

States of Allelic Imbalance on the X Chromosomes in Human Females

by

Katerina Svejcarova Kucera

University Program in Genetics and Genomics  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Huntington F. Willard, Supervisor

\_\_\_\_\_  
Blanche Capel

\_\_\_\_\_  
Terrence Furey

\_\_\_\_\_  
Simon Gregory

Dissertation submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy  
in the University Program in Genetics and Genomics  
in the Graduate School of Duke University

2011

ABSTRACT

States of Allelic Imbalance on the X Chromosomes in Human Females

by

Katerina Svejcarova Kucera

University Program in Genetics and Genomics  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Huntington Willard, Supervisor

\_\_\_\_\_  
Blanche Capel

\_\_\_\_\_  
Terrence Furey

\_\_\_\_\_  
Simon Gregory

An abstract of a dissertation submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy  
in the University Program in Genetics and Genomics  
in the Graduate School of Duke University

2011

Copyright by  
Katerina Svejcarova Kucera  
2011

## Abstract

Allelic imbalance, in which two alleles at a given locus exhibit differences in gene expression, chromatin composition and/or protein binding, is a widespread phenomenon in the human and other complex genomes. Most examples concern individual loci located more or less randomly around the genome and thus imply local and gene-specific mechanisms. However, genomic or chromosomal basis for allelic imbalance is supported by multi-locus examples such as those exemplified by domains of imprinted genes, spanning ~1-2 Mb, or by X chromosome inactivation, involving much of an entire chromosome. Recent studies have shown that genes on the two female X chromosomes exhibit a breadth of expression patterns ranging from complete silencing of one allele to fully balanced biallelic expression. Although evidence for heritability of allele-specific chromatin and expression patterns exists at individual loci, it is unknown whether heritability is also reflected in the chromosome-wide patterns of X inactivation.

The aim of this thesis is to elucidate the extent to which the widespread variable patterns of allelic imbalance on the human X chromosome in females are under genetic control and how access of the transcription machinery to the human inactive X chromosome in females is determined at a genomic level. For the set of variable genes examined in this study, the absence or presence of expression appears to be stochastic with respect to the population rather than abiding by strict genetic rules. Furthermore, variable gene expression that I have detected even among multiple clonal cell lines

derived from a single individual suggests fluctuation in transcriptional machinery engagement. I find that, although expression at most genes on the human inactive X chromosome is repressed as a result of X inactivation, a number of loci are accessible to the transcriptional machinery. It appears that RNA Polymerase II is present at alleles on the inactive X even at the promoters of several silenced genes, indicating a potential for expression.

This thesis embodies a transition in the field of human X chromosome inactivation from gene by gene approaches used in the past to utilizing high-throughput technologies and applying follow-up analytic techniques to draw upon the vast data publicly available from large consortia projects.

## **Dedication**

I dedicate this thesis to my husband Jon, who has been my most patient supporter.

## Contents

Abstract.....	iv
List of Tables .....	xi
List of Figures .....	xii
1. Introduction .....	1
1.1. Allelic imbalance in expression.....	2
1.1.1. Regulation of variation in gene expression levels .....	2
1.1.1.1. Histone modifications.....	6
1.1.1.2. DNA methylation .....	8
1.1.2. X chromosome inactivation .....	10
1.1.2.1. Dosage compensation .....	10
1.1.2.2. Establishment, stability and reversal of the inactive state.....	14
1.1.2.3. Genomic context, evolution and expression profile of the human X chromosome .....	18
1.1.2.4. Heterochromatin of the inactive X chromosome.....	23
1.1.3. Genomic imprinting .....	26
1.1.4. Allelic exclusion.....	28
1.2. Allelic imbalance in transcriptional machinery and other binding proteins ....	30
1.3. Global efforts in genomics .....	36
1.3.1. International HapMap Project .....	37
1.3.2. ENCODE Project .....	37
1.3.3. 1000 Genomes Project.....	38
1.3.4. Cell types in studies of allelic imbalance .....	39

1.4.	Thesis overview.....	40
2.	Genetics of Variable Expression Patterns on the Human Inactive X Chromosome ..	42
2.1.	Introduction .....	44
2.1.1.	Genes on the human inactive X chromosome that exhibit variable expression status .....	44
2.2.	Results.....	48
2.2.1.	Selection of cell lines.....	48
2.2.2.	Derivation of clonal populations from female lymphoblastoid cell lines....	52
2.2.3.	Patterns of gene expression .....	56
2.2.4.	<i>TRAPPC2</i> exhibits unstable inactivation in multiple clones derived from a single female cell line.....	63
2.2.5.	<i>SEPT6</i> is expressed from one, but not the other inactive X chromosome in a single human cell line.....	64
2.3.	Discussion .....	69
2.3.1.	Stochastic inactivation of variable genes .....	69
2.3.2.	Genes with variable inactivation status mostly exhibit low level of expression and low frequency in the population .....	71
2.3.3.	X-inactivation skewing and derivation of homogeneous Xi populations ....	72
2.4.	Materials and methods.....	73
2.4.1.	Cell culture and single cell cloning.....	73
2.4.2.	Nucleic acid purification .....	74
2.4.3.	SNaPshot .....	75
3.	Allele-Specific Distribution of RNA Polymerase II on Female X Chromosomes .....	76
3.1.	Introduction .....	78
3.2.	Results.....	82

3.2.1.	The human X chromosome is relatively PolII poor .....	82
3.2.1.	PolII binding is biased toward the active X chromosome.....	87
3.2.1.	PolII binding on the human inactive X chromosome largely mimics inactivation status of genes .....	91
3.2.1.	PolII occupancy indicates chromatin accessibility and potential for expression .....	101
3.3.	Discussion .....	103
3.3.1.	PolII binding on the Xi is reduced .....	104
3.3.2.	PolII occupancy and gene expression on the Xi.....	105
3.3.3.	Genomic and chromatin features in X chromosome inactivation.....	106
3.3.4.	Some genes on the inactive X chromosome are poised for expression....	107
3.4.	Materials and methods.....	108
3.4.1.	X-inactivation skewing .....	108
3.4.2.	Cell culture and single-cell cloning .....	109
3.4.3.	ChIP-seq .....	109
3.4.4.	Measuring allele-specific PolII occupancy .....	110
3.4.5.	SNaPshot .....	110
4.	Conclusions and Future Work.....	111
4.1.	Dosage compensation of SEPT6.....	114
4.1.1.	Introduction .....	115
4.1.2.	Experimental design .....	116
4.2.	Paired Xi <sup>p</sup> and Xi <sup>m</sup> study in multiple cell lines to identify the level of heritability and cis/trans-acting elements influencing patterns of X inactivation.....	117
4.2.1.	Introduction .....	117
4.2.2.	Experimental Design .....	118

4.2.3. Preliminary data.....	119
Appendix A: Allele-specific dataset of 385 PolIII sites.....	124
Appendix B: Allele-specific expression of genes on the X chromosome.....	134
Appendix C: Primers used in this thesis.....	137
5. References .....	149
Biography .....	179

## List of Tables

Table 1 Selection of CEPH/HapMap family-derived cell lines according to heterozygosity .....	50
Table 2 Derivation of homogeneous cell lines with respect to the Xi .....	55
Table 3 Expression of variable genes in a panel of clonal cell lines.....	60
Table 4 Levels of gene expression in CEU lymphoblastoid cell lines .....	61
Table 5 Analysis of PolII binding sites on the human X chromosome .....	84
Table 6 Relationship of expression and PolII occupancy .....	97
Table 7 Expression and PolII binding on Xi relative to Xa .....	100

## List of Figures

Figure 1 Epigenetic domains of the human Xi .....	20
Figure 2 Models for genes with variable inactivation in the population.....	47
Figure 3 X-inactivation skewing in HapMap cell lines in this study .....	51
Figure 4 Derivation of homogeneous Xi population from randomly inactivated cell lines .....	54
Figure 5 SNaPshot assay at <i>XIST</i> rs1794213 to determine X-inactivation skewing in heterozygous lymphoblastoid cell lines.....	57
Figure 6 Variable expression of <i>CLIC2</i> in lymphoblastoid cell lines derived from a mother and two daughters.....	62
Figure 7 Expression of <i>TRAPPC2</i> in multiple isolates derived from GM12878 cell line... 65	
Figure 8 Expression of <i>SEPT6</i> from Xi <sup>m</sup> and Xi <sup>p</sup> in multiple isolates derived from GM12878 lymphoblastoid cell line .....	66
Figure 9 Expression of <i>SEPT6</i> from Xi <sup>m</sup> and Xi <sup>p</sup> in clonal isolates derived from multiple females.....	68
Figure 10 PolIII peaks per chromosome .....	85
Figure 11 Distribution of the log-ratio of PolIII ChIP-signal versus background signal (horizontal axis) for all PolIII binding sites identified with QuEST .....	86
Figure 12 Allele-specific PolIII occupancy on the X chromosome in GM12878 .....	89
Figure 13 PolIII occupancy on Xi and Xa at the <i>XIST</i> gene .....	90
Figure 14 PolIII occupancy on Xi and Xa at the <i>DDX3X</i> gene .....	93
Figure 15 PolIII occupancy on Xi and Xa at the <i>GNL3L</i> gene.....	94
Figure 16 Expression of genes genomically associated with PolIII occupied heterozygous sites .....	95
Figure 17 Relationship of PolIII levels and gene expression on the Xi.....	98
Figure 18 Xi gene expression in multiple cell lines .....	102

Figure 19 Biased allele-specific expression of X-linked genes in GM12878 Xi <sup>m</sup> and Xi <sup>p</sup> cell populations .....	121
Figure 20 Biased allele-specific PolII occupancy on the X chromosome in GM12878 Xi <sup>m</sup> and Xi <sup>p</sup> cell populations .....	122

# 1. Introduction

Allelic imbalance, in which two alleles at a given locus exhibit differences in gene expression, chromatin composition and/or protein binding, is a widespread phenomenon in the human and other complex genomes. Most examples concern individual loci located more or less randomly around the genome and thus imply local and gene-specific mechanisms. However, a genomic or chromosomal basis for allelic imbalance is supported by multi-locus examples, of which the dosage compensated portion of the human X chromosome is the largest genomic interval in the mammalian genome exhibiting allelic imbalance. In this chapter, I review literature pertaining to allelic imbalance in expression, transcription factor binding, and chromatin features with a special focus on the human X chromosome.

## **1.1. *Allelic imbalance in expression***

### **1.1.1. Regulation of variation in gene expression levels**

Gene expression levels responsible for the extensive variation that exists among individuals are highly variable and genetically regulated (1-3). Some studies have focused on disruption of gene expression as pathological phenomena (4); however, normal variation in gene expression levels also exists among individuals (5). As variation in expression levels is a quantitative trait, linkage and association mapping approaches

have been employed to identify regions of DNA variants influencing levels of gene expression.

Gene expression levels can be regulated by both *cis* and *trans* regulators (5). In gene expression linkage studies, the regulators per se are not always identified; rather, the locations of sequence variation point to regulation sites in the genome that can be further scrutinized for functional elements that could bind or be produced at those sites. It is difficult to estimate the relative contribution of *cis* and *trans* variants influencing gene expression levels as they act in different ways (6). *Trans*-acting regulators are typically produced at distant loci and bind to the target, while *cis*-regulators are typically sites in the proximity of the target that modulate transcription factor binding or function. Methods that do not discriminate between alleles allow for discovery of both *cis*- and *trans*-acting regulators; however, a number of obstacles have been encountered in attempts to estimate proportion of loci influencing gene expression in *cis* and *trans* (6). Studies in yeast, flies and mice have concluded that most polymorphic regulators act in *trans* (7, 8), while the current estimate in humans is that, in normal cells the vast majority of expression phenotypes are regulated by *cis* variants and considerably smaller number by *trans* variants (5, 9-12). However, a recent study by Cheung et al. (13) used a larger sample size of human cell lines than previous studies and showed that many *trans* regulators exist in humans as well, illustrating the necessity

of future work to fully understand the proportion of *trans* and *cis* regulators in the human genome.

Differential allelic expression approaches offer a potential advantage for identifying sites of *cis*-acting elements. As *trans*-acting regulators have the same probability to bind to both alleles, any difference in expression between two alleles of a gene is more simply attributed to a *cis*-acting element(s) that is typically located in the gene's vicinity (9, 11).

Regulation of expression levels can be accomplished by several mechanisms, such as lowering the transcription rate or destabilizing the transcript, resulting in faster degradation rates (6). Typically, elucidating the mechanisms by which the level of gene expression is regulated requires a gene-by-gene approach, and such questions have been answered for only a limited number of loci. The location of a particular *cis*-acting variation site with respect to a gene gives an indication of the likely regulation mode. For example, variation upstream (14) and at the 5' end of a gene is likely to influence RNA Polymerase II (PolII) binding and the assembly of transcriptional machinery (15-18), variation in the body of a gene is more likely to influence the rate of processing (19) or alter splicing (20), and variation toward the 3' end may affect stability of the resulting RNA transcript (21, 22).

Sites of *trans*-acting regulators are more difficult to detect not only due to their distance from the regulated gene but also because they often act in concert with other

factors as a part of a complex. Even though the effect of individual *trans* factors may be small and difficult to detect, together such factors may have great influence on gene expression (5, 23).

In a recent study, Cheung et al. (13) identified *cis* and *trans* regulators for ~1,000 human genes. In this study, the authors validated the *cis* elements by showing differential allelic expression of the associated genes by whole transcriptome shotgun sequencing using next-generation sequencing technologies (RNA-seq) and validated *trans* elements by creating regulator knockdowns, resulting in changes of expression in the target genes. Furthermore, they were able to show the physical relationship of *trans* regulators with the target gene by chromosome conformation capture experiments, by which one can detect long distance DNA interactions by cross linking, isolating and subsequent sequencing of the interacting sequences. They found that a number of *trans* regulators have multiple targets and that a number of genes are regulated by multiple *trans* regulators (13).

Allele-specific expression is often tissue-specific and individual-specific (24). Most recent pilot studies have used lymphoblastoid cell lines from individuals included in the HapMap Project (25, 26) or the 1000 Genomes Project (27) to take advantage of the accumulated knowledge on these cell lines and the renewability of this resource (reviewed in (24)). However, as the groundwork is being accomplished, new studies focus on primary human cell types such as osteoblasts, non-transformed fibroblasts,

keratinocytes, human ES, adipose tissue, normal liver samples and others (24). These studies have found that more than half of all expression-based quantitative trait loci are private to specific tissues, while only up to 30% of loci exhibit allele-specific expression universally (28), with a mean allelic bias of 1.6-fold (29). Estimates in primary tissues have been somewhat lower, detecting on the order of 10-20% of genes that exhibit allelic imbalance (24).

Variation in gene expression and its genetic basis has been thus far studied largely using expression microarrays (6, 30). Recent sequencing technologies have started to enable genome-wide approaches in a more precise manner (31). In addition to RNA-seq that measures the abundance of RNA transcripts based on the number of sequencing reads from each transcribed locus (29, 32, 33), in a complementary manner, ChIP-seq (chromatin immunoprecipitation combined with whole genome shotgun sequencing) allows identification of regulators of gene expression such as transcription factors, enhancers, repressors, or differential chromatin marks, thus allowing for the detection of functional elements directly associated with regulation of gene expression levels (34, 35).

#### **1.1.1.1. *Histone modifications***

The DNA is packaged in chromatin fibers by wrapping around a histone octamer (H2A, H2B, H3 and H4) every 147 bp forming a nucleosome, the fundamental unit of

chromatin (reviewed in (36)). The tails of histones protruding from nucleosome cores can be modified in various ways. Such histone modification can act to either promote chromatin structural changes or allow for repressive or stimulating factors to bind (37-39). Of the known types of histone modifications, acetylation and phosphorylation have been associated with transcriptional activation (40), while sumoylation has been associated with transcriptional repression (41). Histone methylation and ubiquitination can function as both transcriptional activators as well as repressors depending on the modified residue(s) (40). Unlike other modifications, methylation occurs on both lysines and arginines with multiple methylated states on each residue: mono-, di-, and tri-methylation (40). There are at least 30 lysine and arginine methylation sites that have been identified thus providing an astonishing freedom for combinations in residue methylation (40, 42).

Ultimately, nucleosome positioning, histone variants and histone modifications in combination make up the basic structure of chromatin that is further folded into a more complex packaging configuration. The deposition of these features along DNA may be dependent on the underlying sequence and CpG methylation patterns and in concert with transcription factors and other DNA binding proteins regulates transcriptional activity of genes (43).

Pioneering studies (44-48) have associated particular histone modifications with particular genomic features and with transcriptional repression and activation in the

genome. For example, when promoters containing high CpG content are compared to low CpG-content promoters, the associated histone profiles are clearly distinct indicating the influence of DNA sequence on chromatin deposition (48, 49). Such basic partition of promoter classes can be refined by incorporating other characteristics such as DNA motifs or methylation at CpGs to achieve greater understanding of chromatin patterns and their biological meaning. Ultimately, in combination with detecting the presence or absence of specific transcription factors and the resulting transcriptional activity, the precise function of each modification can be uncovered (36).

Allele-specific analyses of histone modifications is currently limited those associated with X-chromosome inactivation (50-52) and genomic imprinting (53). Large-scale studies of histone modifications based on chromatin immunoprecipitation and massive sequencing techniques, however, have instantly shifted our focus from gene-by-gene approaches to whole genome chromatin profiling. The vast genome-wide data for various histone modifications and other chromatin features that are currently being generated, when analyzed in an allele-specific manner, will greatly enhance our understanding of the extent of allelic imbalance in histone modifications.

#### **1.1.1.2. DNA methylation**

DNA methylation at the 5-position of cytosine residues is an essential epigenetic modification associated with and required for regulation of genes (54). Cytosine

methylation occurs most frequently at CpG dinucleotides and its presence is most commonly associated with transcriptional repression.

A number of studies showed that DNA methylation is associated with silencing of imprinted loci (reviewed in (55, 56)). Although imprinting appears to be limited to <1% of mammalian genes (57), non-imprinting differential methylation is more widespread across the genome (58-60). From a number of subsequent studies it has been concluded that most allele-specific methylation sites affect DNA sequences outside CpG islands and that the extent can vary from highly localized sites to long stretches of more sparsely distributed CpGs exhibiting allelic imbalance (24, 58). For the most part, differences in CpG methylation across the genome are subtle, where <8% of allele-specific methylation events exhibit >95% methylation differences and in most cases allele-specific methylation is tightly linked to allele-specific expression of nearby genes (61, 62). Furthermore, differential methylation is genotype dependent and transmitted through families but rarely (<10% of sites) parent of origin dependent (63).

On the human X chromosome, when allele-specific methylation is compared on the two inactivated X chromosomes (maternal origin,  $X_i^m$ , or of paternal origin,  $X_i^p$ ) in a single female, 60% of the detected differentially methylated sites are X inactivation-specific as methylation switches alleles depending on which X chromosome ( $X_i^m$  and  $X_i^p$ ) is inactivated. Surprisingly however, a number of X-specific differential methylation

events are allele-specific rather than resulting from inactivation of a particular X chromosome (63).

### **1.1.2. X chromosome inactivation**

#### **1.1.2.1. Dosage compensation**

In X chromosome inactivation the silencing of one allele is believed to compensate for the amount of gene products originating from unequal gene copy numbers (64). Sex chromosomes present such a gene copy number imbalance between males and females as well as between genes that reside on the sex chromosomes and autosomes.

Generally, organisms are very sensitive to chromosomal monosomies. For example, in *Drosophila*, the upper limit for viability is about 3% of haploidy in the genome (65). In humans, loss of a chromosome or loss of a significant chromosomal segment typically results in spontaneous abortion (66). Turner syndrome (45,X) is the only well-defined disorder associated with a complete loss of a chromosome found in live births. This is most likely because only one fully active X chromosome seems to be required in any individual. Nevertheless, more than 99% of recognized 45,X pregnancies result in spontaneous abortions (67). What allows a single active X chromosome in normal males and females to suffice and Turner syndrome patients to survive while any other monosomies are incompatible with life?

Several distinct dosage compensation mechanisms have arisen independently multiple times throughout evolution, highlighting the importance of balancing expression output in different species. Although clearly discrete, dosage compensation mechanisms in *Drosophila*, *C. elegans* and mammals all result in equal expression levels between the two sexes, but perhaps even more importantly between the sex chromosomes and autosomes (68). Male flies upregulate expression from the single X chromosome twofold to equate autosomal expression levels as well as the female transcriptome (69-71). In worms the expression levels of X-linked genes in both sexes are comparable to autosomal levels (72), suggesting that an X-specific upregulation system must exist from the single male X chromosome (72, 73).

It has been proposed that X chromosome inactivation is an adaptation to the decay of an ancestral Y homologue (74). According to this hypothesis, inactivation of each gene or a gene cluster on the X chromosome has evolved independently in a gradual process and in concert with the acquirement of a dosage compensation mechanism that result in balanced expression levels between the sex chromosomes and autosomes (74). A study that compared multiple mammalian species showed a two-fold expression increase from the single functional X chromosome relative to autosomes in both males and females, presumably a means of matching autosomal expression (72). Thus, it appears that upregulation of gene expression in the heterogametic sex is the common process among worms, flies, and mammals; however, they each have

developed a unique strategy to resolve the consequent overexpression from the two X chromosomes in the homogametic sex.

The mammalian dosage compensation mechanism is the only one known so far that treats the two X chromosomes in the homogametic sex differently. To achieve a balance, one of the X chromosomes in females is rendered inactive (Xi) while the other one remains active (Xa) (75, 76). As a result, both sexes are essentially functionally haploid for the entire X chromosome. Nonetheless, not all genes on the Xi are silenced due to X inactivation (77-82). The genes that are expressed from the Xi are said to “escape” X inactivation. Although there are exceptions, on average, the expression output of biallelically-expressed X-linked genes relative to autosomes is no higher than that of monoallelically expressed genes (72). One explanation for this observation may be that perhaps the expression of particular genes from the Xi has persisted to supplement the not yet sufficiently upregulated Xa homologue.

While compensation for the disparities of genetic material between the homogametic and heterogametic sexes seems more apparent, conflict in levels of gene expression within a single individual may present more immediate biological problem and drive the evolution of dosage compensation mechanisms (64). The varying expression levels of escaping genes and the fact that the total combined expression output from Xi and Xa is on average equivalent to the expression of autosomal genes supports the notion that dosage compensation has occurred gradually and one gene or

gene cluster at a time (64). Nonetheless, it still remains to be uncovered how X inactivation arose, how stable the established features of the Xi are, how certain genes are able to override the silencing process, and what is the significance of biallelic expression from the female X.

A recent study in humans suggested that no dosage compensation exists in the human transcriptome (83). In this study RNA-seq was used rather than expression microarrays to address this question (84). By aligning the obtained RNA-seq reads to the genome, this study detected on average lower expression from the X chromosome than the autosomes in both males and females. The caveat in averaging expressing levels across the entire chromosome lies in the possibility that the X chromosome might have accrued genes during evolution that have relatively low expression levels and that the observed difference is simply due to the presence of lower expressing genes on the X as compared to the autosomes. In this thesis our PolII data indicate that there is no reproducible difference in the level of PolII binding to the X and autosomes (see Chapter 3). However, more analysis is required to account for epitope availability and other potential biases. An approach that compares total expression product for individual genes relative to autosomal genes might be more appropriate to address this controversy (see Chapter 4, future directions).

### **1.1.2.2. Establishment, stability and reversal of the inactive state**

The X-inactivation patterns observed in female-derived human cell lines or tissue samples are the net result of inactivation spreading, establishment and maintenance. Generally, it is difficult to distinguish which of these three stages of X inactivation is responsible for the observed variation in inactivation (77). Elegant studies in mouse embryonic stem cells and in the early stages of mouse development have elucidated a great deal about how X inactivation is initiated, how it spreads along the chromosome and how individual Xi components are established. However, it is unclear whether same principles apply in humans.

One of the key players in the mechanism of X inactivation is the *XIST/Xist* gene (85-87) that produces a large noncoding RNA and is the only gene known that is expressed exclusively from the Xi (86-88). The *XIST/Xist* gene is located within the X-inactivation center region (reviewed in (89, 90)), a genomic interval necessary and sufficient to cause X inactivation (91-94). In the mouse, *Xist* is transcribed from both X chromosomes in undifferentiated female cells (95, 96) as well as from the single X chromosome in male cells (97); however, it is quickly degraded and detected only in low levels around the *Xist* locus (95, 96). The X-inactivation process is triggered by increased expression of *Xist* on one of the X chromosomes one to two days after differentiation (98), generating an RNA product that coats the future Xi in *cis* (95, 96).

In the mouse embryo, the initiation of silencing induced by *Xist* expression is restricted to a specific time interval at the onset of differentiation. The potential for *Xist* to induce silencing diminishes gradually during embryogenesis (99). Initially, imprinted  $X^p$  (paternal) inactivation occurs after the onset of differentiation, when the embryo consists of only a few cells (100). It has been suggested that  $X^p$  is already pre-inactivated in the male germ line and undergoes more global inactivation during early embryogenesis (101). At the 50-100 cell blastocyst stage, just prior to implantation, the  $X^p$  is reactivated in cells of the inner cell mass (ICM) (100), which is destined to become the embryo proper. In cells that will become the extraembryonic tissues (placenta and yolk sac),  $X^p$  remains silent (100). In this brief period with both X chromosomes active in the ICM, one of the two  $X^p$  (paternal) or  $X^m$  (maternal) chromosomes becomes inactivated in random in another round of inactivation (98). The X chromosome that is chosen for inactivation at this time will remain silent throughout all subsequent cell generations (76, 102).

Current data indicate that different regions of the murine *Xist* RNA are responsible for gene silencing and spreading along the X chromosome. Silencing has been attributed to a conserved repeat sequence at the 5' end, while coating is mediated by sequences scattered throughout the rest of the molecule (103). *XIST* RNA in both mice and humans coats only certain chromatin regions and localizes in distinct bands on

the Xi that appear to correlate with gene-rich G-light bands; it is not found at the pseudoautosomal regions (PAR) or at constitutive (centric) heterochromatin (50, 104).

Once established along the chromosome, *Xist* promotes the recruitment and action of Polycomb repressive complexes PRC1 and PRC2 and remains switched on while chromatin changes occur in *cis* (105). At least in mice, some of the early heterochromatin changes include histone modifications such as H3K27 methylation, H2A monoubiquitination, H3K9 deacetylation, and H3K4 demethylation (106-109). Since the H3K9me3 modification is only enriched on human and not murine Xi chromosomes, it remains to be determined when this modification is established.

Later events include global chromatin changes such as histone deacetylation, DNA methylation at cytosine residues and accumulation of the histone variant macroH2A (110). Histone deacetylation occurs 4-6 days after differentiation and it is believed to be involved in maintenance and/or stabilization of the inactive state, rather than initiation (111). DNA methylation is established at gene promoters on the Xi 14-21 days after differentiation (111). DNA methylation is also thought to be involved in maintenance of the inactive state rather than initiation and spreading (112). Methylase deficient mice are able to initiate and establish Xi (113, 114); however, genes on the hypomethylated Xi in these mutants are more easily reactivated than wild-type Xi (113). Overall CpG methylation on the Xi is not significantly higher than the rest of the genome, although specific CpG islands at the promoters of silenced genes are

hypermethylated relative to the Xa (reviewed in (115)); however, because DNA methylation is enriched in gene bodies on the Xa relative to the Xi, its role in X-chromosome inactivation is not completely clear (116).

Attempts to reverse X inactivation experimentally have been consistently unsuccessful. In humans, removal of *XIST* from an established Xi does not result in X reactivation (117). Conditional knockout studies in mice have demonstrated that elimination of *Xist* results in displacement of macroH2A (118) and loss of histone modifications catalyzed by the Polycomb-group protein complex (H3K27 methylation and H2A ubiquitination) (119). Only sporadic reactivation of individual genes has been achieved by inhibiting DNA methylation (120) or histone deacetylation (118). Combination of these treatments resulted in further reactivation of individual genes, but no dramatic chromosome-wide changes (119).

In spite of such a stable inactivation state in somatic cells, reactivation commonly occurs during normal development. Oocytes reverse the inactivation process such that they have two active Xs through meiosis and the single X in the mature, haploid ovum is also active (121-125). In addition, as mentioned above, reactivation occurs after imprinted inactivation in the ICM in the early stages after fertilization (100). Considering the complexity of the aspects required for establishment and maintenance of the Xi, it is astonishing that complete reactivation occurs in such short time windows.

Even though reactivation has been documented, there is no evidence as to how this inherently large-scale process takes place.

### **1.1.2.3. *Genomic context, evolution and expression profile of the human X chromosome***

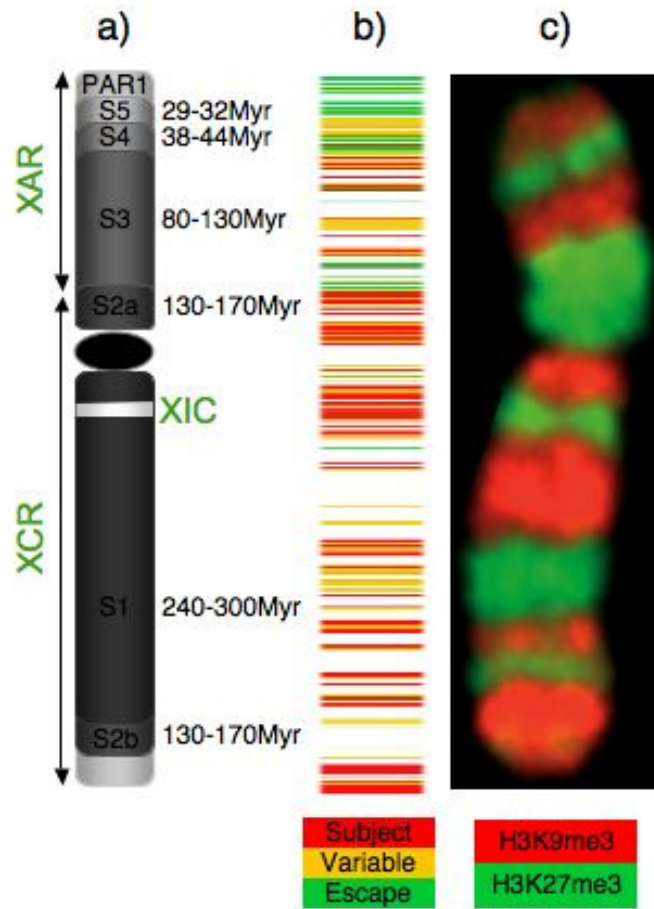
The human X is comprised of approximately 155Mb of DNA, has a low G+C content and is relatively gene poor with only 948 (RefSeq Unique names) annotated genes of below average gene length (126). In contrast, the content of interspersed repetitive elements is above the genome average (126). The most striking enrichment is for type 1 long interspersed nuclear elements (L1) that have been proposed to aid the propagation of X inactivation along the chromosome (127, 128).

The sex chromosomes have many unique features that can be contributed to the evolutionary process(es) they have undergone. The mammalian X and Y chromosomes have evolved from a pair of autosomes in the past 300 million years (Figure 1). The X conserved region (XCR) that encompasses most of the long arm (Xq) and a small portion of the short arm (Xp) is thought to have arisen from the ancestral autosome and shows significant homology to chicken chromosome 4 (129). The bulk of the Xp arm has been added to the proto-X chromosome by a translocation from another autosome (126, 129, 130) and is thus termed the X added region (XAR). The tips of the X chromosome have maintained their homology to the Y chromosome and are utilized for chromosome

pairing during cell division and recombination in meiosis. Both homologues of genes residing in these regions have remained active similarly to autosomal genes; thus these genomic intervals have been termed pseudoautosomal regions (PARs). The combined XCR and XAR, i.e. non-pseudoautosomal portion of the chromosome, can be divided into five evolutionary strata that show decreasing levels of sequence divergence radiating from the X-inactivation center (XIC) toward the end of the Xp arm (Figure 1a) (126, 130). The formation of evolutionary strata is believed to have been caused by suppression of recombination as a result of Y chromosome degradation and consequent stepwise reduction of the PARs shared between the X and Y chromosomes (126).

It has been known for decades that not all genes on the human and mouse Xi are subject to inactivation (reviewed in (131, 132)). In humans, genes in the PARs that have a functional homologue on the Y chromosome were predicted to escape inactivation early on (133). Since then, however, many genes that have no functional homologues on the Y (and thus are expected to be subject to dosage compensation) have been found to escape the inactivation process. While this is an area of ongoing work, the current assessment is that in humans up to 50% of genes may exhibit the capability of being expressed biallelically in all or some females ((77) and this thesis).

The underlying causes of escape from inactivation and their relevance to dosage compensation have not yet been elucidated; however, it is clear that the distribution of biallelically-expressed genes along the human X chromosome is not uniform. Genes that



**Figure 1 Epigenetic domains of the human Xi**

The human X chromosome is composed of the X conserved region (XCR) and X added region (XAR). Based on the level of decreasing divergence, the human X chromosome can also be divided into five evolutionary strata S1-S5 a) (127, 131). The Xi expression profile b) (77) shows genes that escape inactivation (green), genes that are subject to inactivation (red), and genes with variable expression status between individuals (yellow). The Xi heterochromatin c) is organized in alternating domains enriched in H3K9me3 (red) and H3K27me3 (green) (134).

escape the silencing process tend to cluster in multiple locations along the X chromosome, especially on the evolutionarily younger XAR portion of the X (77, 80, 135, 136). Furthermore, there is a direct relationship between the evolutionary age of the sequences comprising each strata and the proportion of expressed genes (77).

A number of studies have been conducted to understand the propensity of some genes or genomic regions to be silenced while others remain active. Many of these studies have focused on analyzing the X chromosome sequence and the enrichment or depletion of specific elements and their association with X chromosome inactivation patterns. Although these studies agree that the X chromosome sequence composition is distinct from autosomes, for example with respect to the amount and distribution of repetitive elements (128), and that specific sequence features associated with silenced and expressed genes exist, the analyses vary in which particular elements are relevant to the inactivation process (137-140). Wang et al. in 2006 analyzed repetitive elements and short sequence features and their association with silenced and expressed genes on the human Xi (140). Using a set of as few as 12 features, their machine learning algorithm was able to correctly predict the inactivation status of more than 80% of all X-linked genes, thus showing that DNA sequence is relevant to the establishment of inactivation at individual loci (140). However, genes located at the borders of the alternating silenced and active domains (77) were often predicted incorrectly, as it was

the case with many escaping genes on the XCR, suggesting that epigenetic regulation is a major player in determining the inactivation status of genes.

Analysis of the interphase Xi appearing as a dark staining Barr body (141) and localization of genes within it detected no difference between the localization of active and inactive genes. All genes appeared to be localized on the periphery of the Barr body separate from repetitive  $\alpha$ -satellite and Cot-1 DNA that reside in the Barr body core, but no spatial distinction between genes that are subject and genes that escape inactivation was observed (142). Similarly, no direct correlation with any particular chromatin type has been uncovered; although, a number of histone modifications has been implicated in X inactivation and Barr body formation (50, 51, 143, 144).

Despite extensive efforts and identification of sequence elements that are clearly relevant to the inactivation status of genes, no particular element or a set of elements with a universal function across the X chromosome has been discovered. This observation suggests that the inactivation of genes or gene domains involve a range of regulatory elements and/or mechanisms. Each gene domain may require only a certain subset of such elements, which is consistent with the hypothesis that X-linked genes are regulated on a domain basis and that recruitment of genes or gene domains into the inactivation system occurred gradually and one at a time (64). It is also likely that only a few dosage sensitive genes have strict inactivation requirements (145, 146) and the existing domains may have been established particularly around these genes. Genes

that do not have strict silencing or expression requirements may have become a part of a particular domain via heterochromatin spreading simply because of their proximity to those genes.

#### **1.1.2.4. Heterochromatin of the inactive X chromosome**

The Xi is packed into a distinct heterochromatic structure termed the Barr body that is visible during interphase on the periphery of the mammalian nucleus (141, 143, 144). Together, *XIST* RNA, histone modifications, histone variants, non-histone proteins, and cytosine methylation form the Xi facultative heterochromatin (105). As a result, the Xi exhibits a unique combination of characteristics including transcriptional silencing, late replication in S phase, condensed appearance and nuclear localization (reviewed in (115)). In this section, I will discuss the Xi heterochromatin in terms of aspects of histones, histone modifications and DNA methylation that were introduced in earlier sections.

Some histone modifications observed in the facultative heterochromatin of the Xi are shared with modifications characteristic of constitutive heterochromatin, while others are specific to the inactive X (147). Methylation of histone H3 at lysine 9 has been associated with heterochromatin and gene silencing in yeast (148), plants (149), worms (150), flies (151) and mammals (152). In humans, H3K9 methylation marks transcriptionally silent genes and centric heterochromatin, as well as the Xi (153).

Markers of transcriptionally active loci such as acetylated isoforms of all four core histones (H2A, H2B, H3 and H4), H3K4me2 and H3K4me3 are depleted in both constitutive (centric) and facultative heterochromatin of the Xi (107, 154, 155). On the other hand, H3K27me3 (134), H4K20me (156), and H2AK119ub are enriched on the Xi, but not in constitutive heterochromatin (52, 157).

H3K9me3 along with H3K27me3 are enriched at distinct and cytologically non-overlapping regions across the human Xi (Figure 1) (50, 51). The H3K9me3 domains are also enriched in heterochromatin protein HP1 (which is known to bind to H3K9me) and H4K20me3. The H3K27me3 human Xi regions are also enriched in *XIST* RNA and the histone variant macroH2A (50, 158). Several sites of the Xi show enrichment of both H3K9me3 and macroH2A relative to the Xa, suggesting some overlap between the two heterochromatin types that had not been observed on a cytological scale (159). Fine resolution analysis revealed that in humans H3K9me3 heterochromatic domains are virtually free of H3K27me3 marks. In contrast a few sites enriched for H3K9me3 are found within the H3K27me3 domains (159), suggesting that the H3K27me3 heterochromatin may be permissive to H3K9me3 while H3K9me3 heterochromatin is rather impenetrable. The more open H3K27me3 chromatin conformation may also be more permissive to other DNA proteins, such as CCCTC-binding factor (CTCF) or PolIII.

Surprisingly, H3K9me3 is not enriched on the mouse Xi while H3K27me3 exhibits a banded pattern similar to that described for the human Xi chromosomes (50, 51, 159,

160). It has been hypothesized that another heterochromatin type exists that is deposited on the mouse Xi in alternating bands with H3K27me3 heterochromatin. These differences indicate that even though X chromosome inactivation occurs in both mice and humans, it may be accomplished and/or maintained in different ways. It has not been determined for either species what purpose any of the identified chromatin types serve.

CTCF acts as a chromatin barrier element by preventing chromatin spreading across the genome (161, 162) and has been implicated as a candidate barrier element at transition zones separating the Xi heterochromatin domains (163, 164) and domains of genes that escape inactivation. The transcription repressor protein YY1 that is involved in activation and repression of a number of promoters (165) associates with CTCF in X chromosome inactivation (166). CTCF also plays role in blocking interactions between enhancers and promoters whereby repressing gene expression and is also involved in regulation of imprinted domains (167, 168). The insulin-like growth factor 2 (IGF2) imprinted gene is controlled by CTCF binding to the imprinting control region of the long non-coding RNA *H19* (169). The involvement of CTCF in regulation of imprinting via long non-coding RNAs indicates another parallel that may exist between X inactivation and imprinting.

### **1.1.3. Genomic imprinting**

In mammals, genomic imprinting refers to an epigenetic marking of genes that results in monoallelic, parent-of-origin specific expression and plays critical roles in embryonic growth and behavioral development (170). Because of their functional haploid state, imprinted genes are more vulnerable to inactivation, overexpression or damage. Much controversy exists over why imprinting evolved and was maintained since the misregulation of imprinted genes can be disastrous.

Since the identification of the first human imprinted gene *H19* in 1992 (171), less than a hundred imprinted genes have been characterized in humans (172). While some bioinformatics studies have predicted thousands of imprinted genes (173-176), the low proportion of imprinted gene candidates that have been experimentally validated (177) suggests that these predictions are largely inflated.

It appears that imprinting occurs more frequently in specific tissues. In a recent genome-wide study ~5% of the assessed genes and noncoding RNAs exhibited imprinted allelic bias in the mouse brain. Most of the identified genes were detected in specific regions within the mouse brain (178) and a number of those genes exhibited a parent of origin allelic bias rather than strict monoallelic expression.

Reminiscent of what has been described for X inactivation on a chromosome wide scale, imprinted genes also commonly occur in clusters and their expression is often co-regulated by an imprinting center (reviewed in (179)) via long noncoding RNAs

(180, 181), differentially methylated regions (reviewed in (24), and allele-specific histone modifications (182, 183). The similarities indicate that the evolutionary processes that drove the evolution of imprinted genes and X inactivation may have been ruled by similar epigenetic and genomic principles (184).

A number of genes that exhibit sex-specific imprinting features have been described. Imprinting of X-linked genes has been suspected because of the differences in anatomical, physiological and behavioral phenotypes in Turner syndrome patients dependent on the parental origin of the single X chromosome (185). A cluster of at least three X-linked paternally imprinted genes (*Xlr3b*, *Xlr4b* and *Xlr4c*) has been identified in the developing brain of a Turner syndrome mouse model (186, 187). An independent locus on the mouse X chromosome, *Rhox5/Pem*, is maternally imprinted in the mouse preimplantation female blastocyst (188) (189). Genomic imprinting of these genes is independent of X-chromosome inactivation because the imprinted alleles maintain their expression regardless of the parental origin of the Xi. In a recent study Gregg *et al.* used a genome-wide RNA-seq based approach to explore sex-specific imprinting effects in various parts of the adult mouse brains (190). This study identified 347 candidate genes that exhibited sex-specific imprinting effect where allelic imbalance was observed in one sex and not the other. Three times more sex-specific imprinted features were found in some regions of the female than in male brain, correlating well with previous observations of high sexual dimorphism (191). Significantly higher gene expression was

detected from the maternally derived X chromosome indicating non-random X inactivation or selection for cells containing the maternally derived active X chromosome. Several further studies in mice have also suggested that X inactivation in other cell lineages favors the maternally derived X chromosome (190, 192-194).

The evolution and effect of such complex parent of origin regulatory designs, some of which are implemented in a sex-specific manner and have to be reconciled with X-chromosome inactivation, remain elusive. Future studies dissecting allele-specific expression and chromatin characteristics may explain the mechanisms of action and effects of these genes and their significance in the development and function of the brain and other tissues.

#### **1.1.4. Allelic exclusion**

Allelic exclusion is a type of monoallelic expression in which one of the two alleles of a gene is selected in random to be silenced. As described above, X-chromosome inactivation in female mammals is an example of allelic exclusion in which an entire chromosome is selected at random for silencing; however, a number of examples of allelic exclusion exist on autosomes as well. Several classes of genes such as those encoding the immunoglobulins, T cell receptors, olfactory receptors, interleukins and natural killer cell receptors exhibit allelic exclusion to ensure that only a single type of receptor is displayed on the surface of each cell (195). New technologies have

enabled whole genome searches for monoallelically-expressed genes. A study by Gimelbrant et al. designed to detect allelic exclusion across the entire human genome has reported that more than 5% of the assessed genes exhibited allelic exclusion (196). While stable within each clonal cell line, the majority of the discovered monoallelically-expressed genes were expressed biallelically in some of the tested clones, indicating that strict monoallelic expression in all clones (as seen in immunoglobulin or odorant receptor genes) is rare. As reported previously (77) and extended in this thesis as well, variability in expression states among individuals and even within an individual on the human X chromosome is also common.

Unlike on the X chromosome, where choice of the entire chromosome dictates whether the maternal or paternal copy of each gene will be expressed or silenced, for autosomal genes this decision appears to be made independently for each gene. Similarly to X-chromosome inactivation, however, asynchronous replication and clonal propagation resulting in a mosaic phenotype with patches of tissues with an identical expression profile are both features of autosomal genes exhibiting allelic exclusion (196, 197), indicating that, epigenetically, regulation of expression on the X chromosome and randomly inactivated autosomal genes may be very similar. Variably expressed genes show lower absolute expression levels in cell lines where they were expressed monoallelically (196), indicating that monoallelic vs. biallelic expression may be a means

for regulating gene product dosage, resulting in human diversity in the population and among clonal tissue patches within an individual.

A disproportionate number of monoallelically-expressed genes encode cell surface proteins (196), suggesting specificity in recognition of cells by other cells or extracellular factors. Furthermore, monoallelically-expressed genes are twice as likely to be located near putative regulatory elements (conserved noncoding sequences) with recent human-specific accelerated evolution frequently found near genes involved in neuronal cell adhesion (198). Taken together, monoallelic expression could be a feature in specialized cell-cell interactions unique to brain development in the human lineage.

## *1.2. Allelic imbalance in transcriptional machinery and other binding proteins*

The ENCODE Project (199) has set to explore the human genome beyond its DNA sequence and define a broad spectrum of chromatin features and their differences among individuals and among cell types. As a part of ENCODE a number of studies have reported on genome-wide distribution of various DNA binding proteins.

The coding regions of genes make up only a small portion of the human genome. Genetic variation that underlies phenotypic diversity in the population extends beyond the transcribed variants to the non-coding regions of the genome that are relevant in

regulation of DNA protein binding. A number of assays, such as DNaseI hypersensitivity (DHS) (200), formaldehyde-assisted isolation of regulatory elements (FAIRE) (201) and chromatin immunoprecipitation (ChIP) (202) have become standard in the analysis of chromatin states and identification of regulatory elements.

The DHS assay detects regions in the genome where nucleosomes have been displaced and the underlying DNA is exposed in order for DNA binding proteins and transcription factors to gain access. At such regions, DNA is 'open' and more sensitive to being digested by DNaseI; therefore, if chromatin is exposed to low DNaseI concentrations, only the 'open' chromatin will be digested. By this approach one can detect the spatial properties of protein binding to various DNA elements such as promoters, enhancers, insulators and other regions; however, the identity of these specific binding proteins cannot be identified by this method alone.

The FAIRE genome-wide approach separates DNA stretches bound by nucleosomes away from regions bound by other proteins such as transcription factors by formaldehyde chemistry. It too serves to enrich for and detect DNA sequences where transcription factors bind; however, it is technologically an independent and thus complementary approach to the DNase hypersensitivity assay.

The ChIP technology targets specific DNA binding proteins, transcription factors and histone modifications via antibody binding and consequent pull-down revealing their precise localization in the genome and thus provides more specific information

about localization of each particular protein. Nonetheless, only one type of binding factor can be targeted in each experiment, requiring a number of assays in order to cover the plethora of factors known to bind DNA in each sample.

All of these methods have now been combined with next-generation sequencing approaches to produce whole-genome short sequence libraries that can be precisely mapped onto a reference genome to detect enrichment of specific sequences at specific locations. The sequence information that these technologies provide can also be utilized to distinguish between the two alleles to which the factors in question bind at each polymorphic site and identify differences in binding that various polymorphisms underlie.

Previously, interactions of proteins have been mapped by ChIP product hybridization to single nucleotide (SNP) genotyping arrays (203, 204). Such approaches are limited to the specific SNPs selected for the microarray, leaving ChIP-seq methodologies vastly better suited for addressing questions of an allele-specific nature. A similar limitation stems from the lack of genotype information available for some cell lines. Many SNPs are not available through the HapMap Project, due to very low frequency in the population or because they are private to specific individuals; however, a number of resequencing efforts that can capture SNPs in a given individual have greatly enhanced SNP information for a number of cell lines (205).

Three recent studies that utilized resequencing and whole-genome chromatin studies have examined protein binding in human cells in an allele-specific manner. McDaniel *et al.* (34) studied open chromatin by DHS-seq (206, 207) and CTCF binding by ChIP-seq (208) in two mother-father-child trios from geographically and ethnically diverse ancestries in the effort to estimate the extent of influence of variation in chromatin structure and transcription factor binding on gene expression and consequent phenotype diversity. In the experimental design, we were able to compare four unrelated individuals (parents of the two trios) to the related the individuals (parent-child pairs). In addition, we also compared the two homologous chromosomes within each individual by the means of allele-specific variation. When both CTCF binding and DHS were possible to measure at one site, there was a strong preferential tendency to bind the same allele in both assays (34). The study found that ~10% of the active chromatin sites exhibited allele-specific differences and ~10% were individual-specific. A small proportion of the individual-specific sites were unique to one or the other family trio, while most (~80%) were present in singletons (only one individual) or multiple individuals regardless of origin.

DHS sites near the transcription start site of genes showed clear positive correlation with gene expression, while CTCF was associated with both active and silenced genes, indicating more complex relationships of CTCF binding and gene expression (34). On the X chromosome, most allele-specific CTCF sites showed a bias

towards Xa alleles (34), indicating greater association of CTCF with active chromatin than with epigenetic silencing. However, several sites with equal binding to both X chromosomes or even sites exhibiting bias towards Xi were also found (34) indicating potential involvement of CTCF in the inactive state of the Xi.

When allele-specific biases in the heterozygous daughters were related to parents homozygous for the opposite alleles, the relative strength of binding at the parental alleles correlated well with the relative strength of binding at the two alleles in the child, such that the more strongly bound allele in the parent was also bound more strongly in the heterozygous child (34). Furthermore, when heterozygous biases at specific alleles were compared in multiple individuals, the bias was nearly in all cases observed in the same direction. And finally, polymorphisms most strongly associated with allelic bias in CTCF binding corresponded to substitutions at highly conserved nucleotide positions in the CTCF-binding motif (34). These findings suggest that allele-specific and individual-specific chromatin states have a genetic basis and are transmitted between generations and that the establishment of chromatin is dependent on the underlying sequences.

A study by Kasowski et al. (35), published concurrently with that of McDaniel et al. (34), examined binding of PolIII and a nuclear factor KB (p65) (NFKB) in ten lymphoblastoid cell lines and one chimpanzee cell line. In this study, similar to McDaniel et al. (34), 25% of PolIII sites and 7.5% of KB sites exhibited allelic imbalance. The

majority of binding regions were occupied in at least two individuals and the number of SNPs located in the binding regions for PolIII and NFκB corresponded to the frequency of allelic differences in binding. When compared with deep RNA-seq, significant correlation was observed between binding biases and mRNA abundance for both factors, indicating a strong effect of chromatin differences on gene expression (35). There were, however, many sites that did not exhibit expression differences despite the observed allelic bias in transcription factor binding, suggesting that these sites may be under different modes of regulation, for some of which a feedback mechanism might be involved to reach the required level of transcripts. As in the previous study, very few unique sites exhibiting allelic imbalance were specific to particular human populations, while extensive divergence was detected between humans and a chimpanzee (35).

The third study, by Reddy et al. (209), examined allelic bias in binding of 24 transcription factors in a single lymphoblastoid cell line. Consistent with the two previous studies, ~8% of sites showed significant biases in binding between alleles and were highly reproducible between replicates. In this study we showed that allelic variation located directly in the transcription factor binding sites explains ~20% of variation in binding differences to the two alleles and that allelic variation near the location of the most intense signal within binding sites has the greatest effect on binding (209). This finding was also supported by the fact that, at binding sites containing allelic variation, the bound alleles were more similar to the consensus motif

than the unbound alleles. The remaining 80% of allelic differences must, therefore, occur by other mechanisms, such as mutations in cofactor binding sites or modifications in heterochromatin elements. While many factors shared the same binding bias at specific sites, no factors had the opposite bias at any sites, indicating widespread cooperation of various transcription factors, and rare, if any antagonistic forces (209). On the X chromosome, all tested factors exhibited strong bias toward the predominantly active maternal X allele (209), consistent with the McDaniel et al. (34) report discussed above.

These studies have catalogued extensive genome-wide allele-specific chromatin information that indicates strong influence of DNA variation on transcription factor binding and chromatin conformation relevant to its accessibility to the transcription machinery. As shown in these studies, allelic imbalance in chromatin affects gene expression and likely plays an important role in human diversity. These studies are essential to understanding the ongoing processes on the Xi, as it is the largest region in the human genome exhibiting allelic imbalance.

### *1.3. Global efforts in genomics*

### **1.3.1. International HapMap Project**

In two phases, the International HapMap Project developed a haplotype map of the human genome using 270 samples from three populations, 30 parent-child trios (90 individuals) from the CEPH cell collection, 90 Yoruba individuals from Ibadan in Nigeria, and 90 Japanese and Han Chinese individuals from Tokyo and Beijing. In Phase I, more than 1 million SNPs (25) and in Phase II more than 3.1 million SNPs (26) were genotyped. The 1000 Genomes Project (below) has been a natural extension of the HapMap project.

### **1.3.2. ENCODE Project**

The ENCODE (Encyclopedia of DNA Elements) Project (210) was launched in 2003 as an international public research effort aimed to catalogue all functional elements within the human DNA sequence. The pilot and technology development phase focused on a 30Mb region representing 1% of the human genome (199). The involved laboratories tested various available approaches and technologies to discover functional regulatory elements within this region that could be later scaled to the whole genome (199). The production phase that aims to evaluate the entire genome and include new pilot studies was initiated in 2007 (210). Ten research groups from the United States and Great Britain have been selected to map functional elements by identifying spatial and temporal characteristics of open chromatin, histone modifications, transcription factor binding sites, DNA methylation sites, and other genomic features

(<http://www.genome.gov/>). Seven cell types including normal and cancerous cells representing the three primary germ cell layers and humane embryonic stem cells were selected for the project. The result of these efforts will be a comprehensive genomic and chromatin profile of a number of normal and disease human cell lines that can serve as a basis for future targeted in depth studies.

### **1.3.3. 1000 Genomes Project**

The 1000 Genomes Project was launched in 2008. It is an international re-sequencing effort to catalogue human genetic variation. The goal was set to find most genetic variants with frequencies of 1% or higher in the population in at least 1000 individuals from various ethnic backgrounds using next generation sequencing technologies. In the pilot phase three approaches were taken: 197 individuals from four populations were sequenced at 2x coverage; two mother-father-child trios were sequenced at 20x coverage with the goal to obtain full assembled genomes for each of these individuals; and 697 individuals from seven populations were sequenced by exon-targeted sequencing to detect coding variation. The first phase was concluded in 2010 (27), and the plan for the full project is to sequence 2,500 samples at 4x coverage from ~30 populations.

#### **1.3.4. Cell types in studies of allelic imbalance**

Gene expression profiles and chromatin signatures vary considerably between different tissue types, primary cell lines and transformed cell lines. Primary cells from human subject that have not been experimentally manipulated could be the ideal source of for obtaining more precise picture of processes occurring in human cells. However, primary cell lines can be influenced by environmental exposures such as diet, medication, or lifestyle that are difficult to unify and have a profound effect on gene expression. Many studies therefore have turned to established human cell lines, in which environmental influences can be minimized. These cell lines are a renewable resource as they can be easily frozen and expanded, offer a high degree of homogeneity of cell types, and the environmental influences and growing conditions can be controlled with standardized culturing protocols. These key features are extremely important in obtaining comparable complementary information from the vast range of experiments that are currently being accomplished at multiple laboratories.

Transformed cells, however, that are subject to experimental manipulation (211), can be influenced by the time in culture and are not typically relevant to studying diseases unrelated to the particular cell type; thus caution has to be exercised when making inferences between the chromatin and expression states of transformed cell lines and those of primary living human cells.

The prevalent source of samples for genetics of gene expression studies is the Centre for Study of Human Polymorphism (CEPH) collection and other cell lines previously selected from the HapMap Project (25, 26). They are human Epstein-Barr-virus-transformed lymphoblastoid cell lines created by immortalization of B cells from a large number of multi generation families from various populations and indeed data collected using these cell lines have been highly concordant (reviewed in (5)).

Cell lines used in this thesis are HapMap CEPH lines. The female GM12878 lymphoblastoid cell line used in Chapters 2 and 3 of this thesis is one of the HapMap cell lines later selected for the 1000 genomes and ENCODE projects. Further, it is one of the 1000 genomes cell lines resequenced at 20X coverage, thus offering the most comprehensive genotype information and parental genotype information currently available.

#### *1.4. Thesis overview*

Profiling of gene expression on the human Xi – both to gain insight into the molecular and chromosomal basis for X-inactivation patterns and to discern their possible implications for medical genetics – has been ongoing since X inactivation was first described 50 years ago. However, the underlying genetic causes of X-inactivation

patterns, including variable expression now recognized at an increasing number of loci on the Xi, have remained largely unexplored.

Current technological capabilities to distinguish alleles between homologous chromosomes at loci along the entire length of the X have now provided us with the tools to study X chromosome inactivation of individual genes directly in human samples, extending to a chromosome-wide scale the concepts illustrated for single loci back in 1960's. These recent technological advances have also instigated the need to develop methods capable of generating appropriate samples for the study of X chromosome inactivation. In Chapter 2, I describe a method of deriving such cell samples from human lymphoblastoid cell lines as well as their direct use in addressing questions regarding genetic influences on X chromosome inactivation. As I show in Chapter 2, the human Xi appears to be more permissive to the transcriptional machinery than previously thought; thus, in Chapter 3, I assess the level of Xi chromatin permissiveness by investigating chromosome-wide PolII binding to the Xi and its relationship to gene expression from the Xi. Finally, in Chapter 4, I propose how the findings presented here might be utilized to answer further important biological questions and how the approaches applied in this thesis can be scaled to a population study using current high-throughput technologies to further investigate the questions raised here.

## 2. Genetics of Variable Expression Patterns on the Human Inactive X Chromosome

Katerina S. Kucera, Dongliang Ge, Jenae E. Logan & Huntington F. Willard

### Collaborators:

*Dongliang Ge* – Center for Human Genome Variation, Duke University – wrote a script to assess the occurrence of expressed heterozygous SNPs in HapMap families

*Jenae Logan* – IGSP, Duke University – undergraduate student, performed assays of *TRAPPC2* and *TIMP1* expression in unrelated females as a part of independent study under my mentorship

In addition to the ~ 15% of genes on the human Xi that consistently escape inactivation, current data indicate that a further 40-50% genes exhibit variable inactivation patterns and are expressed biallelically at various frequencies in the population. In this chapter, in order to test whether variable inactivation of specific genes has a genetic basis, we have selected four novel and seven previously identified genes with variable expression patterns and tested their expression in female lymphoblastoid cell lines derived from a large three-generation family as well as six unrelated females. For the genes examined in this study, the data indicate that their inactivation status as well as relative levels of expression vary widely among individuals and are more easily explained by a stochastic or multi-gene model, rather than by a model of strict inheritance. Furthermore, we found that variation in expression can occur in the population and within an individual as well. For example, while expression of *SEPT6* is extremely stable among clones derived from a single cell line, yet is variable in the population, expression of *TRAPPC2* is variable even among multiple single cell line isolates. Our study suggests that although X chromosome inactivation is a chromosome-wide phenomenon, fine-tuning occurs locally and likely on a gene-by-gene basis.

## 2.1. *Introduction*

### 2.1.1. **Genes on the human inactive X chromosome that exhibit variable expression status**

As discussed in Chapter 1, X chromosome inactivation is established via expression and spreading of the non-coding *XIST* RNA from the X-inactivation center and the consequent deposition of numerous epigenetic elements such as CpG methylation, histone modifications and histone variants along the X chromosome. The heterogeneity of chromatin patterns on the human Xi (50) suggests that silencing along the chromosome might have regional or locus-specific components. Indeed, it has been shown that many genes that escape inactivation are localized in potentially co-regulated domains in humans (77, 212) and more or less dispersed (159) and likely regulated individually in mice. In addition to genes that are consistently expressed from the Xi, a number of genes exhibit expression only in a select portion of the female population (77, 80, 82), suggesting that the regional and/or locus-specific regulation mechanisms might have a genetic basis.

Expression heterogeneity has been observed in mice (78, 79, 159), in rodent/human somatic cell hybrids (77, 80) as well as in primary human cell lines with non-random X chromosome inactivation (77, 80). It was originally estimated that 10% of genes on the human Xi have variable inactivation status with respect to the population (77, 80); however, based on a comparison of somatic cell hybrid and human cell line

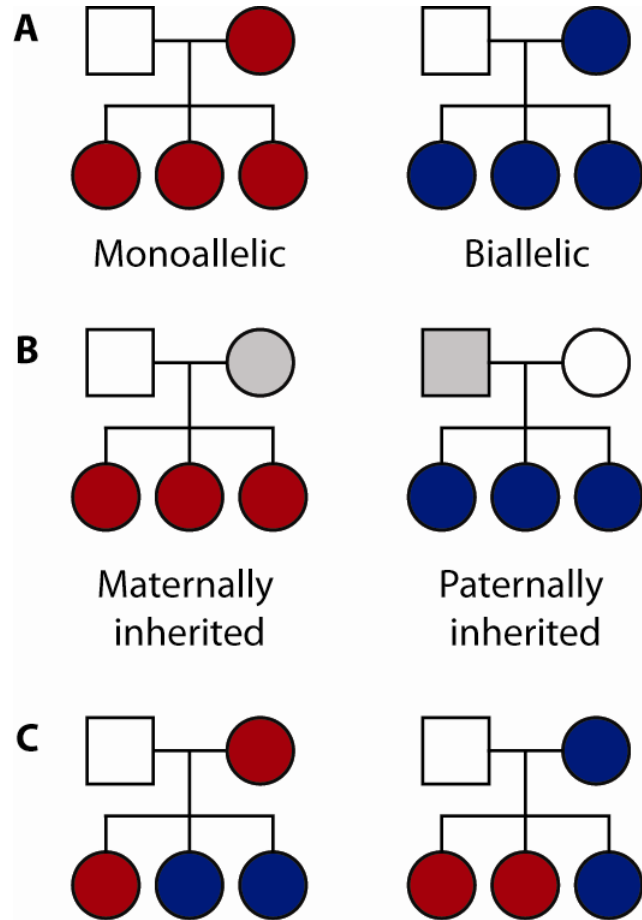
inactivation data it has since been demonstrated that the extent of Xi expression heterogeneity in the population might be even greater (77). Furthermore, a number of genes were originally classified as 'subject' to inactivation, because their expression from the Xi was below an arbitrary statistical lower cutoff set at 5% of that of the Xa allele. However, even genes that show extremely low expression must be accessible to the transcriptional machinery, which might not be the case for truly 'subject' genes that fail to show any transcription from the Xi allele and are strictly silenced in all samples at all times. If we accept this distinction, in the Carrel and Willard data set (77), up to 50% of genes exhibit some level of expression in various frequencies in the population and even more genes might be classified as 'variable' with the use of more sensitive technologies and larger sample sets.

The first two human genes demonstrating a variable inactivation patterns *TIMP1* and *REP1* were identified in 1999 (81, 82). The *TIMP1* gene is surrounded by monoallelically-expressed genes, suggesting that a regulation mechanism specific to *TIMP1* is responsible for its expression pattern (82). Single-cell expression data for the *REP1* (current name *CHM*) gene that exhibits variable patterns of inactivation indicate that, when expressed, it is expressed in reduced levels from all cells rather than being expressed from only a proportion of cells (81). Since then, many more X-linked genes with variable inactivation patterns with respect to the population have been identified (77, 80), and it remains unknown whether these genes also exhibit variable inactivation

with respect to different tissues, whether they are inactivated uniformly in the cell population, or whether they are co-regulated in clusters along the chromosome.

At least three models for variable genes exist that are amenable to study and predict different outcomes in a family (Figure 2). In a *strict inheritance model*, inactivation status is determined by the allele of a given gene and is heritable from generation to generation and among sisters who inherit that allele. A *parent of origin model*, on the other hand, predicts that inactivation status of a gene depends on the allele's parent of origin, and patterns of Xi expression within a family will be determined by whether the allele is inherited from the father or from the mother. Lastly, we consider models in which the inactivation status of variable genes is not consistent from generation to generation or dependent on the parent of origin of the Xi allele, and thus inactivation or expression of variable genes will appear to be random in the population and in a family. Here, we refer to this as a '*stochastic*' model, although formally these patterns could be explained by any number of factors, such as those that are truly stochastic, environmental, or determined by multiple unlinked genetic effects that would thus appear to not co-segregate in families.

In this chapter, we examine the expression of a number of genes on the Xi in a three-generation family in order to address the predictions of each of these three models.



**Figure 2 Models for genes with variable inactivation in the population**

There are at least three different models for variable genes that predict different outcomes in a family: (A) *Strict inheritance* – inactivation status is consistent from generation to generation and among sisters, (B) *Parent of origin* – inactivation status depends on the allele’s parent of origin, (C) *Stochastic* – inactivation status is not consistent from generation to generation. Circles designate females and squares males. Red shading indicates monoallelic expression (i.e. silencing on the Xi), blue biallelic expression (i.e. escape from inactivation on the Xi) and gray shading indicates the parental origin of the inherited allele.

## 2.2. *Results*

### 2.2.1. **Selection of cell lines**

Due to the need to distinguish alleles on the Xi from those on the Xa, expression of genes on the human Xi has previously been analyzed primarily in human/mouse somatic cell hybrid cell lines containing a single human X chromosome (77, 80, 136, 213, 214) and more recently by allele-specific expression methods in human cell lines exhibiting complete nonrandom inactivation (77). Because of the potential influence of the mouse cellular environment on the human X chromosome in human/mouse somatic cell hybrids, allele-specific expression approaches using human cell lines are preferred. Completely non-randomly inactivated cell lines required for X chromosome inactivation studies are rare, and many such cell lines carry chromosomal aberrations that result in inactivation skewing by secondary selection. Furthermore, the human fibroblast and lymphoblastoid cell lines used previously are derived from unrelated individuals and the parental genotypes are mostly not available; thus these cell lines are not ideal for experiments that explore genetic influences on inactivation status of genes.

In this thesis, I utilize cell lines from a multi-generation family and from unrelated females, specifically from the CEU (Utah residents with Northern and Western European ancestry) population, that have been genotyped by the International HapMap Project (25). These cell lines were selected for the HapMap Project from the Centre

d'Etude du Polymorphisme Humain (CEPH) collection, derived from 40 (later 61) large three-generation families (215) in the 1980's.

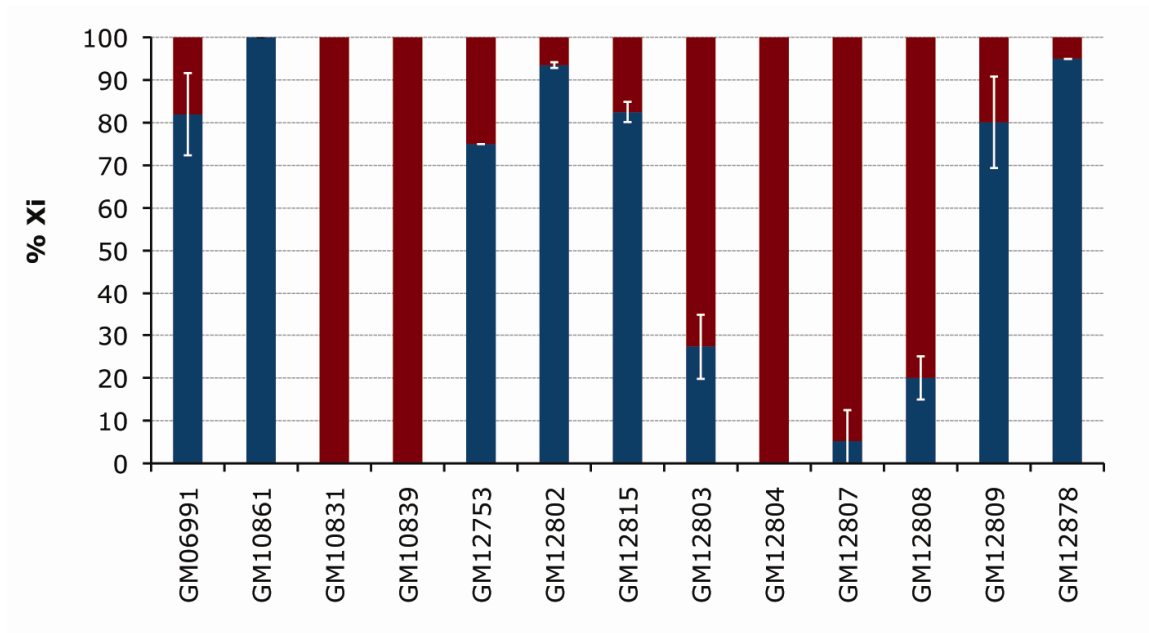
The criteria for selecting a specific family for this thesis are shown in Table 1. I chose for study all females from family 1454 because of the large number of females in the third generation in this family and a high number of genes for which the alleles can be distinguished by expressed heterozygous SNPs. In addition, I selected six unrelated females that were the mothers from families 1341, 1362, 1408, 1420, 1447 and 1463, with the potential to expand this study to include the other members of those families. Cell line GM12878 (mother in family 1463) was also included because of the vast data publically available from large consortia projects (27, 210); however, HapMap data for the father and paternal grandparents in this family are not available.

Lymphoblastoid cell lines are generally poly- or oligo-clonal (216) and are expected to contain both  $Xi^m$  and  $Xi^p$  cells due to the random nature of X chromosome inactivation. We tested each chosen cell line for X chromosome inactivation skewing (Figure 3) and found that all 13 cell lines were >75% skewed toward one Xi chromosome or the other. Six cell lines showed overrepresentation of  $Xi^m$ , six cell lines showed overrepresentation of  $Xi^p$ , and for one cell line GM12815, derived from the grandmother in family 1454, it was not possible to determine the origin of the overrepresented Xi due to unavailability of the parental information.

**Table 1 Selection of CEPH/HapMap family-derived cell lines according to heterozygosity**

CEPH families for which both parents and all four grandparents have been genotyped by the HapMap project are listed. Each cell line was interrogated for expressed heterozygous SNPs by a bioinformatic approach to select an optimal family with respect to the number of daughters and maximum number of expressed genes, for which the transcripts of the two alleles can be distinguished using these SNPs. The highest score for each SNP category is shown in red and the second highest in green. No HapMap genotype data are available for the third generation; thus numbers of heterozygous SNPs for that generation were calculated based on parental genotypes.

		Coriell Family					
		1341	1362	1408	1420	1447	1454
	# of daughters	6	7	5	6	4	5
mother	# of expressed het. SNPs	318	254	297	312	261	195
	# of genes with het. SNPs	166	154	165	177	159	159
Δ mother-father	# of expressed SNPs	116	155	172	136	136	187
	# of genes with parental differences	82	94	93	79	92	119
mother-grandmother shared	# of shared expressed het. SNPs	155	128	135	172	118	166
	# of genes with shared het. SNPs	85	76	89	102	79	96
daughters	# of expected expressed het. SNPs	275	282	321	292	267	335
	# of expected genes with het. SNPs	165	171	176	168	172	199



**Figure 3 X-inactivation skewing in HapMap cell lines in this study**

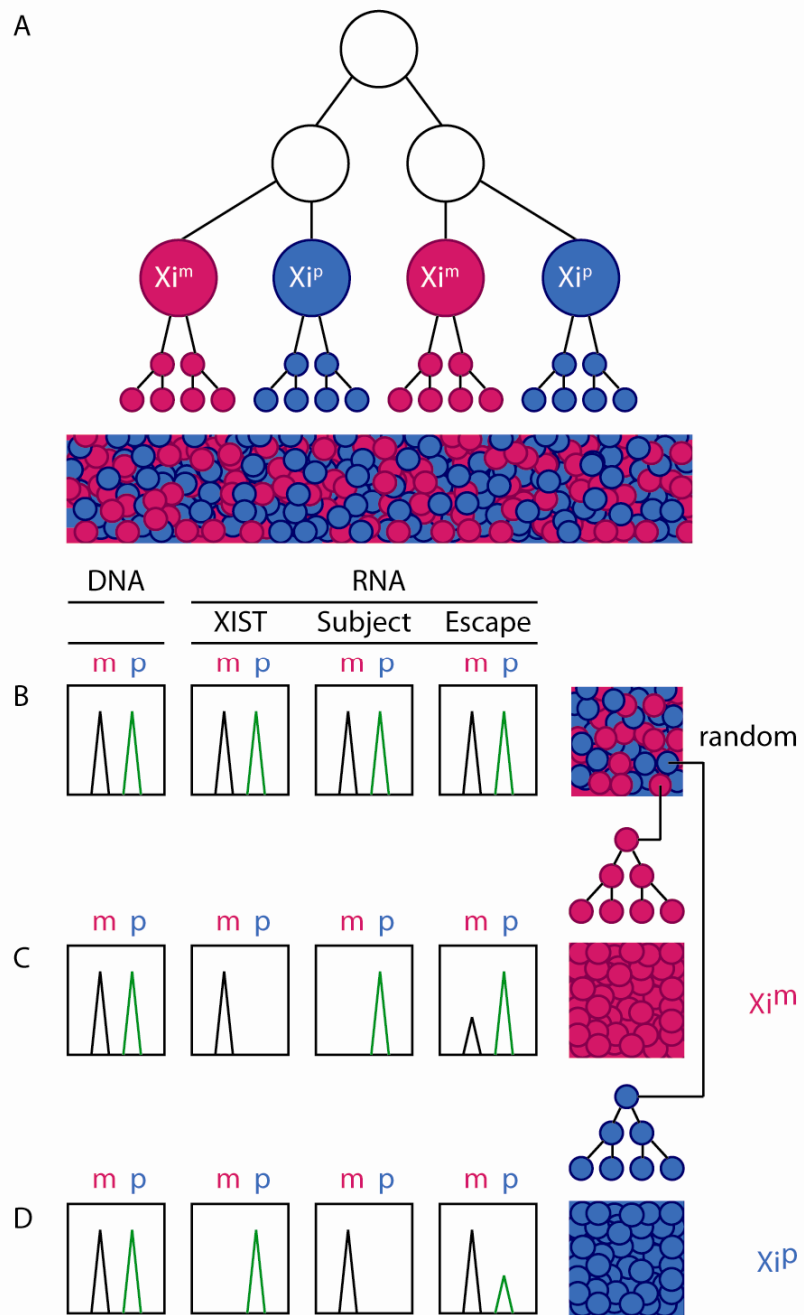
Selected cell lines from the CEU population are shown on the horizontal axis. Vertical axis represents percent of each Xi<sup>p</sup> and Xi<sup>m</sup> in each cell population as determined by allele-specific expression assayed at heterozygous SNPs of monoallelically-expressed genes, including *XIST* (expressed only from the Xi), and genes known to be subject to inactivation; *EBP* and *ATRX* (expressed only from the Xa). See methods. Red bars represent proportion of Xi<sup>m</sup> and blue bars represent proportion of Xi<sup>p</sup> cells in the sample. The level of skewing is the average of Xi<sup>p</sup> and Xi<sup>m</sup> proportions as determined at each informative gene (*XIST*, *EBP*, *ATRX*) in each cell line. Error bars represent standard deviation of the mean.

While it is statistically possible to estimate the inactivation status of genes in severely skewed cell lines, for genes on the Xi that fail to be completely silenced and are expressed in levels that are a fraction of the Xa expression (77) it is important to use pure Xi<sup>m</sup> and Xi<sup>p</sup> cell populations to distinguish low expressing genes from fully silenced genes and to accurately quantitate the level of their expression from the Xi. Furthermore, if both Xi chromosomes (Xi<sup>p</sup> and Xi<sup>m</sup>) can be isolated clonally from a single cell line, one can compare expression profiles of the two different X chromosomes inactivated within the same genomic background.

### **2.2.2. Derivation of clonal populations from female lymphoblastoid cell lines**

Because the propagation of X inactivation is clonal (76), it is possible to achieve complete nonrandom inactivation in a cell culture by expanding single cells (Figure 4). Using this approach, I isolated a number of single cell clones from the selected HapMap cell lines (Table 2) and tested them for X-inactivation skewing, as described below (Figure 5) and in (34).

As described in Chapter 1, *XIST* is the only gene expressed solely from the Xi (85). Its monoallelic expression from the Xi is required for establishment of the inactive state (95, 96) and continues thereafter (85, 86); thus the well-established expression pattern makes *XIST* the optimal marker for X chromosome inactivation skewing (216). Because even classically studied X-linked genes that are subject to inactivation can escape in



**Figure 4 Derivation of homogeneous Xi population from randomly inactivated cell lines**

(A) X chromosome inactivation occurs in early development in random, such that the resulting cell population is a mixture of cells carrying  $Xi^m$  and  $Xi^p$ . Because human lymphoblastoid cell lines are poly- or oligo-clonal, they are also mosaic with respect to the Xi chromosomes. (B) In a randomly inactivated cell population signals from both alleles at heterozygous SNPs are detected in DNA and RNA regardless of the inactivation status of genes. (C, D) In homogeneous cell populations both alleles are detected in DNA; however, expression of *XIST* is only detected from the Xi and expression of genes that are subject to inactivation is only detected from the Xa. Expression levels from the Xi at genes that escape inactivation and are expressed biallelically can be detected and quantitated relative to Xa expression.

**Table 2 Derivation of homogeneous cell lines with respect to the Xi**

Cell line	Isolates (Xi origin)
GM06991	GM06991-pat
GM10861	GM10861-pat
GM10831	GM10831-mat
GM10839	GM10839-mat
GM12753	GM12753-pat
GM12802	GM12802-pat
GM12815	GM12815-unknown
GM12803	GM12803-pat
	GM12803-mat
GM12804	GM12804-mat
GM12807	GM12807-mat
GM12808	GM12808-pat
	GM12808-mat
GM12809	GM12809-pat
	GM12809-mat
GM12878	GM12878-pat
	GM12878-mat

*mother*

*grandmother*

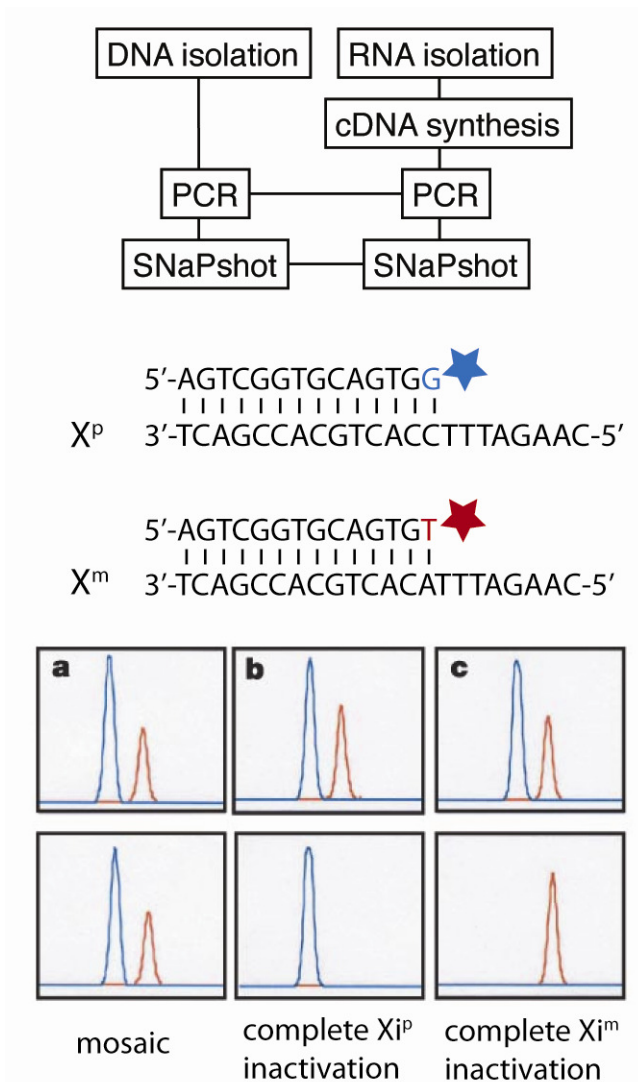
Family

some individuals (77), they are less reliable for detecting X-inactivation skewing; however, when monoallelic expression due to X inactivation is observed, it indicates both that the tested gene is subject to inactivation, as well as that the tested cell population exhibits complete nonrandom inactivation.

To assess the purity of the presumptively clonal cell populations, we tested expression of *XIST* in an allele-specific fashion by the SNaPshot assay (77) (Figure 5) at a heterozygous SNP rs1620574 and selected only those clones that exhibited monoallelic *XIST* expression. We confirmed the observed complete non-random inactivation in these isolates by assaying heterozygous SNPs in two well-established genes that are subject to inactivation *EBP* (rs3048) and *ATRX* (rs3088074) (77). Both of these genes showed monoallelic expression from the opposite X chromosome than *XIST* in the selected cell lines, as expected. The monoallelic expression of *EBP* and *ATRX* thus validates the completeness of skewing as well as the assignment of  $Xi^m$  and  $Xi^p$ . We further verified the purity of each selected isolate upon expansion by the same method.

### **2.2.3. Patterns of gene expression**

Eleven genes with variable inactivation status were included in this study; seven were identified as showing variable patterns of expression previously (77, 80, 82), and I identified four genes as a part of this thesis by screening gene expression in the cell lines



**Figure 5 SNaPshot assay at *XIST* rs1794213 to determine X-inactivation skewing in heterozygous lymphoblastoid cell lines**

In the SNaPshot assay, DNA and RNA are isolated from a cells and cDNA is synthesized from RNA. Further steps (PCR, single fluorescent nucleotide extension and ABI3100 detection) are carried out concurrently for DNA and cDNA. The RNA readout is normalized to DNA and quantitated. Complete non-random inactivation and the direction of skewing is indicated by a single peak in the cDNA sample at *XIST* rs1794213. The parental origin of the Xi is determined by comparison to parental genotypes.

in Table 2. Expression was assayed in cell lines containing heterozygous expressed SNPs in the queried genes. To test the three possible models (Figure 2), I categorized these 11 genes according to the resulting expression status in the assayed cell lines with respect to the alleles as well as parent of origin (Table 3).

For nine of the 11 variable genes I found evidence that speaks against both allelic and parent of origin mode of inheritance of the inactive state. Because two possibilities for parent of origin and two possibilities for genotype exist, one can distinguish between four different combinations of genotype and parent of origin. For 8 of the 11 genes, silencing occurred regardless of the allele and for 7 of the 11 genes silencing occurred regardless of parent of origin. In combination, for 7 of the 11 genes silencing occurred regardless of both allele and parental origin.

Expression of each of the assayed genes is considerably less common in comparison to silencing. For only 3 of the 11 genes did we find more cell lines exhibiting Xi expression rather than silencing. Furthermore, even for genes expressed from the Xi, the levels of expression were often severely reduced. When all 11 genes and all informative cell lines in our study are considered, for 47% of the expressed alleles, expression from the Xi was reduced below 10% of the Xa allele, and only six of the eleven genes showed the capacity of Xi expression above 10% of the active homologue in at least one sample. Over all for 3 of the 11 genes expression occurred regardless of both allele and parent of origin.

For several genes, we had the opportunity to examine expression of the same Xi allele in the mother and a daughter(s) or among multiple daughters. We found that for *ZNF185*, *CLIC2*, *TBL1X*, *ASB11* and *MORF4L2* the same allele does not necessarily exhibit the same inactivation status in related females (Figure 6). In other words, the same allele, passed to two different daughters, may become expressed or silenced depending on in which female the particular Xi allele resides. This observation indicates either *trans* regulation or instability of the inactive state.

Escaping from inactivation occurred in a greater number of genes in one of the five daughters (GM12807) than in her sisters (Table 4). Thus it appears that the Xi in this daughter might be overall more permissive to expression than the others. Interestingly, this daughter carries different alleles (G;C) at *XIST* rs1794213; rs1620574 than her sisters (T;T), suggesting that differences in or proximal to the X-inactivation center may result in different inactivation efficiency of specific genes. However, variation in the *XIST* gene itself is not likely to be responsible for higher escaping rates as another two cell lines with a high proportion of escaping genes carry the (T;T) *XIST* alleles (Table 4). More frequent escaping in specific individuals has been observed previously (77) and it will be interesting to investigate whether specific haplotypes at the X-inactivation center correlate with the level of permissiveness of the inactive state. Only two of the 11

**Table 3 Expression of variable genes in a panel of clonal cell lines**

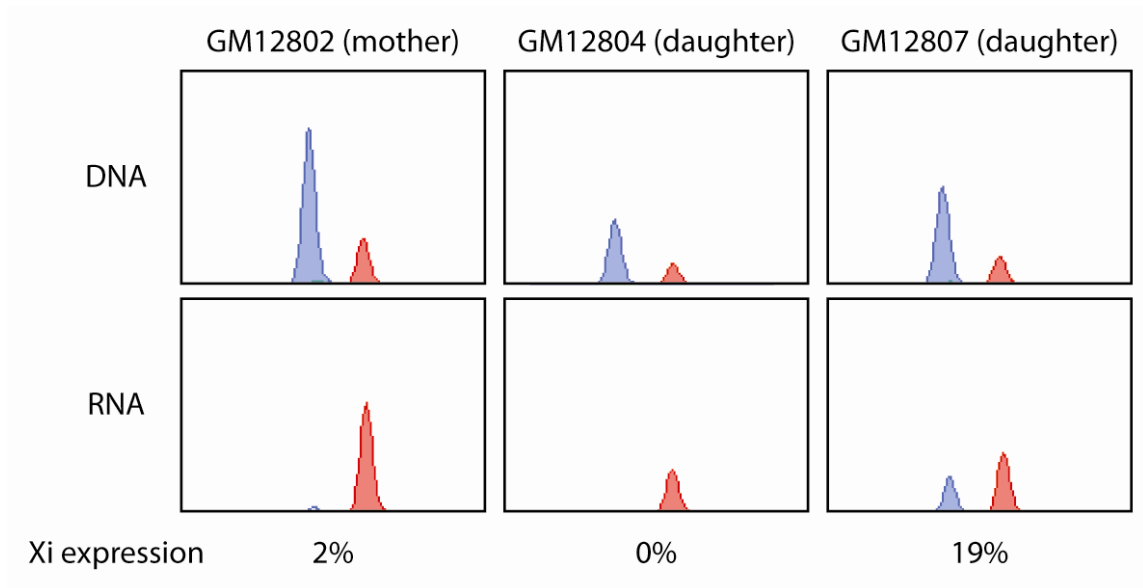
Shown are assayed genes and the proportion of isolates, in which each gene was expressed or silenced. The frequencies of expressed and silenced alleles (nucleotides and parental origin) are shown.

Gene	expressed : silenced	By allele		By mat:pat origin	
		expressed	silenced	expressed	silenced
<i>CLIC2(77)</i>	2:4	2C:0A	2C:2A	1:1	1:2 (1 unknown)
<i>ARHGAP4(77)</i>	3:4	1A:2G	3A:1G	2:1	3:1
<i>DMD</i>	6:1	2G:4A	1G:0A	5:1	1:0
<i>MORF4L2</i>	2:7	0A:2G	1A:6G	2:0	4:3
<i>MOSPD1</i>	0:4	0T:0G	3T:1G	0:0	4:0
<i>ASB11</i>	1:6	0G:1A	1G:5A	1:0	4:2
<i>TBL1X(77)</i>	2:5	0G:2A	2G:3A	2:0	3:2
<i>ZNF185(77)</i>	1:5	1C:0A	5C:1A	1:0	2:3
<i>SEPT6(77)</i>	8:2	6A:2C	2C:0A	4:3 (1 unknown)	0:2
<i>TRAPPC2(77)</i>	3:1	3C:0G	0C:1G	1:2	1:0
<i>TIMP1(77, 82)</i>	1:8	1T:0C	3T:5C	1:0	5:3

**Table 4 Levels of gene expression in CEU lymphoblastoid cell lines**

Cell lines are designated by the last three digits of their identification (for full identification see Table 2). “Proportion escape” ratios greater than 50% are shown in green. Samples that were homozygous (not informative) are indicated by “n.i.”

Gene	Cell line	-991	-861	-831	-839	-753	-802	-815	-803	-804	-807	-808	-809
<i>CLIC2(77)</i>		n.i.	n.i.	n.i.	n.i.	0%	2%	0%	n.i.	0%	19%	n.i.	0%
<i>ARHGAP4(77)</i>		n.i.	n.i.	n.i.	3%	2%	n.i.	n.i.	0%	0%	6%	0%	0%
<i>DMD</i>		n.i.	n.i.	0%	60%	n.i.	9%	n.i.	7%	4%	85%	8%	9%
<i>DMD</i>		n.i.	n.i.	n.i.	65%	n.i.	n.i.	n.i.	49%	n.i.	n.i.	n.i.	0%
<i>MORF4L2</i>		0%	n.i.	0%	3%	0%	n.i.	n.i.	0%	0%	3%	0%	0%
<i>MOSPD1</i>		n.i.	n.i.	0%	n.i.	n.i.	n.i.	n.i.	n.i.	0%	0%	0%	n.i.
<i>ASB11</i>		n.i.	n.i.	0%	n.i.	0%	n.i.	n.i.	0%	11%	0%	0%	0%
<i>TBL1X(77)</i>		n.i.	n.i.	n.i.	1%	0%	n.i.	n.i.	0%	0%	3%	0%	0%
<i>ZNF185(77)</i>		0%	0%	n.i.	n.i.	n.i.	n.i.	n.i.	0%	0%	38%	0%	0%
<i>SEPT6(77)</i>		0%	57%	19%	7%	65%	13%	32%	n.i.	4%	n.i.	n.i.	n.i.
<i>TRAPPC2(77)</i>		54%	n.i.	0%	49%	n.i.	42%	n.i.	n.i.	n.i.	n.i.	n.i.	n.i.
<i>TIMP1(77, 82)</i>		n.i.	0%	0%	5%	n.i.	0%	n.i.	0%	0%	n.i.	n.i.	0%
<i>Proportion escape</i>		1/4	1/3	1/7	8/8	2/6	4/5	1/2	2/8	3/10	6/8	1/7	1/9
<i>XIST rs1794213 (Xi allele)</i>		T	G	T	T	G	T	T	T	T	G	T	T
<i>XIST rs1620574 (Xi allele)</i>		T	C	T	T	C	T	T	T	T	C	T	T



**Figure 6 Variable expression of *CLIC2* in lymphoblastoid cell lines derived from a mother and two daughters**

Shown is DNA and RNA SNaPshot output at *CLIC2* SNP rs559165. In DNA, all three readouts indicate heterozygous heterozygosity (G/T) at the queried SNP. The RNA allelic contribution is variable with respect to the three females. Xi expression is shown as %Xa expression.

tested genes showed the possibility of exhibiting a parent of origin (*MOSPD1*) or allelic effect (*TRAPPC2*). The variable nature of *MOSPD1* expression described previously (77) was not observed in our Xi chromosome panel, potentially due to a parent of origin effect. The inactivation status of *MOSPD1* does not appear to exhibit an allelic effect as three of the tested Xi allele carried a (T) allele and one a (G) allele, all of which were silenced. However, all four tested Xi chromosomes were of maternal origin, and *MOSPD1* was consistently expressed monoallelically from the Xa in all four samples. It will be necessary to test Xi<sup>p</sup> chromosomes for a conclusive result. If *MOSPD1* does in fact exhibit parent of origin effect, alleles of *MOSPD1* residing on Xi<sup>p</sup> would be expressed.

*TRAPPC2* is the only gene from our panel that might exhibit familial mode of inactivation (Figure 2A) where the allele determines the inactivation status of a particular gene. In the available cell lines, the *TRAPPC2* (C) allele was expressed from three Xi chromosomes and the (G) allele was silenced on one Xi chromosome.

#### **2.2.4. *TRAPPC2* exhibits unstable inactivation in multiple clones derived from a single female cell line**

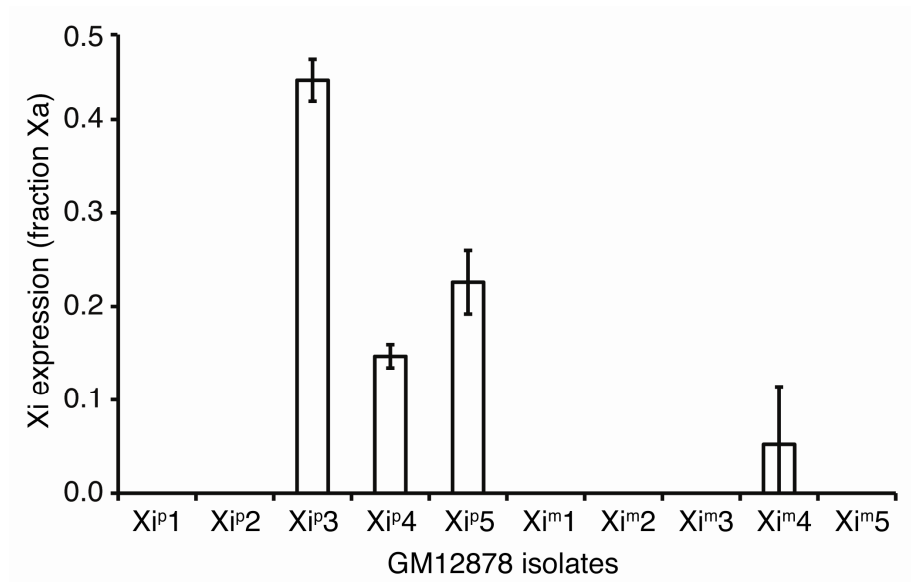
We investigated expression of *TRAPPC2* in four informative cell lines from unrelated individuals. We observed Xi expression from both maternally- and paternally-derived alleles in the four cell lines from unrelated females, suggesting no parent of origin effect on inactivation of the *TRAPPC2* locus. However, the (C) allele when present on the Xi chromosome was always expressed, while the (G) allele was silenced. Although

a greater number of cell lines have to be tested to validate the results, this outcome is consistent with an allelic effect.

We proceeded to assay *TRAPPC2* in multiple cell lines  $Xi^m$  and  $Xi^p$  derived from a single cell line GM12878 in order to assess the stability of expression levels from the  $Xi$  allele. Surprisingly, the expression pattern in these cell lines was highly variable. *TRAPPC2* was expressed in three out of five (G) allele ( $Xi^p$ ) isolates in levels ranging from 17-47% of the  $Xa$  allele expression and in one out of the five (C) allele ( $Xi^m$ ) isolates (Figure 7).

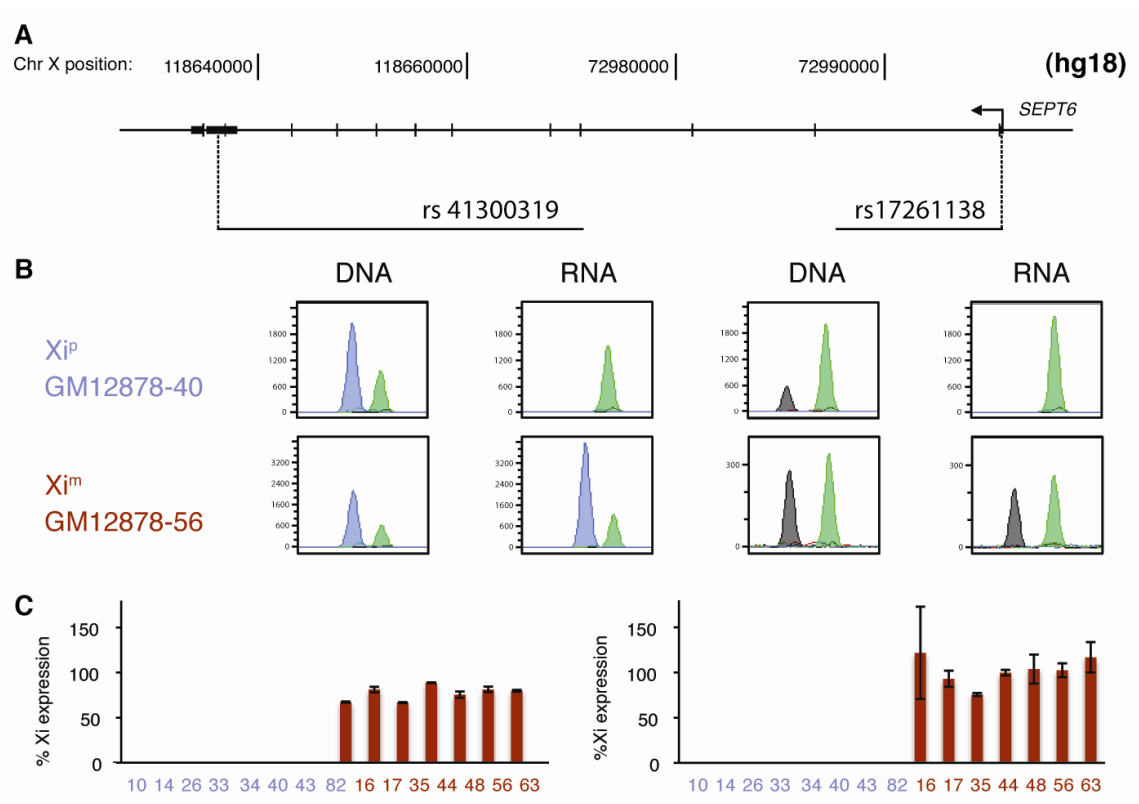
#### **2.2.5. *SEPT6* is expressed from one, but not the other inactive X chromosome in a single human cell line**

*SEPT6* belongs to a class of GTPases required for cytokinesis (217, 218). The *SEPT6* gene (Figure 8A), located at Xq24, has at least six splice isoforms with alternative TSSs or termination sites. To determine the inactivation status of *SEPT6*, we first assayed an expressed SNP in the 3'UTR that is shared by three of the six known isoforms (Figure 8A). By targeting the end of the gene, we aimed to detect the fully elongated transcript. *SEPT6* was expressed monoallelically in all  $Xi^p$  samples and biallelically at 77% (SD=8) relative to the  $Xa$  in all  $Xi^m$  samples (Figure 8B and C). Next, we set to assay a SNP in the 5' UTR that is shared by five of the six *SEPT6* isoforms. We tested the same eight  $Xi^p$  and seven  $Xi^m$  clones at this SNP and found concordant results (Figure 8B and C). There was



**Figure 7** Expression of *TRAPPC2* in multiple isolates derived from GM12878 cell line

GM12878 isolates, five Xi<sup>p</sup> and five Xi<sup>m</sup>, and their expression relative to the Xa are shown on the horizontal axis. Error bars indicate standard deviation of the mean calculated from biological replicates.



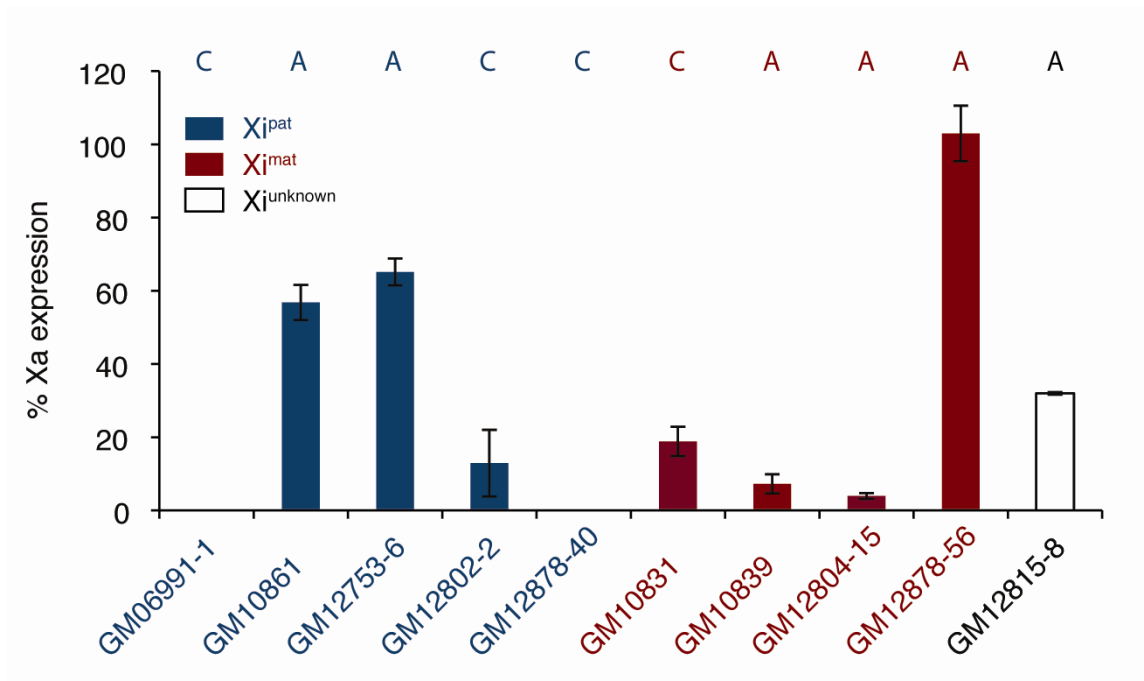
**Figure 8 Expression of *SEPT6* from *Xi<sup>m</sup>* and *Xi<sup>P</sup>* in multiple isolates derived from GM12878 lymphoblastoid cell line**

(A) *SEPT6* is located at ~119Mb (hg 18) in Xq24. Shown are expressed heterozygous SNPs in 5' and 3' UTRs used in SNaPshot expression assay. (B) SNaPshot peaks for DNA and RNA for representative *Xi<sup>P</sup>* and *Xi<sup>m</sup>* GM12878 isolates at rs41300319 and rs17261138 are shown. Peak colors indicate the nucleotide, G-blue, A-green, C-black. (C) Expression levels in each isolate are shown *Xi<sup>P</sup>* (blue bars) and *Xi<sup>m</sup>* (red bars) as %*Xi* expression. Error bars represent standard deviation of the mean in biological replicates.

no detectable expression from  $Xi^p$  and 102% (SD=15) expression from  $Xi^m$  relative to the  $Xa$ .

To assess variability of *SEPT6* in the population, we tested its expression in cell lines from eight additional individuals (Figure 9). We confirmed that *SEPT6* is not only variable with respect to the two  $Xi$  alleles in GM12878 but also with respect to the population. Surprisingly, we did not observe a parent of origin or a familial mode of inheritance as the gene was expressed in all four  $Xi^m$  samples, but also in three of the five  $Xi^p$  samples and was expressed from all (A) alleles, but also from two of the four (C) alleles. In addition, *SEPT6* exhibits a wide range of  $Xi$  expression levels in the population suggesting a more complex mode of regulation.

There were no expressed heterozygous SNPs in the sixth isoform; therefore, it remains unknown whether the expression of all *SEPT6* gene products is regulated in the same manner. *SEPT6* is an example of the ever-expanding group of X-linked genes with variable inactivation status (77, 81). It is the first gene observed to have discordant inactivation status not only in the population but also within a single individual. Further studies will be required to determine the genetic nature and heritability of the *SEPT6* expression. This result is in contrast to the *TRAPPC2* gene, whose expression is variable with respect to the population and with respect to multiple GM12878 isolates (Figure 7).



**Figure 9 Expression of *SEPT6* from Xi<sup>m</sup> and Xi<sup>p</sup> in clonal isolates derived from multiple females**

Expression from the Xi relative to the Xa in each cell line is shown. Names, alleles and corresponding bars of cell lines carrying paternal and maternal Xi alleles are depicted in colors (see figure). Alleles are shown above each expression bar. Error bars represent standard deviation of the mean in biological replicates.

## 2.3. *Discussion*

### 2.3.1. **Stochastic inactivation of variable genes**

In this chapter, I have examined the now commonly observed variable expression pattern of a number of genes residing on the Xi with the aim to explore the possibility of a genetic basis to the expression of Xi alleles. From the data presented here, it is evident that genes with variable expression patterns are commonly expressed and/or silenced regardless of the allele or parent of origin. I observed that a majority of the variably expressed genes are expressed at low frequency in the population and when expressed from the Xi, the levels of their expression are low (<10%). These data suggest that there is a low level of 'leakiness' of the Xi silencing heterochromatin that occurs at a number of X-linked loci.

The heterochromatic Barr Body composed of the silenced human X chromosome was formerly believed to be devoid of expression activity (219). The observable phenotype of 45,X Turner syndrome females, however, illustrates the necessity of presence of both X chromosomes in human female cells and suggests an important functional activity occurring on the Xi chromosome (66). Up to 15% of X-linked genes have been identified that consistently escape X chromosome inactivation and are expressed from both alleles in each female examined thus far (77). Furthermore, variable expression patterns from the Xi have been observed for a number of genes (77-

79, 81, 82) and the relatively low sampling of the previous and current studies has been insufficient to present a complete profile of the population. It is likely that a number of genes currently classified as subject to inactivation might in fact be expressed in some females and, similarly, that genes that appear to consistently escape inactivation so far might be inactivated in rare cases.

Although a number of genes that have the capability of biallelic expression have been identified, it is presently unknown how essential biallelic expression is to each locus and how it is regulated. The extremely low expression levels of most of these genes from the Xi suggests that such small amount of gene product may not make appreciable difference in the context of the expressed Xa allele, as normal fluctuation in expression levels is common. Rather, this result is suggestive of a certain level of permissiveness of the Xi heterochromatin to the transcription machinery allowing low level of expression to occur from many, but not all genes on the Xi.

By comparing *SEPT6* in the two types of clones derived from GM12878, we discovered that *SEPT6* at least in this one cell line is completely silenced when the Xi allele is derived from the father and well expressed when the Xi is derived from the mother. The expression profile of *SEPT6* in the population does not support parent of origin or allelic effect, however. Instead, inactivation of *SEPT6* in the population appears to be random with respect to the parent of origin and the particular allele.

### **2.3.2. Genes with variable inactivation status mostly exhibit low level of expression and low frequency in the population**

We detected low level of transcription from the Xi allele at a number of genes, many times at the limit of detectability. Several explanations for low expression levels exist. The most widespread belief is that whatever silencing mechanism(s) is in play, it is not strong enough to repress all transcription at all times, resulting in leakiness or stochastic expression (220-222), for example in inappropriate tissues for tissue-specific genes (illegitimate transcription) (223). Low abundant transcripts have been detected across various genomes at sites outside currently annotated coding regions such as intergenic and intronic sequences as well as antisense transcripts of currently annotated loci (224).

Another hypothesis suggests that low-level expression is actually a functional process involved in gene silencing (220, 225, 226). It has been recently discovered in yeast and plants that low level of illegitimate transcription is required to repress the gene in question (227). It may also be the case in humans, as illegitimate transcription of human tissue-specific genes has been observed in inappropriate tissues (220, 221, 228, 229).

In the case of inactivation of variable genes, according to the observations reported here, the low-level expression of certain loci is more consistent with the idea that expression at these loci is incompletely repressed, resulting in random low levels occasional expression in the female population, consistent with the stochastic model

proposed in Figure 2. However, it is possible that the observed low level of transcription is involved in silencing of these genes. If fluctuation in the level of expression occurs, the variable expression observed here, might be a result of insufficient detection capability of the technology, thus we may be detecting only the peaks, but not valleys in this fluctuating expression cycle.

Due to this widespread variability across the human Xi and even among multiple clones derived from a single cell line (Figure 7), I postulate that the global action of X chromosome inactivation defined by coating with *XIST* RNA, defined heterochromatin types and hypermethylation is fine tuned at individual loci. It has been proposed that the two alternating heterochromatin types (50) possess different capabilities in terms of gene silencing; thus low level expression may be a feature of one, but not the other heterochromatin type. The spreading and transition between these two types of silencing heterochromatin may also be contributing to the variability in silencing (230). Furthermore, it has been shown that genes that escape inactivation tend to cluster (77, 80, 212), thus it is conceivable that genes with variable expression patterns are more likely to be found in some regions than others.

### **2.3.3. X-inactivation skewing and derivation of homogeneous Xi populations**

A large number of female Xi chromosomes have to be examined before we fully understand the variable patterns of X chromosome inactivation discussed in this

chapter. It is likely, that expression or complete silencing of the Xi allele is essential to only a subset of X-linked genes, while others may be occasionally expressed without a deleterious consequence. Such genes might be identified by screening non-randomly inactivated human female samples that represent a larger portion of the population.

We have shown that single cell cloning is an optimal method to generate female Xi chromosomes suitable for X-inactivation studies in human-derived cell lines. Although all cell lines used in our study exhibited extreme skewing (>75-25%), we were able to derive two distinct Xi homogeneous isolates for four of the 13 cell lines. Our collection of derived cell lines, homogeneous with respect to the origin of the inactivated X chromosome is the first set of samples allowing a study of X chromosome inactivation in a large three-generation family. These samples are prime candidates for further expansion of this relatively low-throughput study pioneered here, using some of the recently developed next-generation sequencing methods.

## **2.4. *Materials and methods***

### **2.4.1. Cell culture and single cell cloning**

GM12878 cells were grown in a humidified incubator at 37°C and 5% CO<sub>2</sub> in RPMI1640 media supplemented with 15% fetal bovine serum and 1% antibiotics according to the ENCODE protocol (34). For single-cell cloning, cells were diluted in 50%

conditioned media in a series of dilutions and seeded in 96-well plates. Cells were grown in 50% conditioned media until expansion and fed fresh media thereafter. To prepare conditioned media, cells were grown in fresh media over night and subsequently removed by centrifugation and filtration; the resulting cell-free conditioned media was diluted 1:1 with fresh media to obtain 50% conditioned media. For each presumptive clone, three biological replicates were grown (two for RNA and one for DNA). To confirm homogeneity of each candidate clone with respect to X inactivation, we tested allele-specific expression in monoallelically-expressed genes, including *XIST* (rs1620574), *ATRX* (rs3088074) and *EBP* (rs3048) (77). We further verified the purity of each selected isolate after expansion by the same method. Two confirmed isolates of each type (100%  $Xi^m$  and 100%  $Xi^p$ ) were chosen for this study. Additional isolates were selected for further *TRAPPC2* expression experiments.

#### **2.4.2. Nucleic acid purification**

DNA from each lymphoblastoid line was isolated from fresh cells using Puregene Tissue Core Kit A (Gentra) and stored at -20°C. RNA was isolated from fresh cells using PerfectPure RNA Tissue Kit (5Prime) and stored at -80°C.

### **2.4.3. SNaPshot**

A quantitative Q-SNaPshot assay (Figure 5) was employed to test the abundance of each allele in the PCR amplicon, using protocols as described previously (51, 77). Briefly, DNA was isolated from fresh cells using a Qiagen Genra Puregene Cell Kit and stored at -20°C. RNA was isolated from fresh cells with a PerfectPure RNA Tissue Kit (5Prime), treated with DNase I (Roche) and stored frozen at -80°C. Quality and concentration of each nucleic acid sample was verified by NanoDrop spectrophotometer and gel electrophoresis. Random primed cDNA was synthesized from 200-300 ng of RNA using an iScript cDNA synthesis kit (BioRad). In each assay, DNA and cDNA was amplified by a standard 25µL protocol using Taq DNA Polymerase (Invitrogen) at 94°C for 2 min, 32-36 cycles of 94°C for 30s, 55°C for 30s, 72°C for 30s, 72°C for 10min, holding at 4°C. PCR products were then purified (EdgeBio). A third primer was used for the Q-SNaPshot single nucleotide extension assay with subsequent ABI 3100 sequencer detection. The cDNA readout was normalized to the DNA signal with known 1:1 ratio of the two alleles to correct for biases in fluorescence output.

### 3. Allele-Specific Distribution of RNA Polymerase II on Female X Chromosomes

Katerina S. Kucera, Timothy E. Reddy, Florencia Pauli, Jason Gertz,  
Jenae E. Logan, Richard M. Myers & Huntington F. Willard

#### Collaborators:

*Timothy E. Reddy and Jason Gertz* – HudsonAlpha Institute for Biotechnology,  
Huntsville, AL – performed bioinformatic analysis of PolII binding

*Florencia Pauli* – coordinated the study

*Jenae E. Logan* – IGSP, Duke University – undergraduate student that performed  
assays of *TRAPPC2* expression in unrelated females as a part of independent study  
under my mentorship

*Jospeh Lucas* – IGSP, Duke University – provided statistical advice

#### This work was published in:

**Kucera KS**, Reddy TE, Pauli F, Gertz J, Logan JE, Myers RM, Willard HF. Allele-specific  
distribution of RNA polymerase II on female X chromosomes. *Hum Mol Genet.*  
2011.

While the distribution of RNA polymerase II (PolII) in a variety of complex genomes is correlated with gene expression, the presence of PolII at a gene does not necessarily indicate active expression. Various patterns of PolII binding have been described genome-wide; however, whether or not PolII binds at transcriptionally inactive sites remains uncertain. The two X chromosomes in female cells in mammals present an opportunity to examine each of the two alleles of a given locus in both active and inactive states, depending on which X chromosome is silenced by X chromosome inactivation. Here, we investigated PolII occupancy and expression of the associated genes across the active (Xa) and inactive (Xi) X chromosomes in human female cells to elucidate the relationship of gene expression and PolII binding. We find that, while PolII in the pseudoautosomal region occupies both chromosomes at similar levels, it is significantly biased toward the Xa throughout the rest of the chromosome. The general paucity of PolII on the Xi notwithstanding, detectable (albeit significantly reduced) binding can be observed, especially on the evolutionarily younger short arm of the X. PolII levels at genes that escape inactivation correlate with the levels of their expression; however, additional PolII sites can be found at apparently silenced regions, suggesting the possibility of a subset of genes on the Xi that are poised for expression. Consistent with this hypothesis, we show that a high proportion of genes associated with PolII-accessible sites, while silenced in GM12878, are expressed in other female cell lines.

### 3.1. *Introduction*

RNA polymerase II (PolII) is an essential component of the eukaryotic transcriptional machinery. While other RNA polymerases transcribe non-coding genes, PolII is involved in transcription of coding genes as well as non-coding RNA genes (231, 232). High throughput CHIP-seq studies have described the genomic distribution of PolII in a number of model organisms (46, 233-235). PolII localization relative to genes has been observed at transcriptional start sites (TSSs) and is less often distributed along the gene body, beyond the 3' end, or can be absent all together (236). In addition to its accumulation within actively expressed genes, PolII has also been found at enhancers (14), in intergenic regions (231), and at TSSs of developmentally regulated genes that are periodically turned off and on (46, 233, 234).

Promoter regions of highly regulated genes that depend on recruitment of PolII for their activation are often covered with nucleosomes (237) that prevent PolII binding until chromatin remodelers allow the preinitiation complex (PIC) to assemble (238). In contrast, constitutively transcribed genes maintain their promoter region uncovered (239), thus allowing permanent access to PolII. Upon gaining access to the promoter, a number of steps must occur for PolII to actively engage in transcription. The PIC assembles on the core promoter and induces DNA unwinding, upon which PolII proceeds to a promoter-proximal pause region. At this position, PolII becomes phosphorylated, escapes the promoter-proximal pause region and continues

transcription (240). The frequently observed bimodal distribution of TSS-associated PolII signals reflects pausing and possibly also poisoning at the TSS, as a result of PolII accumulation due to downstream rate-limiting steps (236, 241-244). Four general classes of genes have been described based on the presence or absence of the PIC at the promoter region as it relates to transcriptional activity: expressed genes with or without PIC detectable at the promoter, and silenced genes with or without PIC detectable at the promoter (245). PolII can also be observed within the bodies of some genes, reflecting active elongation (246). In addition, PolII peaks can be detected several kilobases downstream from the 3' end of some genes (14, 231), likely reflecting the lack of strong termination signals that allow the enzyme to progress beyond the annotated end of genes (14).

High-throughput ChIP-seq and ChIP-chip technologies have provided valuable information on the spatial and temporal characteristics of various chromatin elements, including PolII, throughout the diploid human genome (199, 246). A greater challenge arises when interpreting these data at loci exhibiting allelic imbalance (247), as a number of genes across the human genome are not equally expressed from both alleles due to widespread *cis*- or *trans*-determined influences on transcription (5, 13, 196) and to effects such as genomic imprinting (248), allelic exclusion (196), and, in females, X chromosome inactivation (64, 76). Similarly, associated *cis*-acting regulatory elements, such as various chromatin marks or components of the transcriptional machinery, can themselves exhibit allele-specific patterns (34, 35), and thus, for these genomic loci, the

information pertaining to the two alleles must be separated. Examining the allele-specific basis for such signals allows exploration of the nature of *cis*-regulatory mechanisms on both PolII occupancy and gene expression.

Although the majority of X-linked genes are silenced in mammalian females due to X chromosome inactivation, at least 10% of genes residing on the X chromosome are expressed biallelically in all samples analyzed to date (77, 80, 135, 249) and many more exhibit variable expression patterns among different females (77, 81, 82). It is presently unknown what factors determine the inactivation status of X-linked genes and how important functionally or phenotypically it is for the Xi copy of a given gene to be expressed or silenced. Nonetheless, the evident heterogeneous patterns of gene expression from the Xi provide an opportunity to examine the association among genetic, genomic and/or stochastic signals that may underlie PolII occupancy at different loci.

The facultative heterochromatin of the human Xi consists of at least two distinct types of chromatin marked by a host of components that distinguish them, including various histone modifications or variants, binding proteins and *XIST* RNA (50, 51). While these and other chromatin features have been investigated in mouse and humans (50-52, 250-252) none has yet been tightly correlated with the inactivation status of individual genes. Because PolII is required for transcription, its presence or absence would appear to be better suited as a potential indicator for silencing due to X inactivation. In mouse embryonic stem cells, PolII exclusion occurs soon after the

induction of differentiation, which triggers expression of the *Xist* gene (253) and subsequent inactivation of X-linked genes. PolII is also eliminated from the human Barr body (142); however, because the core of the Barr body is composed of repetitive non-coding DNA, while genes (regardless of their expression status) occupy the periphery (142), exclusion of PolII from the Barr body core does not necessarily imply its exclusion from genes.

The process of X inactivation is initiated early in development and the chromosome to be inactivated is selected at random, such that the resulting female is a mosaic of cells carrying an Xi that is either maternally- ( $Xi^m$ ) or paternally-derived ( $Xi^p$ ) (76, 254, 255). While such clonal mosaicism can obscure the detection of Xi-specific features, a number of investigations have turned either to oligoclonal lymphoblast cell lines that deviate from random distribution and are severely skewed toward one Xi chromosome or the other (216, 256) or to fibroblast cell lines that exhibit complete non-random inactivation and thus are pure populations of cells with either  $Xi^m$  or  $Xi^p$  (77, 80).

In this paper, we describe the PolII landscape on human female X chromosomes utilizing ENCODE PolII ChIP-seq data for the GM12878 cell line, a chromosomally normal female lymphoblast line. We selected this cell line because of the opportunity to build on a number of earlier studies that have characterized GM12878 as part of the ENCODE Project (205) and the 1000 Genomes Project (27, 199, 209). Because of the severe X-inactivation skewing in GM12878 (34), we were able to infer the enrichment of individual alleles on the Xi and Xa and thus analyze these data in an allele-specific

manner to compare PolII distribution on the two X's. Further, for the purpose of elucidating the relationship of PolII and inactivation status of genes, we have focused specifically on PolII sites that could be assigned to particular genes, and, where possible, we have determined their inactivation status in homogeneous GM12878 Xi<sup>p</sup>- and Xi<sup>m</sup>-derived cell lines in order to explore the relationship between PolII binding and X inactivation.

## 3.2. *Results*

### 3.2.1. **The human X chromosome is relatively PolII poor**

We first sought to measure genomic occupancy of PolII on the X chromosome. To do so, we used a genome-wide ChIP-seq approach analyzed by QuEST (257) and, in two replicate samples, detected on average 168 sites of PolII enrichment on the X chromosomes in the GM12878 cell line (Table 5). Notably, this reflects a significantly ( $p < 0.001$ , t test) lower PolII peak occurrence relative to gene density on the X chromosome compared to autosomes (Figure 10A). We observed a similar, albeit less extreme, situation for chromosome 11 ( $p < 0.05$ , t test) that can likely be attributed to the fact that more than 300 olfactory receptor genes are located on chromosome 11 and are likely not expressed in lymphoblasts. The paucity of PolII binding sites could reflect the number of genes, the number of expressed genes, or the frequency of

binding as a function of chromosome size. To explore these differences further, we compared the abundance of PolII peaks to chromosome length (Figure 10B) and to the number of expressed genes for each chromosome in GM12878 (209) (Figure 10C). While PolII binding is loosely correlated with chromosome length (Figure 10B), it is better correlated with overall gene density (Figure 10A) and is particularly well correlated with the number of genes expressed in GM12878 (Figure 10C). Thus, the relative scarcity of PolII peaks on the X chromosome can be at least partially explained by the lower proportion of genes that are expressed from the X chromosome in the GM12878 cell line as compared to autosomes (Figure 10).

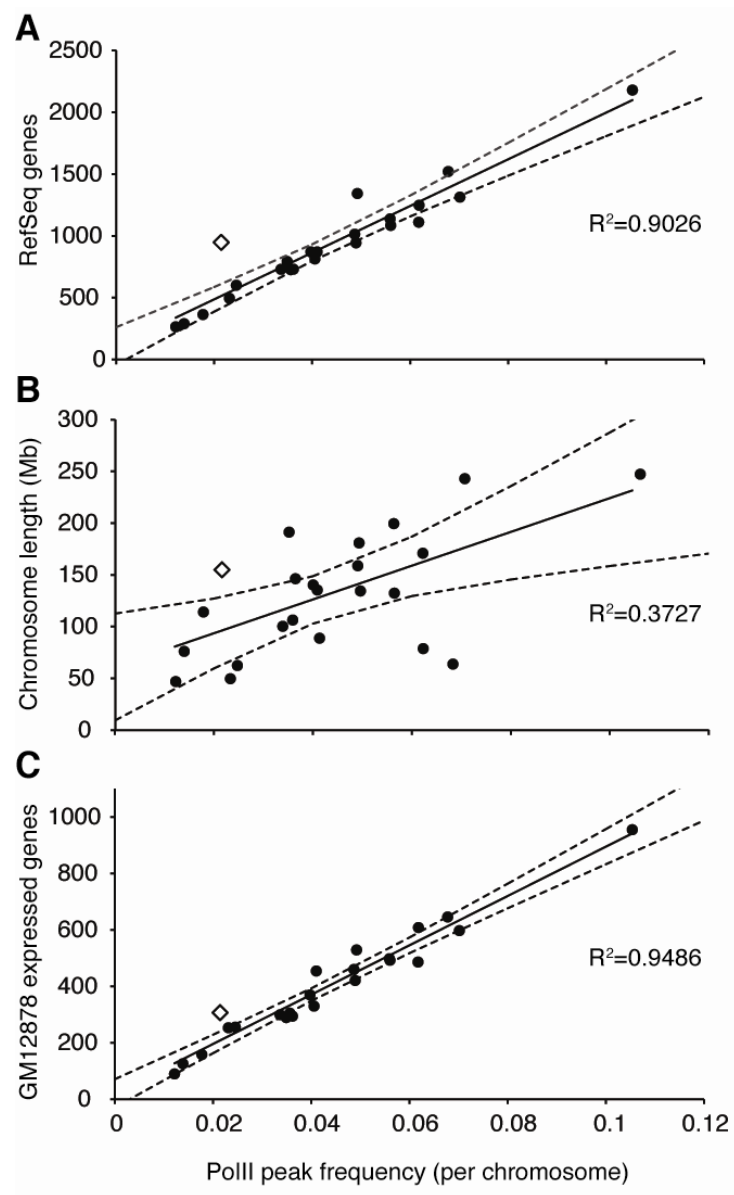
Though we observed fewer PolII binding sites on the X, the apparent intensity of the ChIP-seq signal (based on the number of mapped reads) on the X chromosome is comparable to those on autosomes (Figure 11), suggesting either that PolII occupies both X chromosomes or that increased binding of PolII on X<sub>a</sub> compensates for the lack of binding on the X<sub>i</sub>, e.g. (72). To ensure that the result was not specific to the QuEST algorithm (257), we repeated the analysis with a different method, based on MACS (258) (see Methods). The MACS analysis resulted in about ten times as many called peaks compared to QuEST; however, the overall conclusions were upheld, as we detected only marginally weaker peaks on the non-pseudoautosomal portion of the X chromosome as compared to the autosomes (data not shown).

**Table 5 Analysis of PolII binding sites on the human X chromosome**

	<b>PolII peaks called (QuEST)<sup>a</sup></b>	<b>PolII occupied heterozygous SNPs<sup>b</sup></b>
Total	168	385
TSS +/- 500bp	120	37
Intragenic	138	257
Extragenic	30	128
< 5Kb from genes	9	60
5-30Kb from genes	15	17
> 30Kb from genes	6	51

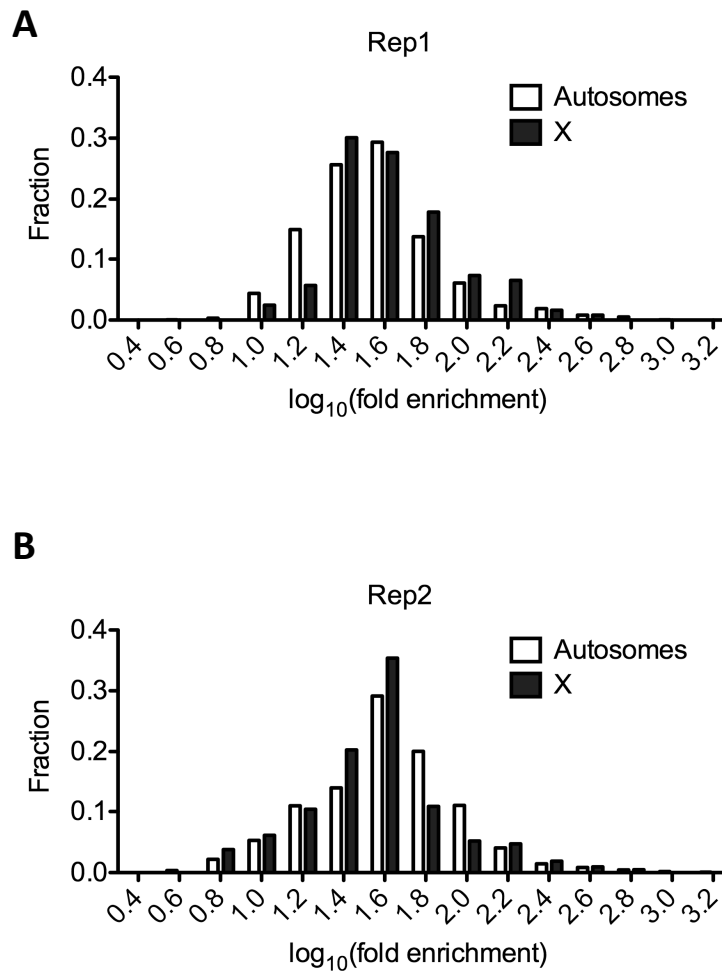
<sup>a</sup> Non-allele-specific PolII peaks called by Quest software (257) as a composite of the two X chromosomes

<sup>b</sup> PolII occupancy at heterozygous sites; sites with minimum of 5 reads at one or the other allele were considered



**Figure 10 PolII peaks per chromosome**

Each human chromosome is represented by a circle (autosomes) or an open diamond (X chromosome). Linear regression (solid line) and 97% confidence interval (dashed lines) are shown. Number of PolII peaks is shown relative to (A) RefSeq genes; (B) chromosome length; and (C) expressed genes in GM12878 (expression data from (209)).



**Figure 11 Distribution of the log-ratio of PolII ChIP-signal versus background signal (horizontal axis) for all PolII binding sites identified with QuEST**

Ratio of signal intensity was calculated by QuEST based on kernel density estimates of the binding site (257). The median signal for an X chromosomal binding site had stronger signal than the median autosomal binding site in the first replicate (A) (median fold enrichment 36 vs 43, respectively;  $p = 0.0031$ ; two-sided Wilcoxon test) but weaker signal in the second replicate (B) (median fold enrichment 41 vs. 36, respectively;  $p = 0.0006$ ; two-sided Wilcoxon test). As such, there does not appear to be a reproducible bias of PolII ChIP-seq signal intensity differentiating the X from autosomes.

### 3.2.1. PolII binding is biased toward the active X chromosome

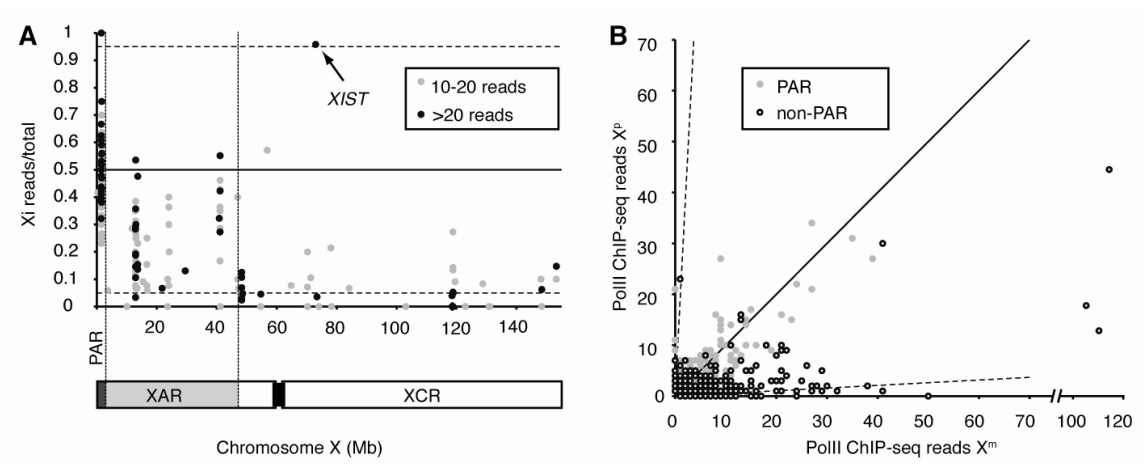
In a previous study (34), we reported 92% skewing of X inactivation in the GM12878 cell line toward the  $X_i^p$ . Because X-inactivation skewing can drift in cell culture over time (216, 256), we verified the extent of skewing in GM12878 in RNA samples harvested concurrently with the PolII ChIP-seq samples generated for this study. We detected 95% (SD=1%) skewing in the same direction as previously described (34), by testing allele-specific expression at *XIST* (rs1620574) and *EBP* (rs3048) by SNaPshot assays (77).

To determine whether PolII binding on the X chromosome is influenced by X-chromosome inactivation, we focused on the subset of individual PolII sites that contain heterozygous SNPs and employed an allele-specific alignment approach to distinguish binding on the  $X_a$  and  $X_i$  chromosomes. We identified 385 heterozygous SNPs (Table 5) that had at least five mapped reads on at least one of the two X chromosomes in the PolII ChIP-seq dataset (Figure 12, Appendix A). We chose this threshold to limit artifacts due to spurious alignment or non-specific immunoprecipitation of DNA. This allele-specific dataset of 385 sites forms the basis for the analysis in this section.

PolII binding in the pseudoautosomal region on the distal short arm (PAR1) is especially dense as compared to the non-pseudoautosomal region (Figure 12A). Furthermore, as expected, we observed no clear bias in occupancy toward  $X_a$  or  $X_i$  in PAR1, with 47% of the total PolII reads that mapped to the PAR1 aligning to the  $X_i$  alleles. A few individual heterozygous PolII binding sites showed significant bias towards

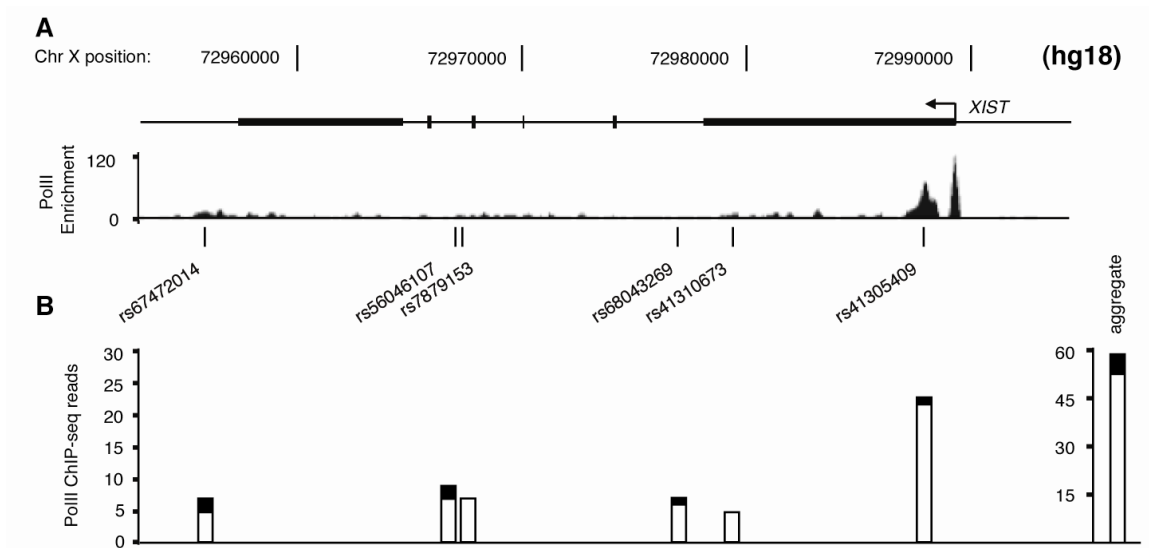
one or the other X (Figure 12), and these likely represent occasional allele-specific bias similar to that seen throughout the genome (209) or a currently unrecognized influence of X inactivation on certain regions within PAR1. It remains unknown what significance these PolII sites have and whether they are functionally associated with the neighboring genes.

The remaining majority of PolII sites at heterozygous positions on the X were located in the non-pseudoautosomal region of the X chromosome. Two noteworthy observations indicate relative depletion of PolII across the non-pseudoautosomal portion of the Xi. First, considering this region as a whole, we detected overall about five times more total ChIP-seq reads mapping to the Xa than to Xi in this region (Appendix A), consistent with reduced PolII binding being associated with X inactivation. Second, of the 268 non-pseudoautosomal PolII sites examined, 39% exhibited significant bias ( $p < 0.01$ ) in PolII occupancy toward one chromosome or the other (Figure 12B, Appendix A). Among these, all but two showed bias toward the Xa allele. Not unexpectedly, the two sites with significant PolII binding bias toward the Xi were located in the *XIST* gene (Figure 13) (85, 86). At a PolII site near the 5' end of *XIST*, 23 of the 24 sequences (96%) aligned to the paternally-derived (Xi) allele, which correlates well with the extent of X-inactivation skewing in the cell line, determined by allele-specific expression analysis (see Methods) (34) and implies absence of PolII on the Xa. As PolII is completely depleted from the Xa near the *XIST* TSS, the relative abundance of PolII at this site is an expression-independent indicator of X-inactivation skewing.



**Figure 12 Allele-specific PolIII occupancy on the X chromosome in GM12878**

(A) Distribution of heterozygous sites with at least 10 mapped PolIII ChIP-seq reads on the X chromosome, displayed as the proportion of sites on the Xi. XAR = X-added region; XCR = X-conserved region. (B) PolIII occupancy on the X<sup>m</sup> versus X<sup>p</sup> chromosomes. Dashed axes represent 95% skewing (i.e. monoallelic occupancy) observed in the GM12878 cell line. Solid diagonal represents equal PolIII occupancy on Xi and Xa.



**Figure 13 PolII occupancy on Xi and Xa at the *XIST* gene**

(A) PolII ChIP-seq enrichment throughout the gene is shown with the highest enrichment near the TSS. Locations and rs numbers of heterozygous SNPs in GM12878 are indicated. (B) Relative PolII occupancy on the two X chromosomes in GM12878, as measured by the number of aligned ChIP-seq reads at each heterozygous SNP, mapping to the X<sup>p</sup> (white bars) or X<sup>m</sup> (black bars). The aggregated signal from the *XIST* locus is shown on the far right, indicating significant PolII binding bias toward X<sup>p</sup>, consistent with the reported (28) and measured (see Methods) X-inactivation bias in this cell line.

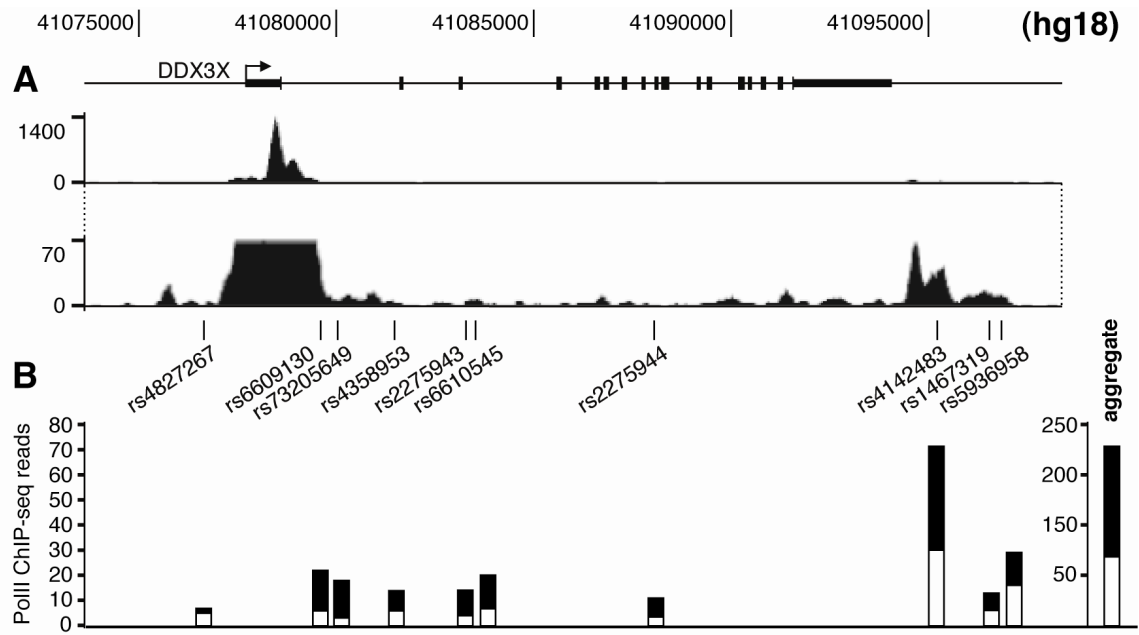
The observed PolII profile on Xa and Xi (Figure 12) appears to reflect in part the evolutionary origins of the X chromosome, as the bias is nearly complete in the portion of the X that corresponds to the ancestral conserved region of the X chromosome (XCR), but is decidedly less extreme in the portion of the X that was added more recently during evolution (XAR) (126, 129, 259). PolII occupancy also mirrors the reported patterns of X inactivation, as a much greater proportion of genes in XAR escape inactivation than in XCR (77) and thus might be expected to be associated with PolII on both Xa and Xi.

### **3.2.1. PolII binding on the human inactive X chromosome largely mimics inactivation status of genes**

Although PolII is generally depleted from the human Xi, many sites deviate from the predicted level of X-inactivation skewing (Figure 12, Appendix A). In the non-pseudoautosomal region of the X chromosome, over 45% of the informative PolII binding sites showed more than 10% PolII occupancy on the Xi relative to the Xa, suggesting at least some PolII located on the Xi allele. Strikingly, at 12% of sites, the level of Xi PolII occupancy exceeded 50% of the PolII levels on Xa. Genes with high PolII occupancy on the Xi were of special interest for considering relative gene expression from Xi and Xa, as they could reflect either biallelic expression or the presence of PolII at silenced genes.

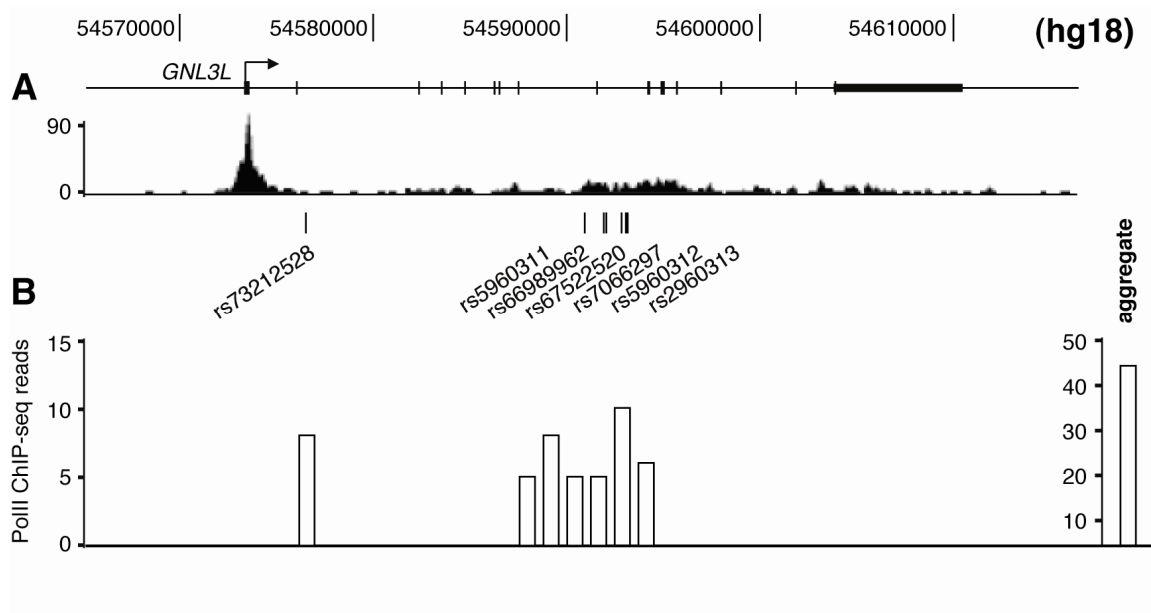
To assess the relationship of PolIII binding and gene expression from the Xi, we aggregated PolIII binding across annotated genes and their flanking regions (30Kb upstream and 5Kb downstream), considering only those sites that could be unambiguously associated with assayable genes (Appendix B, Figures 13, 14, 15). While sites upstream from TSSs seem likely to reflect association at regulatory regions for those genes (14) and downstream sites likely result from PolIII progression beyond the 3' end (14, 231), we cannot rule out the possibility that there is no functional association of these sites with the assigned genes.

We identified 31 genes with robust and informative PolIII binding across the X chromosome (including PAR1), that is, genes containing expressed heterozygous SNPs that could be studied with allele-specific expression assays. We assayed allele-specific expression at these 31 genes in a series of apparently clonal cell lines derived from GM12878 that were homogeneous with respect to the Xi<sup>p</sup> or Xi<sup>m</sup> inactive X chromosomes (Appendix B). All seven pseudoautosomal genes, as expected, exhibited biallelic expression, with mostly similar levels of expression from both alleles corresponding to the PolIII occupancy detected at the associated heterozygous binding sites (Figure 16A). We noted two pseudoautosomal genes *IL3RA* and *CD99* that were relatively less expressed from the paternally-derived copy, regardless of the Xi origin (Appendix 2), which may reflect either a parent-of-origin effect or imbalance due to the particular allelic variants in these genes; further studies in other cell lines would be required to distinguish these possibilities. This situation is not unprecedented on the X



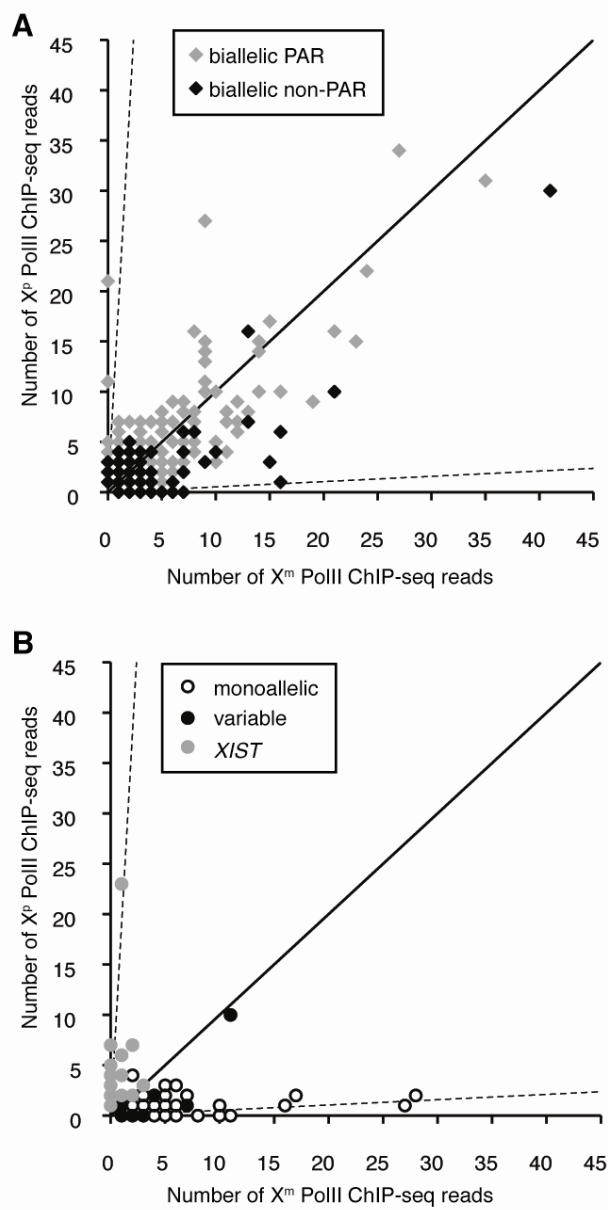
**Figure 14 PolII occupancy on Xi and Xa at the *DDX3X* gene**

(A) PolII ChIP-seq enrichment throughout the gene is shown with the highest enrichment near the TSS. Locations and rs numbers of heterozygous SNPs in GM12878 are indicated. (B) Relative PolII occupancy on the two X chromosomes in GM12878, as measured by the number of aligned ChIP-seq reads at each heterozygous SNP, mapping to the  $X^p$  (white bars) or  $X^m$  (black bars). The aggregated signal from the *DDX3X* locus is shown on the far right, indicating high PolII occupancy on the Xi, consistent with *DDX3X* biallelic expression (Appendix B).



**Figure 15 PolII occupancy on Xi and Xa at the *GNL3L* gene**

(A) PolII ChIP-seq enrichment throughout the gene is shown with the highest enrichment near the TSS. Locations and rs numbers of heterozygous SNPs in GM12878 are indicated. (B) Relative PolII occupancy on the two X chromosomes in GM12878, as measured by the number of aligned ChIP-seq reads at each heterozygous SNP, mapping to the  $X^P$  (white bars) or  $X^m$  (black bars). The aggregated signal from the *GNL3L* locus is shown on the far right, indicating complete lack of PolII binding at the  $X^P$ , consistent with the monoallelic expression detected at this gene (Appendix B).



**Figure 16 Expression of genes genomically associated with PolII occupied heterozygous sites**

(A) Allele-specific binding of PolII sites associated with biallelically-expressed genes (as designated in the key). (B) Allele-specific binding of PolII sites associated with monoallelically-expressed or variable genes (as designated in the key). Dashed axes and diagonal axis are as in Figure 2.

chromosome, as such allelic imbalance has been described at the *SYBL1* locus in PAR2 (260).

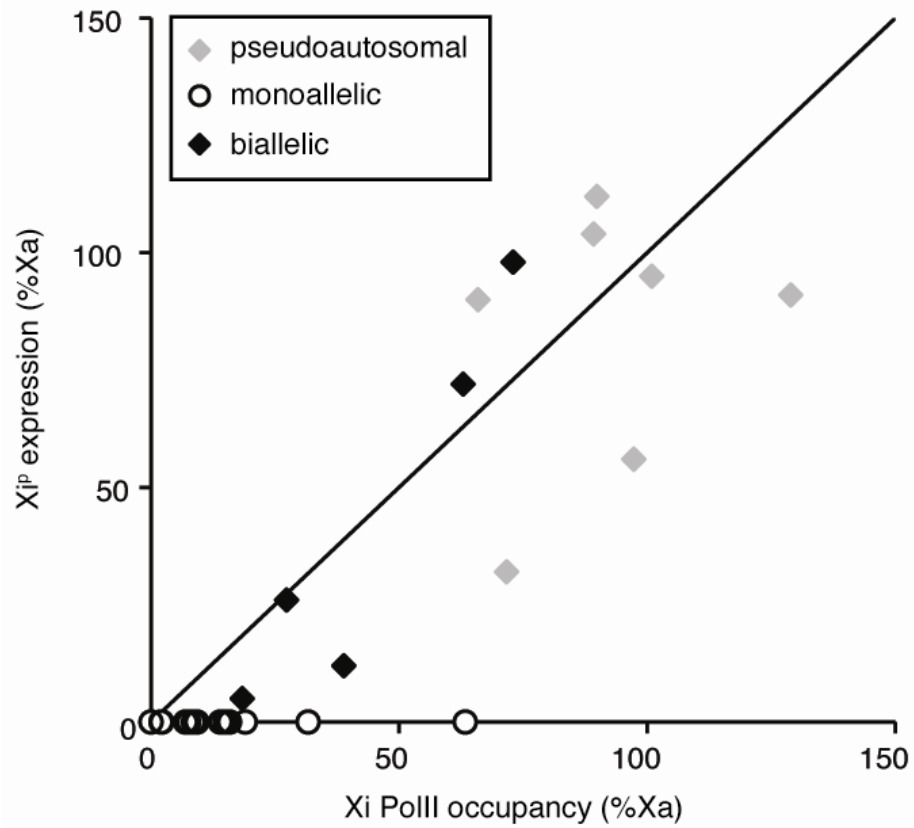
We detected several patterns of expression among the remaining 24 tested genes located on the non-pseudoautosomal portion of the X chromosome (Table 6). As expected, the majority (75%) of the tested genes were expressed monoallelically in both types of Xi isolates (all expressed exclusively from the Xa allele, other than *XIST*, which was expressed only from the Xi). In addition, five genes escaped inactivation (21%) at varying levels of expression, and one gene (*SEPT6*) exhibited variable expression between the two types of isolates (see Chapter 2).

When the gene expression profiles were related back to the individual PolIII-occupied heterozygous sites (allele-specific dataset of 385 PolIII sites, Appendix A), as well as to PolIII occupancy aggregated across annotated genes and the flanking regions (Figure 17, Appendix B), gene silencing on the Xi was found to correlate with PolIII depletion at most sites. However, several regions on the Xi<sup>P</sup> showed PolIII binding well above the expected ~5% level reflecting the X-inactivation skewing ratio. The relationship of PolIII binding at a specific site to a particular gene is difficult to establish, especially for sites located in intergenic regions; nevertheless, it appears that PolIII can bind at low levels even in regions where silenced genes reside, indicating higher accessibility of the transcriptional machinery to the Xi than previously thought.

**Table 6 Relationship of expression and PolII occupancy**

	<b>Genes</b>	<b>PolII at Xi (# of genes)</b>
Total	31	
Pseudoautosomal	7	7/7
Non-pseudoautosomal	24	
Monoallelic Xa	17	8 <sup>a</sup> /17
Monoallelic Xi	1	1/1
Biallelic	5	5/5
Variable	1	0/1

<sup>a</sup> PolII at Xi associated with monoallelically-expressed loci is defined as at least three reads mapping to Xi and >15% of Xa PolII occupancy



Among the five biallelically-expressed genes, Xi expression ranged from 5% to 98% relative to levels detected from the Xa allele (Table 7), showing strong correlation with levels of PolII occupancy at sites associated with those genes ( $R^2=0.9$ ) (Figure 17). In addition, the levels of expression for these five genes are consistent in the derivative cell lines containing the two different Xi chromosomes ( $Xi^p$  and  $Xi^m$ ), as well as in duplicate lines that were cultured separately for several weeks, indicating that the regulation of expression levels of these genes is largely constant and thus presumably reflects stability of the epigenomic environment of the alleles being compared. It is yet to be uncovered whether similar levels of expression are recapitulated in different individuals and whether they are biologically or phenotypically relevant.

Our analysis of gene expression is limited by the occurrence of transcribed informative SNPs in the cell line under study. Nonetheless, there were a number of PolII sites associated with genes that lacked expressed heterozygous SNPs, some of which have been studied previously and thus allow for some interpretation. For example, the *TIMM17B* and *RBM3* genes are inactivated in multiple human diploid cell lines (77), which is in agreement with the 95% Xa PolII occupancy in this study, suggesting that both genes are also inactivated in GM12878. In contrast, the *UBA1* gene contains three heterozygous PolII sites within its gene body, and, in the aggregate, 33% of reads at these sites map to the Xi allele. This finding is consistent with previous reports that *UBA1* escapes inactivation in all samples tested thus far (77, 135, 261).

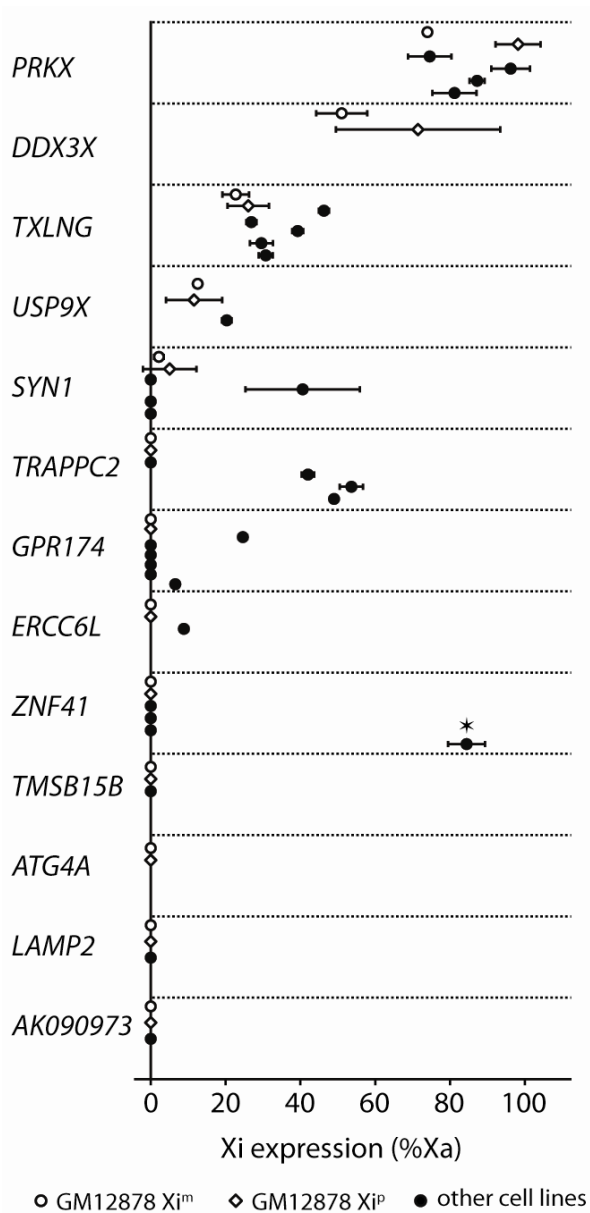
**Table 7 Expression and PolII binding on Xi relative to Xa**

	<b>Function</b>	<b>X<sup>P</sup> PolII (%Xa)</b>	<b>Xi<sup>P</sup> Expression (%Xa)</b>	<b>Y homology</b>
<i>SYN1</i>	Synaptogenesis and neurotransmitter release	19%	5%	---
<i>USP9X</i>	Peptidase C19 family, similar to ubiquitin-specific proteases	36%	12%	<i>USP9Y</i>
<i>TXLNG</i>	Intracellular vesicle trafficking, cell cycle	28%	26%	<i>CYorf15</i>
<i>DDX3X</i>	Alteration of RNA secondary structure	66%	72%	<i>DDX3Y</i>
<i>PRKX</i>	Serine-threonine protein kinase	75%	98%	<i>PRKY</i>

### 3.2.1. PolII occupancy indicates chromatin accessibility and potential for expression

As described above, PolII binding at genes that are not being actively transcribed indicates a level of openness of the local chromatin environment. We hypothesized that such genes have the potential to be expressed in other cell lines either consistently or variably with respect to the population. To test this hypothesis, we generated Xi-homogeneous cell populations (Table 2) from seven additional HapMap CEU cell lines (25, 26) and tested expression of 11 genes that were informative for study in these cell lines; four genes that were expressed biallelically in GM12878 and seven genes that were silenced in GM12878, yet showed significant PolII levels at the Xi allele (see Table 6).

In the case of the four genes that are biallelically-expressed in GM12878, we confirmed biallelic expression of *PRKX*, *TXLNG* and *USP9X* in all informative female lines tested, indicating that all three genes consistently escape inactivation (Figure 18). Furthermore, the relative levels of Xi expression of these three genes observed in GM12878 (Table 7) are recapitulated in the additional cell lines, with *PRKX* exhibiting highest Xi expression (~85% of Xa levels), *TXLNG* intermediate (~35% of Xa) and *USP9X* lowest expression (~20% of Xa). In contrast, *SYN1* is a gene with variable expression from the Xi. It escapes inactivation in GM12878 at a very low level, but is expressed robustly from Xi in another female cell line and is completely silenced in three other females (Figure 18).



**Figure 18 Xi gene expression in multiple cell lines**

Genes associated with PolII in GM12878 (shown on vertical axis), were tested for expression in GM12878 (open symbols) and additional heterozygous cell lines (closed circles). Error bars represent standard deviation of the mean between biological replicates. Expression of *ZNF41* has been tested previously in ten human cell lines. It was subject to inactivation in nine informative cell lines (not shown in the figure) and escaped inactivation in one cell line (designated by an asterisk).

Of the eight silenced genes associated with PolIII in GM12878 (Table 2), we show that three (*ERCC6L*, *GPR174* and *TRAPPC2*) are expressed from the Xi in other female cell lines (Figure 18), and one additional gene (*ZNF41*) has been reported to escape inactivation previously in at least one female cell line (77). Combined, these data demonstrate that PolIII not only binds to the Xi, but also can become engaged and produce a transcript in some individuals.

### 3.3. Discussion

In this study, we have examined the profile of PolIII binding on human female X chromosomes. Overall the frequency of PolIII binding on the X chromosome and chromosome 11 is reduced compared to other chromosomes, largely because of the lower proportion of expressed genes in GM12878 EBV-immortalized lymphoblastoid cell line (Figure 10). Although the low proportion of genes expressed from the X chromosome relative to autosomes largely explains the reduced frequency of PolIII binding, other factors will have to be explored to explain this phenomenon fully. As is evident here, many X-linked genes are not expressed in lymphoblastoid cell lines. These genes might have inactivation profiles relevant to tissue-specific diseases that cannot therefore be addressed in blood-derived cell lines and will have to be explored in other types of cells.

### **3.3.1. PolII binding on the Xi is reduced**

While the frequency of PolII binding on the X is lower than on nearly all autosomes (Figure 10A), interestingly, we detected no significant overall difference in the sequence read depth of the diploid PolII signals between the X chromosome and autosomes (Figure 11). Considering that PolII binding to the Xi is greatly reduced (Figure 12), this finding suggests overall increased PolII occupancy on Xa relative to autosomes. Furthermore, this finding confirms that the observed low number of PolII binding sites on the X chromosome relative to the autosomes (Figure 10A) is a biological phenomenon rather than simply reflecting decreased efficiency in peak detection. Although the low proportion of genes expressed from the X chromosome relative to autosomes largely explains the reduced frequency of PolII binding, other factors will have to be explored to explain this phenomenon fully.

Although overall PolII binding is reduced across the non-pseudoautosomal region of the X chromosome (Figure 12), the evolutionarily younger XAR spanning most of the Xp arm exhibits higher PolII occupancy than the Xq arm. This is consistent with the previously documented higher occurrence of genes that escape X inactivation and are expressed from the human Xi in the XAR (77). Here we present another piece of evidence that X inactivation is influenced by the evolutionary origins of the underlying sequences and show that PolII binds more readily at formerly autosomal sequences that were introduced to the X-inactivation system more recently (126, 129, 130, 259).

### 3.3.2. PolII occupancy and gene expression on the Xi

Our allele-specific analysis of PolII binding has identified five novel biallelically-expressed X-linked genes (Table 7 and Appendix B), some of which had been hypothesized previously to escape inactivation in studies performed in somatic cell hybrids or predicted by male/female dosage comparisons (262). For three of these genes b

, we were able to verify biallelic expression in multiple individuals. The proportion of biallelically-expressed genes in our study (21%) is consistent with previous estimates (77). Also consistent with previous observations is that all five of the biallelically-expressed genes are located on Xp, where escape from inactivation occurs more frequently than on the Xq (77). Our allele-specific analysis of these five genes expressed from the Xi suggests that, although the engaged phosphorylated form of PolII may be a more precise indicator for expression levels (246), the unphosphorylated PolII queried in our study (see Methods) is a suitable indicator of inactivation status of genes and relative gene expression levels.

We detected increased PolII levels at several sites along the Xi, indicating a potential for expression despite the apparent association with silenced loci. In all cases, the Xa homologue was actively transcribed, while there was no detectable transcription product from the Xi. PolII occupancy at the Xi allele was particularly striking at the *TRAPPC2* gene, where 67% of PolII reads aggregated across the locus mapped to the transcriptionally silenced Xi allele. Interestingly, as I have shown in Chapter 2,

inactivation of *TRAPPC2* is unstable and the gene is indeed occasionally expressed from the Xi in additional clones derived from GM12878 (Figure 7). Because the strongest PolII signal at the *TRAPPC2* locus was detected downstream from the gene (albeit within <1Kb from the 3' end and thus likely resulting from PolII progression beyond the 3' end), an alternative hypothesis that this PolII peak may be associated with another nearby gene should be considered. The *RAB9A* gene is transcribed in the opposite direction to *TRAPPC2*, and the 3' ends of the two genes are located <3Kb apart. *RAB9A* has a Y-linked homologue and has been observed to escape inactivation in hybrid cell lines (77); thus it is plausible that it escapes inactivation in GM12878 cell line as well. Because progression of PolII beyond the 3' end could result from either or both of these neighboring genes, it is difficult to assign this particular PolII signal to one gene or the other with certainty.

### **3.3.3. Genomic and chromatin features in X chromosome inactivation**

Because CpG methylation and heterochromatin markers have been associated with X-inactivation patterns (50-52, 250-252), it was of interest to compare our expression and PolII data with other chromatin and genomic features of the X publicly available through the ENCODE database (<http://genome.ucsc.edu/ENCODE/>). We failed to detect significant association of any particular class of genes with chromatin or genomic features such as presence of CpG islands or CTCF binding. We did, however, note a slight increase in association of PolII with H3K27me3 in gene bodies, indicating a

potential role that this modification could play in marking particular Xi genes for expression in the portions of the Xi that are enriched for this heterochromatin mark. It has been speculated previously that H3K27me<sub>3</sub>-marked heterochromatin might be less effective at silencing gene expression than H3K9me<sub>3</sub>-marked heterochromatin (50, 51, 108). To facilitate our understanding of X chromosome inactivation, further studies of multiple human females could take advantage of the vast datasets that are being generated by the ENCODE (199) and Human Epigenome (263) Projects. In combination with 1000 Genomes Project (27), future studies will be able to distinguishing features of the two homologous chromosomes in each individual to understand, for example, the role of H3K9me<sub>3</sub> and H3K27me<sub>3</sub> as well as many other histone modifications and chromatin features potentially implicated in X chromosome inactivation.

#### **3.3.4. Some genes on the inactive X chromosome are poised for expression**

While the finding of sites of PolII occupancy on the Xi near genes that are subject to inactivation may appear surprising, the large number of previously identified genes with variable expression patterns (77, 80-82) is consistent with the idea that some genes on the Xi are poised for transcription in some individuals and/or cell lineages. Sites with increased PolII binding on the Xi identified in our study were candidate loci for variable expression patterns and indeed, when assayed in additional informative cell lines, we

found that a high proportion of genes with Xi PolII association are occasionally expressed from the Xi in the population (Figure 18).

Most genes that are expressed from the Xi generate RNA output that is reduced relative to the Xa (77), an indication that, although expression is generally suppressed chromosome-wide, the Xi is permissive to the transcriptional machinery. However, it remains unclear whether these low expression levels result from leakiness of the inactive state, with little or no phenotypic consequence in the context of full expression from the Xa copy, or, more provocatively, they reflect, at least in some cases, purposeful turning on the Xi allele to finely tune the overall genetic output. It is likely that there are numerous gene-specific control mechanisms among the ~1,200 genes on the X, superimposed upon the chromosome-wide and/or regional landscapes anticipated by the process of X inactivation. Indeed, the contrast in patterns of PolII distribution and gene expression between the XAR and XCR point to the interplay of evolutionary, genomic, and local effects that remain to be fully explored in this context.

### **3.4. *Materials and methods***

#### **3.4.1. X-inactivation skewing**

In a previous study (34), we reported 92% skewing of X inactivation in the GM12878 cell line toward the Xi<sup>P</sup>. Because X-inactivation skewing can drift in cell culture

over time (216, 256), we verified the extent of skewing in GM12878 in RNA samples harvested concurrently with the PolII ChIP-seq samples generated for this study. We detected 95% (SD=1%) skewing in the same direction as previously described (34), by testing allele-specific expression at *XIST* (rs1620574) and *EBP* (rs3048) by SNaPshot assays (77).

### **3.4.2. Cell culture and single-cell cloning**

As described in Chapter 2

### **3.4.3. ChIP-seq**

PolII ChIP was performed on GM12878 cells, and Illumina-based sequencing was carried out as described previously (23, 264) with minor modifications. GM12878 cells were fixed, lysed, and nuclei were collected. After lysing nuclei, chromatin was sonicated and immunoprecipitated with 8WG16 antibody to PolII (Covance, MMS-126R). Afterward, protein:DNA crosslinks were reversed overnight at 65°C, and DNA was isolated using a Qiagen PCR Cleanup column. DNA was prepared for sequencing on an Illumina Genome Analyzer as described previously, modified to eliminate PCR prior to size selection and to use 15 cycles of PCR after size selection. Libraries were then sequenced to a minimum depth of 12 million 36-nucleotide reads per biological replicate.

#### **3.4.4. Measuring allele-specific PolII occupancy**

To map allele-specific occupancy of PolII sites, we constructed a GM12878 parent-specific version of the human reference genome (hg18), as described in (209). We aligned ChIP-seq reads to the modified reference genome using the Bowtie aligner (265) and removed any alignments that mismatched at a known heterozygous SNP position.

To detect allelic bias of PolII occupancy at each SNP, we counted the number of reads mapping specifically to a paternal or maternal allele at all heterozygous positions and calculated a p-value according to a binomial model that assumes equal likelihood of each allele.

#### **3.4.5. SNaPshot**

A quantitative Q-SNaPshot assay was employed to test the abundance of each allele in the PCR amplicon, using protocols as described previously (51, 77). Modifications specific to this thesis are described in Chapter 2.

## **4. Conclusions and Future Work**

The aim of this thesis work was to explore genetic influences on inactivation patterns on the human X chromosome. I have employed two approaches to address this question. First, I investigated genes with variable inactivation patterns among females in a large family and in the population (Chapter 2); and second, I identified genes associated with PolII binding in the GM12878 human female lymphoblastoid cell line and compared expression of these genes on the two Xi chromosomes in that cell line (Chapter 3).

No strictly heritable or parent of origin expression patterns were uncovered in this study (Table 4); however, I found that individual genes exhibit a number of distinct expression characteristics that distinguish them. For example, relative expression levels of genes that consistently escape inactivation in the population are remarkably stable among females (Figure 18), suggesting that a precise contribution of the Xi transcript might be required for proper function of these genes. In the case of genes with variable expression, some such as *SEPT6* (Figure 8, 9), *ZNF185* or *ZNF41* are highly expressed from the Xi in some individuals and completely silenced in others, while other genes, such as *CLIC2* (Figure 6), *TBL1X*, *MORF4L2*, *ARHGAP4* and *ASB11* (Table 4) escape inactivation only rarely and at very low levels. These patterns of expression activity are also recapitulated in the PolII profile across the Xi (Figure 12). Low level of PolII occupancy can be observed at a number of sites on the Xi, correlating well with occasional low-level gene expression. Several genes however, are associated with

significant PolII binding that is indicative of high relative expression levels seen either consistently or variably with respect to the population (Figure 18).

I conclude that these patterns are indicative of a multi-level regulation system, in which the chromosome-wide inactivation, facilitated by *XIST* and subsequent chromatin changes, is fine-tuned at the local level. At some sites X inactivation appears to be 'leaky', resulting in occasional low levels of expression (~5-10% of the level of the Xa homologue) that might not be biologically relevant. At other loci however, expression from the Xi appears to be tightly regulated, potentially reflecting a genetic effect, where some variants might be expressed from the Xi while others are not, or reflecting insufficient upregulation of the Xa homologue that might necessitate expression from the Xi to achieve proper dosage. Both of these hypotheses can be addressed with the system and techniques developed here in broader population and family studies (see below).

This thesis illustrates how the recent genome, transcriptome and chromatin projects (27, 210), and the vast data produced by these efforts, can also be utilized and serve as a basis for studies aimed to better understand the dynamics of the two X chromosomes in human females.

In studies of X chromosome inactivation over the past decade, we have transitioned from exploring the human Xi in human-mouse somatic cell hybrids (80) to detecting allele-specific inactivation of individual genes (77) in human cell lines derived from individuals with completely skewed inactivation ratios, to deriving pure lineages

with respect to the inactivated X chromosome from any female cell line (266). The latest approach offers significant advantages, as one can select samples for which extensive genomic, transcriptome and chromatin information is available, allowing one to maximize the amount of information that can be obtained from such cell lines. The advent of rapid and cost-efficient genome sequencing to detect all variation in any genome is essential for allele-specific approaches necessary in studies of such phenomena as X-chromosome inactivation, because of the ability to distinguish between the two alleles at a maximum number of loci. Here, we have taken advantage of the extensive sequencing 1000 Genomes Project (27) and a whole-genome PolII ChIP-seq data set (209) that allowed us to compare the two Xi chromosomes in a normal female cell line at high resolution. In this thesis, to complement this whole-chromosome PolII ChIP-seq approach, we have employed a gene-by gene expression profiling to answer specific biological questions. Because of the recently developed RNA-seq and ChIP-seq technologies, this approach can now be scaled to population and family studies addressing the differences between the Xi and Xa not only in expression, but also in various chromatin features, to better understand the genetic, genomic and stochastic factors that influence X chromosome inactivation in humans.

#### *4.1. Dosage compensation of SEPT6*

#### 4.1.1. Introduction

Expression levels from the Xi relative to Xa for genes that consistently escape inactivation appear to be stable from individual to individual, yet different from gene to gene (Figure 18). This observation is consistent with the idea that inactivation has occurred on a gene-by-gene basis as a response to degradation to the Y chromosome, where the rate of upregulation of the Xa homologue corresponds to the rate of downregulation of the Xi homologue, such that the resulting level of RNA gene product is unchanged and therefore dosage compensated (64). If dosage compensation were required for genes with variable expression in the population and if expression from the Xi were to supplement insufficient expression levels from the Xa, one could expect that expression from the Xi might supplement lower expression from the Xa in some individuals.

The level of *SEPT6* expression from the Xi (Figure 6) varies among individuals, yet it is stable in multiple cell populations derived from a single individual; thus the level of *SEPT6* expression appears to be stable within a cell line, yet the levels of expression from each allele may vary in the population, potentially as a result of genetic variation or *trans* factors affecting the expression of *SEPT6*. Because *SEPT6* is on the X chromosome, this potential *trans* regulation might be involved in dosage compensation of the gene or potentially of a cluster of genes.

Septins are a family of GTP-binding proteins that self-assemble into higher-order oligomeric structures that can further assemble into filaments in a head to head fashion (267), suggesting that expression of septins may be under strict regulation in order to generate a precise dosage of septin protein molecules. Expression of *SEPT6* is coordinated with that of *SEPT2* and *SEPT7*, producing protein units that eventually form uniformly sized perfect palindrome heteromers (268-270). Mammalian septins are involved in chromosome segregation, cytokinesis, DNA damage response, cell migration, membrane dynamics, exocytosis and apoptosis (271). Humans have 13 functional septin genes categorized into four subgroups (272). Although expression of *SEPT6* is coordinated with other complex members, monomeric SEPT6 protein is quickly degraded (273), and therefore regulated posttranscriptionally, indicating that precise dosage compensation at the transcript level may not be necessary.

#### **4.1.2. Experimental design**

In future experiments, one could compare total levels of *SEPT6* transcript in the ten cell lines with known relative levels of expression from the Xi (Figure 6) by real time RT-PCR. The level of transcript could be normalized to three or four reference genes with high expression stability, such as *GAPD*, *SDHA*, *YWHAZ*, *B2M*, *INTS4*, *ZNF410*, *BUD13* (274). Comparison of the 10 samples, some of which show various expression levels from the Xi and others of which are completely silenced, would allow one to

determine whether the variation in expression of the Xi *SEPT6* homologue is due to dosage compensation in different individuals.

#### 4.2. *Paired Xi<sup>p</sup> and Xi<sup>m</sup> study in multiple cell lines to identify the level of heritability and cis/trans-acting elements influencing patterns of X inactivation*

##### 4.2.1. Introduction

In this work, I have described two distinct cell populations derived from a single cell line (GM12878) that are pure with respect to the inactive X chromosome, Xi<sup>p</sup> and Xi<sup>m</sup>, that they clonally propagate. In addition to the GM12878 cell line, I have derived paired pure Xi populations from another three cell lines (Table 2) that can now be utilized to perform pairwise comparative studies of Xi chromosomes existing in identical genomic backgrounds. Comparison of the two Xi chromosomes in a single cell line at the level of DNA sequence, expression and chromatin features allows for exploration and discovery of the *cis*-acting elements responsible for the unique features of two Xi chromosomes while eliminating the influences of heterogeneous genomic backgrounds. Furthermore, because the three additional cell lines that I have derived from three sisters in family 1454 are unrelated to GM12878, it is possible to examine the level of heritability in X chromosome inactivation patterns. If necessary, further pure Xi<sup>p</sup> or Xi<sup>m</sup> cell populations derived from unrelated females (Table 2) can also be used in the study.

#### 4.2.2. Experimental Design

Allele-specific expression and chromatin experiments to detect inactivation patterns on both Xi chromosomes could be performed on paired cell populations of (Xi<sup>m</sup> and Xi<sup>p</sup>) established from cell lines derived from families. For example, I have already derived paired cell lines for three sisters in family 1454 and GM12878 cell line. Chromatin studies could include, but are not limited to PolIII and CTCF, histone modifications such as H3K9me3 and H3K27me3 and others. The experiments proposed here would allow one to:

- a) Catalogue new genes as subject, escape or variable with respect to their inactivation status and thus further improve the current Xi gene expression profile
- b) Refine the boundaries of escape and subject domains
- c) Detect Xi-specific chromatin characteristics associated with genes and gene domains
- d) Compare the two Xi chromosomes in each cell line to assess the level of X inactivation variability in the same genetic background
- e) Compare inactivation on Xi<sup>p</sup> chromosomes in the three sisters to assess the level of X inactivation variability on the same X chromosome in different genetic backgrounds

### 4.2.3. Preliminary data

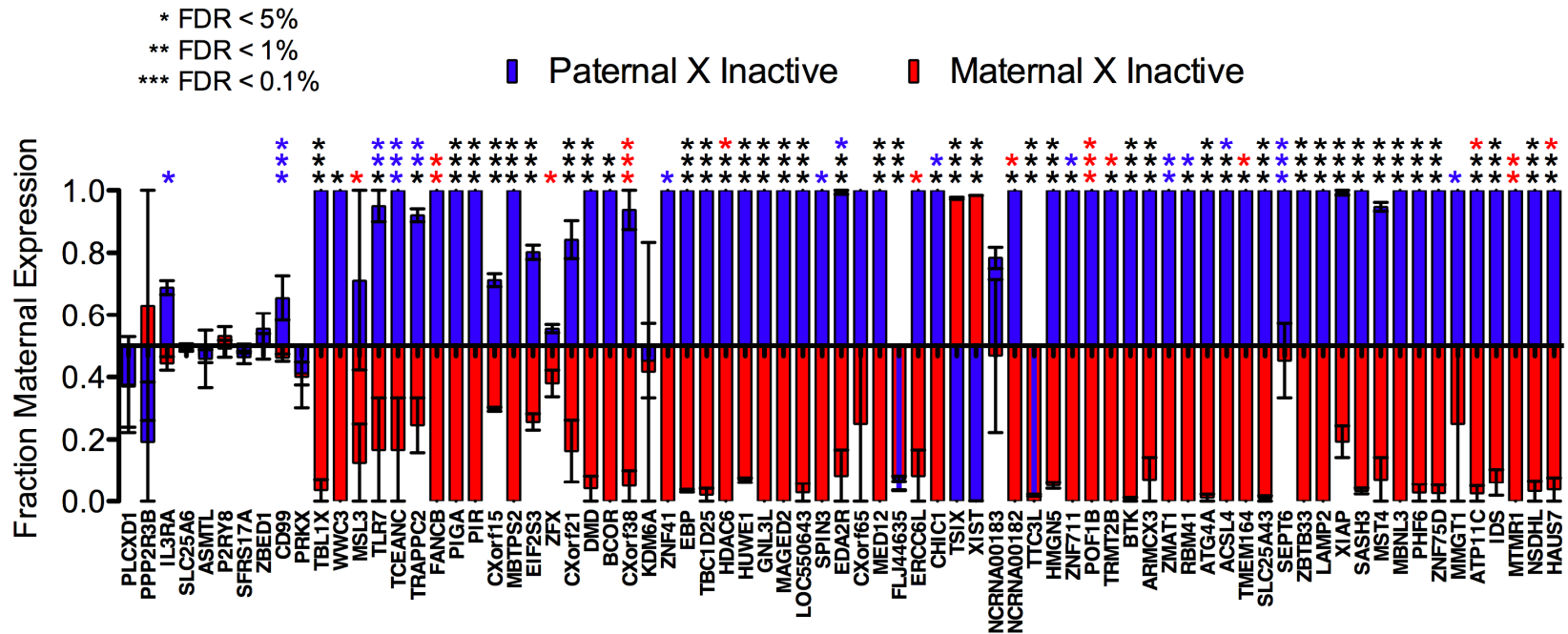
Relevant to the potential future experiments, we have already performed several pilot experiments in GM12878-Xi<sup>m</sup> and GM12878-Xi<sup>p</sup> derived cell populations comparing gene expression (Figure 19), RNA PolIII occupancy (Figure 20), DNA methylation and CTCF binding, in collaboration with the Myers Laboratory at Hudson Alpha Institute in Huntsville AL, and experiments that detect sites of open chromatin in collaboration with the Crawford/Furey group in the IGSP at Duke. The bioinformatic analysis was performed by Dr. Tim Reddy, currently at Hudson Alpha Institute.

It is evident from these experiments that all levels of chromatin, DNA methylation and expression have features specific to X chromosome inactivation, as the bias observed at a large number of sites is to the opposite alleles in the two distinct Xi<sup>m</sup> and Xi<sup>p</sup> cell populations. Interestingly, however, a number of sites exist that appear to show incomplete bias in expression (Figure 19), demonstrating either consistent biallelic expression from both Xi alleles or variance between the two alleles. This incomplete inactivation and/or asymmetric inactivation between the two Xi alleles is even more pronounced in the clonal PolIII X chromosome profiles (Figure 20), further supporting our conclusion in Chapter 3 that the chromatin of the Xi is accessible to PolIII, allowing the enzyme to bind at a number of sites and remain poised for expression.

Caution has to be taken when aligning sequence reads to the genome, such that they are assigned to the correct location and allele. For example, some X-linked genes have pseudogenes with high sequence similarity and a failure to distinguish between

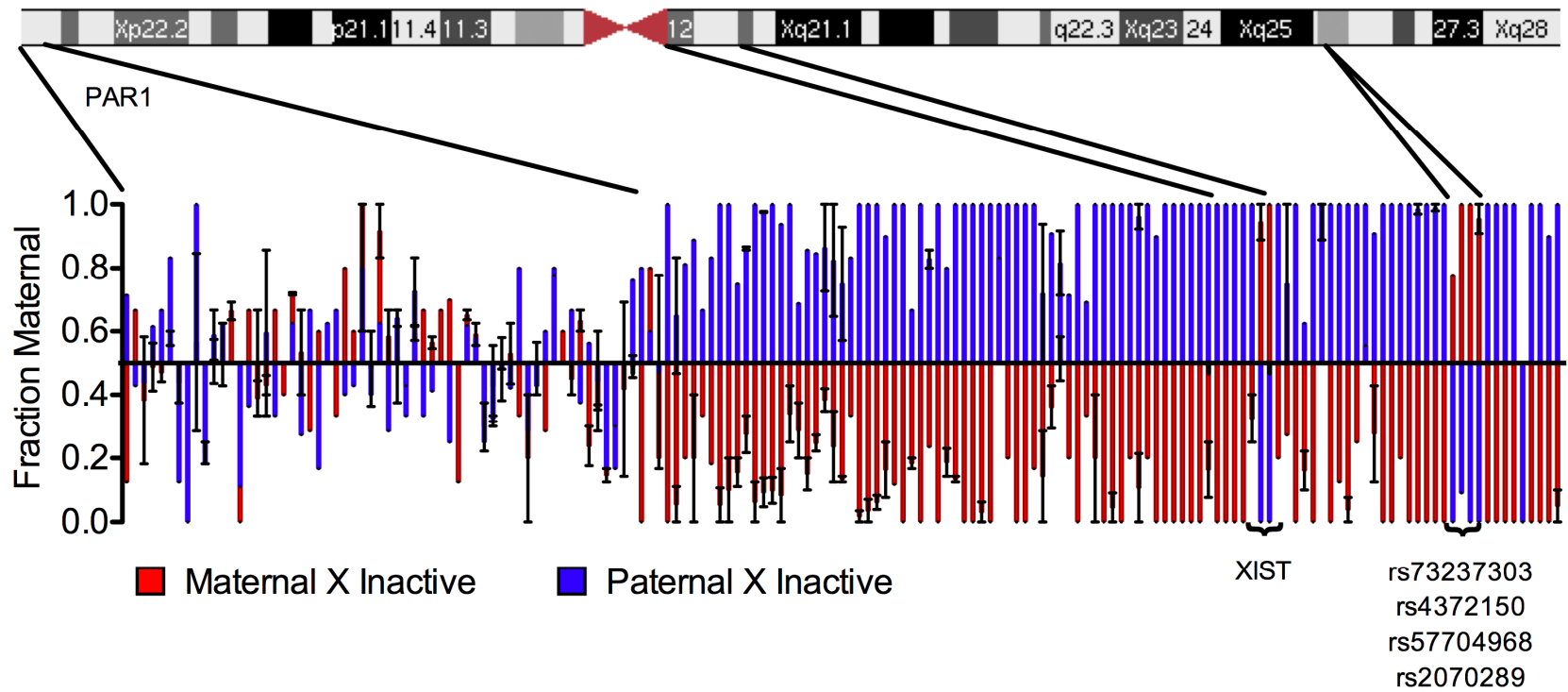
those might distort the reported allelic bias. SNaPshot validation of allelic bias might be necessary to confirm any unexpected or questionable results.

It is unknown whether these allelic differences have a genetic basis or are a result of normal stochastic processes occurring in the cell. By comparing X inactivation profiles in a sufficient number of related and unrelated individuals, it should be possible now to better understand the genetic contribution to the variable features of the human Xi.



**Figure 19** Biased allele-specific expression of X-linked genes in GM12878 Xi<sup>m</sup> and Xi<sup>p</sup> cell populations

Genes containing expressed heterozygous SNPs with significant RNA expression ( $\geq 7$  RNA-seq reads/SNP site) are included in the figure. Allelic bias is shown for two replicates of each Xi<sup>p</sup> (blue) and Xi<sup>m</sup> (red) isolates in the relative order from left to right as they appear on the X chromosome. Significance of bias expression as indicated by false discovery rate (FDR) is shown by number of stars above each expression bar color-coded according to the tested X chromosome, Xi<sup>p</sup> (blue stars) and Xi<sup>m</sup> (red stars). Black stars indicate significant allelic bias for both Xi chromosomes.



**Figure 20** Biased allele-specific PolII occupancy on the X chromosome in GM12878  $Xi^m$  and  $Xi^p$  cell populations

Sites with significant PolII binding ( $\geq 7$  PolII ChIP-seq reads/SNP site) are included in the figure. Allelic bias is shown for two replicates of each  $Xi^p$  (blue) and  $Xi^m$  (red) isolates in the relative order on the X chromosome from left to right. Sites within the *XIST* genes are highlighted. Rs numbers for sites with  $Xi$  bias (other than *XIST*) are indicated.

With the ever decreasing cost and increasing efficiency of genome and transcriptome sequencing, the next logical step in the effort to refine the profile of the human Xi is to generate a population size pure Xi sample resource, using the single cell cloning method developed here, that can then be interrogated for allele-specific expression to obtain a complete profile of the human Xi across population.

Furthermore, when single cell RNA-seq becomes feasible the burden of single-cell cloning of transformed cell lines can be eliminated and X inactivation studies can be performed in human tissue samples to refine not only the population profile of the Xi but also detect tissue-specific X inactivation phenomena. Because it appears that only a subset of genes have strict inactivation requirements in all females and/or in all tissues, such population studies are essential to determine which genes are strictly inactivated and which X-linked genes are strictly expressed at a specific level from the Xi as aberrant inactivation patterns at those genes are more likely to result in phenotypic consequences.

## Appendix A: Allele-specific dataset of 385 PolII sites

Heterozygous PolII sites that had at least five mapped reads on at least one of the two X chromosomes in the PolII ChIP-seq dataset are included. Columns indicate in order: (A) chromosome X, (B) genomic position on chromosome X, (C) SNP identification, number of reads mapping to the SNP aligning to the paternal (D) (pat reads) and maternal (E) (mat reads) allele.

chr	Position	SNP ID	pat reads	mat reads	gene
chrX	152125	rs28490293	5	1	
chrX	268312	rs28485241	5	7	
chrX	909384	NA12878.395306	0	5	
chrX	1431754	rs17880161	4	7	IL3RA
chrX	1432118	NA12878.395433	3	5	IL3RA
chrX	1457133	rs6645279	5	1	IL3RA
chrX	1457644	rs17883366	4	5	IL3RA
chrX	1457687	rs17883130	7	8	IL3RA
chrX	1460009	rs17883884	10	10	IL3RA
chrX	1460398	rs28585015	15	9	IL3RA
chrX	1461471	rs11480	22	24	IL3RA
chrX	1464278	rs28540518	21	27	
chrX	1466681	rs6644952	10	16	SLC25A6
chrX	1466719	rs35187644	14	14	SLC25A6
chrX	1467305	rs4548374	15	23	SLC25A6
chrX	1467447	NA12878.395438	21	0	SLC25A6
chrX	1467976	rs5989744	9	19	SLC25A6
chrX	1468324	rs7205	10	14	SLC25A6
chrX	1468583	rs14005	14	9	SLC25A6
chrX	1469574	rs5989758	16	8	SLC25A6
chrX	1469993	rs35040302	10	9	SLC25A6
chrX	1471328	rs35907463	27	39	
chrX	1472101	rs61174754	2	6	
chrX	1505209	NA12878.395459	5	8	ASMTL
chrX	1505215	rs28528112	5	7	ASMTL
chrX	1505291	rs5949085	5	2	ASMTL
chrX	1505299	NA12878.395460	5	3	ASMTL
chrX	1506051	rs4292893	3	10	ASMTL
chrX	1506198	rs5949095	3	5	ASMTL
chrX	1506370	rs5989906	3	5	ASMTL
chrX	1506601	rs35193769	8	8	ASMTL
chrX	1506876	rs5948863	6	8	ASMTL

chrX	1510697	rs4500242	3	7	ASMTL
chrX	1510714	rs28623006	3	7	ASMTL
chrX	1510760	rs5989937	6	7	ASMTL
chrX	1510779	rs5989939	2	6	ASMTL
chrX	1513872	rs45465401	4	7	ASMTL
chrX	1514142	rs45498107	2	5	ASMTL
chrX	1515085	rs6644941	8	8	ASMTL
chrX	1517304	rs28653052	6	8	ASMTL
chrX	1517307	rs28537349	7	12	ASMTL
chrX	1518206	rs28626010	9	12	ASMTL
chrX	1530513	rs4933150	3	6	ASMTL
chrX	1536319	NA12878.395464	8	8	
chrX	1536383	NA12878.395465	9	0	
chrX	1543443	rs4933152	7	4	P2RY8
chrX	1545844	rs28450615	1	5	P2RY8
chrX	1546301	NA12878.395468	3	6	P2RY8
chrX	1546426	rs28680859	2	6	P2RY8
chrX	1547212	rs28391357	5	4	P2RY8
chrX	1550657	rs5949204	7	5	P2RY8
chrX	1550833	rs4303028	4	7	P2RY8
chrX	1551249	rs55691527	0	6	P2RY8
chrX	1551751	NA12878.395471	1	5	P2RY8
chrX	1552606	rs55710683	0	5	P2RY8
chrX	1553071	NA12878.395474	5	1	P2RY8
chrX	1554606	rs6644956	5	0	P2RY8
chrX	1557142	rs5990010	5	4	P2RY8
chrX	1557721	rs6644960	4	8	P2RY8
chrX	1559276	rs5948907	3	5	P2RY8
chrX	1559339	rs5948909	6	12	P2RY8
chrX	1561382	rs3922990	17	15	P2RY8
chrX	1562967	rs34487068	11	9	P2RY8
chrX	1572820	rs28567826	15	14	P2RY8
chrX	1573095	rs6645235	16	21	P2RY8
chrX	1573434	rs6645256	5	7	P2RY8
chrX	1573590	rs6644718	6	1	P2RY8
chrX	1574248	rs5948914	0	6	P2RY8
chrX	1578403	NA12878.395480	5	4	P2RY8
chrX	1580151	rs28668537	10	10	P2RY8
chrX	1583417	rs6645262	7	11	P2RY8
chrX	1586210	NA12878.395482	11	0	P2RY8
chrX	1586244	rs4073115	13	9	P2RY8
chrX	1586271	rs4073116	5	6	P2RY8
chrX	1586541	rs60352406	9	6	P2RY8
chrX	1586839	rs62603032	8	11	P2RY8

chrX	1586845	rs62603050	8	13	P2RY8
chrX	1600371	rs6644727	9	7	P2RY8
chrX	1600573	rs6644728	27	9	P2RY8
chrX	1600746	NA12878.395486	5	6	P2RY8
chrX	1603897	NA12878.395488	3	9	P2RY8
chrX	1604333	rs28623387	7	3	P2RY8
chrX	1604843	rs5948937	0	5	P2RY8
chrX	1605215	NA12878.395489	7	1	P2RY8
chrX	1606126	NA12878.395491	1	6	P2RY8
chrX	1606136	NA12878.395492	2	5	P2RY8
chrX	1606744	NA12878.395494	6	3	P2RY8
chrX	1607199	NA12878.395496	6	5	P2RY8
chrX	1607697	NA12878.395497	4	11	P2RY8
chrX	1607809	NA12878.395498	7	2	P2RY8
chrX	1608516	rs5948941	8	7	P2RY8
chrX	1608518	NA12878.395501	8	8	P2RY8
chrX	1610295	NA12878.395506	4	5	P2RY8
chrX	1610532	rs60235956	6	3	P2RY8
chrX	1610593	NA12878.395507	4	7	P2RY8
chrX	1610663	NA12878.395508	6	7	P2RY8
chrX	1610702	NA12878.395509	8	5	P2RY8
chrX	1610754	NA12878.395510	5	10	P2RY8
chrX	1612018	rs4418527	0	5	P2RY8
chrX	1613639	NA12878.395514	34	27	P2RY8
chrX	1613651	NA12878.395515	31	35	P2RY8
chrX	1671165	rs7881311	3	9	AKAP17A
chrX	1673021	rs6644621	6	7	AKAP17A
chrX	1674613	rs6644765	6	3	AKAP17A
chrX	1677271	rs6644622	7	6	AKAP17A
chrX	1681163	rs11553959	2	6	AKAP17A
chrX	1682047	rs28567840	6	7	
chrX	1682055	NA12878.395524	7	8	
chrX	1683199	rs28610575	14	11	
chrX	1685002	rs5989833	5	2	
chrX	1685134	rs28506771	2	5	
chrX	1689417	rs60374201	0	5	
chrX	1689420	rs59448123	0	5	
chrX	1689429	rs57272468	0	5	
chrX	2402295	NA12878.395798	0	5	
chrX	2449652	rs5982723	3	5	ZBED1
chrX	2641035	rs311067	1	5	CD99
chrX	3531878	rs11152533	5	2	PRKX
chrX	3628132	rs741422	1	16	PRKX
chrX	9643443	rs2238876	0	5	

chrX	9950396	NA12878.396714	1	5
chrX	10026068	rs7879307	0	6
chrX	10039196	rs756827	0	15
chrX	10058377	rs12846942	0	6
chrX	10075697	rs5979264	0	5
chrX	11687212	NA12878.396835	0	6
chrX	11693212	NA12878.396836	4	10
chrX	11721135	rs858078	0	5
chrX	12817579	rs3853839	2	8
chrX	12819487	rs850632	7	13
chrX	12877811	rs1947953	45	114
chrX	12878568	rs882814	9	22
chrX	12879145	rs883812	8	20
chrX	12884527	rs936666	2	15
chrX	12886456	rs9284569	1	7
chrX	12899349	rs1483192	0	5
chrX	12901928	rs936668	1	29
chrX	12904171	rs955279	18	105
chrX	12906233	rs12690381	13	110
chrX	12907390	rs1483191	6	26
chrX	12912828	rs5934048	1	12
chrX	12927659	NA12878.396914	5	21
chrX	12928298	rs5935449	2	12
chrX	12933250	rs850639	2	13
chrX	12933645	rs850638	2	10
chrX	12933662	rs850637	3	8
chrX	12933745	rs850636	5	11
chrX	12934000	rs850635	10	18
chrX	12936432	rs2074111	15	13
chrX	12937556	rs5979772	5	8
chrX	13004350	rs5935497	0	5
chrX	13004524	rs5935498	1	8
chrX	13004525	rs5935499	1	7
chrX	13010453	rs5935501	9	21
chrX	13012031	rs4830486	1	14
chrX	13017248	NA12878.396918	2	5
chrX	13019909	rs2013673	2	5
chrX	13020171	rs12687333	5	14
chrX	13020193	rs12688657	5	15
chrX	13020688	rs4456042	2	17
chrX	13331560	rs6632816	5	0
chrX	13331569	rs12849398	5	0
chrX	13401796	NA12878.396951	1	14
chrX	13416299	rs5934106	1	7

chrX	13597764	rs4830886	3	19	
chrX	13598825	rs5979930	4	22	
chrX	13599405	rs5979931	0	7	
chrX	13599666	rs5979934	2	11	
chrX	13618099	rs4528031	3	10	
chrX	13620654	NA12878.396968	1	5	
chrX	13620854	rs4830888	1	7	
chrX	13623983	rs12558341	0	6	
chrX	13626075	rs7063528	0	5	
chrX	13639413	rs5979951	10	11	TRAPPC2
chrX	14602126	rs6630812	0	5	
chrX	15086954	rs4830925	0	5	
chrX	15263701	NA12878.397109	0	9	
chrX	15421333	rs2271550	1	10	
chrX	16699576	rs6632891	1	12	
chrX	16699849	rs6632893	0	8	
chrX	16699864	rs7883420	2	11	
chrX	16714376	rs12401124	3	9	Cxorf15
chrX	16723023	rs7888607	0	5	Cxorf15
chrX	16764039	rs4828560	0	6	Cxorf15
chrX	16770055	rs3747366	0	7	Cxorf15
chrX	16773510	rs4828535	0	5	RBBP7
chrX	16776662	rs5924560	0	5	RBBP7
chrX	16778261	rs55938758	0	6	RBBP7
chrX	16797576	NA12878.397205	1	16	RBBP7
chrX	17665102	NA12878.397238	0	5	
chrX	19515372	NA12878.397321	0	6	
chrX	19671711	NA12878.397326	0	5	
chrX	19702351	rs6633305	0	5	
chrX	19703322	NA12878.397327	0	6	
chrX	21767512	rs6528055	2	28	MBTPS2/YY2
chrX	21991618	rs5904505	0	5	
chrX	23595885	rs496067	0	15	
chrX	23596108	rs557914	1	9	
chrX	23596986	rs528683	0	5	
chrX	23628019	rs11797996	0	6	
chrX	23671924	rs11798220	0	5	
chrX	23709659	rs6526342	2	5	
chrX	23984502	rs4969555	5	4	
chrX	23986566	rs7059750	5	0	
chrX	23990190	rs5949272	4	6	
chrX	23991534	rs7065853	1	12	
chrX	23995031	rs6526367	3	7	
chrX	23996845	rs12556742	2	7	

chrX	23998873	rs16997670	3	5	
chrX	24002625	rs2318792	3	7	
chrX	24007459	rs12833088	2	5	
chrX	24007990	rs11094948	4	7	
chrX	24080557	rs2704816	1	6	
chrX	24114631	rs2704829	2	8	
chrX	24143830	rs13679	0	5	
chrX	24144295	rs5990013	2	6	
chrX	24171506	rs4285645	1	7	
chrX	24440863	rs5986552	0	5	
chrX	29482007	rs12014473	1	5	
chrX	29483476	rs4829224	3	20	
chrX	30177470	rs5973801	0	5	
chrX	30485609	rs2532872	0	5	
chrX	30487767	rs887369	0	5	
chrX	31271211	rs5927009	0	5	
chrX	38545987	rs41305747	1	8	
chrX	39081343	rs10126579	0	5	
chrX	39236771	rs5963661	0	5	
chrX	40048111	rs5963176	0	6	
chrX	40754577	rs4827249	2	7	
chrX	40829245	rs35884638	10	21	USP9X
chrX	40854617	NA12878.399024	0	5	USP9X
chrX	41076460	rs4827267	5	2	DDX3X
chrX	41079317	rs6609130	6	16	DDX3X
chrX	41079767	NA12878.399040	3	15	DDX3X
chrX	41081361	rs4358953	6	8	DDX3X
chrX	41083400	rs2275943	4	10	DDX3X
chrX	41083659	rs6610545	7	13	DDX3X
chrX	41088063	rs2275944	4	7	DDX3X
chrX	41095667	rs4142483	30	41	DDX3X
chrX	41096980	rs1467319	6	7	DDX3X
chrX	41097265	rs5963958	16	13	DDX3X
chrX	41187349	rs1794672	0	10	
chrX	45501025	rs17310797	0	5	
chrX	46938674	rs56269549	4	6	
chrX	46951274	rs4529579	0	5	
chrX	46955899	rs5953012	1	5	
chrX	46971291	rs6417923	1	9	
chrX	46991051	rs5906360	1	7	
chrX	47226829	rs2071779	0	8	ZNF41
chrX	47300801	rs5906428	1	14	
chrX	47308129	rs2854420	0	8	
chrX	47308426	rs56014229	0	6	

chrX	47315401	rs2854412	0	5	
chrX	47326992	rs55990337	1	6	SYN1
chrX	47331597	rs6609534	2	7	SYN1
chrX	47395742	NA12878.399421	0	8	
chrX	48252484	rs12559784	1	6	
chrX	48252679	rs3810678	0	7	
chrX	48252685	rs3810679	0	6	
chrX	48318443	rs2249583	3	25	
chrX	48318473	rs2249585	2	27	
chrX	48321962	rs235827	1	41	
chrX	48322706	rs7879722	3	21	
chrX	48322846	rs7879764	1	32	
chrX	48323323	rs235826	0	7	
chrX	48639881	rs55721510	1	20	
chrX	49015967	rs4824747	0	6	
chrX	53760841	rs6638366	0	5	
chrX	54576333	NA12878.399720	0	8	GNL3L
chrX	54590911	rs5960311	0	5	GNL3L
chrX	54591912	NA12878.399724	0	8	GNL3L
chrX	54592045	NA12878.399725	0	5	GNL3L
chrX	54592832	rs7066297	0	5	GNL3L
chrX	54593074	rs5960312	0	10	GNL3L
chrX	54593131	rs5960313	0	6	GNL3L
chrX	54682460	rs5960325	0	10	
chrX	54682675	rs4275347	1	21	
chrX	56616407	rs5960739	0	6	
chrX	56619535	rs2130428	0	7	
chrX	56773650	rs1967450	0	5	
chrX	56808436	rs5960793	8	6	
chrX	57330082	rs2516023	0	6	
chrX	64835059	rs56204907	1	12	
chrX	67650509	rs4827580	1	5	
chrX	68302236	NA12878.400442	0	7	
chrX	70204114	rs12851337	0	5	
chrX	70238356	rs5980742	0	8	
chrX	70241122	rs12857595	0	11	
chrX	70254843	rs12840573	0	5	
chrX	70255124	rs11796153	1	13	
chrX	70262408	NA12878.400604	2	8	
chrX	70375437	NA12878.400614	0	7	
chrX	70670733	rs35011801	2	5	
chrX	70754779	rs2280964	0	5	
chrX	71375807	rs17302362	2	17	ERCC6L
chrX	72956221	NA12878.400806	5	2	XIST

chrX	72966996	rs56046107	7	2	XIST
chrX	72967289	rs7879153	7	0	XIST
chrX	72976876	NA12878.400813	6	1	XIST
chrX	72979419	rs41310673	5	0	XIST
chrX	72987866	rs41305409	23	1	XIST
chrX	73081208	NA12878.400824	4	5	
chrX	73137427	NA12878.400838	1	5	
chrX	73419591	rs174149	0	6	AK311345/AK05 7701
chrX	73427129	rs174143	2	7	AK311345/AK05 7701
chrX	73428082	rs174141	3	6	AK311345/AK05 7701
chrX	73428610	rs174140	1	27	AK311345/AK05 7701
chrX	73429420	rs174139	0	8	AK311345/AK05 7701
chrX	74060882	NA12878.400901	0	15	
chrX	75310492	rs5981411	0	8	
chrX	75310494	rs6647955	0	8	
chrX	78090057	rs2251819	5	3	
chrX	78099816	NA12878.401242	3	11	
chrX	78249375	NA12878.401250	2	5	
chrX	78286594	rs5912225	1	6	GPR174
chrX	78288640	rs5912226	0	10	GPR174
chrX	78289838	rs937082	0	11	GPR174
chrX	78305988	rs5912229	2	5	GPR174
chrX	78324314	rs5959256	0	5	
chrX	84145499	rs3747422	1	14	
chrX	100534215	rs3027594	1	5	
chrX	100535510	NA12878.403425	0	6	
chrX	100770043	rs1046929	0	6	
chrX	100770548	rs9887270	1	7	
chrX	100771502	rs5951274	0	5	
chrX	101073045	rs34507465	2	5	
chrX	103060764	rs1180803	0	11	AK026512/AK09 0973
chrX	103073336	rs5945723	3	5	AK090973/TMSB 15B
chrX	115040693	NA12878.404341	2	5	
chrX	118487578	rs5910591	0	16	
chrX	118488564	rs5910592	0	11	
chrX	118489695	rs5910593	1	24	
chrX	118583426	NA12878.404564	0	24	
chrX	118684677	NA12878.404570	1	7	SEPT6

chrX	118774302	rs2782220	1	7	
chrX	118774631	rs2782221	3	8	
chrX	118802595	rs11797768	0	7	
chrX	118803456	rs5910671	2	12	
chrX	118803959	rs7889762	2	13	
chrX	118805699	rs45601335	0	6	
chrX	118805765	rs5910674	0	7	
chrX	118806771	rs2429002	0	9	
chrX	118806800	NA12878.404576	0	10	
chrX	118806926	NA12878.404577	0	11	
chrX	118809478	rs28384478	0	50	
chrX	118809733	rs2240448	4	72	
chrX	118810317	rs28384432	0	12	
chrX	118889733	rs708463	2	38	
chrX	118960476	rs5910712	0	5	
chrX	118960703	rs194318	0	8	
chrX	118961041	rs2239964	0	6	
chrX	119327443	rs5910826	0	8	
chrX	119503839	rs173915	1	10	LAMP2
chrX	122557022	rs2498044	0	7	
chrX	122822403	rs28382702	0	9	
chrX	122826230	rs12849057	0	8	
chrX	122922483	rs34620385	0	17	
chrX	122947249	NA12878.404916	1	6	
chrX	123058588	rs6655784	0	5	
chrX	128752837	rs5932684	1	11	
chrX	128755571	rs859577	0	8	
chrX	129067999	NA12878.405369	0	5	
chrX	130703617	rs4372150	7	2	
chrX	130792830	rs5977559	0	11	
chrX	131018819	rs17879773	0	8	
chrX	131040193	rs5977623	0	10	
chrX	131177992	rs5977658	0	7	
chrX	131443550	rs2770689	1	5	
chrX	133877301	rs4830307	1	6	
chrX	135777849	NA12878.405669	1	8	
chrX	142173608	rs5953824	0	7	
chrX	148146779	rs760049	1	9	
chrX	148367071	rs5936290	0	5	IDS
chrX	148370504	rs1064463	1	5	IDS
chrX	148386632	rs4844026	0	10	IDS
chrX	148388397	rs10779253	0	5	IDS
chrX	148401966	rs605115	0	6	IDS
chrX	148402143	NA12878.406801	0	5	IDS

chrX	148430445	rs41302150	2	30	
chrX	148719742	NA12878.406813	0	7	
chrX	151951580	rs1029279	1	5	
chrX	152703342	rs41299124	0	8	
chrX	153263273	rs41299120	1	9	
chrX	153293600	rs62617809	5	29	
chrX	153358392	NA12878.407164	0	5	

## Appendix B: Allele-specific expression of genes on the X chromosome

Genes include genes that are associated with PolIII signal and were assayed for allele-specific expression. Columns indicate in order: (A) heterozygous SNP assayed by SNaPshot for allele-specific expression; (B) gene containing the assayed SNP; (C) SNP alleles; (D) genomic position of that SNP on chrX; (E) SNP location relative to the tested gene; (F) expression in  $X_i^p$  clones relative to  $X_a$ ; (G) expression in  $X_i^m$  clones relative to  $X_a$ ; (H) expression status; (I) PolIII reads aggregated across each gene and flanking regions (30Kb up and 5Kb downstream), paternal reads per gene; (J) PolIII reads aggregated across each gene and flanking regions (30Kb up and 5Kb downstream), maternal reads per gene; (K) specific RefSeq or UCSC transcript name containing that SNP; (L) transcription start and (M) transcription end of each transcript. Primers for SNaPshot assays for each SNP are listed in appendix C.

assayed expressed SNP	gene name	ambiguity	position	SNP location	Expression % $X_a$		status	Pol2 reads*		Transcript name	Transcription	
					$X_i^p$	$X_i^m$		pat/gene	mat/gene		start	end
rs17879004	<b>IL3RA</b>	A/C	1427404	exon	56	114	biallelic - PAR	84	82	NM_002183	1415508	1461582
rs7205	<b>SLC25A6</b>	C/T	1468324	exon	95	108	biallelic - PAR	119	112	NM_001636	1465044	1471039
rs5948863	<b>ASMTL</b>	A/G	1506876	exon	90	109	biallelic - PAR	111	160	NM_004192	1482031	1531870
				exon						NM_001173473	1482031	1532655
				exon						NM_001173474	1482031	1531870
rs4548373	<b>P2RY8</b>	C/G	1543540	3'UTR	104	107	biallelic - PAR	523	557	NM_178129	1541465	1616037
rs6644621	<b>AKAP17A</b>	C/T	1673021	exon	112	88	biallelic - PAR	35	37	NM_005088	1670485	1681411
				exon						NR_027383	1670485	1681407
rs4892932	<b>ZBED1</b>	A/G	2416530	3'UTR	91	73	biallelic - PAR	19	14	NM_001171136	2414454	2429008
				3'UTR						NM_004729	2414454	2429008
				3'UTR						NM_001171135	2414454	2428580
rs17849631	<b>CD99</b>	C/T	2642482	exon	32	195	biallelic - PAR	43	57	NM_002414	2619227	2669350
				exon						NM_001122898	2619227	2669350
rs10871864	<b>PRKX</b>	A/G	3602725	exon	98	74	biallelic	96	125	NM_005044	3532383	3641675
rs7716	<b>TRAPPC2</b>	C/G	13640708	3'UTR	0	0	monoallelic	14	21	NM_014563	13640281	13662675

				3'UTR						NM_001011658	13640281	13662675
				3'UTR						NM_001128835	13640281	13662675
rs5924530	<b>TXLNG</b>	A/G	16769549	exon	26	23	biallelic	23	80	NM_018360	16714475	16772563
				exon						NM_001168683	16714475	16772563
rs67984110	<b>RBBP7</b>	C/T	16797576	exon	0	0	monoallelic	10	68	NM_002893	16772695	16798455
rs3213451	<b>MBTPS2</b>	A/G	21771355	exon	0	0	monoallelic	4	41	NM_015884	21767576	21813462
rs5951642	<b>YY2</b>	A/G	21786033	3'UTR	0	0	monoallelic	3	41	uc010nfq.1	21767674	21785627
rs73203659	<b>USP9X</b>	A/G	40979832	3'UTR	12	13	biallelic	18	44	NM_001039591	40829831	40980776
				3'UTR						NM_001039590	40829831	40980776
rs5963957	<b>DDX3X</b>	A/C	41093254	3'UTR	72	51	biallelic	98	148	NM_001356	41077594	41094468
				3'UTR						NM_001193417	41077594	41094468
				3'UTR						NM_001193416	41077594	41094468
rs5905607	<b>ZNF41</b>	C/T	47212055	5'UTR	0	0	monoallelic	4	25	NM_153380	47190504	47227289
				exon						NM_007130	47190504	47227289
rs1142636	<b>SYN1</b>	C/T	47351305	exon	5	2	biallelic	6	31	NM_006950	47316243	47364200
				exon						NM_133499	47316243	47364200
rs56393649	<b>GNL3L</b>	A/G	54605525	3'UTR	0	0	monoallelic	2	87	NM_019067	54573368	54610445
				3'UTR						NM_001184819	54573368	54610445
rs6625979	<b>ERCC6L</b>	A/G	71341443	3'UTR	0	0	monoallelic	4	24	NM_017669	71341231	71375583
rs174151	<b>NCRNA00182 AK311345</b>	C/T	73418287	exon	0	0	monoallelic	20	134	NR_028379	73164695	73430134
rs174154	<b>AK057701</b>	A/T	73416613	3'UTR	0	0	monoallelic	6	75	uc004ebr.1	73414933	73429160
rs3827440	<b>GPR174</b>	A/G	78313644	exon	0	0	monoallelic	12	76	NM_032553	78313124	78314382
				5'UTR						NM_001142525	101862297	101894025
				5'UTR						NM_001142526	101862297	101894025
				5'UTR						NM_001142527	101862297	101894025
				5'UTR						NM_001142528	101862297	101894025
				5'UTR						NM_030639	101862297	101894025
rs1180803	<b>AK026512</b>	C/T	10306076 4	5'UTR	0	0	monoallelic	0	11	uc004elm.1	103058895	103060787
rs5945723	<b>AK090973</b>	A/G	10307333	3'UTR	0	0	monoallelic	3	19	uc004eln.1	103060134	103074141

			6									
rs56362431	<b>TMSB15B</b>	C/G	10310452 7	5'UTR	0	0	monoallelic	6	18	uc004elq.2	103103897	103107219
				3'UTR						NM_178270	107221554	107284557
rs9887690	<b>ATG4A</b>	A/C	10728427 8	3'UTR	0	0	monoallelic	4	19	NM_178270	107221555	107284557
				3'UTR						NM_052936	107221555	107284557
rs41300319	<b>SEPT6</b>	A/G	11863428 4	3'UTR	0	75	variable	2	16	NM_145802	118633715	118711361
				3'UTR						NM_145800	118633715	118711361
rs73219144	<b>LAMP2</b>	A/G	11946048 3	exon	0	0	monoallelic	3	18	NM_001122606	119444030	119487232
				exon						NM_002294	119444030	119487232
				exon						NM_013995	119454376	119487232
rs3747461	<b>MMGT1</b>	G/A	13487448 4	3'UTR	0	0	monoallelic	1	5	NM_173470	134871897	134883800
rs4844025	<b>IDS</b>	C/T	14837105 5	3'UTR	0	0	monoallelic	8	89	NM_001166550	148368200	148394789
				3'UTR						NM_000202	148368200	148394789
				3'UTR						uc004fcw.2	148368202	148423359
				3'UTR						uc010nsu.1	148368202	148422479
rs1620574	<b>XIST</b>	C/T	72960542				monoallelic	87	15	NR_001564	72957220	72989313

\* Pol2 aggregated over genes + flanks (30Kb upstream, 5Kb downstream)

## Appendix C: Primers used in this thesis

Primers are named according to the following scheme

Gene name : (assayed SNP rs number) : primer direction

F = forward; if multiple primers were designed, number follows

R = reverse; if multiple primers were designed, number follows

S, S1, S2 = SNaPshot primer adjacent to the assayed SNP; numbers designate SNaPshot primer directions where 1 = forward and 2 = reverse

g = designates primers specific to genomic DNA

Primer/gene name	Primer sequence
XISTrs1620574F	CAGTAAGCCAATAGTTCATTCC
XISTrs1620574R	TGGGGATGCAAAGATAAACC
XISTrs1620574S	GTATAGAACTGTAGGCTT
XISTrs1794213F	TCAAATGTAACTGCATGATTGC
XISTrs1794213R	GACTGTGCCAACGCTACTCC
XISTrs1794213S	CCTGGGACTGTTGAGCATGT
EBPrs3048F	CCTATACACACGCAGCCATCC
EBPrs3048R	GACCTGACAACAGCCATGTG
EBPrs3048S	CACAAAGACATGACTACCAACGC
CXorf6rs2070779F	CAGGAATGATCCCCTCACC
CXorf6rs2070779R	GGCATTCTTGGCACACTAGG
CXorf6rs2070779S	TGCCAGCATATGCAGAG
CXorf6rs567517F	AGGAACAGCTCGCCTTAACC
CXorf6rs567517R	TTGCGAACCTTTCAGTCTCC
CXorf6rs567517S	GTGGGGCACTTAACGATAAAAC
FHL1rs9018F	TTTATGGGTTTGAAACTTGC
FHL1rs9018R	CTCCCCCTCTAGAGTTTTGC
FHL1rs9018S	GCAGTGCTGAAATTCATCCTAC
EBPrs3048F2	CCTATACACACGCAGCCATC
FHL1rs9018F2	TAGCCCCCTCAGATGTTCC
FHL1rs9018R2	TCACAACAGAAGGGACTTTGC
CXorf6rs2070779F(2)	CCCATCCTCTCAGCTCAGG
CXorf6rs2070779R(2)	GTGTTTTCCCGAAACTGC

PIGArs3087965F	TTTCCATATACCGACCAGTGC
PIGArs3087965R	TGGCCATTATTCTATTCAACAGG
PIGArs3087965S	AAGAAAAATAATAATTTGCAAATCAC
CD99L2rs6877F	CGCCCAGCTTTATCTTTCC
CD99L2rs6877R	TTTCAGGAGCTTCTCCTTCC
CD99L2rs6877S	ACTGGTTGAAAGTGGCCAATCTCT
FHL1rs9018F3	AAACGAGCCTGTTTCAGAGG
FHL1rs9018R3	CATATCCATGTTGATGACAGC
FHL1rs9018F4	GGAACATGCAGGTGATTTGG
FHL1rs9018R4	CATTTCAGGTAAGCGGTAGGT
HDHD1Ars10458F1	TCTCAGAAGTCAAGCTGATGGA
HDHD1Ars10458R1	GCATGTAATGCCTACCTGCAT
HDHD1Ars10458S1	GCATATTACCAAACCTGATACACAC
HDHD1Ars10458S2	CACACACACCTCCATATATACA
HDHD1Ars2379207F1	TAGGAAGATGTCCGGGTCTG
HDHD1Ars2379207R1	AAACATGGCATCCCCTTTG
HDHD1Ars2379207S1	GCTTGTCTTCATATCGAACGA
HDHD1Ars2379207S1a	TTGTCTTCATATCGAACGA
SYTL4rs8780F1	ATGAACCGAATGGAGGAATG
SYTL4rs8780R1	GGGAAAAGCAACAGTGAGGA
SYTL4rs8780S1	ATTAAAGCACACATACATGTCAG
SYTL4rs8780R2	TTGACAAGTGAGAAAGCTTTATTTAA
SYTL4rs8780F2	CCAAAGTTGTGTGTGGAAGG
ARHGAP4rs2070097F1	AGCTCGCCGAACAGGTCT
ARHGAP4rs2070097R1	ACTGCCCATGACCTGGACT
ARHGAP4rs2070097S1	GGCCACCGAGTCCAGGTC
ARHGAP4rs2070097R2	CAGCATGAAGGCATCTTC
ARHGAP4rs2070097R3	CTGCAGCATGAAGGCATCT
ARHGAP4rs2070097R4	CATTCGCTTCATCAACCTCA
ARHGAP4rs2070097F2	CAGCACTCACTGGTTGAGGA
ARHGAP4rs2070097S2	TGGGCAGTGCAGC
ARHGAP4rs2070097R5	GACCCACTGGTGGAGG
ARHGAP4rs2070097F3	AGAAGCCAGCAGCTCG

ARHGAP4rs2070097F4	CCGAAGAAGCCAGCAG
ARHGAP4rs2070097R6	TGGTGGAGGGCTGCACTGCCCA
ARHGAP4rs2070097F5	CCGAAGAAGCCAGCAGCTCG
GPC4rs1048369F1	ACCTGTTTGCAGTGACAGGA
GPC4rs1048369R1	CACGTTCGTTCCCATTGTATG
GPC4rs1048369S1	CAGTGACAGGAAATGGATTAG
GPC4rs1048369S2	GGTTGTTGCCCTGGTTG
CLIC2rs559165F1	GCTAAACAGAAGAGTTAGGAGAGC
CLIC2rs559165R1	CAGTGGACAGATTATTTATGGC
CLIC2rs559165S1	CTGATCACAAATATAGCCTT
CLIC2rs559165S2	GTTAGGAGAGCTCTTACAGGAGA
CLIC2rs559165S2a	GGAGAGCTCTTACAGGAGA
GPC4rs1129980F1	CAGCAGCTGGCACTAGTTTG
GPC4rs1129980S1	TTCTTGGCCTGTTTCAGTTT
NAP1L3rs1045686F1	CCTTGCCTTAGCCTCTGTTG
NAP1L3rs1045686R1	ACCCAGAGGTGAAAGCTGAA
NAP1L3rs1045686S1	CTTTACCTCAGGAATTTCTTTAG
NAP1L3rs1045686S2	AAGGATGAAGAAAAGGAAGTT
ASB11rs5935944F1	CTCCCTGAAATGTGTGACAA
ASB11rs5935944R1	AGAAATAAGACCTCCAAGGACGA
ASB11rs5935944S1	ATGGTCCGGACCTTCTGAT
ASB11rs5935944S2	GAAGAATTGTGAAGTTTCAGAACG
CD99L2rs6877F1	CAGTCCCCATTTCTTCTCA
CD99L2rs6877R1	GAGAGGGGAGACACAGTGGA
CD99L2rs6877S1	ATTTTCTCTGCCAAATTAAGCTGA
CD99L2rs6877S2	GTTGAAAGTGGCCAATCTCT
COL4A6rs2295912F1	TAATCCGGGGGATCCTAGAG
COL4A6rs2295912R1	GGTGAAGATGGAAAAGTTGGTG
COL4A6rs2295912S1	CTGGTTCTCCTCTCATGCC
COL4A6rs2295912S2	CAGGATTTCCAGGAGTTGC
DMDrs3361F1	TGCAAAGGATGGAAACACAG
DMDrs3361R1	GGGTGGTTTGGTTTTTGGTG
DMDrs3361S1	CAAATGTGATGGGGCTACTGT

DMDrs3361S2	TAAACTTTGGGAAAAGGTGTAA
DMDrs1801187F1	TGTGTGAAATGGCTGCAAAT
DMDrs1801187R1	GCCCAAAGGTGGACTCTACA
DMDrs1801187S1	CCTGCAGTGGTCACCG
DMDrs1801187S2	GCAAACCTTGATGGCAAACC
DMDrs1800280F1	TCCCTTGATCACCTCAGCTT
DMDrs1800280R1	ACTCGGCTTCTACGAAAGCA
DMDrs1800280S1	CGTGGCCTCTTGAAGTTCC
DMDrs1800280S2	GATGAGACCCTTGAAAGACTCC
DMDrs5927163F1	TGCAAAACAGCATCTTTCTCC
DMDrs5927163R1	TGGTTACCACACCGACGTT
DMDrs5927163S1	AGGTGGGCATGCCTAACA
DMDrs5927163S2	GCTGCATAATAAATGACTGAAAGAATC
KCND1rs2238977F1	AAGGAGTCTGGGGGACATTT
KCND1rs2238977R1	CAGAGAATCCCCACGTGTT
KCND1rs2238977S1	GCCAAGAAGGAGCTGAGG
KCND1rs2238977S2	CATCTCAGTCTCTCTAAACCCT
MORF4L2rs874F1	AGGAACCACCAGCATTCAAG
MORF4L2rs874R1	AGCGTCTACAGACAGCTCACC
MORF4L2rs874S1	ATTTTCTGTTTAAAACAGAACGG
MORF4L2rs874S2	GGGCTTATGTTTCAGTTTGTTT
MOSPD1rs6529647F1	GTTTTAGGCCGGTCAGTTTG
MOSPD1rs6529647R1	GCAAGGATGCTTCTGAGTGA
MOSPD1rs6529647S1	GTGAATTAGAAAAAAGGTGAGAAAATAAC
MOSPD1rs6529647S2	CCTGTTGTGTAAGAAACTTTAAACATT
ATRXrs3088074F1	TTTGGTTTTGAGATGCTTGC
ATRXrs3088074R1	GCTTCCACTGATGGTGTCCG
ATRXrs3088074S1	TTCCAAAGAAGTAAAACCTCT
ATRXrs3088074S2	GATAAGCTTTCTGGGAAAGAG
AL833609rs5916825F1	GAATGGCTCTGAATGACAGG
AL833609rs5916825R1	CCTAGAAACCGTTTGAGTTTCC
AL833609rs5916825S1	AGAGATATGGTCACTTCA
AL833609rs5916825S2	TCCCTAAGAAACCTGGAA

CXorf6rs567517F1	TTGCGAACCTTTCAGTCTCC
CXorf6rs567517R1	ACTGTCTCTGCGCCTCTCC
CXorf6rs567517S1	AATGAGCTTTACAGCAGAAGC
BGNrs4833F1	AGGCTTCTGGGACTTCACC
BGNrs4833R1	AGCCGAAAGGACACATGG
BGNrs4833S1	GGGCGCTGACACCTC
BGNrs4833S2	CGGGTCCAGGACGCC
MYCL2rs5962813F1	CAAGAGACGGAATGATCAAC
MYCL2rs5962813R1	TTTCTGCAATTGCCGTC
MYCL2rs5962813S1	CAAGGCCACGGAATACTTAC
MYCL2rs5962813S2	GGCTTCCGCCAGTTC
REPS2rs11864F1	ATTTGCTAAGCATTGGGAAC
REPS2rs11864R1	ACCCAGTTTCACATCATG
REPS2rs11864S1	CTCTTTTTCACACTTGTTGAAA
REPS2rs11864S2	TAACTATGTTCTTCACAGAC
TBL1Xrs16985675F1	AAAATGCTGTGATAAACCAAAC
TBL1Xrs16985675R1	GAATTCCAGCCGACTGAA
TBL1Xrs16985675S1	TTCCCTAACAATTTGGACACTAC
TBL1Xrs16985675S2	ACCTCCTTTGTGAGAGCAAT
XPNPEP2rs3747343F1	GCAGTGGTGACTATGAAGAAAG
XPNPEP2rs3747343R1	ATGGACAAGAGGAAGGGG
XPNPEP2rs3747343S1	GCTCCTCACCGAGATTCC
XPNPEP2rs3747343S2	CACACGCCCTCCAGC
ZNF185rs11582F1	GTCTTTCTCCAGTTTCTGAGC
ZNF185rs11582R1	ACTACAGCTGAGGAAGAGCAG
ZNF185rs11582S1	GAGAGGAACATTCCATTTATTTGTA
ZNF185rs11582S2	CCTCCAGAGGAAATCCA
GAB3rs17281349F1	GATTTCTCTACTCCAGATCAGG
GAB3rs17281349R1	CATTTGATGATGTTTTTGTGAC
GAB3rs17281349S1	CTGCCATGGCATGAGG
GAB3rs17281349S2	CTCCAGTCATTTGGTCCAC
SEPT6F1	GCGAATGCTGAAGAAGATAC
SEPT6R1	ACCCAAGTGTAAGTGACTGG

SEPT6S1	CAGAAATGTTTAAACAGTGGCT
SEPT6S2	TCGATGGGTTGACTGTCTA
IL3RAF1	AGTGAACAATAGCTATTGCCAG
IL3RAR1	GTTCTCAGGGAAGAGGATCCACG
IL3RAS1	CCGAGTGGCCAACCC
IL3RAS2	ATCCACGTGGAGAATGG
SLC25A6F1	GCTCTGTGCCTGACTTTC
SLC25A6R1	TTCGCCTTCAAGGATAAGTA
SLC25A6S1	ccaGGCGGGTTCTGGC
SLC25A6S2	GTGTACCCGCTGGATTT
ASMTLF1	CTGTGACACTCAGCCTCTG
ASMTLR1	GTACACGGGGACTTTCTGA
ASMTLS1	ccGCAGGTCCTCCGG
ASMTLS2	CTACTACCCGCCCCG
P2RY8F1	TCCATGGGGTAAAAGGAC
P2RY8R1	ATTTTTGGCACATTTGTTCT
P2RY8S1	GGTCCTTGCCTGAGTCA
P2RY8S2	GAACACACTCAGGCTTCC
SFRS17AF1	GGTCAAGGTGTTTGAGAAGT
SFRS17AR1	GCCCTTGACATGAGTTTCA
SFRS17AS1	GCCTATGTGCAGTACCG
SFRS17AS2	GGATGAAGCCCATGTACTC
ZBED1F1	GCTGCGATTATAGACAGGAG
ZBED1R1	GAGGAAGCGTGTTGTCTTAC
ZBED1S1	CGACCCTCTCTCACTTCTC
ZBED1S2	AGGTGGAAAAAGGAAAGAGATT
PRKXF1	CCTTCAGGACAGACTTCTCA
PRKXR1	AGAAGACAGCCAAGCATTT
PRKXS1	CTTTAGGCGGATGACGTC
PRKXS2	CAAGGTGATGAGCATTCC
TLR7F1	CTCAGTCAGCTTCTTAACCAA
TLR7R1	GGTGGACCATATGCATTTAT
TLR7S1	AGTGCTTCCTGCTCTTTTT

TLR7S2	CAGAAGCAGGCCCAAG
TLR8F1	GGGCAAATATGTGACAGAAC
TLR8R1	GGGTAACTGGTTGTCTTCAA
TLR8S1	AAATGGCTTGAATATCACAGA
TLR8S2	GTTGAGGAATGCCCC
IDSF1	TACCTGGAAACACTGGAGAC
IDSR1	CATTGTCAATCTCCCAGAAT
IDSS1	AGAAGATTCAACATGTTATAAATATGAG
IDSS2	CAAAGGGATGGAATATTTACA
CXorf15F1	TACAAGGCCCTTCAAATAAA
CXorf15R1	ACAGCGCTTCTCTGACTCT
CXorf15S1	TGTGCAGGGCTCTTCA
CXorf15S2	ATTGAGCTCATTCTTTCTGT
RBBP7F1	CTGTTCTAAGATGACGACCTG
RBBP7R1	GATGGGGATTGGAGAGAC
RBBP7S1	CGTACAGGGGCTGCG
RBBP7S2	TCTTCCCGACTGGGTC
REPS2F1	TTTTTGAAGCAGTTTTGGAT
REPS2R1	AACTTACTCCCAACCCTTC
REPS2S1	GAAAAACAAATTGTTGAATCTATTC
REPS2S2	TTCAAAATCCTTAAATATGCAC
USP9XF1	GCTTTTGCCCTATTGTATCT
USP9XR1	CCAGGAGAAAAATCAATGAA
USP9XS1	TGTTAAACCATGTTGCTGCT
USP9XS2	TGGTGGAGGGCTTAGAAA
ZNF41F1	AGATCCTGCAGTTTCCACTA
ZNF41R1	GGCACCACCAGAAAACAT
ZNF41S1	GTCATCACTCCACGTTTAC
ZNF41S2	ACCACCAGAAAACATTTGTG
SYN1F1	ACCAAATGCGGTAGTCTC
SYN1R1	CCGAATTCTCTGATCTCAAC
SYN1S1	CACGACCTTCACCCC
SYN1S2	GATATGGAAGTTCTTCGGAA

SYN1gF1	AAAAGTTGTCTCCCCACATT
ERCC6LF1	AACTTTCCAAGAAACAACACA
ERCC6LR1	TTGTTCCCATAATTGGATTC
ERCC6LS1	CTCAGAATACCAGACTATGGAGT
ERCC6LS2	CCCTGCAACGCCCCC
AK057701F1	AGTCCCATTGAAAGGAAAAT
AK057701R1	ACTCCTGCAGAAGATCAGAA
AK057701aS1	GTAATATTTCAAACCTATGGTCCAT
AK057701aS2	GAGTAGTTGGGAGAGAAACCA
AK057701bS1	CAATAATGGATGAACACAACCTG
AK057701bS2	CTTTTGTAATGTTTTGTTGAAAC
BHLHB9F1	GAGGAGTGTTTTCTTGCATC
BHLHB9R1	AGTCTCAGTCACCACCTCTC
BHLHB9S1	CTCCGCAGAGCAGAAC
BHLHB9S2	CCGCCGTTCTCCCA
AK026512F1	GGGAAACAGCATTTACTGAG
AK026512R1	CTTTCCTGTTAAGTGGGC
AK026512S1	GCCCTGTGAAGGAGGTAG
AK026512S2	CCTGTTAAGTGGGCACACT
AK090973F1	CATAAAATCCCCAAGGAACT
AK090973R1	GTCAGAAAGGCTCACAGAAC
AK090973S1	GCAATATCCTTTATAATAAACT
AK090973S2	AGGGAAATACCTAACGTTGAC
TMSB15BF1	GCCCTGCTATTATTTATTGG
TMSB15BR1	AGGTAGAGCTTTCATGAGCA
TMSB15BS1	TAGGGTAACCTCGACAGAATG
TMSB15BS2	AGAAGTGTCGAACATGCCT
AK094280F1_BC028211F1	GCAAGATCTTTGTCTGAACTG
AK094280R1	AACGCTGGACCCTGAGA
AK094280S1_BC028211S1	CTGGCTCCAGGACAGC
AK094280S2_BC028211S2	CTTCCCCTCTAGACCGTG
BC028211R1	CTTGAGGTCTTCGGTAAGAA
LAMP2F1	ATCTGAAATGCTCCAGACAC

LAMP2R1	AAGGAAGTGAACATCAGCAT
LAMP2S1	GTTATTTGCAATGCTGAAAAC
LAMP2S2	CATGTATTTGGTTAATGGCTC
PPP2R3BF1	AGGTGCTCTCGATCTTGC
PPP2R3BR1	G TTCAGACACGGAAAGAAGAG
PPP2R3BS1	CCTGCGGGCGTCCTCT
PPP2R3BS2	ATTCCGACCTTCTACTTCCC
CD99F1	GAGTATCTGTCCTGCCGC
CD99R1	GTTTCTTGGGGATTGCAG
CD99S1	ATGGTGGTTTCGATTTATC
CD99S2	ATTGTCAGGAAGGGCATC
CD99gF1	TCACAAAAGAAAATCGCATA
CD99gR1	CCCCACTTCCTGTCTCTGTA
CD99gS2	CACCAGGAAGGGCATC
ARSD-longF1	CAACCTCCAGAAAAGTCCA
ARSD-longR1	GCACATAGAGTTTGCCTCA
ARSD-longS1	CAACTCCCTGACCTCCTTC
ARSD-longS2	GAAAAAGGAAGGAGCACTTAAC
ARSD-shortF1	TTGAACTCCTGGTCTCAAGC
ARSD-shortR1	CCGAAAGTGAAGCGATAAAG
ARSD-shortS1	CCATAGAAGGAAGACAGCGT
ARSD-shortS2	TAATGTTTCAGGACAATGCAGA
TRAPPC2F1	TTACAAGTACGAGTATCAACAGTTTACT
TRAPPC2R1	GAAACACTTGTGCAGTGCTA
TRAPPC2S1	GCAACTGCTCCCAGC
TRAPPC2S2	CCAAATCATGTTTATAAAATAGGA
MBTPS2F1	CTTATGAAGACTGGCTGGAA
MBTPS2R1	GAGAGTCAGCCATCATTGT
MBTPS2S1	AAGCAAGGATGCTTTACCA
MBTPS2S2	CCATTCCAAAATTGAACCA
MBTPS2gR1	CCACAATGATACAATCCCAT
MBTPS2gS2	AAACAAAAGAAGATGAAGAATACCA
YY2F1	CCAACAGGAGAAAAATTCGT

YY2R1	TTTAGTTCTAACTTCTGTAAGTATAGGC
YY2S1	GTTTTTAAAAAACTTGTTAAAAAATTC
YY2S2	TCCTAAAAGTATAAACATGAACACTT
DDX3XF1	GTTGGAAATATGTACATAACTGCAC
DDX3XR1	GCCACTGTAAAAAGATGGAG
DDX3XS1	CCACTTTGAATTCTGTGCTA
DDX3XS2	GCATTCTGGCCACAAAA
RBM3F1	CTCGCTACGTACTCTTTATCAATC
RBM3R1	GCTCGTCGGTGTTAAAGTT
RBM3S1	CTTCCGTCTCGCTATTTT
RBM3S2	GAGAAAGAAAATGGAGATGTGA
GNL3LF1	GATTATAGGCGTGAGCCACT
GNL3LR1	TTAACGGTGGTAATGGGTCT
GNL3LS1	CACAGCTGCATCTTAACCTT
GNL3LS2	AGAGGCAGAGGCAAAGG
FAAH2F1	CGCATTGAGTTGTTCTCTT
FAAH2R1	TTTCTCTGTCGGATCAGCTT
FAAH2S1	GCAGCTTTAGTCTTAGGGGG
FAAH2S2	GTCTTTGAGGCAAACCTTTGG
AK311345_AK057701F1	CCAAGACTTCACAATGGTTTG
AK311345_AK057701R1	TGCCACACTCAATCAGAAAC
AK311345_AK057701S1	AAGACTTCACAATGGTTTGATAAC
AK311345_AK057701S2	AATCAGCATCCTGCACCT
APOOLF1	GGCCTCCATACCTGCTTAAT
APOOLR1	TGCTTGAGCTCAGGAGTTC
APOOLS1	CCTCCCAAGTAGCTGGG
APOOLS2	GTGGCGGGTGCCTATAG
BEX5F1	GGGGTCAGCTTCTACCACT
BEX5R1	ATGGTAATTGGAAATCTGAGCT
BEX5S1	CTCTGGGTCTCCTAAGGG
BEX5S2	CTGAGCTCACAGTGACGTG
PSMD10F1	CTCCTACCTCAGCCTCTCT
PSMD10R1	TGGAGTGCCAGTGAATGATA

PSMD10S1	CAGTCATAAAGGTTACAATTTTCTTTAT
PSMD10S2	GAAGCCTGTGCATTAAGCT
ATG4AF1	TTGATCTCCCTTCTGTTTGC
ATG4AR1	TGCATGTCAACAGCAGAGT
ATG4AS1	AATCATGATCACTTAAATCAGGG
ATG4AS2	CCTTTACAGAAAAAGTTTGCTGA
SLC25A5F1	GTTCTGTCCTTCTGGCG
SLC25A5R1	TCAAGAGGGTACACAAAACAC
SLC25A5S1	AGAGAACCCAGTTTTGGC
SLC25A5S2	CAGATTCCTGCAAAGTAG
MMGT1F1	TTGCTCACTGCACACAGAAT
MMGT1R1	CCTTGAAAAATGTTGGTGGT
MMGT1S1	AACACAGCAGTCACACAGTATTTTC
MMGT1S2	CACTACAATTTTTGGAATCCTTTT
PNMA5F1	TTCGGTCTTCTGAGCCTATG
PNMA5R1	CTGGAGTTAATGAAGCTCATTC
PNMA5S1	CGCCTGCGTTTCACTAA
PNMA5S2	AGGGAGGCTTACCCCC
DNASE1L1F1	CCTACAGCTGCTTCTGGAC
DNASE1L1R1	TACTCTGGCGGGGAAG
DNASE1L1S1	CGGAGCGCCTGAC
DNASE1L1S2	GCGGGGAAGGAGA
DKFZp686L07201F1	CCCCAGTTTATTGAATCCTG
DKFZp686L07201R1	CAAACCTTTATTGGGGCCTTA
DKFZp686L07201S1	GCCATTCTTATGATGTAATCAGC
DKFZp686L07201S2	GAAATTCAGGAAGTGAAATCG
DKFZp686L07201gF1	ATAGGGAAATTGGGGTGTTT
DKFZp686L07201gS1	CTTTCAGTCTTATGATGTAATCAGC
GPR174F1	GTTCTTTCCTTGCCACTGAG
GPR174R1	CAGAAGCGGAGTTACAAACC
GPR174S1	CTCAGAACCAGTGATGATACC
GPR174S2	CATTTGGTCTTATTGCCAG
SEPT6rs17261138F1	AGCTCCTCCGACAAAGAG

SEPT6rs17261138R1	TTGCAGGAAGAACTGGAA
SEPT6rs17261138S1	TTCCCCCTATGGCCAG
SEPT6rs17261138S2	CAGGAAGAACTGGAAGCTGG
SEPT6rs17261138F2	GGAGCTACTGCACAGGAAAC
SEPT6rs17261138F3	CACCTGGCGAGCTATATC
SEPT6rs17261138R2	GCAGGAAGAACTGGAAGCTG
XIAPrs17334746F1	TTCATAGAACGTCCAGGGTTT
XIAPrs17334746R1	TCAACAGAAGTTAGGAGAACATAACAA
XIAPrs17334746S1	ATTACAAGATTCTCACAACAAACC
XIAPrs17334746S2	ATGCCTTACTCACCTCTACAAT

## 5. References

1. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743-747.
2. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavaré, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet*, **1**, e78.
3. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422-425.
4. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423-428.
5. Cheung, V.G. and Spielman, R.S. (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat. Rev. Genet.*, **10**, 595-604.
6. Cheung, V.G. and Spielman, R.S. (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews Genetics*, **10**, 595-604.
7. Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.*, **35**, 57-64.
8. Wang, H.Y., Fu, Y., McPeck, M.S., Lu, X., Nuzhdin, S., Xu, A., Lu, J., Wu, M.L. and Wu, C.I. (2008) Complex genetic interactions underlying expression differences between *Drosophila* races: analysis of chromosome substitutions. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 6362-6367.

9. Cheung, V.G., Bruzel, A., Burdick, J.T., Morley, M., Devlin, J.L. and Spielman, R.S. (2008) Monozygotic twins reveal germline contribution to allelic expression differences. *Am. J. Hum. Genet.*, **82**, 1357-1360.
10. Pant, P.V., Tao, H., Beilharz, E.J., Ballinger, D.G., Cox, D.R. and Frazer, K.A. (2006) Analysis of allelic differential expression in human white blood cells. *Genome Res.*, **16**, 331-339.
11. Pastinen, T., Ge, B. and Hudson, T.J. (2006) Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet.*, **15 Spec No 1**, R9-16.
12. Price, A.L., Patterson, N., Hancks, D.C., Myers, S., Reich, D., Cheung, V.G. and Spielman, R.S. (2008) Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genet.*, **4**, e1000294.
13. Cheung, V.G., Nayak, R.R., Wang, I.X., Elwyn, S., Cousins, S.M., Morley, M. and Spielman, R.S. (2010) Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.*, **8**.
14. De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L. and Natoli, G. (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.*, **8**, e1000384.
15. Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845-1848.
16. Core, L.J. and Lis, J.T. (2008) Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, **319**, 1791-1792.
17. Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M. and Burdick, J.T. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365-1369.

18. Knight, J.C., Keating, B.J., Rockett, K.A. and Kwiatkowski, D.P. (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat. Genet.*, **33**, 469-475.
19. Brodsky, A.S., Meyer, C.A., Swinburne, I.A., Hall, G., Keenan, B.J., Liu, X.S., Fox, E.A. and Silver, P.A. (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.*, **6**, R64.
20. Knight, J.C. (2005) Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.*, **83**, 97-109.
21. Fritsche, L.G., Loenhardt, T., Janssen, A., Fisher, S.A., Rivera, A., Keilhauer, C.N. and Weber, B.H. (2008) Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA. *Nat. Genet.*, **40**, 892-896.
22. Mio, F., Chiba, K., Hirose, Y., Kawaguchi, Y., Mikami, Y., Oya, T., Mori, M., Kamata, M., Matsumoto, M., Ozaki, K. *et al.* (2007) A functional polymorphism in COL11A1, which encodes the alpha 1 chain of type XI collagen, is associated with susceptibility to lumbar disc herniation. *Am. J. Hum. Genet.*, **81**, 1271-1277.
23. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497-1502.
24. Tycko, B. (2010) Allele-specific DNA methylation: beyond imprinting. *Hum. Mol. Genet.*, **19**, R210-220.
25. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299-1320.
26. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-861.

27. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
28. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246-1250.
29. Ge, B., Pokholok, D.K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D.J., Le, J., Koka, V., Lam, K.C., Gagne, V. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.*, **41**, 1216-1222.
30. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics*, **33**, 422-425.
31. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621-628.
32. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768-772.
33. Lalonde, E., Ha, K.C., Wang, Z., Bemmo, A., Kleinman, C.L., Kwan, T., Pastinen, T. and Majewski, J. (2011) RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.*, **21**, 545-554.
34. McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235-239.

35. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232-235.
36. Zhou, V.W., Goren, A. and Bernstein, B.E. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, **12**, 7-18.
37. Yun, M., Wu, J., Workman, J.L. and Li, B. (2011) Readers of histone modifications. *Cell Res.*, **21**, 564-578.
38. Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D. and Patel, D.J. (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol*, **14**, 1025-1040.
39. Fischle, W., Wang, Y. and Allis, C.D. (2003) Histone and chromatin cross-talk. *Curr. Opin. Cell Biol.*, **15**, 172-183.
40. Bannister, A.J. and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381-395.
41. Yang, X.J. and Gregoire, S. (2006) A recurrent phospho-sumoyl switch in transcriptional repression and beyond. *Mol. Cell*, **23**, 779-786.
42. Goldberg, A.D., Allis, C.D. and Bernstein, E. (2007) Epigenetics: a landscape takes shape. *Cell*, **128**, 635-638.
43. Margueron, R. and Reinberg, D. (2010) Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.*, **11**, 285-296.
44. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.

45. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897-903.
46. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. and Young, R.A. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77-88.
47. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311-318.
48. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553-560.
49. Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457-466.
50. Chadwick, B.P. and Willard, H.F. (2004) Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 17450-17455.
51. Valley, C.M., Pertz, L.M., Balakumaran, B.S. and Willard, H.F. (2006) Chromosome-wide, allele-specific analysis of the histone code on the human X chromosome. *Hum. Mol. Genet.*, **15**, 2335-2347.
52. Brinkman, A.B., Roelofsen, T., Pennings, S.W., Martens, J.H., Jenuwein, T. and Stunnenberg, H.G. (2006) Histone modification patterns associated with the human X chromosome. *EMBO Rep*, **7**, 628-634.
53. McEwen, K.R. and Ferguson-Smith, A.C. (2010) Distinguishing epigenetic marks of developmental and imprinting regulation. *Epigenetics Chromatin*, **3**, 2.

54. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33 Suppl**, 245-254.
55. Sasaki, H., Allen, N.D. and Surani, M.A. (1993) DNA methylation and genomic imprinting in mammals. *EXS*, **64**, 469-486.
56. Wilkins, J.F. (2005) Genomic imprinting and methylation: epigenetic canalization and conflict. *Trends Genet.*, **21**, 356-365.
57. Weaver, J.R., Susiarjo, M. and Bartolomei, M.S. (2009) Imprinting and epigenetic changes in the early embryo. *Mamm. Genome*, **20**, 532-543.
58. Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V.V., Schupf, N., Vilain, E. *et al.* (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.*, **40**, 904-908.
59. Shoemaker, R., Deng, J., Wang, W. and Zhang, K. (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.*, **20**, 883-889.
60. Zhang, Y., Rohde, C., Reinhardt, R., Voelcker-Rehage, C. and Jeltsch, A. (2009) Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biol.*, **10**, R138.
61. Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S. and Liu, C. (2010) Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.*, **86**, 411-419.
62. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*, **6**, e1000952.

63. Gertz, J., Varley, K.E., Reddy, T.E., Bowling, K.M., Pauli, F., Parker, S.L., Kucera, K.S., Willard, H.F. and Myers, R.M. (2011) Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. *PLoS Genet*, **7**, e1002228.
64. Heard, E. and Disteché, C.M. (2006) Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev.*, **20**, 1848-1867.
65. Lindsley, D.L., Sandler, L., Baker, B.S., Carpenter, A.T., Denell, R.E., Hall, J.C., Jacobs, P.A., Miklos, G.L., Davis, B.K., Gethmann, R.C. *et al.* (1972) Segmental aneuploidy and the genetic gross structure of the *Drosophila* genome. *Genetics*, **71**, 157-184.
66. Nussbaum, R.L., McInnes, R.R. and Willard, H.F. (2001) *Thompson & Thompson Genetics in Medicine, Chapter 10 Clinical Cytogenetics: Disorder of the Autosomes and Sex Chromosomes*. Saunders, Philadelphia.
67. Nussbaum, R.L., McInnes, R.R. and Willard, H.F. (2001) *Thompson & Thompson Genetics in Medicine, Chapter 9 Principles of Clinical Cytogenetics*. Saunders, Philadelphia.
68. Cheng, M.K. and Disteché, C.M. (2006) A balancing act between the X chromosome and the autosomes. *J Biol*, **5**, 2.
69. Straub, T. and Becker, P.B. (2007) Dosage compensation: the beginning and end of generalization. *Nat Rev Genet*, **8**, 47-57.
70. Hamada, F.N., Park, P.J., Gordadze, P.R. and Kuroda, M.I. (2005) Global regulation of X chromosomal genes by the MSL complex in *Drosophila melanogaster*. *Genes Dev*, **19**, 2289-2294.
71. Straub, T., Gilfillan, G.D., Maier, V.K. and Becker, P.B. (2005) The *Drosophila* MSL complex activates the transcription of target genes. *Genes Dev*, **19**, 2284-2288.

72. Nguyen, D.K. and Disteche, C.M. (2006) Dosage compensation of the active X chromosome in mammals. *Nat. Genet.*, **38**, 47-53.
73. Gupta, V., Parisi, M., Sturgill, D., Nuttall, R., Doctolero, M., Dudko, O.K., Malley, J.D., Eastman, P.S. and Oliver, B. (2006) Global analysis of X-chromosome dosage compensation. *J Biol*, **5**, 3.
74. Jegalian, K. and Page, D.C. (1998) A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature*, **394**, 776-780.
75. Ohno, S., Kaplan, W.D. and Kinoshita, R. (1959) Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. *Exp. Cell Res.*, **18**, 415-418.
76. Lyon, M.F. (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature*, **190**, 372-373.
77. Carrel, L. and Willard, H.F. (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**, 400-404.
78. Sheardown, S., Norris, D., Fisher, A. and Brockdorff, N. (1996) The mouse *Smcx* gene exhibits developmental and tissue specific variation in degree of escape from X inactivation. *Hum. Mol. Genet.*, **5**, 1355-1360.
79. Carrel, L., Hunt, P.A. and Willard, H.F. (1996) Tissue and lineage-specific variation in inactive X chromosome expression of the murine *Smcx* gene. *Hum. Mol. Genet.*, **5**, 1361-1366.
80. Carrel, L., Cottle, A.A., Goglin, K.C. and Willard, H.F. (1999) A first-generation X-inactivation profile of the human X chromosome. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 14440-14444.
81. Carrel, L. and Willard, H.F. (1999) Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X inactivation in some

human cell lines but is inactivated in others. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 7364-7369.

82. Anderson, C.L. and Brown, C.J. (1999) Polymorphic X-chromosome inactivation of the human TIMP1 gene. *Am. J. Hum. Genet.*, **65**, 699-708.
83. Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Zhang, J. and He, X. (2010) RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat. Genet.*, **42**, 1043-1047.
84. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509-1517.
85. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R. and Willard, H.F. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38-44.
86. Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J. and Willard, H.F. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527-542.
87. Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C. *et al.* (1991) Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, **351**, 325-329.
88. Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P., Smith, S., McCabe, V.M., Norris, D.P., Penny, G.D., Patel, D. and Rastan, S. (1991) Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature*, **351**, 329-331.
89. Willard, H.F. and Carrel, L. (2001) Making sense (and antisense) of the X inactivation center. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 10025-10027.

90. Augui, S., Nora, E.P. and Heard, E. (2011) Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.*, **12**, 429-442.
91. Brown, C.J., Lafreniere, R.G., Powers, V.E., Sebastio, G., Ballabio, A., Pettigrew, A.L., Ledbetter, D.H., Levy, E., Craig, I.W. and Willard, H.F. (1991) Localization of the X inactivation centre on the human X chromosome in Xq13. *Nature*, **349**, 82-84.
92. Lafreniere, R.G., Brown, C.J., Rider, S., Chelly, J., Taillon-Miller, P., Chinault, A.C., Monaco, A.P. and Willard, H.F. (1993) 2.6 Mb YAC contig of the human X inactivation center region in Xq13: physical linkage of the RPS4X, PHKA1, XIST and DXS128E genes. *Hum. Mol. Genet.*, **2**, 1105-1115.
93. Rastan, S. (1983) Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre. *J. Embryol. Exp. Morphol.*, **78**, 1-22.
94. Rastan, S. and Robertson, E.J. (1985) X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation. *J. Embryol. Exp. Morphol.*, **90**, 379-388.
95. Panning, B., Dausman, J. and Jaenisch, R. (1997) X chromosome inactivation is mediated by Xist RNA stabilization. *Cell*, **90**, 907-916.
96. Sheardown, S.A., Duthie, S.M., Johnston, C.M., Newall, A.E., Formstone, E.J., Arkell, R.M., Nesterova, T.B., Alghisi, G.C., Rastan, S. and Brockdorff, N. (1997) Stabilization of Xist RNA mediates initiation of X chromosome inactivation. *Cell*, **91**, 99-107.
97. Lee, J.T., Davidow, L.S. and Warshawsky, D. (1999) Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.*, **21**, 400-404.
98. Kay, G.F., Penny, G.D., Patel, D., Ashworth, A., Brockdorff, N. and Rastan, S. (1993) Expression of Xist during mouse development suggests a role in the initiation of X chromosome inactivation. *Cell*, **72**, 171-182.

99. Wutz, A. and Jaenisch, R. (2000) A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Mol. Cell*, **5**, 695-705.
100. Mak, W., Nesterova, T.B., de Napoles, M., Appanah, R., Yamanaka, S., Otte, A.P. and Brockdorff, N. (2004) Reactivation of the paternal X chromosome in early mouse embryos. *Science*, **303**, 666-669.
101. Huynh, K.D. and Lee, J.T. (2003) Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature*, **426**, 857-862.
102. Davidson, R.G., Nitowsky, H.M. and Childs, B. (1963) Demonstration of Two Populations of Cells in the Human Female Heterozygous for Glucose-6-Phosphate Dehydrogenase Variants. *Proc. Natl. Acad. Sci. U. S. A.*, **50**, 481-485.
103. Wutz, A., Rasmussen, T.P. and Jaenisch, R. (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.*, **30**, 167-174.
104. Duthie, S.M., Nesterova, T.B., Formstone, E.J., Keohane, A.M., Turner, B.M., Zakian, S.M. and Brockdorff, N. (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.
105. Masui, O. and Heard, E. (2006) RNA and protein actors in X-chromosome inactivation. *Cold Spring Harb. Symp. Quant. Biol.*, **71**, 419-428.
106. de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M. *et al.* (2004) Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Dev Cell*, **7**, 663-676.
107. Heard, E., Rougeulle, C., Arnaud, D., Avner, P., Allis, C.D. and Spector, D.L. (2001) Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell*, **107**, 727-738.

108. Rougeulle, C., Chaumeil, J., Sarma, K., Allis, C.D., Reinberg, D., Avner, P. and Heard, E. (2004) Differential histone H3 Lys-9 and Lys-27 methylation profiles on the X chromosome. *Mol. Cell. Biol.*, **24**, 5475-5484.
109. Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T.B., Webster, Z., Peters, A.H., Jenuwein, T., Otte, A.P. and Brockdorff, N. (2003) Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev Cell*, **4**, 481-495.
110. Nusinow, D.A., Sharp, J.A., Morris, A., Salas, S., Plath, K. and Panning, B. (2007) The histone domain of macroH2A1 contains several dispersed elements that are each sufficient to direct enrichment on the inactive X chromosome. *J. Mol. Biol.*, **371**, 11-18.
111. Keohane, A.M., O'Neill L, P., Belyaev, N.D., Lavender, J.S. and Turner, B.M. (1996) X-Inactivation and histone H4 acetylation in embryonic stem cells. *Dev. Biol.*, **180**, 618-630.
112. Migeon, B.R. (1994) X-chromosome inactivation: molecular mechanisms and genetic consequences. *Trends Genet.*, **10**, 230-235.
113. Sado, T., Fenner, M.H., Tan, S.S., Tam, P., Shioda, T. and Li, E. (2000) X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. *Dev. Biol.*, **225**, 294-303.
114. Sado, T., Okano, M., Li, E. and Sasaki, H. (2004) De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. *Development*, **131**, 975-982.
115. Heard, E., Clerc, P. and Avner, P. (1997) X-chromosome inactivation in mammals. *Annu. Rev. Genet.*, **31**, 571-610.
116. Hellman, A. and Chess, A. (2007) Gene body-specific methylation on the active X chromosome. *Science*, **315**, 1141-1143.

117. Brown, C.J. and Willard, H.F. (1994) The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature*, **368**, 154-156.
118. Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R. and Jaenisch, R. (1999) Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. *Nat. Genet.*, **22**, 323-324.
119. Plath, K., Talbot, D., Hamer, K.M., Otte, A.P., Yang, T.P., Jaenisch, R. and Panning, B. (2004) Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *J. Cell Biol.*, **167**, 1025-1035.
120. Mohandas, T., Sparkes, R.S. and Shapiro, L.J. (1981) Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science*, **211**, 393-396.
121. Andina, R.J. (1978) A study of X chromosome regulation during oogenesis in the mouse. *Exp. Cell Res.*, **111**, 211-218.
122. Gartler, S.M., Rivest, M. and Cole, R.E. (1980) Cytological evidence for an inactive X chromosome in murine oögonia. *Cytogenet. Cell Genet.*, **28**, 203-207.
123. McMahon, A., Fosten, M. and Monk, M. (1981) Random X-chromosome inactivation in female primordial germ cells in the mouse. *J. Embryol. Exp. Morphol.*, **64**, 251-258.
124. Monk, M. and McLaren, A. (1981) X-chromosome activity in foetal germ cells of the mouse. *J. Embryol. Exp. Morphol.*, **63**, 75-84.
125. Johnston, P.G. and Cattanach, B.M. (1981) Controlling elements in the mouse. IV. Evidence of non-random X-inactivation. *Genet. Res.*, **37**, 151-160.
126. Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P. *et al.* (2005) The DNA sequence of the human X chromosome. *Nature*, **434**, 325-337.

127. Lyon, M.F. (1998) X-chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.*, **80**, 133-137.
128. Lyon, M.F. (2003) The Lyon and the LINE hypothesis. *Semin. Cell Dev. Biol.*, **14**, 313-318.
129. Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J.A. and Hameister, H. (2004) Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.*, **20**, 598-603.
130. Lahn, B.T. and Page, D.C. (1999) Four evolutionary strata on the human X chromosome. *Science*, **286**, 964-967.
131. Disteche, C.M. (1995) Escape from X inactivation in human and mouse. *Trends Genet.*, **11**, 17-22.
132. Disteche, C.M., Filippova, G.N. and Tsuchiya, K.D. (2002) Escape from X inactivation. *Cytogenet Genome Res*, **99**, 36-43.
133. Lyon, M.F. (1962) Sex chromatin and gene action in the mammalian X-chromosome. *Am. J. Hum. Genet.*, **14**, 135-148.
134. Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B. and Zhang, Y. (2003) Role of histone H3 lysine 27 methylation in X inactivation. *Science*, **300**, 131-135.
135. Brown, C.J., Carrel, L. and Willard, H.F. (1997) Expression of genes from the human active and inactive X chromosomes. *Am. J. Hum. Genet.*, **60**, 1333-1343.
136. Brown, C.J. and Willard, H.F. (1990) Localization of a gene that escapes inactivation to the X chromosome proximal short arm: implications for X inactivation. *Am. J. Hum. Genet.*, **46**, 273-279.

137. Li, Z.J., Song, S.X., Zhai, Y., Hou, J., Han, L.Z. and Wang, X.F. (2005) Genome sequence comparative analysis of long arm and short arm of human X chromosome. *Yi Chuan Xue Bao*, **32**, 1-10.
138. McNeil, J.A., Smith, K.P., Hall, L.L. and Lawrence, J.B. (2006) Word frequency analysis reveals enrichment of dinucleotide repeats on the human X chromosome and [GATA]<sub>n</sub> in the X escape region. *Genome Res.*, **16**, 477-484.
139. Tsuchiya, K.D., Grealley, J.M., Yi, Y., Noel, K.P., Truong, J.P. and Disteché, C.M. (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.
140. Wang, Z., Willard, H.F., Mukherjee, S. and Furey, T.S. (2006) Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput Biol*, **2**, e113.
141. Barr, M.L. and Bertram, E.G. (1949) A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, **163**, 676.
142. Clemson, C.M., Hall, L.L., Byron, M., McNeil, J. and Lawrence, J.B. (2006) The X chromosome is organized into a gene-rich outer rim and an internal core containing silenced nongenic sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 7688-7693.
143. Chadwick, B.P., Valley, C.M. and Willard, H.F. (2001) Histone variant macroH2A contains two distinct macrochromatin domains capable of directing macroH2A to the inactive X chromosome. *Nucleic Acids Res.*, **29**, 2699-2705.
144. Chadwick, B.P. and Willard, H.F. (2001) Histone H2A variants and the inactive X chromosome: identification of a second macroH2A variant. *Hum. Mol. Genet.*, **10**, 1101-1113.

145. Deys, B.F., Grzeschick, K.H., Grzeschick, A., Jaffe, E.R. and Siniscalco, M. (1972) Human phosphoglycerate kinase and inactivation of the X chromosome. *Science*, **175**, 1002-1003.
146. Gartler, S.M., Vullo, C. and Gandini, E. (1962) Glucose-6-phosphate dehydrogenase deficiency in an XO individual. *Cytogenetics*, **1**, 1-4.
147. Jenuwein, T. (2001) Re-SET-ting heterochromatin by histone methyltransferases. *Trends Cell Biol.*, **11**, 266-273.
148. Wheeler, B.S., Blau, J.A., Willard, H.F. and Scott, K.C. (2009) The impact of local genome sequence on defining heterochromatin domains. *PLoS Genet*, **5**, e1000453.
149. Tariq, M. and Paszkowski, J. (2004) DNA and histone methylation in plants. *Trends Genet.*, **20**, 244-251.
150. Liu, T., Rechtsteiner, A., Egelhofer, T.A., Vielle, A., Latorre, I., Cheung, M.S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P. *et al.* (2011) Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.*, **21**, 227-236.
151. Ebert, A., Lein, S., Schotta, G. and Reuter, G. (2006) Histone modification and the control of heterochromatic gene silencing in *Drosophila*. *Chromosome Res.*, **14**, 377-392.
152. Hublitz, P., Albert, M. and Peters, A.H. (2009) Mechanisms of transcriptional repression by histone lysine methylation. *Int. J. Dev. Biol.*, **53**, 335-354.
153. Lachner, M., O'Sullivan, R.J. and Jenuwein, T. (2003) An epigenetic road map for histone lysine methylation. *J. Cell Sci.*, **116**, 2117-2124.
154. O'Neill, L.P., Randall, T.E., Lavender, J., Spotswood, H.T., Lee, J.T. and Turner, B.M. (2003) X-linked genes in female embryonic stem cells carry an epigenetic mark prior to the onset of X inactivation. *Hum. Mol. Genet.*, **12**, 1783-1790.

155. Boggs, B.A., Cheung, P., Heard, E., Spector, D.L., Chinault, A.C. and Allis, C.D. (2002) Differentially methylated forms of histone H3 show unique association patterns with inactive human X chromosomes. *Nat. Genet.*, **30**, 73-76.
156. Kohlmaier, A., Savarese, F., Lachner, M., Martens, J., Jenuwein, T. and Wutz, A. (2004) A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biol.*, **2**, E171.
157. Baarends, W.M., Wassenaar, E., van der Laan, R., Hoogerbrugge, J., Sleddens-Linkels, E., Hoeijmakers, J.H., de Boer, P. and Grootegoed, J.A. (2005) Silencing of unpaired chromatin and histone H2A ubiquitination in mammalian meiosis. *Mol. Cell. Biol.*, **25**, 1041-1053.
158. Costanzi, C. and Pehrson, J.R. (1998) Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature*, **393**, 599-601.
159. Valley, C.M. (2007), In *Department of Molecular Genetics and Microbiology* Duke University, Durham.
160. Rougeulle, C., Chaumeil, J., Sarma, K., Allis, C.D., Reinberg, D., Avner, P. and Heard, E. (2004) Differential histone H3 Lys-9 and Lys-27 methylation profiles on the X chromosome. *Mol. Cell. Biol.*, **24**, 5475-5484.
161. Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M. and Lander, E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 7145-7150.
162. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkova, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231-1245.
163. Filippova, G.N., Cheng, M.K., Moore, J.M., Truong, J.P., Hu, Y.J., Nguyen, D.K., Tsuchiya, K.D. and Disteche, C.M. (2005) Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell*, **8**, 31-42.

164. Xu, N., Donohoe, M.E., Silva, S.S. and Lee, J.T. (2007) Evidence that homologous X-chromosome pairing requires transcription and Ctf protein. *Nat. Genet.*, **39**, 1390-1396.
165. Shi, Y., Seto, E., Chang, L.S. and Shenk, T. (1991) Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A protein. *Cell*, **67**, 377-388.
166. Donohoe, M.E., Zhang, L.F., Xu, N., Shi, Y. and Lee, J.T. (2007) Identification of a Ctf cofactor, Yy1, for the X chromosome binary switch. *Mol. Cell*, **25**, 43-56.
167. Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.*, **17**, 520-527.
168. Dunn, K.L. and Davie, J.R. (2003) The many roles of the transcriptional regulator CTCF. *Biochem. Cell Biol.*, **81**, 161-167.
169. Bell, A.C. and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482-485.
170. Reik, W. and Walter, J. (2001) Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, **2**, 21-32.
171. Zhang, Y. and Tycko, B. (1992) Monoallelic expression of the human H19 gene. *Nat. Genet.*, **1**, 40-44.
172. Morcos, L., Ge, B., Koka, V., Lam, K.C., Pokholok, D.K., Gunderson, K.L., Montpetit, A., Verlaan, D.J. and Pastinen, T. (2011) Genome-wide assessment of imprinted expression in human cells. *Genome Biol.*, **12**, R25.
173. Luedi, P.P., Dietrich, F.S., Weidman, J.R., Bosko, J.M., Jirtle, R.L. and Hartemink, A.J. (2007) Computational and experimental identification of novel human imprinted genes. *Genome Res.*, **17**, 1723-1730.

174. Luedi, P.P., Hartemink, A.J. and Jirtle, R.L. (2005) Genome-wide prediction of imprinted murine genes. *Genome Res.*, **15**, 875-884.
175. Sha, K. (2008) A mechanistic view of genomic imprinting. *Annu Rev Genomics Hum Genet*, **9**, 197-216.
176. Nikaido, I., Saito, C., Mizuno, Y., Meguro, M., Bono, H., Kadomura, M., Kono, T., Morris, G.A., Lyons, P.A., Oshimura, M. *et al.* (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res.*, **13**, 1402-1409.
177. Daelemans, C., Ritchie, M.E., Smits, G., Abu-Amero, S., Sudbery, I.M., Forrest, M.S., Campino, S., Clark, T.G., Stanier, P., Kwiatkowski, D. *et al.* (2010) High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. *BMC Genet*, **11**, 25.
178. Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D. and Dulac, C. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643-648.
179. Spahn, L. and Barlow, D.P. (2003) An ICE pattern crystallizes. *Nat. Genet.*, **35**, 11-12.
180. Lyle, R., Watanabe, D., te Vrugte, D., Lerchner, W., Smrzka, O.W., Wutz, A., Schageman, J., Hahner, L., Davies, C. and Barlow, D.P. (2000) The imprinted antisense RNA at the *Igf2r* locus overlaps but does not imprint *Mas1*. *Nat. Genet.*, **25**, 19-21.
181. Lee, M.P., DeBaun, M.R., Mitsuya, K., Galonek, H.L., Brandenburg, S., Oshimura, M. and Feinberg, A.P. (1999) Loss of imprinting of a paternally expressed transcript, with antisense orientation to *KVLQT1*, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 5203-5208.

182. Carr, M.S., Yevtodiyenko, A., Schmidt, C.L. and Schmidt, J.V. (2007) Allele-specific histone modifications regulate expression of the Dlk1-Gtl2 imprinted domain. *Genomics*, **89**, 280-290.
183. Delaval, K., Govin, J., Cerqueira, F., Rousseaux, S., Khochbin, S. and Feil, R. (2007) Differential histone modifications mark mouse imprinting control regions during spermatogenesis. *EMBO J.*, **26**, 720-729.
184. Wutz, A. (2011) RNA-mediated silencing mechanisms in mammalian cells. *Prog Mol Biol Transl Sci*, **101**, 351-376.
185. Mannens, M. and Alders, M. (1999) Genomic imprinting: concept and clinical consequences. *Ann. Med.*, **31**, 4-11.
186. Raefski, A.S. and O'Neill, M.J. (2005) Identification of a cluster of X-linked imprinted genes in mice. *Nat. Genet.*, **37**, 620-624.
187. Davies, W., Isles, A., Smith, R., Karunadasa, D., Burrmann, D., Humby, T., Ojarikre, O., Biggin, C., Skuse, D., Burgoyne, P. *et al.* (2005) Xlr3b is a new imprinted candidate for X-linked parent-of-origin effects on cognitive function in mice. *Nat. Genet.*, **37**, 625-629.
188. Kobayashi, S., Isotani, A., Mise, N., Yamamoto, M., Fujihara, Y., Kaseda, K., Nakanishi, T., Ikawa, M., Hamada, H., Abe, K. *et al.* (2006) Comparison of gene expression in male and female mouse blastocysts revealed imprinting of the X-linked gene, RhoX5/Pem, at preimplantation stages. *Curr. Biol.*, **16**, 166-172.
189. Maclean, J.A., Bettegowda, A., Kim, B.J., Lou, C.H., Yang, S.M., Bhardwaj, A., Shanker, S., Hu, Z., Fan, Y., Eckardt, S. *et al.* (2011) The rhoX homeobox gene cluster is imprinted and selectively targeted for regulation by histone H1 and DNA methylation. *Mol. Cell. Biol.*, **31**, 1275-1287.
190. Gregg, C., Zhang, J., Butler, J.E., Haig, D. and Dulac, C. (2010) Sex-specific parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 682-685.

191. Keverne, E.B. (2007) Genomic imprinting and the evolution of sex differences in mammalian reproductive strategies. *Adv. Genet.*, **59**, 217-243.
192. Chadwick, L.H. and Willard, H.F. (2005) Genetic and parent-of-origin influences on X chromosome choice in Xce heterozygous mice. *Mamm. Genome*, **16**, 691-699.
193. Fowles, D.J., Ansell, J.D. and Micklem, H.S. (1991) Further evidence for the importance of parental source of the Xce allele in X chromosome inactivation. *Genet. Res.*, **58**, 63-65.
194. Falconer, D.S., Isaacson, J.H. and Gauld, I.K. (1982) Non-random X-chromosome inactivation in the mouse: difference of reaction to imprinting. *Genet. Res.*, **39**, 237-259.
195. Chess, A. (1998) Expansion of the allelic exclusion principle? *Science*, **279**, 2067-2068.
196. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. and Chess, A. (2007) Widespread monoallelic expression on human autosomes. *Science*, **318**, 1136-1140.
197. Mostoslavsky, R., Singh, N., Tenzen, T., Goldmit, M., Gabay, C., Elizur, S., Qi, P., Reubinoff, B.E., Chess, A., Cedar, H. *et al.* (2001) Asynchronous replication and allelic exclusion in the immune system. *Nature*, **414**, 221-225.
198. Prabhakar, S., Noonan, J.P., Paabo, S. and Rubin, E.M. (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science*, **314**, 786.
199. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799-816.

200. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.
201. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877-885.
202. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669-680.
203. Kadota, M., Yang, H.H., Hu, N., Wang, C., Hu, Y., Taylor, P.R., Buetow, K.H. and Lee, M.P. (2007) Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. *PLoS Genet*, **3**, e81.
204. Maynard, N.D., Chen, J., Stuart, R.K., Fan, J.B. and Ren, B. (2008) Genome-wide mapping of allele-specific protein-DNA interactions in human cells. *Nat. Methods*, **5**, 307-309.
205. Pennisi, E. (2010) Genomics. 1000 Genomes Project gives new map of genetic diversity. *Science*, **330**, 574-575.
206. Gross, D.S. and Garrard, W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, **57**, 159-197.
207. Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**, 123-131.
208. Filippova, G.N. (2008) Genetics and epigenetics of the multifunctional protein CTCF. *Curr. Top. Dev. Biol.*, **80**, 337-360.

209. Reddy, T.E., Gertz, J., Pauli, F., Newberry, K.M., Kucera, K.S., Wold, B., Willard, H.F. and Myers, R.M. ([in preparation]) Effects of sequence variation on differential allelic transcription factor occupancy and gene expression.
210. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636-640.
211. Choy, E., Yelensky, R., Bonakdar, S., Plenge, R.M., Saxena, R., De Jager, P.L., Shaw, S.Y., Wolfish, C.S., Slavik, J.M., Cotsapas, C. *et al.* (2008) Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet*, **4**, e1000287.
212. Miller, A.P. and Willard, H.F. (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.
213. Brown, C.J. and Willard, H.F. (1989) Noninactivation of a selectable human X-linked gene that complements a murine temperature-sensitive cell cycle defect. *Am. J. Hum. Genet.*, **45**, 592-598.
214. Brown, C.J., Flenniken, A.M., Williams, B.R. and Willard, H.F. (1990) X chromosome inactivation of the human TIMP gene. *Nucleic Acids Res.*, **18**, 4191-4195.
215. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M. and White, R. (1990) Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, **6**, 575-577.
216. Rupert, J.L., Brown, C.J. and Willard, H.F. (1995) Direct detection of non-random X chromosome inactivation by use of a transcribed polymorphism in the XIST gene. *Eur. J. Hum. Genet.*, **3**, 333-343.
217. Surka, M.C., Tsang, C.W. and Trimble, W.S. (2002) The mammalian septin MSF localizes with microtubules and is required for completion of cytokinesis. *Mol. Biol. Cell*, **13**, 3532-3545.

218. Kinoshita, M., Field, C.M., Coughlin, M.L., Straight, A.F. and Mitchison, T.J. (2002) Self- and actin-templated assembly of Mammalian septins. *Dev Cell*, **3**, 791-802.
219. Beutler, E., Yeh, M. and Fairbanks, V.F. (1962) The normal human female as a mosaic of X-chromosome activity: studies using the gene for C-6-PD-deficiency as a marker. *Proc. Natl. Acad. Sci. U. S. A.*, **48**, 9-16.
220. Chelly, J., Concordet, J.P., Kaplan, J.C. and Kahn, A. (1989) Illegitimate transcription: transcription of any gene in any cell type. *Proc. Natl. Acad. Sci. U. S. A.*, **86**, 2617-2621.
221. Kimoto, Y. (1998) A single human cell expresses all messenger ribonucleic acids: the arrow of time in a cell. *Mol. Gen. Genet.*, **258**, 233-239.
222. Kuznetsov, V.A., Knott, G.D. and Bonner, R.F. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, **161**, 1321-1332.
223. Chelly, J., Hugnot, J.P., Concordet, J.P., Kaplan, J.C. and Kahn, A. (1991) Illegitimate (or ectopic) transcription proceeds through the usual promoters. *Biochem. Biophys. Res. Commun.*, **178**, 553-557.
224. Lee, S., Bao, J., Zhou, G., Shapiro, J., Xu, J., Shi, R.Z., Lu, X., Clark, T., Johnson, D., Kim, Y.C. *et al.* (2005) Detecting novel low-abundant transcripts in *Drosophila*. *RNA*, **11**, 939-946.
225. Buratowski, S. and Moazed, D. (2005) Gene regulation: expression and silencing coupled. *Nature*, **435**, 1174-1175.
226. Vaughn, M.W. and Martienssen, R.A. (2005) Finding the right template: RNA Pol IV, a plant-specific RNA polymerase. *Mol. Cell*, **17**, 754-756.
227. Niu, D.K. (2005) Low-level illegitimate transcription of genes may be to silence the genes. *Biochem. Biophys. Res. Commun.*, **337**, 413-414.

228. Sarkar, G. and Sommer, S.S. (1989) Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity. *Science*, **244**, 331-334.
229. Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.*, **23**, 387-388.
230. Scott, D., McLaren, A., Dyson, J. and Simpson, E. (1991) Variable spread of X inactivation affecting the expression of different epitopes of the Hya gene product in mouse B-cell clones. *Immunogenetics*, **33**, 54-61.
231. Koch, F., Jourquin, F., Ferrier, P. and Andrau, J.C. (2008) Genome-wide RNA polymerase II: not genes only! *Trends Biochem. Sci.*, **33**, 265-273.
232. Brannan, C.I., Dees, E.C., Ingram, R.S. and Tilghman, S.M. (1990) The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, **10**, 28-36.
233. Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J. and Adelman, K. (2007) RNA polymerase is poised for activation across the genome. *Nat. Genet.*, **39**, 1507-1511.
234. Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M. and Young, R.A. (2007) RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat. Genet.*, **39**, 1512-1516.
235. Lee, C., Li, X., Hechmer, A., Eisen, M., Biggin, M.D., Venters, B.J., Jiang, C., Li, J., Pugh, B.F. and Gilmour, D.S. (2008) NELF and GAGA factor are linked to promoter-proximal pausing at many genes in *Drosophila*. *Mol. Cell. Biol.*, **28**, 3290-3300.
236. Fuda, N.J., Ardehali, M.B. and Lis, J.T. (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, **461**, 186-192.

237. Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., Gao, Y., Li, L. and Adelman, K. (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, **143**, 540-551.
238. Cairns, B.R. (2009) The logic of chromatin architecture and remodelling at promoters. *Nature*, **461**, 193-198.
239. Tirosh, I. and Barkai, N. (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome Res.*, **18**, 1084-1091.
240. Peterlin, B.M. and Price, D.H. (2006) Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell*, **23**, 297-305.
241. Lis, J. (1998) Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 347-356.
242. Adelman, K., Kennedy, M.A., Nechaev, S., Gilchrist, D.A., Muse, G.W., Chinenov, Y. and Rogatsky, I. (2009) Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 18207-18212.
243. Baugh, L.R., Demodena, J. and Sternberg, P.W. (2009) RNA Pol II accumulates at promoters of growth genes during developmental arrest. *Science*, **324**, 92-94.
244. Gilmour, D.S. (2009) Promoter proximal pausing on genes in metazoans. *Chromosoma*, **118**, 1-10.
245. Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D. and Ren, B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876-880.
246. Gilchrist, D.A., Fargo, D.C. and Adelman, K. (2009) Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods*, **48**, 398-408.

247. Zakharova, I.S., Shevchenko, A.I. and Zakian, S.M. (2009) Monoallelic gene expression in mammals. *Chromosoma*, **118**, 279-290.
248. Wilkinson, L.S., Davies, W. and Isles, A.R. (2007) Genomic imprinting effects on brain development and function. *Nat. Rev. Neurosci.*, **8**, 832-843.
249. Berletch, J.B., Yang, F. and Disteché, C.M. (2010) Escape from X inactivation in mice and humans. *Genome Biol.*, **11**, 213.
250. Ke, X. and Collins, A. (2003) CpG islands in human X-inactivation. *Ann. Hum. Genet.*, **67**, 242-249.
251. Marks, H., Chow, J.C., Denissov, S., Francoijs, K.J., Brockdorff, N., Heard, E. and Stunnenberg, H.G. (2009) High-resolution analysis of epigenetic changes associated with X inactivation. *Genome Res.*, **19**, 1361-1373.
252. Mietton, F., Sengupta, A.K., Molla, A., Picchi, G., Barral, S., Heliot, L., Grange, T., Wutz, A. and Dimitrov, S. (2009) Weak but uniform enrichment of the histone variant macroH2A1 along the inactive X chromosome. *Mol. Cell. Biol.*, **29**, 150-156.
253. Chaumeil, J., Le Baccon, P., Wutz, A. and Heard, E. (2006) A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.*, **20**, 2223-2237.
254. Amos-Landgraf, J.M., Cottle, A., Plenge, R.M., Friez, M., Schwartz, C.E., Longshore, J. and Willard, H.F. (2006) X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *Am. J. Hum. Genet.*, **79**, 493-499.
255. Wutz, A. and Gribnau, J. (2007) X inactivation Xplained. *Curr. Opin. Genet. Dev.*, **17**, 387-393.
256. Plagnol, V., Uz, E., Wallace, C., Stevens, H., Clayton, D., Ozcelik, T. and Todd, J.A. (2008) Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS One*, **3**, e2966.

257. Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829-834.
258. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
259. Graves, J.A. (1995) The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding. *Bioessays*, **17**, 311-320.
260. D'Esposito, M., Ciccodicola, A., Gianfrancesco, F., Esposito, T., Flagiello, L., Mazzarella, R., Schlessinger, D. and D'Urso, M. (1996) A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X inactivation. *Nat. Genet.*, **13**, 227-229.
261. Goto, Y. and Kimura, H. (2009) Inactive X chromosome-specific histone H3 modifications and CpG hypomethylation flank a chromatin boundary between an X-inactivated and an escape gene. *Nucleic Acids Res.*, **37**, 7416-7428.
262. Bondy, C.A. and Cheng, C. (2009) Monosomy for the X chromosome. *Chromosome Res.*, **17**, 649-658.
263. (2008) Moving AHEAD with an international human epigenome project. *Nature*, **454**, 711-715.
264. Reddy, T.E., Pauli, F., Sprouse, R.O., Neff, N.F., Newberry, K.M., Garabedian, M.J. and Myers, R.M. (2009) Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.*, **19**, 2163-2171.
265. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

266. Kucera, K.S., Reddy, T.E., Pauli, F., Gertz, J., Logan, J.E., Myers, R.M. and Willard, H.F. (2011) Allele-Specific Distribution of RNA Polymerase II on Female X Chromosomes. *Hum. Mol. Genet.*
267. Weirich, C.S., Erzberger, J.P. and Barral, Y. (2008) The septin family of GTPases: architecture and dynamics. *Nat Rev Mol Cell Biol*, **9**, 478-489.
268. Mendoza, M., Hyman, A.A. and Glotzer, M. (2002) GTP binding induces filament assembly of a recombinant septin. *Curr. Biol.*, **12**, 1858-1863.
269. Sheffield, P.J., Oliver, C.J., Kremer, B.E., Sheng, S., Shao, Z. and Macara, I.G. (2003) Borg/septin interactions and the assembly of mammalian septin heterodimers, trimers, and filaments. *J. Biol. Chem.*, **278**, 3483-3488.
270. Sirajuddin, M., Farkasovsky, M., Zent, E. and Wittinghofer, A. (2009) GTP-induced conformational changes in septins and implications for function. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 16592-16597.
271. Peterson, E.A. and Petty, E.M. (2010) Conquering the complex world of human septins: implications for health and disease. *Clin. Genet.*, **77**, 511-524.
272. Kinoshita, M. (2003) Assembly of mammalian septins. *J Biochem*, **134**, 491-496.
273. Sellin, M.E., Sandblad, L., Stenmark, S. and Gullberg, M. (2011) Deciphering the rules governing assembly order of mammalian septin complexes. *Mol. Biol. Cell*.
274. Hruz, T., Wyss, M., Docquier, M., Pfaffl, M.W., Masanetz, S., Borghi, L., Verbrugghe, P., Kalaydjieva, L., Bleuler, S., Laule, O. *et al.* (2011) RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics*, **12**, 156.

## Biography

Kate Kucera was born in Decin, Czech Republic. She moved to the United States in 1998 and obtained a Bachelors of Science in Chemistry degree from University of North Florida in 2003. She is married to Jonathan Jesse Kucera Jr. and has two sons, Garrett and Krystof.

### Publications related to this thesis:

Reddy TE, Gertz J, Pauli F, Newberry K, **Kucera KS**, Wold B, Willard HF, Myers RM. Allele biased transcription factor occupancy is prevalent across the human genome and associated with allele biased gene expression. Submitted.

Gertz, J. Varley KE, Reddy TE, Bowling KM, Pauli F, Parker, SL, **Kucera KS**, Willard HF, Myers RM. Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. PLoS Genet, 2011, e1002228.

**Kucera KS**, Reddy TE, Pauli F, Gertz J, Logan JE, Myers RM, Willard HF. Allele-specific distribution of RNA polymerase II on female X chromosomes. Hum Mol Genet. 2011.

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, **Kucera KS**, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, Birney E. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010.

### Previous publications:

Murphy SK, Nolan CM, Huang Z, **Kucera KS**, Freking BA, Smith TP, Leymaster KA, Weidman JR, Jirtle RL. Callipyge mutation affects gene expression in cis: a potential role for chromatin structure. Genome Res. 2006.

Zhang J, Bao S, Furumai R, **Kucera KS**, Ali A, Dean NM, Wang XF. Protein phosphatase 5 is required for ATR-mediated checkpoint activation. Mol Cell Biol. 2005.