in intervals of size $K_S = 0.01$ to derive the instantaneous rate of K_S according to [35].

Duplicate gene copies in which one of the pair has one exon and the duplicate copy has multiple exons were called retrotransposed duplicate gene copies. In addition, duplicate gene pairs were classified as tandem duplicates if there were no genes separating them.

Expression data analysis

Expression data for 61 non-redundant and nonpathogenic human tissues in U133A and GNF1H Affymetrix arrays were obtained from [28]. To validate mapping between probe sets and genes, we aligned the transcripts of consensus coding sequences, known genes, and novel genes downloaded from Ensembl (release 38 of NCBI build 36) with the exemplar and consensus sequences for each array using BLAST [70] with $E < 10^{-20}$. According to the criteria described in [71,72], the acceptable alignments were selected if: the identity was 100% and the length was greater than 49 bp; or the identity was higher than 94% and the length was at least either 99 bp or 90% of the length of the query. We considered three scenarios for mapping relationships: a single probe set hitting one gene (9,508 probe sets); multiple probe sets hitting one gene (13,186 probe sets and 4,997 genes); and a single probe set hitting multiple genes (4,493 probe sets and 6,764 genes). All genes following the first two scenarios were utilized in the present study. For each gene following the second scenario, the probe set with the highest expression value (defined by average difference) was selected. All genes following the third scenario were removed from the analysis due to potential cross-hybridization. Following [28], genes with average difference >200 in a particular tissue were considered to be expressed in this tissue.

Identification of putative TSSs

The putative TSSs were identified using the method described in the ENCODE pilot project [73]. Briefly, we utilized tag clusters from two sets of 5′-end-tag-capture technologies: CAGE [20] and PETs [31]. If two tag clusters were located on the same strand and within 60 bp (which was derived from analyzing the distribution of distances between tag clusters in [73]) of each other, they were considered as one tag cluster. To map tag clusters to genes, the following two criteria were considered. First, the strand of a tag cluster was required to be identical to the strand of a gene. Second, a tag cluster was required to be located in the 5′ upstream region from the most upstream start codon of a gene. Because we constructed artificial coding regions of genes by including all their exons, our analysis is not affected by alternative start codons. To confirm the reliability of the tag data, RefSeq [74], H-Invitational [75] and human ESTs [76] RNA data from the UCSC Genome Browser [77] were utilized. We excluded tag clusters with a single tag as well as those whose coordinates did not overlap with the genomic coordinates of the 5′ end of cDNAs or ESTs. To define a representative tag site (to be used as a putative TSS) for each tag cluster, we selected the tag site that was supported by the highest number of 5′ start sites. Otherwise, if several sites in a tag cluster had the same number of 5′ start sites, the central coordinate of this tag cluster was defined as the representative tag site.

Analysis of turnover of TSSs between human-mouse orthologous gene pairs

To evaluate conservation of TSSs between human-mouse orthologous genes, we obtained two distinct classes of orthologous genes from [23]. Briefly, ‘conserved promoter regions’ means that upstream sequences of TSSs between human and mouse orthologous genes were aligned; otherwise, ‘non-conserved promoter regions’ means there were no significant alignments. We

excluded orthologous genes that were classified into both classes because alternatively spliced variants of each gene had different conservation patterns of promoter regions. As a result, 1,610 orthologous gene pairs that were classified into just one class in a mutually exclusive manner were retained. We downloaded human and mouse protein sequences from Ensembl (release 38 of NCBI build 36). All genes were aligned using CLUSTALW [67], and the yn00 module [68] of PAML [69] was used to calculate K_S between orthologous genes.

Classification of the type of gene duplication into structural categories

Structural categorization of duplicate genes was performed using reconstructed full-length coding sequences. We downloaded annotated human genome data from Ensembl (release 38 of NCBI build 36). Alternatively spliced variants lacking start or stop codons or lacking canonical exon boundaries (5′-GT…AG-3′, 5′-GC…AG-3′, or 5′-AT…AC-3′) were excluded. For each gene with several alternatively spliced variants, all exons were aligned against each other, and, if some exons overlapped, they were merged in a single exon. Next, exons were sorted by their genomic coordinates and were reassembled to form reconstructed full-length coding sequences.

The reconstructed full-length coding sequences were aligned using AVID [78] with default parameters. Each pair of duplicate genes was classified into one of the four structural categories: completely similar, 5′ similar, 3′ similar, and neither 5′ nor 3′ similar. If the proportion of aligned sequences was greater than 0.9, duplicate gene pairs were categorized as completely similar. The other duplicate gene pairs were exclusively classified in just one category of 5′ similar, 3′ similar, or neither 5′ nor 3′ similar. If alignments between the two copies started at the start codons of both copies, then such duplicates were classified as 5′ similar. Alternatively, if the

alignments ended at the stop codons of both copies, we classified the duplicate genes into 3′ similar. Finally, the remaining duplicate gene pairs were labeled as neither 5′ nor 3′ similar.

Cis-regulatory regions analysis

To detect homologous sequences in *cis*-regulatory regions, we used a modified version of REALIGNER [37]. Using BL2SEQ (part of the Blast suite [70]) with mismatch penalty equal to -2 and word size equal to 7, we constructed alignments of 2-kb (-1.5 kb to +0.5 kb) genomic regions surrounding putative TSSs between copies in each duplicate gene pair. We selected alignments satisfying three criteria: hit length >7 bp; identity >70%; and identical hit strand. If two local alignments overlapped, an alignment with the higher bit score was retained. If the bit scores of the two overlapping alignments were identical, a longer alignment or the one closest to TSS was retained. If the two local alignments were not syntenic (the order of blocks in each alignment was inconsistent), an alignment with the lower bit score was removed. Finally, all local alignments ordered by their genomic coordinates were used as a conserved *cis*-regulatory region for a duplicate gene pair.

TEs within *cis*-regulatory regions were classified into two sets: with the insertion occurring in the ancestral sequence before duplication of a genomic region; with the insertion in only one duplicate copy after the duplication event. We used the Repeatmasker [79] tables at the UCSC Genome Browser [77] to map the coordinates of TEs into *cis*-regulatory regions.

Multiple regression analysis

Linear multiple regression analysis was performed in the R statistical package. The original model included all seven predictors and their interaction terms, but was pruned to include

only significant predictors (and significant interaction terms). RCVE [80,81] was utilized to assess the contribution of each predictor to explaining the total variability:

$$RCVE = \frac{R^2_{full} - R^2_{reduced}}{R^2_{full}}$$

where R^2_{full} and $R^2_{reduced}$ are the R^2 for the full model and the model except for the predictor of interest, respectively. In addition, variance inflation factors [82] were calculated for each predictor to diagnose multicollinearity. All predictors and their interaction terms included in the final model had variance inflation factors below 2 (data not shown), suggesting that multicollinearity was not adversely affecting the model.

References

1. Ohno S: *Evolution by Gene Duplication*. New York: Springer Verlag; 1970.
2. Taylor JS, Raes J: **Duplication and divergence: the evolution of new genes and old ideas.** *Annu Rev Genet* 2004, **38:**615-643.
3. Wagner A: **Selection and gene duplication: a view from the genome.** *Genome Biol* 2002, **3:**reviews1012.
4. Zhang J: **Evolution by gene duplication: an update.** *Trends Ecol Evol* 2003a, **18:**292-298.
5. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
6. Shiu SH, Borevitz JO: **The next generation of microarray research: applications in evolutionary and ecological genomics.** *Heredity* 2008, **100:**141-149.
7. Conant GC, Wagner A: **Asymmetric sequence divergence of duplicate genes.** *Genome Res* 2003, **13:**2052-2058.
8. Gu X, Zhang Z, Huang W: **Rapid evolution of expression and regulatory divergences after yeast gene duplication.** *Proc Natl Acad Sci USA* 2005, **102:**707-712.
9. Gu Z, Nicolae D, Lu HH, Li WH: **Rapid divergence in expression between duplicate genes inferred from microarray data.** *Trends Genet* 2002a, **18:**609-613.
10. Makova KD, Li WH: **Divergence in the spatial pattern of gene expression between human duplicate genes.** *Genome Res* 2003, **13:**1638-1645.
11. Wagner A: **Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate.** *Proc Natl Acad Sci USA* 2000, **97:**6579-6584.
12. Zhang P, Gu Z, Li WH: **Different evolutionary patterns between young duplicate genes in the human genome.** *Genome Biol* 2003b, **4:**R56.

13. Lynch M, Katju V: **The altered evolutionary trajectories of gene duplicates.** *Trends Genet* 2004, **20:**544-549.
14. Hurles M: **Gene duplication: the genomic trade in spare parts.** *PLoS Biol* 2004, **2:**E206.
15. Katju V, Lynch M: **The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome.** *Genetics* 2003, **165:**1793-1803.
16. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends Genet* 2003, **19:**68-72.
17. Thornburg BG, Gotea V, Makalowski W: **Transposable elements as a significant source of transcription regulating signals.** *Gene* 2006, **365:**104-110.
18. Kamat A, Hinshelwood MM, Murry BA, Mendelson CR: **Mechanisms in tissue-specific regulation of estrogen biosynthesis in humans.** *Trends Endocrinol Metab* 2002, **13:**122-128.
19. van de Lagemaat LN, Landry JR, Mager DL, Medstrand P: **Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions.** *Trends Genet* 2003, **19:**530-536.
20. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, *et al*.: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
21. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: **A high-resolution map of active promoters in the human genome.** *Nature* 2005, **436:**876-880.
22. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, Kondo H, Wagatsuma M, Murakawa K, Ishida S, Ishibashi T, Takahashi-Fujii A, Tanase T, Nagai K, Kikuchi H, Nakai K, *et al*.: **Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes.** *Genome Res* 2006, **16:**55-65.
23. Tsuritani K, Irie T, Yamashita R, Sakakibara Y, Wakaguri H, Kanai A, Mizushima-Sugano J, Sugano S, Nakai K, Suzuki Y: **Distinct class of putative "non-conserved" promoters in humans: comparative studies of alternative promoters of human and mouse genes.** *Genome Res* 2007, **17:**1005-1014.
24. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16:**1-10.
25. Landry JR, Mager DL, Wilhelm BT: **Complex controls: the role of alternative promoters in mammalian genomes.** *Trends Genet* 2003, **19:**640-648.
26. Strausberg RL, Levy S: **Promoting transcriptome diversity.** *Genome Res* 2007, **17:**965-968.
27. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM: **Identification and functional analysis of human transcriptional promoters.** *Genome Res* 2003, **13:**308-312.
28. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the**

mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004, **101:**6062-6067.

29. Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T: **Genome cluster database. A sequence family analysis platform for Arabidopsis and rice.** *Plant Physiol* 2005, **138:**47-54.
30. Yang J, Lusk R, Li WH: **Organismal complexity, protein complexity, and gene duplicability.** *Proc Natl Acad Sci USA* 2003, **100:**15661-15665.
31. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.** *Nat Methods* 2005, **2:**105-111.
32. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S: **Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions.** *Genome Res* 2004, **14:**1711-1718.
33. Jin VX, Singer GA, Agosto-Perez FJ, Liyanarachchi S, Davuluri RV: **Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs.** *BMC bioinformatics* 2006, **7:**114.
34. Yi S, Ellsworth DL, Li WH: **Slow molecular clocks in Old World monkeys, apes, and humans.** *Mol Biol Evol* 2002, **19:**2191-2198.
35. Hughes T, Liberles DA: **The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.** *J Mol Evol* 2007, **65:**574-588.
36. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CA: **Heterotachy in mammalian promoter evolution.** *PLoS Genet* 2006, **2:**e30.
37. Iwama H, Gojobori T: **Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network.** *Proc Natl Acad Sci USA* 2004, **101:**17156-17161.
38. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev* 2006, **7:**98-108.
39. Li WH, Yang J, Gu X: **Expression divergence between duplicate genes.** *Trends Genet* 2005, **21:**602-607.
40. Scannell DR, Wolfe KH: **A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast.** *Genome Res* 2008, **18:**137-147.
41. Semon M, Wolfe KH: **Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*.** *Proc Natl Acad Sci USA* 2008, **105:**8333-8338.
42. Ha M, Li WH, Chen ZJ: **External factors accelerate expression divergence between duplicate genes.** *Trends Genet* 2007, **23:**162-166.
43. Prince VE, Pickett FB: **Splitting pairs: the diverging fates of duplicated genes.** *Nat Rev* 2002, **3:**827-837.
44. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154:**459-473.
45. Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH: **Role of positive selection in the retention of duplicate genes in mammalian genomes.** *Proc Natl Acad Sci USA* 2006, **103:**2232-2236.
46. Papp B, Pal C, Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast.** *Trends Genet* 2003, **19:**417-422.
47. Castillo-Davis CI, Hartl DL, Achaz G: **cis-Regulatory and protein evolution in orthologous and duplicate genes.** *Genome Res* 2004, **14:**1530-1536.

48. Leach LJ, Zhang Z, Lu C, Kearsey MJ, Luo Z: **The role of cis-regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes.** *Mol Biol Evol* 2007, **24:**2556-2565.
49. Zhang L, Li WH: **Mammalian housekeeping genes evolve more slowly than tissue-specific genes.** *Mol Biol Evol* 2004, **21:**236-239.
50. Subramanian S, Kumar S: **Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome.** *Genetics* 2004, **168:**373-381.
51. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158:**927-931.
52. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17:**68-74.
53. Drummond DA, Raval A, Wilke CO: **A single determinant dominates the rate of yeast protein evolution.** *Mol Biol Evol* 2006, **23:**327-337.
54. Rocha EP, Danchin A: **An analysis of determinants of amino acids substitution rates in bacterial proteins.** *Mol Biol Evol* 2004, **21:**108-116.
55. Hirsh AE, Fraser HB: **Protein dispensability and rate of evolution.** *Nature* 2001, **411:**1046-1049.
56. Zhang J, He X: **Significant impact of protein dispensability on the instantaneous rate of protein evolution.** *Mol Biol Evol* 2005, **22:**1147-1155.
57. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296:**750-752.
58. Drummond DA, Wilke CO: **Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.** *Cell* 2008, **134:**341-352.
59. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proc Natl Acad Sci USA* 2005, **102:**5483-5488.
60. Cusack BP, Wolfe KH: **Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates.** *Mol Biol Evol* 2007, **24:**679-686.
61. Liang H, Lin YS, Li WH: **Fast evolution of core promoters in primate genomes.** *Mol Biol Evol* 2008, **25:**1239-1244.
62. Rossnes R, Eidhammer I, Liberles DA: **Phylogenetic reconstruction of ancestral character states for gene expression and mRNA splicing data.** *BMC Bioinformatics* 2005, **6:**127.
63. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH: **Extent of gene duplication in the genomes of Drosophila, nematode, and yeast.** *Mol Biol Evol* 2002b, **19:**256-262.
64. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30:**1575-1584.
65. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85:**2444-2448.
66. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12:**85-94.
67. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
68. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.** *Mol Biol Evol* 2000, **17:**32-43.
69. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.

70. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
71. Chung WY, Albert R, Albert I, Nekrutenko A, Makova KD: **Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network.** *BMC Bioinformatics* 2006, **7:**46.
72. Huminiecki L, Lloyd AT, Wolfe KH: **Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases.** *BMC Genomics* 2003, **4:**31.
73. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, *et al*.: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.
74. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35:**D61-65.
75. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, *et al*.: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones.** *PLoS Biol* 2004, **2:**e162.
76. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update.** *Nucleic Acids Res* 2004, **32:**D23-26.
77. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 2008, **36:**D773-779.
78. Bray N, Dubchak I, Pachter L: **AVID: A global alignment program.** *Genome Res* 2003, **13:**97-102.
79. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16:**418-420.
80. Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD: **A macaque's-eye view of human insertions and deletions: differences in mechanisms.** *PLoS Comput Biol* 2007, **3:**1772-1782.
81. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD: **The genome-wide determinants of human and chimpanzee microsatellite evolution.** *Genome Res* 2008, **18:**30-38.
82. Kutner MH, Nachtsheim CJ, Neter J, Li W: *Applied Linear Statistical Models*. New York: McGraw-Hill; 2005.

Chapter 7

Conclusions

Concluding remarks

The importance of understanding the expression divergence of duplicate genes

Because of the importance of gene duplication in biological evolution, it is crucial to know how duplicate genes have diverged and which factors determine their destiny. Genome-wide gene expression analyses have contributed to the understanding of evolution of gene expression in duplicate genes. Although it has been shown that duplicate genes diverge rapidly in their expression, little has been known about which factors influence their expression divergence at the genomic level (Wagner 2000; Li et al. 2005; Scannell and Wolfe 2008; Semon and Wolfe 2008). The research presented in this dissertation (Chapter 6) represents the first report of the effects of structural differences in coding region and of TSS turnover on the divergence of duplicate gene expression.

Recently, additional factors also have been shown to influence the expression divergence pattern of duplicate genes at the genomic level: promoter types (Taylor et al. 2006), genomic neighborhood effects (De et al. 2009), histone modifications (Zheng 2008; Li et al. 2010), and external environmental factors (Ha et al. 2007). A multiple regression analysis, proposed in the Chapter 6, performed using these new factors as well as factors used in Chapter 6, can lead to a much broader understanding of the patterns of duplicate gene expression divergence.

Although most studies of expression divergence between paralogs have utilized microarray data (*i.e.*, they were obtained by hybridization of RNA to oligonucleotide arrays) (Wagner 2000; Gu et al. 2002; Makova and Li 2003; Blanc and Wolfe 2004; Gu et al. 2005; He

and Zhang 2005), these data are likely to be quite noisy (Makova and Li 2003). Indeed, we also observed that many human duplicate gene copies are likely to be cross-hybridized in the oligonucleotide array, and this tends to underestimate the degree of expression divergence (Park and Makova 2009). It is now possible to use high-volume transcriptome data (e.g., RNA-seq) based on next-generation sequencing technologies (reviewed in (Wang et al. 2009)). This not only will enable scientists to reinvestigate the expression patterns and genomic factors contributing to expression divergence of duplicate genes, but also will provide important information for understanding human transcriptome heterogeneity, complexity, and evolution.

Application of new methods to classify genes with distinct biological features

The research presented in this dissertation introduced a LDA classifier and program modules to compute classification accuracy (Chapter 2). This classifier was used to distinguish escape *vs.* inactivated genes, and achieved high classification accuracy for them (Chapter 2). The same statistical approaches were also successfully applied to the XLMR study (Chapter 5). Hopefully, other researchers in computational biology will utilize these tools for use in their genomic research. For convenience to the users, they are freely available over the Internet at Galaxy (Blankenberg et al. 2010) (http://g2.bx.psu.edu/).

A pilot study of escape gene evolution

To achieve dosage balance of X-linked genes between mammalian males and females, one female X chromosome becomes inactivated. However, approximately 15% of genes on this inactivated chromosome escape X chromosome inactivation. This indicates that they are expressed from both the active and inactive X chromosomes, and frequently cause dosage imbalances between males and females. In the research presented in this dissertation (Chapter 4), we first observed that escape genes are subject to stronger purifying selection than inactivated

genes and that positive selection does not significantly affect the evolution of these genes. Intriguingly, escape genes possessing Y homologs evolve under the strongest purifying selection. We also found evidence of stronger conservation in gene expression levels in escape than inactivated genes. We hypothesize that divergence in function and expression between X and Y gametologs is driving such strong purifying selection for escape genes.

Recently, Yang and colleagues (Yang et al. 2010) generated a genome-wide inactivation profile of the mouse X chromosome by high-throughput RNA sequencing. As expected, given the different number and identity of escape genes between human and mouse (Disteche 1995), only a few mouse genes escape X inactivation. There was no clustering of escape genes and no preference for evolutionary young strata as defined in human. Interestingly, some escape genes in mouse are subject to inactivation in human (Yang et al. 2010). This means that despite conservation of X inactivation mechanisms in eutherian mammals, orthologus genes on the X can have different inactivation statuses (e.g., (Jegalian and Page 1998; Yen et al. 2007)), and thus may have distinct selective constraints. My co-adviser, Dr. Laura Carrel, also observed some exceptional cases in the same XCI profile among eutherian mammals (unpublished). In future work, I hope to employ the comparative genomic X inactivation data in this study and apply it to understand the evolutionary fates of genes on the X chromosome.

Progress in the understanding of the impact of genomic influences on inactive X chromosome and escape gene expression

Mechanisms to coordinately control gene expression over large genomic distances are complex and must involve, at some level, the underlying DNA sequences. Mammalian X chromosome inactivation represents one of the most fascinating examples of such mechanisms in action. Indeed, most genes on one X chromosome are silenced, while ~15% of X-linked genes escape inactivation and yet another ~10% exhibit variable patterns of inactivation among

individuals (Carrel and Willard 2005). Even given the several hallmarks described above (see Chapter 1), the mechanism for the regulation of inactivated genes is still not clearly understood (Carrel and Willard 2005; Prothero et al. 2009). In this dissertation (Chapter 2 and Chapter 4), we focused on two models (*i.e.*, the way station and the boundary elements models) to understand the spread and/or maintenance of XCI and identified X chromosome primary DNA sequence features that can discriminate between the X inactivation statuses of genes. The resulting data provide valuable insights into the regulation of escape gene expression and understanding the XCI mechanism.

Many studies have indicated that the inactive X chromosome tends to be clearly distinguished from the active X chromosome by genome-wide epigenetic modifications (Valley and Willard 2006; Payer and Lee 2008; Prothero et al. 2009). These epigenetic marks also partly differentiate escape genes from inactivated ones on the inactive X chromosome (Marks et al. 2009; Mietton et al. 2009; Yang et al. 2010). To provide better understanding of genomic and epigenomic influences on the XCI mechanism, three further efforts can be addressed. First, identifying more correct motifs may further help to validate current XCI models. In this dissertation we used one locus intensively in order to detect candidate boundary elements. If more loci with well-studied boundary regions are available, highly confident motifs can be identified and further improve the prediction of XCI status. Second, incorporating more complex patterns of inactive X expression (i.e., heterogeneous genes that escape inactivation in some females) may provide valuable insights into the XCI regulation and evolution of escape genes. For example, studies of XCI models at the boundary region between escape and heterogeneous genes can help to reveal sources of the XCI mechanism. Using population and comparative genomic data, if available, we can understand why some genes have shown heterogeneous expression patterns with an evolutionary point of view (e.g., they were inactivated genes, and will become escape genes or vice versa). Third, considering additional parameters including epigenetic marks may

provide substantial predictive contributions to distinguish between escape and inactivated genes. Candidate features to be investigated should include CpG islands, gene density, transposable elements density, location within an escape domain particularly with respect to domain boundaries, distance from the XIST locus, CTCF sites, histone modifications and variants, and nucleosome occupancy. Identifying the genomic and epigenomic features that control XCI will further aid in our understanding of long-range control of gene expression on the X chromosome.

Synthesis

Although we do not yet have a complete understanding of what genomic factors determine XCI status of genes on the inactive X chromosome and what fundamental factors drive the expression divergence of duplicate genes, the research included in this dissertation provides novel insights into the mechanisms of XCI and evolution of duplicate genes. Through investigating these issues, this dissertation also produced important genomic data regarding the mechanisms of escaping XCI and for the expression divergence of duplicate genes. These data will be a valuable resource for other researchers investigating the mechanisms and evolution of XCI and duplicate genes.

Major contributions of dissertation

In summary, this dissertation has provided the following major contributions:

1. We have successfully addressed the two models proposed so far to understand the molecular mechanisms for controlling genes escaping X inactivation. According to the way stations model (Chapter 2), we have revealed that the majority of the sequences enriched in the vicinity of inactivated genes were found within L1 repeats. This result strongly supports an involvement of

L1 repeats in X chromosome inactivation (Gartler and Riggs 1983; Lyon 2006) and additionally suggested that *Alu* repetitive elements can function as predictors for escape genes. We also identified a set of short sequences and demonstrated that these sequences capture most of the genomic signal determining X inactivation. According to the boundary elements model (Chapter 4), we found some unique or overrepresented motifs in boundary regions, even though the resulting data set needs to be evaluated in future experimental studies, indicating that they are the candidates for the boundary elements separating genes with different XCI profiles. Our investigation of these genomic sequence landmarks contributing to regulation of the expression status of genes on the inactive X chromosome provides further aid in our understanding of long-range control of gene expression and the impact of repetitive elements throughout the genome.

2. We have first shown that escape genes experience stronger purifying selection than inactivated genes at both the protein-coding and gene expression levels (Chapter 3). This effect largely results from the importance of function and dosage of escape genes, as observed when their gene dosage is altered in individuals with Turner syndrome (Good et al. 2003; Bondy 2006). Understanding the reasons why some genes escape X inactivation is very important for identifying the molecular mechanisms of dosage compensation and sexual antagonism, and thus, our observations may provide valuable insights into the understanding of sex chromosome evolution and X-linked diseases.

3. We developed a statistical classifier for discriminating escape genes from inactivated genes (Chapter 2) and XLMR genes from non-XLMR genes (Chapter 5). This tool is freely available at the Galaxy web site (http://g2.bx.psu.edu/).

4. We have revealed that the turnover of transcription start sites, structural heterogeneity of coding sequences, and the divergence of *cis*-regulatory regions between human paralogous genes play a pivotal role in determining the expression divergence of duplicate genes (Chapter 6). This study represents the first report of the effects of structural differences in coding region and of unique TSSs on the divergence of duplicate gene expression. Our observations of frequent turnover of TSSs between duplicate genes and a high proportion of young duplicate genes with incompletely similar structures contradict the assumptions of classic gene duplication models, according to which duplicate genes are considered to be equal both structurally and functionally at the point of duplication (Zhang 2003; Hurles 2004). Although several future studies are needed to understand the evolution of gene duplication, our investigation of factors contributing to expression divergence of duplicate genes provides important information for understanding human transcriptome heterogeneity, complexity, and evolution.

References

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant cell* **16**(7): 1679-1691.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al* **Chapter 19**: Unit 19 10 11-21.

Bondy CA. 2006. Turner's syndrome and X chromosome-based differences in disease susceptibility. *Gend Med* **3**(1): 18-30.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

De S, Teichmann SA, Babu MM. 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome research* **19**(5): 785-794.

Disteche CM. 1995. Escape from X inactivation in human and mouse. *Trends Genet* **11**(1): 17-22.

Gartler SM, Riggs AD. 1983. Mammalian X-chromosome inactivation. *Annual review of genetics* **17**: 155-190.

Good CD, Lawrence K, Thomas NS, Price CJ, Ashburner J, Friston KJ, Frackowiak RS, Oreland L, Skuse DH. 2003. Dosage-sensitive X-linked locus influences the development of amygdala and orbitofrontal cortex, and fear recognition in humans. *Brain* **126**(Pt 11): 2431-2446.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences of the United States of America* **102**(3): 707-712.

Gu Z, Nicolae D, Lu HH, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **18**(12): 609-613.

Ha M, Li WH, Chen ZJ. 2007. External factors accelerate expression divergence between duplicate genes. *Trends Genet* **23**(4): 162-166.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2): 1157-1164.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS biology* **2**(7): E206.

Jegalian K, Page DC. 1998. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**(6695): 776-780.

Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet* **21**(11): 602-607.

Li J, Yuan Z, Zhang Z. 2010. Revisiting the contribution of cis-elements to expression divergence between duplicated genes: the role of chromatin structure. *Molecular biology and evolution* **27**(7): 1461-1466.

Lyon MF. 2006. Do LINEs have a role in X-chromosome inactivation? *Journal of biomedicine & biotechnology* **2006**(1): 59746.

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research* **13**(7): 1638-1645.

Marks H, Chow JC, Denissov S, Francoijs KJ, Brockdorff N, Heard E, Stunnenberg HG. 2009. High-resolution analysis of epigenetic changes associated with X inactivation. *Genome research* **19**(8): 1361-1373.

Mietton F, Sengupta AK, Molla A, Picchi G, Barral S, Heliot L, Grange T, Wutz A, Dimitrov S. 2009. Weak but uniform enrichment of the histone variant macroH2A1 along the inactive X chromosome. *Molecular and cellular biology* **29**(1): 150-156.

Park C, Makova KD. 2009. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. *Genome biology* **10**(1): R10.

Payer B, Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. *Annual review of genetics* **42**: 733-772.

Prothero KE, Stahl JM, Carrel L. 2009. Dosage compensation and gene expression on the mammalian X chromosome: one plus one does not always equal two. *Chromosome Res* **17**(5): 637-648.

Scannell DR, Wolfe KH. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome research* **18**(1): 137-147.

Semon M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences of the United States of America* **105**(24): 8333-8338.

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CA. 2006. Heterotachy in mammalian promoter evolution. *PLoS genetics* **2**(4): e30.

Valley CM, Willard HF. 2006. Genomic and epigenomic approaches to the study of X chromosome inactivation. *Current opinion in genetics & development* **16**(3): 240-245.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proceedings of the National Academy of Sciences of the United States of America* **97**(12): 6579-6584.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews* **10**(1): 57-63.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome research* **20**(5): 614-622.

Yen ZC, Meyer IM, Karalic S, Brown CJ. 2007. A cross-species comparison of X-chromosome inactivation in Eutheria. *Genomics* **90**(4): 453-463.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**(6): 292-298.

Zheng D. 2008. Asymmetric histone modifications between the original and derived loci of human segmental duplications. *Genome biology* **9**(7): R105.

Appendix A

Supporting Material for Chapter 2

Supplemental Tables

Table S2-1: Gene and Contig Information for the Xp22 *E* and *I* Subgenomes

Table S2-2: List of Overrepresented Oligomers

Table S2-3: Gene Lists for Training and Test Datasets

Table S2-4: Results of Classification When Only $(GATA)_n$ Was Used

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.0020151#s5

Appendix B

Supporting Material for Chapter 3

Supplemental Figure Legends

Figure S3-1: The number of orthologous genes considered. The number of genes excluded at each step is indicated in the parentheses. *: genes IKBKG, PLXNB3, and SH3BGRL were excluded because they had completely different XCI profile between two assays (e.g., escape vs. inactivated and vice versa; see Table S1). **: Candidate pseudogenes were identified if there were stop codons within the inferred CDS regions of at least one non-human species.

Figure S3-2: A comparison of K_A/K_S ratios among escape, heterogeneous, and inactivated orthologous genes from 5-way (human-chimpanzee-orangutan-macaque-marmoset) alignments. The number of genes considered is given below each plot. In the box plots, edges and vertical dashed lines represent quartiles and range, respectively. Notches indicate standard deviations of the median. Non-overlapping notches are evidence that the medians are different. Outliers are not shown.

Figure S3-3: A comparison of median expression divergence ratios between escape and inactivated genes. The number of genes available for each tissue is given below each plot. In the box plots, edges and vertical dashed lines represent quartiles and range, respectively. Notches indicate standard deviations of the median. Non-overlapping notches are evidence that the medians are different. Outliers are not shown.

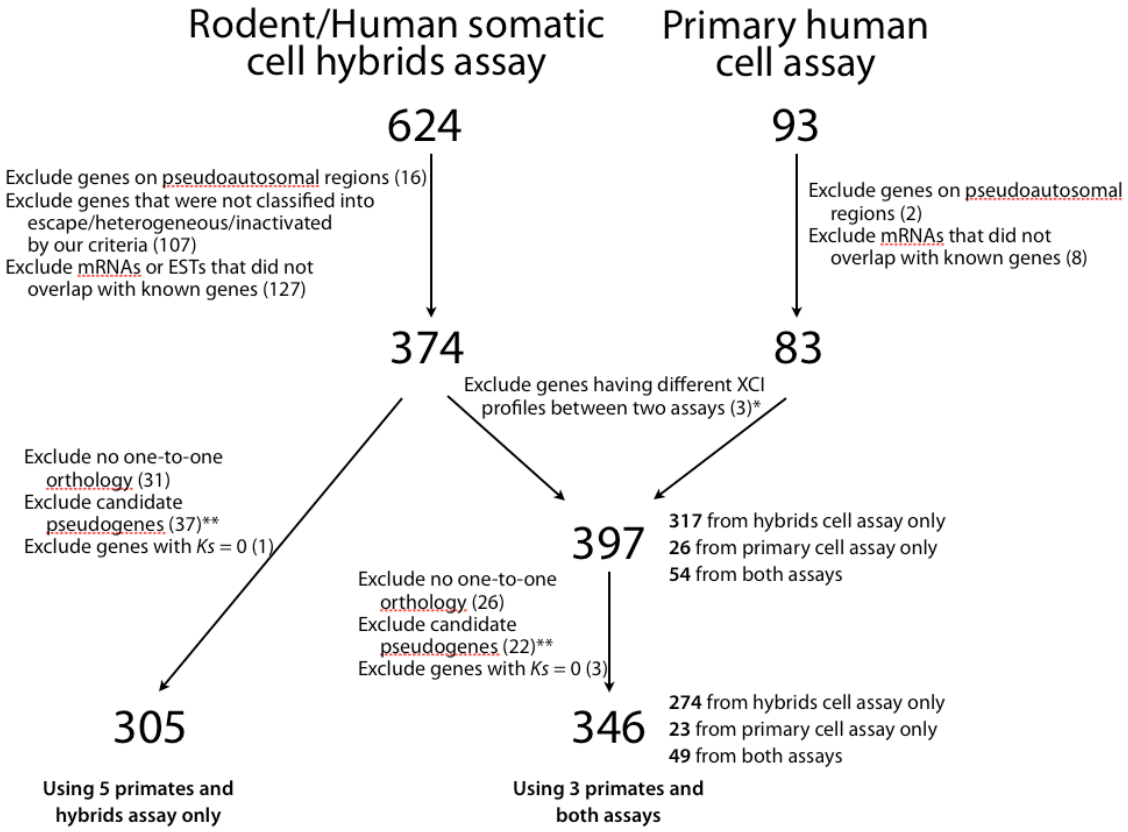

Rodent/Human somatic cell hybrids assay
Primary human cell assay
624
93
Exclude genes on pseudoautosomal regions (16)
Exclude genes that were not classified into escape/heterogeneous/inactivated by our criteria (107)
Exclude mRNAs or ESTs that did not overlap with known genes (127)
Exclude genes on pseudoautosomal regions (2)
Exclude mRNAs that did not overlap with known genes (8)
374
83
Exclude genes having different XCI profiles between two assays (3)*
Exclude no one-to-one orthology (31)
Exclude candidate pseudogenes (37)**
Exclude genes with Ks = 0 (1)
397
317 from hybrids cell assay only
26 from primary cell assay only
54 from both assays
Exclude no one-to-one orthology (26)
Exclude candidate pseudogenes (22)**
Exclude genes with Ks = 0 (3)
305
346
274 from hybrids cell assay only
23 from primary cell assay only
49 from both assays
Using 5 primates and hybrids assay only
Using 3 primates and both assays

Figure S3-1

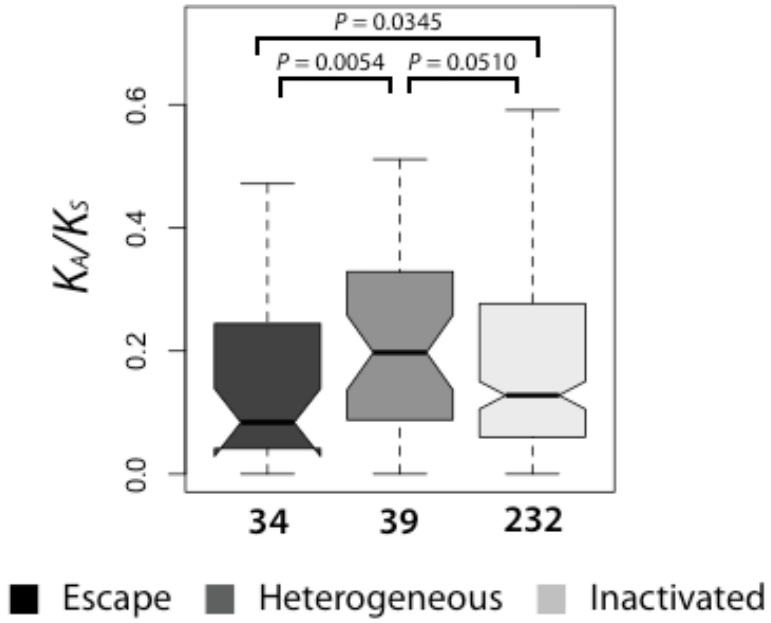


Figure S3-2

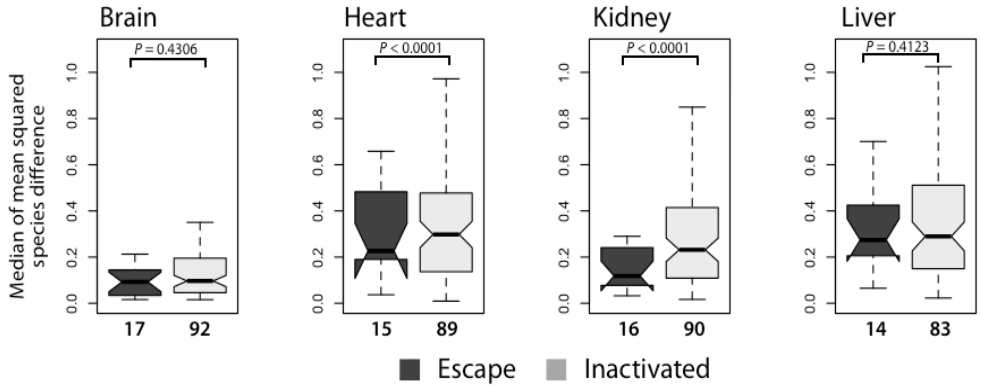


Figure S3-3

Supplemental Tables

Table S3-1. Genes that were classified differently by the two assays. Nine genes were classified as heterogeneous in one assay, but as escape or inactivated according to the other assay. While these discordant results could simply reflect the relatively low number of samples tested in either system, we decided that the primary cell line results would take precedence, since some epigenetic features of X inactivation are not fully maintained in the hybrid cells (Kahan and DeMars 1975; Clemson et al. 1998; Hansen et al. 1998).

	XCI status	
Gene Name	**Primary human cell assay**	**Somatic cell hybrids assay**
ARSE	Heterogeneous	Escape
ATP7A	Inactivated	Heterogeneous
CLIC2	Inactivated	Heterogeneous
GRPR	Inactivated	Heterogeneous
GYG2	Heterogeneous	Escape
HCFC1	Escape	Heterogeneous
IKBKG*	Inactivated	Escape
MAOA	Heterogeneous	Escape
MSL3L1	Inactivated	Heterogeneous
PIR	Heterogeneous	Escape
PLXNB3*	Inactivated	Escape
SH3BGRL*	Inactivated	Escape

*: Genes that were excluded from this study because they had completely different XCI profile between two assays (e.g., escape vs. inactivated and vice versa).

Table S3-2. Genes having K_A/K_S ratio greater than 1.

Gene Name	XCI status	K_A/K_S ratio	Log likelihood value: Model A (M0; free parameter)	Log likelihood value: Model B (M0; fixed ratio = 1)	χ^2	*P* (d.f. = 1)
ACRC	Inactivated	1.072	-3260.256017	-3260.34037	0.168706	0.6813
FLJ31752 (LOC401588)	Inactivated	1.151	-2621.265494	-2621.358608	0.186228	0.6661
LOC159091 (FAM122C)	Inactivated	1.107	-609.685978	-609.704317	0.036678	0.8481
LOC340602	Inactivated	1.244	-1711.550479	-1711.843581	0.586204	0.4439
NXT2	Inactivated	2.289	-814.876059	-815.882881	2.013644	0.1559
SCML1	Inactivated	1.122	-1690.934446	-1691.017822	0.166752	0.6830
XEDAR (EDA2R)	Heterogeneous	1.183	-1358.736364	-1358.810139	0.14755	0.7009

Gene names in RefSeq are indicated in parentheses.

Table S3-3. Genes with positively selected codons. "X" indicates that a gene has positively selected codons with a significant *P* value after adopting a false discovery rate (FDR) for multiple tests (q value < 0.05).

Gene Name	XCI status	Site models		Improved branch-site model		
		M1[a] vs. M2[b]	M7[c] vs. M8[d]	H[e]	C[f]	M[g]
ACRC	Inactivated				X	
APG4A (ATG4A)	Inactivated				X	
ARSD	Escape					X
ATRX	Inactivated		X		X	
BIRC4	Inactivated				X	
CXCR3	Inactivated	X	X			X
DMD	Inactivated				X	
FGD1	Inactivated				X	
MBNL3	Inactivated				X	
MGC39350 (CXorf38)	Escape				X	
MTMR1	Inactivated	X	X			
OATL1	Inactivated				X	
ODZ1	Inactivated				X	
OGT	Inactivated				X	
PCYT1B	Inactivated		X		X	
PJA1	Inactivated				X	
PLP2	Inactivated				X	
PNMA6A	Inactivated		X			
PRDX4	Inactivated				X	
PRPS2	Inactivated				X	
RAI2	Inactivated				X	
THOC2	Inactivated		X		X	
TRPC5	Heterogeneous				X	
UREB1 (HUWE1)	Inactivated				X	
UTX	Escape			X		
ZCWCC2 (MORC4)	Inactivated				X	

[a]M1: Nearly neutral model
[b]M2: Positive selection model
[c]M7: β-distribution, neutral model
[d]M8: β-distribution, selection model
[e]H: Human lineage was the foreground branch, where positive selection was allowed; all the other branches were the background branches, where purifying or neutral selection was allowed.
[f]C: Chimpanzee lineage was the foreground branch, where positive selection was allowed; all the other branches were the background branches, where purifying or neutral selection was allowed.
[g]M: Macaque lineage was the foreground branch, where positive selection was allowed; all the other branches were the background branches, where purifying or neutral selection was allowed.
Gene name in RefSeq is indicated in parentheses.

Table S3-4

Multiple regression models for K_A/K_S ratio in X-linked genes		
Predictors	*P*-value	RCVE*
Evol[†]	0.929	0.001
XCI[§]	0.082(MS[‡])	0.575
Recomb[¶]	0.298	0.205
R^2		0.015

*RCVE: relative contribution to the variability explained. [†]Evol: XAR versus XCR. [§]XCI: escape versus non-escape (i.e., heterogeneous and inactivated genes). [‡]MS: marginally significant. [¶]Recomb: local recombination rate (Kong et al. 2002)

Table S3-5. Median difference in the K_A/K_S ratio between an escape gene with a functional Y homolog and other escape genes from the same cluster or from other clusters.

Cluster	Genes	Genes with Y homologs	Median Δ K_A/K_S in the same cluster	Median Δ K_A/K_S outside of the cluster	*P* value
1	XG, ARSD, FAM16AX (HDHD1A), STS, PNPLA4	NLGN4X	0.2113	0.0325	0.0125
2	TMEM27, CA5B, AP1S2, CTPS2, CALB3 (S100G), RBBP7	CXORF15	0.2666	0.3278	0.3653
3	EIF2S3	ZFX	0.0074	0.0745	0.0967
4	MGC39350 (CXorf38)	USP9X, DDX3X	0.1497	0.0564	0.3333
5	FUNDC1	UTX	0.0644	0.1046	0.2257
6	SMC1L1	JARID1C (SMCX)	0.0558	0.0647	0.3868

Gene name in RefSeq is indicated in parentheses.

Table S3-6. Functional differences between the studied XCR/YCR gametologs not included in the study of Wilson & Makova (Wilson and Makova 2009). The unique functions reported for either the X copy or the Y copy are listed in each respective column. Functions similar for both the X and the Y copy are listed across both columns.

Gametologs	X copy	Y copy
SMCX/Y	Required for normal morphogenesis of the neural tube (Takeuchi et al. 1995) Histone demethylases specific for tri- and di-methylated H3K4 (Christensen et al. 2007)	Involved in H-Y transplantation antigene leading to rejection of male organ and bone marrow grafts (Wang et al. 1995)
RPS4X/Y	Turner syndrome gene (Zinn et al. 1994)	Fewer functional constraint than X copy (Bergen et al. 1998)

Table S3-7. K_A and K_S values of all genes tested.

Gene Name	RefSeq ID	Loc*	K_A	K_S	K_A/K_S	XCI Status	E†	S§
XG	NM_001141919	2.7	0.03009	0.11238	0.26773	Escape	XAR	5
GYG2	NM_003918	2.8	0.02593	0.11042	0.23482	heterogeneous	XAR	5
ARSD	NM_001669	2.8	0.12813	0.53327	0.24027	Escape	XAR	5
ARSE	NM_000047	2.9	0.02481	0.15157	0.1637	heterogeneous	XAR	5
NLGN4X	NM_020742	5.8	0.00331	0.1144	0.02898	Escape	XAR	4
FAM16AX (HDHD1A)	NM_001135565	7.0	0.02443	0.10938	0.22338	Escape	XAR	4
STS	NM_000351	7.1	0.01553	0.06938	0.22382	Escape	XAR	4
PNPLA4	NM_001142389	7.8	0.01582	0.05751	0.27513	Escape	XAR	4
TBL1X	NM_005647	9.4	0.00632	0.10985	0.05756	Inactivated	XAR	3
KIAA1280(WWC3)	NM_015691	9.9	0.00971	0.08344	0.11643	Inactivated	XAR	3
CLCN4	NM_001830	10.1	0.00092	0.08499	0.01086	Inactivated	XAR	3
MID1	NM_000381	10.4	0.00072	0.0538	0.01348	Inactivated	XAR	3
HCCS	NM_001122608	11.0	0.01057	0.07271	0.14534	Inactivated	XAR	3
ARHGAP6	NM_013427	11.1	0.01346	0.08014	0.16797	Inactivated	XAR	3
MSL3L1	NM_078629	11.7	0.00303	0.05565	0.05443	Inactivated	XAR	3
M62076(FRMPD4)	NM_014728	12.1	0.0077	0.04656	0.16532	Inactivated	XAR	3
PRPS2	NM_001039091	12.7	0.00399	0.11267	0.03538	Inactivated	XAR	3
RAB9A	NM_004251	13.6	0.00216	0.05462	0.03961	Escape	XAR	3
SEDL	NM_014563	13.6	0	0.02972	0.0001	Escape	XAR	3
GPM6B	NM_001001995	13.7	0	0.02965	0.0001	Escape	XAR	3
FAM51A1 (GEMIN8)	NM_001042479	13.9	0.02499	0.09322	0.26804	Escape	XAR	3
GLRA2	NM_001118886	14.5	0.00106	0.0495	0.02148	Inactivated	XAR	3
FANCB	NM_001018113	14.8	0.01998	0.04981	0.40103	Inactivated	XAR	3
MOSPD2	NM_152581	14.8	0.01403	0.05542	0.25317	Inactivated	XAR	3
ASB11	NM_080873	15.2	0.00842	0.06099	0.13809	heterogeneous	XAR	3
PIGA	NM_002641	15.2	0.02229	0.03145	0.70858	Inactivated	XAR	3
PIR	NM_003662	15.3	0.00728	0.05254	0.13855	heterogeneous	XAR	3
TMEM27	NM_020665	15.6	0.01921	0.03763	0.51055	Escape	XAR	3
CA5B	NM_007220	15.7	0.00922	0.06993	0.1318	Escape	XAR	3
AP1S2	NM_003916	15.8	0.00374	0.012	0.31169	Escape	XAR	3
GRPR	NM_005314	16.1	0.00609	0.05363	0.11357	Inactivated	XAR	3
CTPS2	NM_019857	16.5	0.00401	0.04926	0.08135	Escape	XAR	3
CALB3(S100G)	NM_004057	16.6	0.00564	0.05266	0.10704	Escape	XAR	3
CXORF15	NM_018360	16.7	0.01703	0.04411	0.38602	Escape	XAR	3
RBBP7	NM_002893	16.8	0	0.03198	0.0001	Escape	XAR	3
NHS	NM_001136024	17.3	0.01429	0.05465	0.26146	heterogeneous	XAR	3
SCML1	NM_001037540	17.7	0.07367	0.06565	1.12216	Inactivated	XAR	3
RAI2	NM_021785	17.7	0.02208	0.08679	0.25444	Inactivated	XAR	3
SCML2	NM_006089	18.2	0.01423	0.05226	0.27229	Inactivated	XAR	3
DKFZp686C0388 (FLJ13808)	NM_001079858	18.9	0.01607	0.0367	0.43791	Inactivated	XAR	3
PDHA1	NM_000284	19.3	0.00001	0.06694	0.0001	Inactivated	XAR	3
SH3KBP1	NM_031892	19.5	0.00497	0.05607	0.08862	heterogeneous	XAR	3
CXorf23	NM_198279	19.8	0.00985	0.045	0.21882	Inactivated	XAR	3
EIF1AX	NM_001412	20.1	0.00001	0.12149	0.0001	Escape	XAR	3
RPS6KA3	NM_004586	20.1	0.00066	0.04249	0.01547	Inactivated	XAR	3
YY2	NM_206923	21.8	0.05083	0.10531	0.48266	Inactivated	XAR	3
PHEX	NM_000444	22.0	0.0083	0.05392	0.1539	heterogeneous	XAR	3
PRDX4	NM_006406	23.6	0.01445	0.05298	0.27267	Inactivated	XAR	3
ACATE2(ACOT9)	NM_001037171	23.6	0.01003	0.06225	0.1611	Inactivated	XAR	3
SAT	NM_002970	23.7	0.00825	0.03597	0.22946	Inactivated	XAR	3

AA601738(APOO)	NM_024122	23.8	0.00533	0.03434	0.15523	Inactivated	XAR	3
KLHL15	NM_030624	23.9	0.00083	0.04913	0.01683	Inactivated	XAR	3
EIF2S3	NM_001415	24.0	0.00115	0.06377	0.01804	Escape	XAR	3
ZFX	NM_003410	24.1	0.00113	0.04433	0.02541	Escape	XAR	3
PDK3	NM_001142386	24.4	0	0.04356	0.0001	Inactivated	XAR	3
PCYT1B	NM_004845	24.5	0.00629	0.06152	0.10219	Inactivated	XAR	3
POLA	NM_016937	24.6	0.01514	0.0412	0.36752	Inactivated	XAR	3
GK	NM_001128127	30.6	0.00171	0.03625	0.04719	Inactivated	XAR	3
DMD	NM_004006	31.0	0.00989	0.05057	0.19555	Inactivated	XAR	3
PRRG1	NM_001142395	37.1	0.00233	0.03317	0.07038	Inactivated	XAR	3
FLJ42925 (LANCL3)	NM_198511	37.3	0.00472	0.08502	0.05549	Inactivated	XAR	3
TCTE1L(DYNLT3)	NM_006520	37.6	0.00001	0.05347	0.0001	Inactivated	XAR	3
SYTL5	NM_138780	37.8	0.01474	0.06831	0.21585	Inactivated	XAR	3
SRPX	NM_006307	37.9	0.00346	0.04988	0.06933	Inactivated	XAR	3
Hs.61438(RPGR)	NM_000328	38.0	0.03224	0.07415	0.43484	Inactivated	XAR	3
OTC	NM_000531	38.1	0.01046	0.0314	0.33302	Inactivated	XAR	3
TM4SF2(TSPAN7)	NM_004615	38.3	0.00372	0.02622	0.14197	Inactivated	XAR	3
FLJ43479(BCOR)	NM_001123385	39.8	0.00783	0.05896	0.13273	Inactivated	XAR	3
ATP6AP2	NM_005765	40.3	0.00739	0.03611	0.20462	Inactivated	XAR	3
MGC39350 (CXorf38)	NM_144970	40.4	0.01082	0.05761	0.18778	Escape	XAR	3
USP9X	NM_001039590	40.8	0.00255	0.04512	0.05653	Escape	XAR	3
DDX3X	NM_001356	41.1	0.00107	0.05481	0.01957	Escape	XAR	3
FLJ22219(CASK)	NM_003688	41.3	0	0.03763	0.0001	Inactivated	XAR	3
GPR34	NM_001097579	41.4	0.0063	0.05439	0.11585	Inactivated	XAR	3
MAOA	NM_000240	43.4	0.04326	0.33333	0.12977	heterogeneous	XAR	3
FUNDC1	NM_173794	44.3	0.00307	0.01771	0.17359	Escape	XAR	3
UTX	NM_021140	44.6	0.00212	0.01942	0.10918	Escape	XAR	3
FLJ31752 (LOC401588)	NM_001039891	46.2	0.03226	0.02801	1.15142	Inactivated	XAR	3
CHST7	NM_019886	46.3	0.01632	0.14733	0.11079	Inactivated	XAR	3
SLC9A7	NM_032591	46.4	0.00099	0.06016	0.01645	Inactivated	XAR	3
RP2	NM_006915	46.6	0.00559	0.03393	0.16486	Inactivated	XAR	3
PHF16	NM_014735	46.7	0.00556	0.03354	0.16587	Inactivated	XAR	3
RGN	NM_152869	46.8	0.01494	0.07071	0.2113	Inactivated	XAR	3
P17.3(NDUFB11)	NM_019056	46.9	0.04055	0.0822	0.49327	Inactivated	XAR	3
UBE1	NM_003334	46.9	0.00531	0.05715	0.09285	Escape	XAR	3
ZNF41	NM_153380	47.2	0.02008	0.04404	0.45585	Inactivated	XAR	3
SYN1	NM_006950	47.3	0.00964	0.0604	0.15964	Inactivated	XCR	2
PFC(CFP)	NM_002621	47.4	0.0341	0.08279	0.41185	Inactivated	XCR	2
UXT	NM_153477	47.4	0.00401	0.00761	0.52681	Inactivated	XCR	2
ZNF81	NM_007137	47.6	0.00902	0.04556	0.19809	Inactivated	XCR	2
ZNF21(ZNF182)	NM_001007088	47.7	0.00634	0.04321	0.14676	Inactivated	XCR	2
SLC38A5	NM_033518	48.2	0.01333	0.06333	0.2105	Inactivated	XCR	2
FTSJ1	NM_177439	48.2	0.00001	0.06051	0.0001	Inactivated	XCR	2
PORCN	NM_203473	48.3	0.00258	0.09383	0.02749	Inactivated	XCR	2
OATL1	NM_002536	48.3	0.00825	0.05777	0.14288	Inactivated	XCR	2
RBM3	NM_006743	48.3	0.00001	0.05214	0.0001	Inactivated	XCR	2
WDR13	NM_017883	48.3	0.00347	0.10175	0.03411	Inactivated	XCR	2
GATA1	NM_002049	48.5	0.01677	0.06709	0.24993	Inactivated	XCR	2
PCSK1N	NM_013271	48.6	0.0597	0.07889	0.75672	Inactivated	XCR	2
TIMM17B	NM_005834	48.6	0.00344	0.11527	0.02985	Inactivated	XCR	2
SLC35A2	NM_005660	48.6	0.00618	0.06713	0.09207	Inactivated	XCR	2
PIM2	NM_006875	48.7	0.00632	0.07764	0.08135	Inactivated	XCR	2
DKFZp761A052 (OTUD5)	NM_017602	48.7	0.00501	0.03795	0.13202	Inactivated	XCR	2

KCND1	NM_004979	48.7	0.00397	0.10052	0.03945	Inactivated	XCR	2
GRIPAP1	NM_207672	48.7	0.00625	0.04554	0.13726	Inactivated	XCR	2
TFE3	NM_006521	48.8	0.0054	0.04709	0.11478	Inactivated	XCR	2
JM11(CCDC120)	NM_033626	48.8	0.00784	0.06562	0.1194	Inactivated	XCR	2
JM4	NM_007213	48.8	0.01075	0.11123	0.09665	Inactivated	XCR	2
WDRX1(WDR45)	NM_007075	48.8	0	0.04697	0.0001	Inactivated	XCR	2
GPKOW	NM_015698	48.9	0.02207	0.04548	0.4852	Inactivated	XCR	2
FLJ21687(MAGIX)	NM_024859	48.9	0.05259	0.0849	0.61943	Inactivated	XCR	2
PLP2	NM_002668	48.9	0.0128	0.06513	0.19648	Inactivated	XCR	2
CACNA1F	NM_005183	48.9	0.0081	0.06511	0.12435	Inactivated	XCR	2
JM1(CCDC22)	NM_014008	49.0	0.00943	0.07463	0.12633	Inactivated	XCR	2
FOXP3	NM_014009	49.0	0.0109	0.06987	0.15601	Inactivated	XCR	2
PPP1R3F	NM_033215	49.0	0.02374	0.08477	0.28007	Inactivated	XCR	2
CLCN5	NM_001127899	49.6	0.00186	0.02809	0.06613	Inactivated	XCR	2
KIAA1202 (SHROOM4)	NM_020717	50.4	0.01873	0.04273	0.43829	Inactivated	XCR	2
NUDT10	NM_153183	51.1	0.00768	0.05204	0.14764	Inactivated	XCR	2
LOC340602	NM_203407	51.2	0.06647	0.05342	1.24434	Inactivated	XCR	2
NUDT11	NM_018159	51.2	0.00267	0.06331	0.04214	Inactivated	XCR	2
SREB3(GPR173)	NM_018969	53.1	0.00487	0.06627	0.07347	Inactivated	XCR	2
JARID1C	NM_004187	53.2	0.00265	0.04094	0.06479	Escape	XCR	2
SMC1L1	NM_006306	53.4	0.00039	0.04288	0.00898	Escape	XCR	2
FLJ32783	NM_001031745	53.5	0.01816	0.04695	0.38677	heterogeneous	XCR	2
HADH2 (HSD17B10)	NM_004493	53.5	0.01317	0.04787	0.27515	Inactivated	XCR	2
UREB1(HUWE1)	NM_031407	53.6	0.00265	0.03498	0.07573	Inactivated	XCR	2
PHF8	NM_015107	54.0	0.00491	0.02634	0.18655	Inactivated	XCR	2
PRKWNK3(WNK3)	NM_020922	54.2	0.01059	0.03386	0.3129	Inactivated	XCR	2
DT1P1A10(TSR2)	NM_058163	54.5	0	0.00001	0.41753	Inactivated	XCR	2
FGD1	NM_004463	54.5	0.00975	0.06072	0.16051	Inactivated	XCR	2
FLJ10613(GNL3L)	NM_019067	54.6	0.01609	0.03975	0.40471	Inactivated	XCR	2
MAGED2	NM_177433	54.9	0.00848	0.03754	0.2259	Inactivated	XCR	2
PFKFB1	NM_002625	55.0	0.00668	0.05557	0.12018	Inactivated	XCR	2
APEX2	NM_014481	55.0	0.02158	0.04723	0.45689	Inactivated	XCR	2
ALAS2	NM_001037968	55.1	0.01083	0.04663	0.23218	Inactivated	XCR	2
LOC90736 (FAM104B)	NM_138362	55.2	0.03491	0.0482	0.72419	Inactivated	XCR	2
MAGEH1	NM_014061	55.5	0.18834	0.61268	0.3074	Inactivated	XCR	2
KLF8	NM_007250	56.3	0.01121	0.01972	0.56826	Inactivated	XCR	2
SPIN3	NM_001010862	57.0	0.00961	0.04423	0.21726	Inactivated	XCR	2
SPIN2	NM_019003	57.2	0.01433	0.05087	0.28171	Inactivated	XCR	2
ZXDB	NM_007157	57.6	0.02252	0.09139	0.24641	Inactivated	XCR	2
ZXDA	NM_007156	57.9	0.04365	0.08675	0.50311	Inactivated	XCR	2
ASB12	NM_130388	63.4	0.00915	0.08639	0.10587	Inactivated	XCR	1
MTMR8	NM_017677	63.4	0.01456	0.031	0.46961	Inactivated	XCR	1
HCA127	NM_018684	64.1	0.00001	0.05322	0.0001	Inactivated	XCR	1
FLJ34366 (ZC3H12B)	NM_001010888	64.6	0.00399	0.02801	0.14239	Inactivated	XCR	1
FLJ12525(LAS1L)	NM_031206	64.6	0.00719	0.02578	0.27885	Inactivated	XCR	1
MSN	NM_002444	64.8	0.00001	0.07367	0.0001	Inactivated	XCR	1
XEDAR(EDA2R)	NM_021783	65.7	0.03585	0.03029	1.18334	heterogeneous	XCR	1
OPHN1	NM_002547	67.2	0.00177	0.03967	0.04471	Inactivated	XCR	1
MGC21416(YIPF6)	NM_173834	67.6	0.00267	0.02703	0.09874	Inactivated	XCR	1
STARD8	NM_014725	67.8	0.00838	0.06458	0.1298	Inactivated	XCR	1
EFNB1	NM_004429	68.0	0.0149	0.06585	0.22634	Inactivated	XCR	1
PJA1	NM_022368	68.3	0.03804	0.13621	0.27927	Inactivated	XCR	1
ED1(EDA)	NM_001399	68.8	0.00724	0.04369	0.16577	Inactivated	XCR	1

IGBP1	NM_001551	69.3	0.0149	0.06684	0.22295	Inactivated	XCR	1
ARR3	NM_004312	69.4	0.01291	0.04828	0.2674	Inactivated	XCR	1
DLG3	NM_021120	69.6	0.00296	0.04853	0.06092	Inactivated	XCR	1
SNX12	NM_013346	70.2	0	0.04063	0.0001	Inactivated	XCR	1
MLLT7	NM_005938	70.2	0.00779	0.03888	0.20023	heterogeneous	XCR	1
NLGN3	NM_018977	70.3	0.00254	0.03479	0.07314	Inactivated	XCR	1
ITGB1BP2	NM_012278	70.4	0.00525	0.0189	0.27754	Inactivated	XCR	1
OGT	NM_181672	70.7	0.00285	0.05716	0.04979	Inactivated	XCR	1
ACRC	NM_052957	70.7	0.12584	0.11734	1.07238	Inactivated	XCR	1
CXCR3	NM_001504	70.8	0.03844	0.1265	0.30388	Inactivated	XCR	1
PIN4	NM_006223	71.3	0.00501	0.0714	0.0701	heterogeneous	XCR	1
RPS4X	NM_001007	71.4	0	0.02744	0.0001	Escape	XCR	1
CITED1	NM_004143	71.4	0.0158	0.06124	0.25799	Inactivated	XCR	1
PHKA1	NM_002637	71.7	0.00766	0.04367	0.17535	Inactivated	XCR	1
NAP1L2	NM_021963	72.3	0.01403	0.06036	0.23242	Inactivated	XCR	1
CHIC1	NM_001039840	72.7	0.00403	0.04885	0.0824	Inactivated	XCR	1
RNF12	NM_016120	73.7	0.00356	0.07476	0.04763	Inactivated	XCR	1
CXorf26	NM_016500	75.3	0.0036	0.04064	0.08867	Inactivated	XCR	1
MAGEE1	NM_020932	75.6	0.02403	0.0656	0.36639	heterogeneous	XCR	1
ATRX	NM_000489	76.6	0.00747	0.05414	0.13797	Inactivated	XCR	1
DKFZp564K142	NM_032121	77.0	0.00872	0.04062	0.21462	Inactivated	XCR	1
COX7B	NM_001866	77.0	0.02529	0.04318	0.58568	heterogeneous	XCR	1
ATP7A	NM_000052	77.1	0.01135	0.04225	0.26856	Inactivated	XCR	1
PGK1	NM_000291	77.2	0.00151	0.05396	0.0279	Inactivated	XCR	1
GPR23	NM_005296	77.9	0.00479	0.09495	0.05041	Inactivated	XCR	1
ITM2A	NM_004867	78.5	0.00001	0.05539	0.0001	heterogeneous	XCR	1
RPS6KA6	NM_014496	83.2	0.00392	0.0173	0.22639	Inactivated	XCR	1
UNQ8193(APOOL)	NM_198450	84.1	0.02389	0.05122	0.46634	Inactivated	XCR	1
SATL1	NM_001012980	84.2	0.06404	0.083	0.77161	Inactivated	XCR	1
CHM	NM_000390	85.0	0.01688	0.05669	0.29773	heterogeneous	XCR	1
NAP1L3	NM_004538	92.8	0.01783	0.03356	0.53122	Inactivated	XCR	1
HSU24186 (DIAPH2)	NM_006729	95.8	0.00943	0.06293	0.14985	heterogeneous	XCR	1
TM4SF6(TSPAN6)	NM_003270	99.8	0.01049	0.02508	0.41828	Inactivated	XCR	1
SRPX2	NM_014467	99.8	0.00572	0.05556	0.10296	heterogeneous	XCR	1
SYTL4	NM_080737	99.8	0.00694	0.04531	0.15314	Inactivated	XCR	1
CSTF2	NM_001325	100.0	0.00174	0.053	0.03274	Inactivated	XCR	1
FLJ12687(CXorf34)	NM_024917	100.2	0.02493	0.03576	0.69709	Inactivated	XCR	1
FLJ14084 (TMEM35)	NM_021637	100.2	0.00001	0.06722	0.0001	Inactivated	XCR	1
FSHPRH1(CENPI)	NM_006733	100.2	0.01625	0.04337	0.37467	Inactivated	XCR	1
DRP2	NM_001939	100.4	0.00278	0.05217	0.05337	Inactivated	XCR	1
HNRPH2	NM_019597	100.5	0	0.01884	0.0001	Inactivated	XCR	1
ARMCX1	NM_016608	100.7	0.00751	0.04753	0.15792	Inactivated	XCR	1
FLJ20811 (ARMCX6)	NM_001009584	100.8	0.0255	0.05527	0.46132	Inactivated	XCR	1
ALEX3(ARMCX3)	NM_177947	100.8	0.00469	0.04231	0.1108	Inactivated	XCR	1
ARMCX3	NM_016607	100.8	0.00469	0.04231	0.1108	Inactivated	XCR	1
ARMCX2	NM_014782	100.8	0.01265	0.05093	0.24842	Inactivated	XCR	1
my048(TCEAL2)	NM_080390	101.3	0.03599	0.05327	0.67565	Inactivated	XCR	1
TMSNB(TMSL8)	NM_021992	101.7	0.0124	0.05644	0.2198	Inactivated	XCR	1
FLJ12969 (ARMCX5)	NM_022838	101.7	0.02584	0.04051	0.63788	Inactivated	XCR	1
GPRASP2	NM_138437	101.9	0.01969	0.04591	0.42901	Inactivated	XCR	1
BEX1	NM_018476	102.2	0.01147	0.03613	0.31741	Inactivated	XCR	1
NXF3	NM_022052	102.2	0.03588	0.0762	0.47087	heterogeneous	XCR	1
MGC45400	NM_153333	102.4	0.01703	0.02472	0.68896	Inactivated	XCR	1

(TCEAL8)								
BEX2	NM_032621	102.5	0.02105	0.07571	0.27804	Inactivated	XCR	1
MGC23947 (TCEAL7)	NM_152278	102.5	0.0042	0.03147	0.13332	Inactivated	XCR	1
WBP5	NM_001006613	102.5	0.00407	0.04622	0.08796	heterogeneous	XCR	1
NGFRAP1	NM_206915	102.5	0	0.03375	0.0001	Inactivated	XCR	1
RAB40A	NM_080879	102.6	0.01907	0.11779	0.16186	heterogeneous	XCR	1
FLJ21174 (TCEAL4)	NM_001006935	102.7	0.01536	0.03625	0.42372	Inactivated	XCR	1
TCEAL1	NM_001006640	102.8	0.00318	0.02457	0.12938	Inactivated	XCR	1
MORF4L2	NM_001142424	102.8	0.00164	0.04204	0.03901	heterogeneous	XCR	1
PLP1	NM_000533	102.9	0.00173	0.01687	0.10234	Inactivated	XCR	1
RAB9B	NM_016370	103.0	0.00266	0.02399	0.1108	Inactivated	XCR	1
MGC39900 (TMSB15B)	NM_194324	103.1	0.00999	0.06524	0.1531	Inactivated	XCR	1
DKFZp686O1267 (MCART6)	NM_001012755	103.2	0.0042	0.07076	0.05941	Inactivated	XCR	1
CXORF39	NM_207318	103.3	0	0.02698	0.0001	Inactivated	XCR	1
FLJ33516 (MUM1L1)	NM_152423	105.3	0.03145	0.05812	0.54118	Inactivated	XCR	1
FLJ20298 (TBC1D8B)	NM_017752	105.9	0.00862	0.04227	0.20395	Inactivated	XCR	1
ZCWCC2(MORC4)	NM_024657	106.1	0.01111	0.04607	0.2412	Inactivated	XCR	1
FLJ11016(RBM41)	NM_018301	106.2	0.00565	0.02849	0.19826	heterogeneous	XCR	1
PRPS1	NM_002764	106.8	0	0.04528	0.0001	Inactivated	XCR	1
DSIPI(TSC22D3)	NM_198057	106.8	0.01195	0.10096	0.11839	Inactivated	XCR	1
MID2	NM_012216	107.0	0.00072	0.04526	0.01583	Inactivated	XCR	1
PSMD10	NM_002814	107.2	0.00208	0.03783	0.05506	Inactivated	XCR	1
APG4A(ATG4A)	NM_052936	107.2	0.00997	0.0303	0.32898	Inactivated	XCR	1
COL4A6	NM_001847	107.3	0.01813	0.0502	0.36108	heterogeneous	XCR	1
NXT2	NM_018698	108.7	0.05392	0.02355	2.28936	Inactivated	XCR	1
KCNE1L	NM_012282	108.8	0.07046	0.31832	0.22136	Inactivated	XCR	1
CHRDL1	NM_145234	109.8	0.00235	0.05985	0.03926	Inactivated	XCR	1
MDS031(ALG13)	NM_018466	110.8	0.00986	0.02767	0.35621	heterogeneous	XCR	1
TRPC5	NM_012471	110.9	0.00265	0.03521	0.07532	heterogeneous	XCR	1
AMOT	NM_001113490	111.9	0.00993	0.04691	0.21158	Inactivated	XCR	1
LRCH2	NM_020871	114.3	0.00593	0.05539	0.10713	Inactivated	XCR	1
PLS3	NM_001136025	114.7	0	0.02809	0.0001	heterogeneous	XCR	1
KLHL13	NM_033495	116.9	0.00081	0.04812	0.0168	Inactivated	XCR	1
DKFZp686L20145 (WDR44)	NM_019045	117.4	0.00945	0.04074	0.23206	Inactivated	XCR	1
DOCK11	NM_144658	117.5	0.00218	0.04585	0.04754	heterogeneous	XCR	1
IL13RA1	NM_001560	117.7	0.01559	0.04151	0.37552	Inactivated	XCR	1
RNF127(LONRF3)	NM_001031855	118.0	0.01209	0.05577	0.21671	Inactivated	XCR	1
PGRMC1	NM_006667	118.3	0.01388	0.07968	0.17425	Inactivated	XCR	1
LOC203427 (SLC25A43)	NM_145305	118.4	0.02789	0.03775	0.73872	heterogeneous	XCR	1
UBE2A	NM_003336	118.6	0	0.02043	0.0001	Inactivated	XCR	1
NRF	NM_017544	118.6	0.00342	0.03938	0.08688	Inactivated	XCR	1
SEPT6	NM_015129	118.6	0.00137	0.06068	0.02256	Inactivated	XCR	1
RPL39	NM_001000	118.8	0.00784	0.13389	0.05856	Inactivated	XCR	1
ZNF183(RNF113A)	NM_006978	118.9	0.01139	0.02661	0.4281	Inactivated	XCR	1
AKAP28(AKAP14)	NM_178813	118.9	0.02595	0.06548	0.3963	Inactivated	XCR	1
ZBTB33	NM_006777	119.3	0.00334	0.03798	0.08798	Inactivated	XCR	1
FLJ20716 (FAM70A)	NM_017938	119.3	0.00372	0.03233	0.11507	Inactivated	XCR	1
LAMP2	NM_001122606	119.4	0.01482	0.04081	0.36318	Inactivated	XCR	1
CUL4B	NM_003588	119.5	0.00215	0.03683	0.05843	heterogeneous	XCR	1

C1GALT2 (C1GALT1C1)	NM_001011551	119.6	0.00147	0.03961	0.0371	Inactivated	XCR	1
THOC2	NM_001081550	122.6	0.00193	0.0397	0.04863	Inactivated	XCR	1
BIRC4	NM_001167	122.8	0.0198	0.04139	0.47826	Inactivated	XCR	1
STAG2	NM_001042750	122.9	0.00071	0.03938	0.018	Inactivated	XCR	1
ODZ1	NM_014253	123.3	0.00484	0.05799	0.08339	Inactivated	XCR	1
SMARCA1	NM_003069	128.4	0.00193	0.05182	0.03721	Inactivated	XCR	1
APLN	NM_017413	128.6	0.03698	0.06398	0.57793	Inactivated	XCR	1
XPNPEP2	NM_003399	128.7	0.01054	0.08331	0.1265	Inactivated	XCR	1
CXorf9	NM_018990	128.7	0.00468	0.05549	0.08434	Inactivated	XCR	1
ZDHHC9	NM_016032	128.8	0.00157	0.05344	0.0294	Inactivated	XCR	1
UTP14A	NM_006649	128.9	0.01719	0.05959	0.28846	heterogeneous	XCR	1
ELF4	NM_001127197	129.0	0.00922	0.0776	0.11883	Inactivated	XCR	1
Hs.424932(AIFM1)	NM_004208	129.1	0.00954	0.03619	0.26375	Inactivated	XCR	1
SLC25A14	NM_003951	129.3	0	0.04403	0.0001	Inactivated	XCR	1
RBMX2	NM_016024	129.4	0.01581	0.06961	0.2271	Inactivated	XCR	1
COVA1(ENOX2)	NM_182314	129.6	0.00407	0.05543	0.07333	heterogeneous	XCR	1
MST4	NM_016542	131.0	0.00274	0.03386	0.08085	Inactivated	XCR	1
Hs.119889(RAP2C)	NM_021183	131.2	0	0.03629	0.0001	Inactivated	XCR	1
MBNL3	NM_018388	131.3	0.01758	0.03407	0.51598	Inactivated	XCR	1
GPC4	NM_001448	132.3	0.00615	0.06015	0.10227	Inactivated	XCR	1
PLAC1	NM_021796	133.5	0.02905	0.07966	0.36472	heterogeneous	XCR	1
LOC159090 (FAM122B)	NM_145284	133.7	0.03335	0.09874	0.33779	heterogeneous	XCR	1
LOC159091 (FAM122C)	NM_138819	133.8	0.06023	0.05443	1.10653	Inactivated	XCR	1
T91371(MOSPD1)	NM_019556	133.8	0.0042	0.04092	0.10258	Inactivated	XCR	1
DKFZP564B147 (FAM127B)	NM_001078172	134.0	0.02088	0.33464	0.0624	Inactivated	XCR	1
CXX1(FAM127A)	NM_001078171	134.0	0.02368	0.31359	0.07551	Inactivated	XCR	1
FLJ23614(ZNF449)	NM_152695	134.3	0.00613	0.04994	0.1228	Inactivated	XCR	1
DDX26B	NM_182540	134.5	0.00888	0.03274	0.27132	Inactivated	XCR	1
SLC9A6	NM_001042537	134.9	0.00255	0.0296	0.08602	Inactivated	XCR	1
FHL1	NM_001449	135.1	0	0.02002	0.0001	Inactivated	XCR	1
BRS3	NM_001727	135.4	0.01643	0.03346	0.49109	heterogeneous	XCR	1
HTATSF1	NM_014500	135.4	0.01281	0.05553	0.23066	Inactivated	XCR	1
TNFSF5(CD40LG)	NM_000074	135.6	0.0087	0.03866	0.22518	heterogeneous	XCR	1
ARHGEF6	NM_004840	135.6	0.01061	0.05545	0.19141	Inactivated	XCR	1
RBMX	NM_002139	135.8	0.00267	0.05824	0.04579	Inactivated	XCR	1
ATP11C	NM_173694	138.6	0.00626	0.03383	0.18499	Inactivated	XCR	1
FMR1	NM_002024	146.8	0.00459	0.0304	0.15097	Inactivated	XCR	1
IDS	NM_000202	148.4	0.01987	0.0539	0.3687	Inactivated	XCR	1
CXORF40~ (CXORF40A)	NM_178124	148.4	0.02494	0.0331	0.75359	Inactivated	XCR	1
MAGEA8	NM_005364	148.8	0.02218	0.11173	0.19854	heterogeneous	XCR	1
CXORF6	NM_005491	149.3	0.01397	0.07254	0.1926	Inactivated	XCR	1
MTM1	NM_000252	149.5	0.00392	0.0465	0.08428	Inactivated	XCR	1
MTMR1	NM_003828	149.6	0.01661	0.1122	0.14801	Inactivated	XCR	1
CD99L2	NM_031462	149.7	0.02991	0.05255	0.56914	Inactivated	XCR	1
CD99L2	NM_134446	149.7	0.02434	0.05474	0.44468	Inactivated	XCR	1
HMGB3	NM_005342	149.9	0.00001	0.06159	0.0001	Inactivated	XCR	1
LOC203547	NM_001017980	150.3	0.00001	0.05576	0.0001	Inactivated	XCR	1
GABRE	NM_004961	150.9	0.00567	0.0675	0.08406	Inactivated	XCR	1
CALT(CETN2)	NM_004344	151.7	0.00256	0.07561	0.03385	Inactivated	XCR	1
NSDHL	NM_001129765	151.8	0.01717	0.05725	0.29988	Inactivated	XCR	1
PNMA5	NM_001103151	151.9	0.03861	0.09193	0.41998	Inactivated	XCR	1
PNMA6A	NM_032882	152.0	0.48855	1.33264	0.3666	Inactivated	XCR	1

TREX2	NM_080701	152.4	0.01683	0.2731	0.06164	Inactivated	XCR	1
BGN	NM_001711	152.4	0.00276	0.25167	0.01095	Inactivated	XCR	1
ATP2B3	NM_001001344	152.4	0.00349	0.12487	0.02797	Inactivated	XCR	1
DUSP9	NM_001395	152.6	0.01515	0.13791	0.10982	Inactivated	XCR	1
ABCD1	NM_000033	152.6	0.01206	0.15572	0.07747	heterogeneous	XCR	1
STK23(SRPK3)	NM_014370	152.7	0.00902	0.20297	0.04445	Inactivated	XCR	1
SSR4	NM_006280	152.7	0.00513	0.15327	0.03349	Inactivated	XCR	1
PDZK4(PDZD4)	NM_032512	152.7	0.00789	0.16303	0.04841	Inactivated	XCR	1
L1CAM	NM_000425	152.8	0.00311	0.13503	0.02304	Escape	XCR	1
ARHGAP4	NM_001666	152.8	0.00883	0.1517	0.05819	Escape	XCR	1
ARD1	NM_003491	152.8	0.00849	0.05264	0.16134	heterogeneous	XCR	1
CXORF12 (TMEM187)	NM_003492	152.9	0.03271	0.20993	0.15583	heterogeneous	XCR	1
IRAK1	NM_001569	152.9	0.01909	0.09606	0.1987	Inactivated	XCR	1
MECP2	NM_001110792	152.9	0.00369	0.04876	0.07564	Inactivated	XCR	1
TKTL1	NM_012253	153.2	0.02406	0.09096	0.26454	Inactivated	XCR	1
FLNA	NM_001110556	153.2	0.00364	0.12678	0.02872	Inactivated	XCR	1
RPL10	NM_006013	153.3	0.01211	0.08112	0.14928	Inactivated	XCR	1
DNASE1L1	NM_001009933	153.3	0.02204	0.20409	0.108	Inactivated	XCR	1
TAZ	NM_000116	153.3	0.00648	0.05696	0.11376	Inactivated	XCR	1
ATP6AP1	NM_001183	153.3	0.01237	0.12171	0.1016	Inactivated	XCR	1
GDI	NM_001493	153.3	0.00198	0.09435	0.02099	Inactivated	XCR	1
DXS9928E(FAM50A)	NM_004699	153.3	0.00116	0.1546	0.00751	Inactivated	XCR	1
PLXN3(PLXNA3)	NM_017514	153.3	0.00737	0.18192	0.04049	Inactivated	XCR	1
DXS9879E (LAGE3)	NM_006014	153.4	0.04371	0.1161	0.37647	Inactivated	XCR	1
UBL4	NM_014235	153.4	0.0054	0.1012	0.0534	Inactivated	XCR	1
SLC10A3	NM_019848	153.4	0.02385	0.12847	0.18567	Inactivated	XCR	1
FAM3A	NM_021806	153.4	0.00441	0.17449	0.0253	Inactivated	XCR	1
G6PD	NM_000402	153.4	0.00234	0.14479	0.01619	Inactivated	XCR	1
GAB3	NM_001081573	153.6	0.0116	0.04762	0.24358	Escape	XCR	1
DKC1	NM_001363	153.6	0.00369	0.05962	0.06194	Inactivated	XCR	1
MPP1	NM_002436	153.7	0.0062	0.05149	0.12041	Inactivated	XCR	1
F8	NM_000132	153.7	0.01722	0.04583	0.37582	Inactivated	XCR	1
HCBP6(FUNDC2)	NM_023934	153.9	0.0178	0.02399	0.74196	Inactivated	XCR	1
C6.1A(BRCC3)	NM_001018055	154.0	0.0035	0.02493	0.14029	heterogeneous	XCR	1
VBP1	NM_003372	154.1	0.00856	0.05833	0.14682	Inactivated	XCR	1
CLIC2	NM_001289	154.2	0.00743	0.05321	0.13965	Inactivated	XCR	1
TMLHE	NM_018196	154.4	0.00751	0.046	0.16335	Inactivated	XCR	1

Gene name in RefSeq is indicated in parentheses.

Loc[*]: Genomic coordinate of genes (megabase pair: Mbp)

E[†]: Evolutionary strata (XAR: X-added region, XCR: X-conserved region)

S[§]: Five evolutionary layers; their assignments are based on (Lahn and Page 1999; Skaletsky et al. 2003; Ross et al. 2005); because of unequal representation of genes in each layer, the five evolutionary layers were not used in this study.

Appendix C

Supporting Material for Chapter 4

Supplemental Figure Legends

Figure S4-1: Schematic diagram of showing boundary and non-boundary regions. Blue and yellow arrows depict escape and inactivated genes, respectively.

Figure S4-2: Schematic diagram of how to identify motifs overrepresented in the boundary regions, regardless of their uniqueness. Blue and yellow arrows depict escape and inactivated genes, respectively

Figure S4-3: Distribution of LTR (A) and L1 (B) transposable elements. Blue and yellow arrows depict escape and inactivated genes, respectively.

Figure S4-4: Positions of CpG islands on human and mouse. Blue and yellow arrows depict escape and inactivated genes, respectively. Black arrows depict genes whose XCI profile have not been assayed yet. Green bars indicate CpG islands within promoter regions of genes with clear XCI profile.

Figure S4-5: Positions of CTCF binding sites. Blue and yellow arrows depict escape and inactivated genes, respectively.

Figure S4-6: Distribution of H3K27me3 histone modification. Blue and yellow arrows depict escape and inactivated genes, respectively. Gray arrow depicts a gene whose XCI profile has not been assayed yet.

Figure S4-7: Frequencies of overrepresented oligomers found in (Carrel et al. 2006). Blue and yellow arrows depict escape and inactivated genes, respectively.

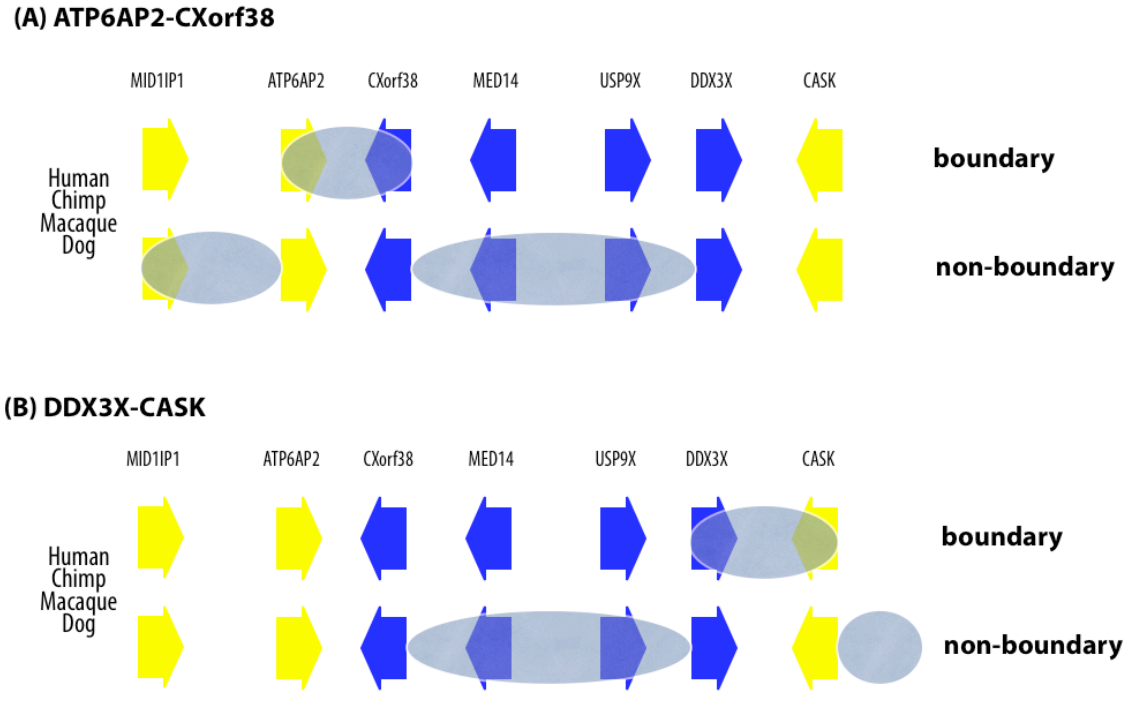


Figure S4-1

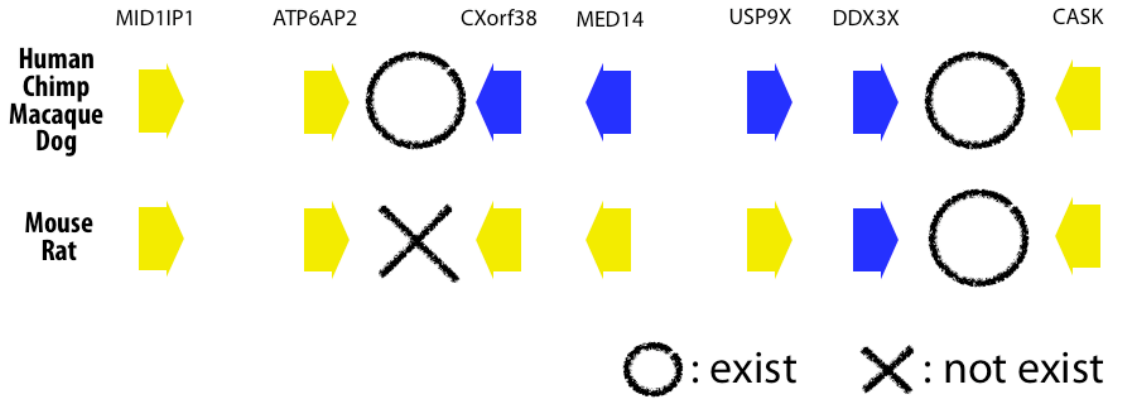


Figure S4-2

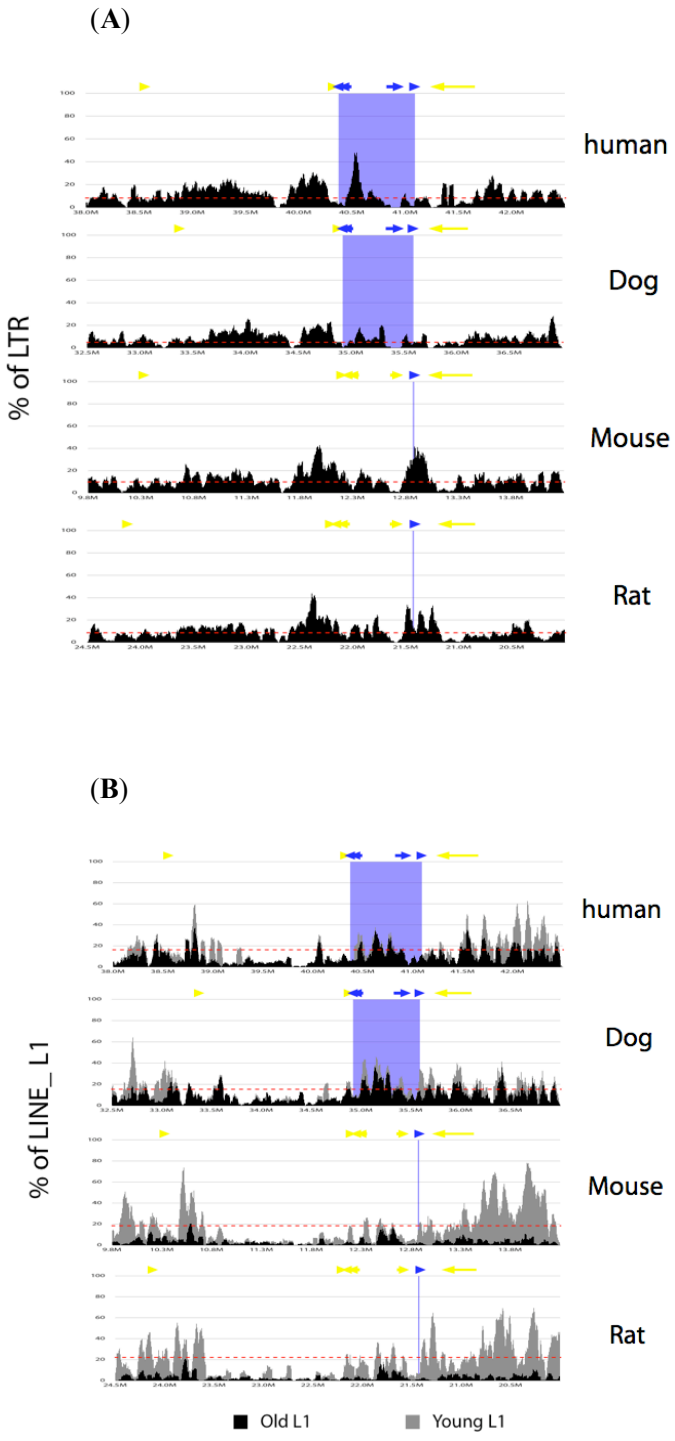

(A)
% of LTR
human
Dog
Mouse
Rat
(B)
% of LINE_L1
human
Dog
Mouse
Rat
Old L1
Young L1

Figure S4-3

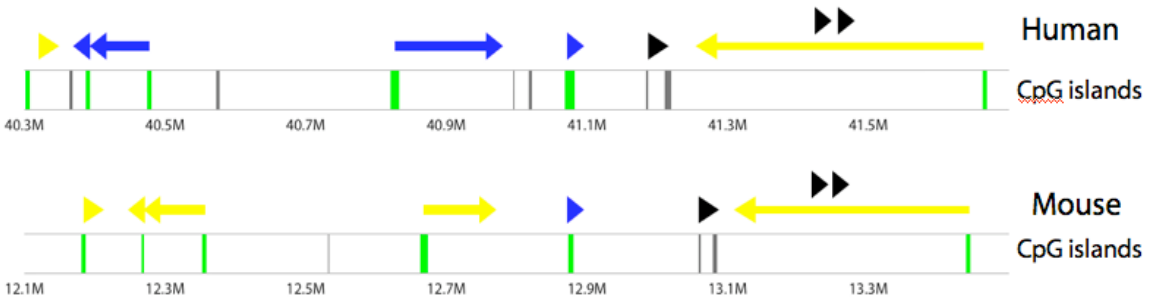


Figure S4-4

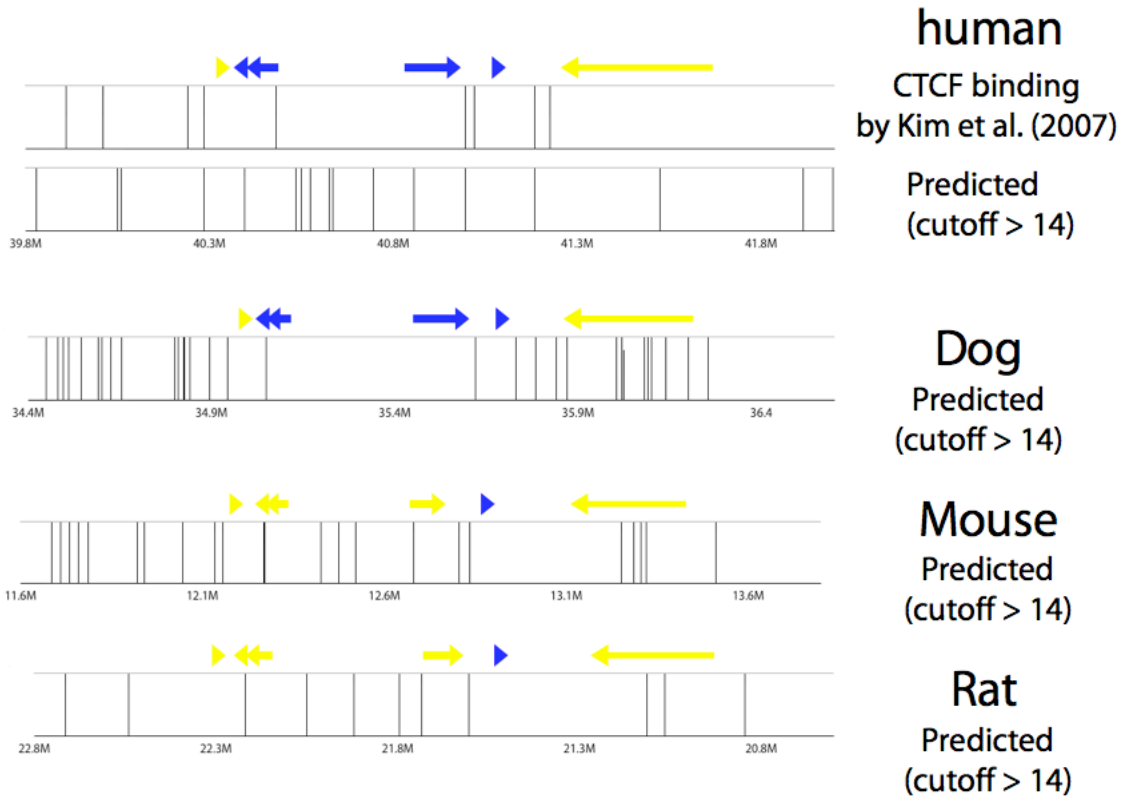


Figure S4-5

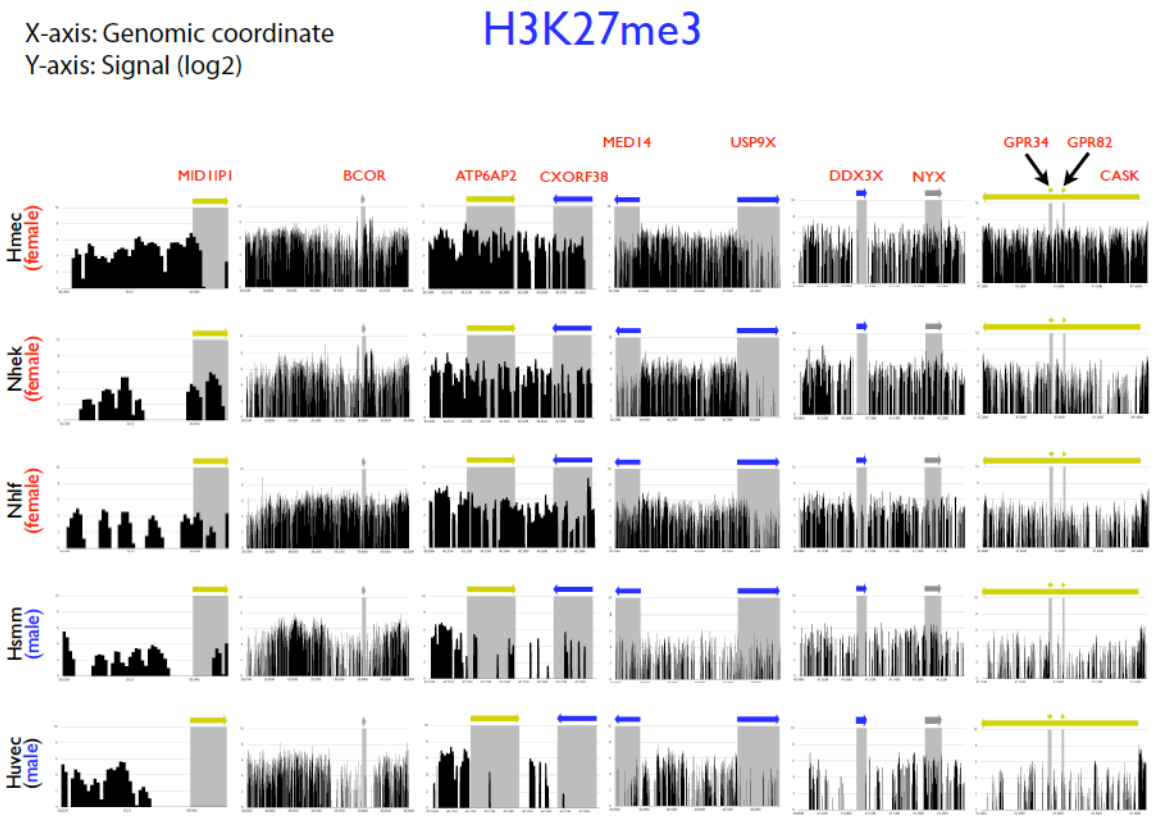


Figure S4-6

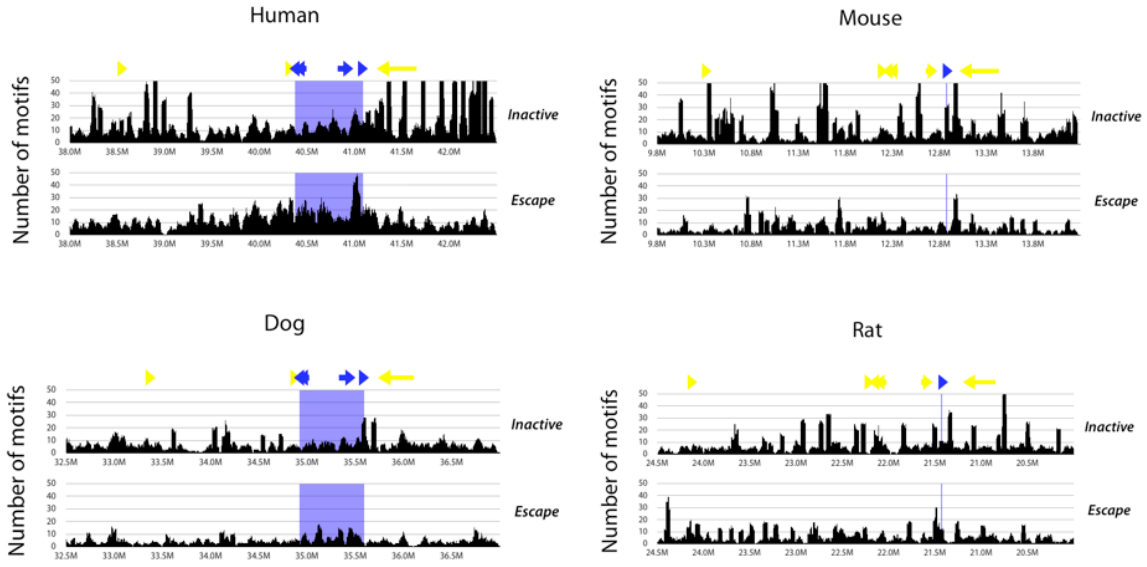


Figure S4-7

Supplemental Tables

Table S4-1. Evolutionarily conserved sequences in the boundary regions

Escape gene cluster*	Length of evolutionarily conserved sequences (upstream boundary / downstream boundary)
FAM9C	16.6 kb (4.3 / 12.3)
TMEM27	18.8 kb (3.3 / 15.5)
EIF1AX	0.5 kb (0.5 / 0)
EIF2S3	5.7 kb (0.2 / 5.5)
CXorf38	3.6 kb (0.6 / 3)
UBA1	1.6 kb (0.1 / 1.5)
JARID1C	1.1 kb (1.1 / 0)
L1CAM	3 kb (1.8 / 1.2)
GAB3	0.8 kb (0.7 / 0.1)
Total	51.7 kb (12.6 / 39.1)

* - first gene of each escape gene cluster

Table S4-2. Significantly overrepresented non-redundant motifs

Algorithm	AlignACE	MEME	Weeder	Total
Number of significantly overrepresented motifs	69	26	24	119
Number of redundant motifs	47 (9)*	0	0	47
Number of non-redundant motifs	31	26	24	81

* -number of motifs after clustering similar motifs is shown in parentheses

Appendix D

Supporting Material for Chapter 5

Supplemental Figure Legends

Figure S5-1: LDA classification success rates for different values of the turning parameter τ. (A) Test set of XAR genes, with training performed on XCR genes. (B) Test set of XCR genes, with training performed on XAR genes. Leave-one-out cross-validation was utilized to calculate correct classification rates. Dots indicate optimal values of τ (see Table S5-6).

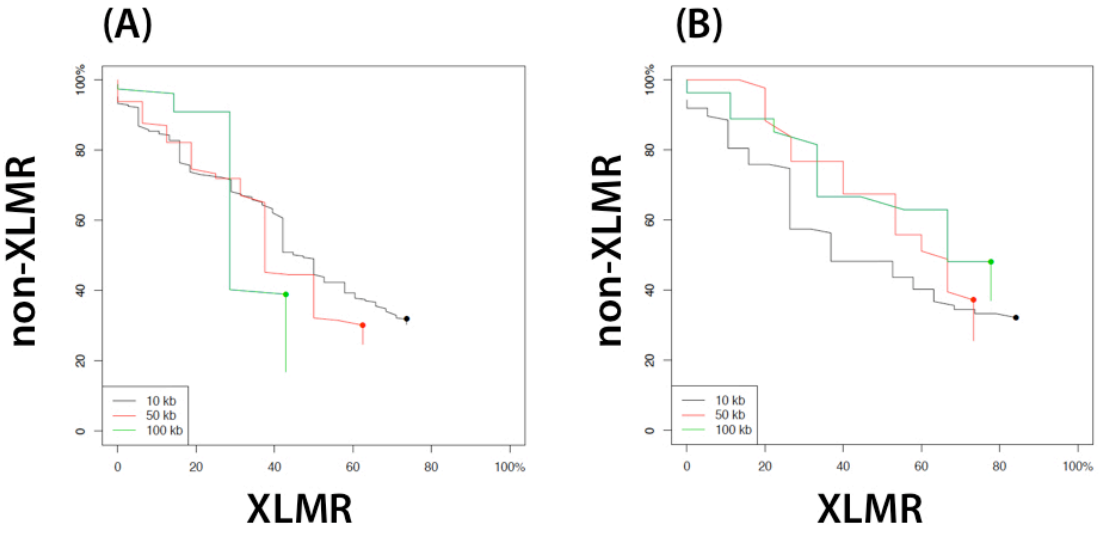

(A)
non-XLMR
XLMR
10 kb
50 kb
100 kb
0
20
40
60
80
100%
(B)
non-XLMR
XLMR
10 kb
50 kb
100 kb
0
20
40
60
80
100%

Figure S5-1

Supplemental Tables

Table S5-1. Overview of the prioritization of genes on the X chromosome as XLMR candidates using a binary filtering process

Nr of annotation terms matched (/40)	Total number of genes in category	Known XLMR genes in category	Non-XLMR genes in category	Category enrichment of XLMR genes (%)
27	1	1	0	100
24	2	1	1	50
23	3	1	2	33
22	2	2	0	100
21	7	3	4	43
20	6	3	3	50
19	16	4	12	25
18	17	3	14	18
17	15	3	12	20
16	21	3	18	14
15	17	4	13	24
14	35	6	29	17
13	30	9	21	30
12	25	3	22	12
11	35	6	29	17
10	23	2	21	9
9	36	2	34	6
8	22	1	21	5
7	40	7	33	18
6	49	3	46	6
5	55	4	51	7
4	81	6	75	7
3	93	3	90	3
2	182	2	180	1
0	1	0	1	0
TOTAL	**814**	**82**	**732**	

Genes were ranked as likely XLMR genes based on the number of annotation terms matched (more matches = more likely to be an XLMR gene). The dashed line indicates the top third of the rank list at which point 68% of known XLMR genes appear (top 16 categories).

Table S5-2. A list of loci that have been linked to different forms of XLMR (listed in order of their location on the X chromosome).

0	OMIM ID	PubMed ID
P ARM		
Xp22.3	300406%	(Dessay et al., 2002)
Xp22.1-p21.3		(Van Esch et al., 2005)
Xp22.13	#302350	(Zhu et al., 1990)
Xp22.13-p21.1	300148%	(Steinmuller et al., 1998)
Xp22	304050%	(Ballabio and Andria, 1992)
Xp22	300421%	(Wittwer et al., 1996)
Xp21.1-p11.22	309610%	(Watty et al., 1991)
Xp11.3-4	300422%	(Piluso et al., 2003)
Xp11.22	309545%	(Wilson et al., 1991)
Q ARM		
Xq11-q21	300519%	(Martin et al., 2000)
Xq12-q21	300262%	(Abidi et al., 1999)
Xq12-q21.31		(Shrimpton et al., 2000; Shrimpton et al., 1999)
Xq13-q22	309605%	(Miles and Carpenter, 1991)
Xq13.2-q21.2		(Stevenson et al., 1997)
Xq21.33-q23		(Chudley et al., 1999)
Xq22.3	300581%	(Jehee et al., 2005)
Xq23-q24		(Carpenter et al., 2000)
Xq24	*300360	(Vitale et al., 2001)
Xq25-q27		(Cilliers et al., 2007)
Xq26-q27	300238%	(Shashi et al., 2000)10677307
Xq26-q27	307700%	(Trump et al., 1998)
Xq26-q27.1	304340%	(Huang et al., 1991)
Xq27.3-q28	302000%	(Wijker et al., 1995)
Xq27-q28	309800%	(Graham et al., 1991)
Xq27-q28	301590%	(Graham et al., 1991)
Xq27-q28	309620%	(Dlouhy et al., 1987)
Xq28	300261%	(Armfield et al., 1999)
SPANNING		
Xp21.1-q22	309585%	(Wilson et al., 1991)
Xp21.2-q13		(Turner et al., 1994)
Xp11.3-q23	300218%	(Ahmad et al., 1999)
Xp11.1-q21.2		(Johnson et al., 1998)
Xp11.4-q24		(Oosterwijk et al., 1999)

No causative gene(s) have as yet been identified for these linked regions. Syndromes that are listed as an entry on the OMIM database have the OMIM ID listed and where available the key reference mentioning the linkage analysis is listed.

Table S5-3. Number of genes in each subgenome.

	X-added region				**X-conserved region**			
	XLMR		**Non-XLMR**		**XLMR**		**Non-XLMR**	
Contig size	# Genes	Length*	# Genes	Length	# Genes	Length	# Genes	Length
5 kb	21	0.1	96	0.5	45	0.2	291	1.5
10 kb	19	0.2	84	0.9	38	0.4	257	2.7
50 kb	15	0.8	40	2.2	16	0.8	141	7.3
100 kb	9	0.9	26	2.7	7	0.7	74	7.7

*: The length of each subgenome is given in megabases (Mb).

Table S5-4. Number of overrepresented oligomers in each subgenome.

	X-added Region								**X-conserved Region**							
	XLMR				**Non-XLMR**				**XLMR**				**Non-XLMR**			
	Contig (kb)				**Contig (kb)**				**Contig (kb)**				**Contig (kb)**			
Oligomers	**5**	**10**	**50**	**100**	**5**	**10**	**50**	**100**	**5**	**10**	**50**	**100**	**5**	**10**	**50**	**100**
8-mers	9	4	6	22	757	411	160	136	0	2	2	1	753	819	889	1518
12-mers	3	6	72	68	167	314	686	860	3	29	25	9	700	3112	12660	20522
16-mers	0	0	69	71	157	247	309	536	1	8	8	14	317	2255	3710	6171
20-mers	0	0	56	56	89	199	188	357	0	0	2	10	199	1987	2804	4228
24-mers	0	0	37	37	64	124	120	230	0	0	1	4	145	1805	2044	3206

Table S5-5. Number of significantly overrepresented 12-mers with $P < 0.05$

	X-added region		X-conserved region	
	XLMR	**Non-XLMR**	**XLMR**	**Non-XLMR**
10 kb	3	44	6	625
50 kb	2	259	1	288
100 kb	2	259	9	622

Table S5-6. Number of genes for test sets and success rates of LDA.

Set Analyzed	**Parameter**	**10 kb (Genes)**	**50 kb (Genes)**	**100 kb (Genes)**
Test set of genes in XCR using training set derived from XAR genes	τ	0.95	0.93	0.8
	Success in XLMR	74% (38)	63% (16)	43% (7)
	Success in non-XLMR	32% (257)	30% (141)	39% (74)
Test set of genes in XAR using training set derived from XCR genes	τ	0.98	0.96	0.94
	Success in XLMR	84% (19)	73% (15)	78% (9)
	Success in non-XLMR	32% (84)	37% (40)	48% (26)

τ is a tuning parameter, which was selected to maximize the sum of correct classification rates for XLMR and non-XLMR sets.

Table S5-7. Correctly and incorrectly classified genes using the identified oligomer signature

A

	X-added Region				
	Genes	**Classes**	**10 kb**	**50 kb**	**100 kb**
	BCOR	XLMR	O	O	O
	DMD	XLMR	O	O	O
	HCCS	XLMR	O	O	O
	IL1RAPL1	XLMR	O	O	O
	MAOA	XLMR	O	O	O
	PDHA1	XLMR	O	O	O
	RPS6KA3	XLMR	O	O	O
	TSPAN7	XLMR	O	O	O
	VCX3A	XLMR	O	O	O
	CD99	Non-XLMR	O	O	O
	CNKSR2	Non-XLMR	X	O	O
	DDX53	Non-XLMR	X	O	O
	EGFL6	Non-XLMR	X	O	O
	FAM47B	Non-XLMR	O	O	O
	FAM9C	Non-XLMR	X	O	O
	FUNDC1	Non-XLMR	O	O	O
	GEMIN8	Non-XLMR	X	O	O
	GLRA2	Non-XLMR	X	O	O
	GPR64	Non-XLMR	O	O	O
	GRPR	Non-XLMR	X	O	O
	MAGEB10	Non-XLMR	O	O	O
	MAGEB18	Non-XLMR	X	O	O
	MAGEB2	Non-XLMR	X	O	O
	MAP3K7IP3	Non-XLMR	X	O	O
	MID1IP1	Non-XLMR	X	O	O
	PDK3	Non-XLMR	O	O	O
	PPP2R3B	Non-XLMR	O	X	O
	PRDX4	Non-XLMR	X	O	O
	PTCHD1	Non-XLMR	X	O	O
	RAI2	Non-XLMR	X	O	O
	TMEM47	Non-XLMR	X	O	O
	USP9X	Non-XLMR	O	O	O
	VCX	Non-XLMR	X	O	O
	VCX2	Non-XLMR	X	O	O
	VCX3B	Non-XLMR	X	O	O

B

X-Conserved Region				
Genes	**Classes**	**10 kb**	**50 kb**	**100 kb**
AGTR2	XLMR	O	O	O
ARHGEF9	XLMR	O	O	X
KIAA2022	XLMR	O	O	O
KLF8	XLMR	O	O	O
PAK3	XLMR	O	O	O
SOX3	XLMR	O	O	X
ZDHHC15	XLMR	O	O	O
ACTRT1	Non-XLMR	O	X	O
APOOL	Non-XLMR	O	O	O
AR	Non-XLMR	X	O	O
C1GALT1C1	Non-XLMR	O	O	O
CDR1	Non-XLMR	O	O	O
CHIC1	Non-XLMR	X	O	O
CHM	Non-XLMR	X	O	O
CHRDL1	Non-XLMR	X	O	O
CPXCR1	Non-XLMR	O	O	O
CXorf26	Non-XLMR	O	O	O
CXorf40B	Non-XLMR	X	O	O
CXorf57	Non-XLMR	O	O	O
CXorf61	Non-XLMR	O	O	O
CYSLTR1	Non-XLMR	O	O	O
DACH2	Non-XLMR	X	O	O
DIAPH2	Non-XLMR	X	O	O
ENOX2	Non-XLMR	X	O	O
ESX1	Non-XLMR	X	X	O
FGF13	Non-XLMR	X	O	O
FGF16	Non-XLMR	X	X	O
FXYD8	Non-XLMR	X	O	O
GABRE	Non-XLMR	O	O	O
GPC4	Non-XLMR	X	O	O
GSPT2	Non-XLMR	O	X	O
HDX	Non-XLMR	X	O	O
HEPH	Non-XLMR	O	O	O
IL1RAPL2	Non-XLMR	X	O	O
ITM2A	Non-XLMR	O	O	O
ZC4H2	Non-XLMR	X	X	O
KLHL4	Non-XLMR	O	O	O
LAS1L	Non-XLMR	X	O	O
LOC203547	Non-XLMR	O	O	O
LONRF3	Non-XLMR	X	O	O
MAGEA4	Non-XLMR	O	X	O
MAGEE2	Non-XLMR	O	O	O
MAGEH1	Non-XLMR	O	O	O
MAMLD1	Non-XLMR	X	X	O
MSN	Non-XLMR	X	O	O

NAP1L2	Non-XLMR	O	O	O
NKAP	Non-XLMR	O	O	O
NUDT10	Non-XLMR	X	O	X
ODZ1	Non-XLMR	X	X	O
PABPC5	Non-XLMR	O	O	O
PAGE1	Non-XLMR	X	O	O
PAGE4	Non-XLMR	X	X	O
PASD1	Non-XLMR	O	O	O
PGRMC1	Non-XLMR	X	O	O
PLS3	Non-XLMR	O	O	O
POF1B	Non-XLMR	O	O	O
RAP2C	Non-XLMR	O	X	O
RPS6KA6	Non-XLMR	O	O	O
SERPINA7	Non-XLMR	O	O	O
SLC25A43	Non-XLMR	X	O	O
SPANXB1	Non-XLMR	X	O	O
SPANXN1	Non-XLMR	O	O	O
SPANXN2	Non-XLMR	X	O	O
SPIN2A	Non-XLMR	O	O	O
SPIN3	Non-XLMR	O	O	O
SPIN4	Non-XLMR	O	O	O
STARD8	**Non-XLMR**	**X**	**X**	**X**
TAF9B	Non-XLMR	O	O	O
TBX22	Non-XLMR	X	O	O
TCEAL2	Non-XLMR	O	O	O
THOC2	Non-XLMR	X	O	O
TMEM164	Non-XLMR	O	O	O
TMEM28	Non-XLMR	X	X	O
TMLHE	Non-XLMR	O	O	O
TRO	Non-XLMR	O	O	X
UBQLN2	Non-XLMR	O	O	O
WDR44	Non-XLMR	O	O	O
ZIC3	Non-XLMR	X	O	O
ZNF275	Non-XLMR	O	O	O
ZNF711	Non-XLMR	O	O	O
ZXDA	Non-XLMR	O	O	O

"O" describes correctly classified genes. "X" described incorrectly classified genes. Only one gene was wrongly classified at all distances – shown in bold.

Appendix E

Supporting Material for Chapter 6

Additional data file Legends

Additional data file 1: A table listing the classification of duplicate gene pairs based on the absence or presence of shared TSSs and different duplication mechanisms.

Additional data file 2: A Venn diagram depicting the number of duplicate gene pairs that were identified by the FASTA and TRIBE-MCL methods.

Additional data file 3: Average sequence identity between TSS regions of duplicate genes.

Additional data file 4: Number of duplicate gene pairs with shared TSSs (A) and without shared TSSs (B) plotted against the instantaneous rate of K_S.

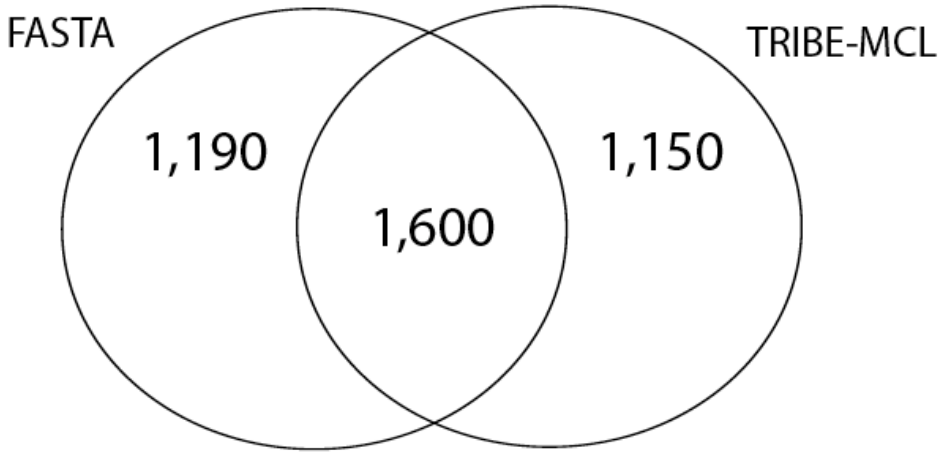
Additional data file 5: Proportions of group A duplicate gene pairs with shared TSSs depending on different identity thresholds.

Additional data file 6: Number of duplicate gene pairs in different structure categories plotted against the instantaneous rate of K_S

Table S1. Classification of duplicate gene pairs based on the absent or present of shared TSSs and different duplication mechanisms

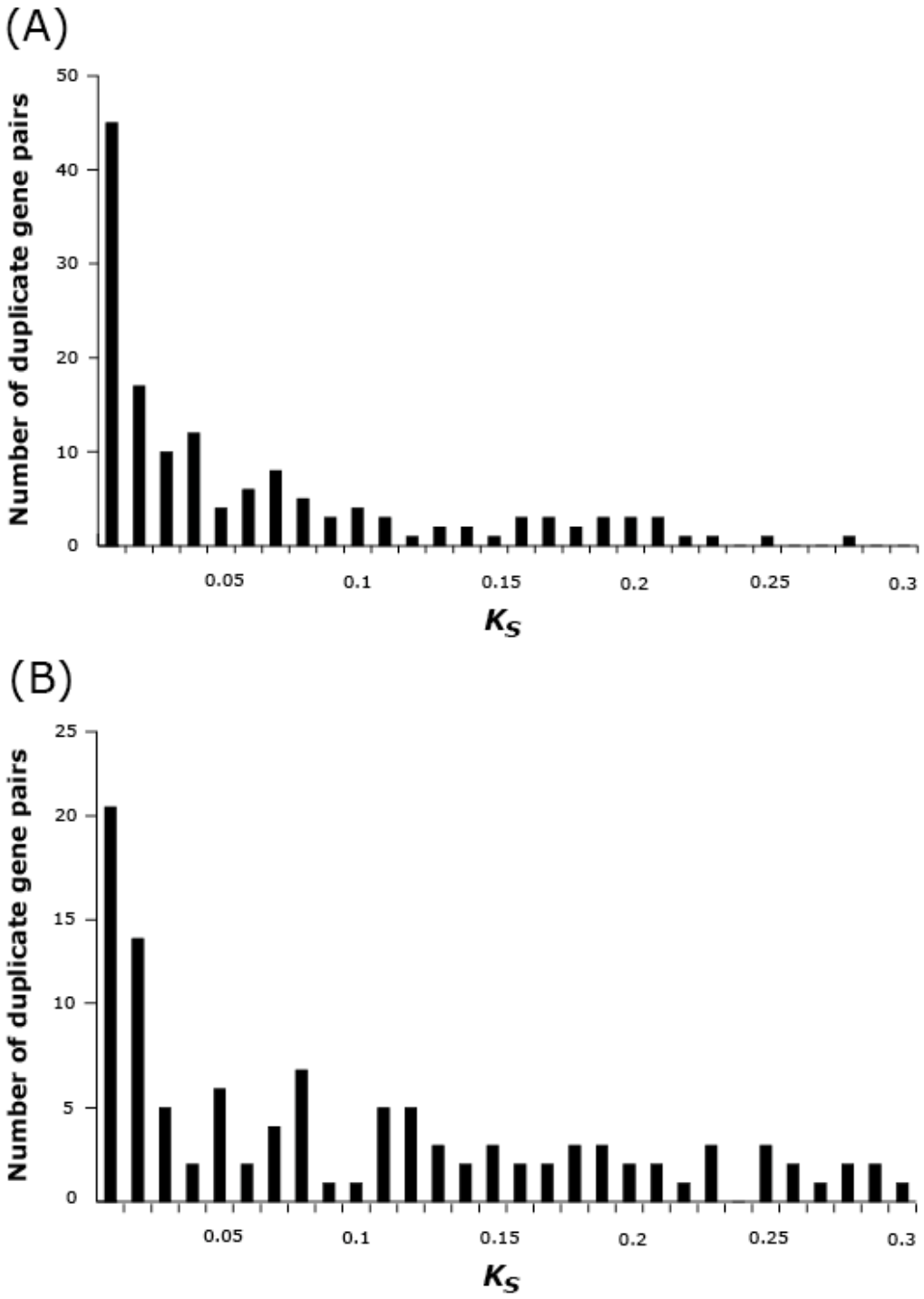
	Tandem duplication	Nontandem duplication	Retrotransposition
Number of duplicate genes with shared TSSs	119 (31.5%)	34 (4.6%)	67 (30.5%)
Number of duplicate genes without shared TSSs	259 (68.5%)	712 (95.4%)	153 (69.5%)
Total	378	746	220

Additional data file 1

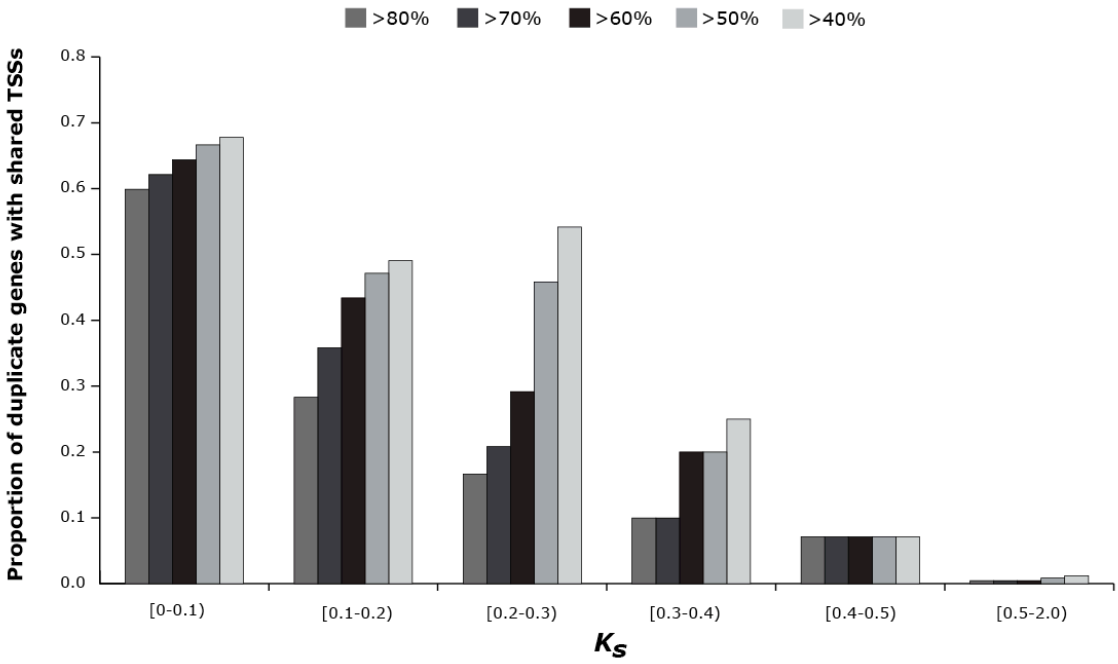


Additional data file 2

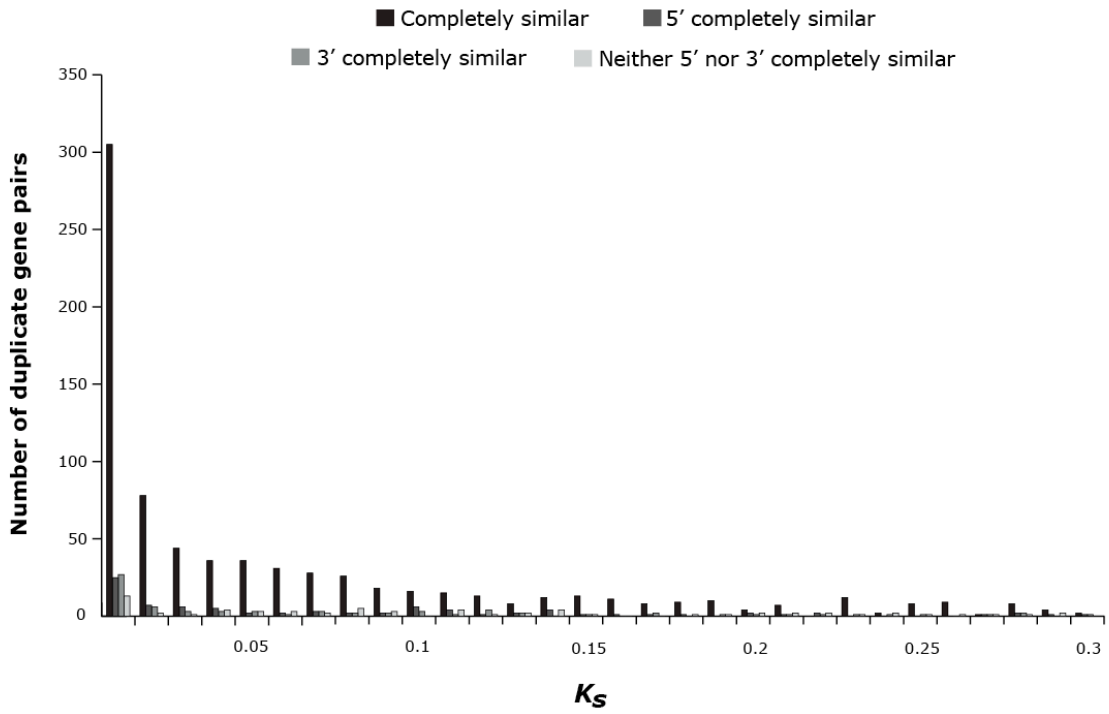
Additional data file 3



Additional data file 4



Additional data file 5



Additional data file 6

VITA Chungoo Park

EDUCATION

Ph.D., Biology with option in Molecular Evolutionary Biology (Anticipated 2010)
The Pennsylvania State University,
University Park, PA, USA

M.S., Information and Communications (2002)
Gwangju Institute of Science and Technology (GIST)
Gwangju, South Korea

B.S., Computer Science & Engineering (2000)
Inha University
Incheon, South Korea

HONORS AND AWARDS

ISMB Travel Fellowship award sponsored by NSF (2008)
The Penn State University Braddock Graduate Fellowship (2008)
The Penn State University Department of Biology Travel Award (2006 - 2009)
Institute for Molecular Evolutionary Genetics Travel Award (2006 - 2009)
The J. Ben and Helen D. Hill Memorial Fund Award (2005 - 2006)
The Penn State University Graduate Fellowship (2004 - 2005)
The Penn State University Braddock Graduate Fellowship (2004 - 2005)
Gwangju Institute of Science and Technology Scholarship (2000 - 2002)
Inha University Scholarship (1997 - 1999)

SELECTED PUBLICATIONS

Park, C., L. Carrel, K. D. Makova (2010) Strong purifying selection at genes escaping X chromosome inactivation. *Mol Biol Evol* Jun 9

Park, C., K. D. Makova (2009) Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. *Genome Biol* 10(1):R10.

Carrel, L., **C. Park**, S. Tyekucheva, J. Dunn, F. Chiaromonte, and K. D. Makova (2006) Genomic environment predicts expression patterns on the human inactive X chromosome. *PLoS Genet* 2:(9) e151.

INVITED SEMINARS

Department of Ecology and Evolutionary Biology, University of Michigan, April 2010

SELECTED PRESENTATIONS

"What factors contribute to rapid divergence in expression between duplicate genes?", Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, PA, February 2008.

"Genomic environment predicts expression patterns on the inactive X chromosome", Evolution Meeting. Stony Brook, NY, June 2006.

"Unique genomic landscape underlies human X inactivation profile", Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, PA, April 2006.