Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M.Sc. Tania Studer

born in Lausanne, Switzerland

Oral examination: 5th of December 2018

The developmental sex-biased expression of genes escaping X chromosome inactivation across mammals

Referees: Prof. Dr. Henrik Kaessmann

Prof. Dr. Gudrun Rappold

ABSTRACT

In mammals, X-chromosome inactivation (XCI) re-establishes the dosage balance between male and female gene expression levels. However, up to a third of the X-linked genes escape this phenomenon with varying degrees of consistency across tissues, cell lines, or individuals. Here, I take advantage of a new extensive developmental dataset for several organs and species to explore the developmental dimension of sex chromosomes expression levels. I find that only a small fraction of genes escaping XCI show a consistent female overexpression across development, organs, and species. The consistently sex-biased genes are almost exclusively either directly involved in the establishment of XCI, or are protein-coding genes with a broadly expressed Y homolog. The conservation of the sex-bias of these genes across species suggests that they might be responsible for the evolution of escape from XCI. I also present and test a model of allelic contribution to the total expression levels of X-linked genes using marsupial bulk tissue RNA sequencing data. Finally, I describe my contribution to a study of sex-biased micro-RNAs in mammals.

ZUSAMMENFASSUNG

Die X-Chromosom-Inaktivierung (XCI) in Säugetieren führt zur Wiederherstellung des Gleichgewichts zwischen männlicher und weiblicher Genexpression. Ein Drittel aller Gene entzieht sich jedoch diesem Phänomen, mit dabei variierender Stärke und Abhängigkeit von Gewebe, Zelllinie oder Individuum. In dieser Studie benutze ich einen neuen und umfangreichen entwicklungsbiologischen Datensatz mehrerer Organe und Arten, um die entwicklungsbiologischen Expressionslevel von Sexchromosomen zu studieren. Ich kann zeigen, dass nur ein kleiner Anteil der Gene, die der XCI entkommen, dauerhaft und über Entwicklungsstatus, Organ und Art hinweg im weiblichen Geschlecht exprimiert ist. Der Großteil dieser dauerhaft geschlechtsabhängigen Gene hat entweder einen direkten Einfluss auf die Etablierung der XCI, oder gehört zu einer Gruppe proteinkodierender Gene mit einem großflächig exprimierten, männlichen Y-Homolog. Die Konservierung der Geschlechtsspezifität dieser Gene über Arten hinweg lässt vermuten, dass sie für die Evolution des XCI-Entkommens verantwortlich sein könnten.
Im Weiteren stelle und teste ich ein Modell vor welches sich mit dem Einfluss von Allelen auf das Gesamtexpressionslevel X-gebundener Gene beschäftigt. Dieses Model beruht auf einem RNA-Sequenzierungsdatensatz der aus Beuteltiergewebe gewonnen wurde. Abschließend beschreibe ich meine Mitwirkung an einer Studie, die sich mit der geschlechtsabhängigen Verteilung von mircoRNAs in Säugetieren befasst.

TABLE OF CONTENTS

INTRODUCTION

The emergence of sex chromosomes throughout evolution can profoundly impact the gene expression landscape of organisms. In mammals, where sex is determined by heteromorphic sex chromosomes, gene products from the female X chromosome (XX) could be twice as abundant as in males (XY). To counterbalance this, there is a mechanism of dosage compensation, which inactivates one of the female X chromosomes, effectively re-establishing a similar expression level between the sexes. But some genes are not subject to this dosage compensation. This project studies the genes that escape this compensation mechanism, and uses development to shed light on the evolutionary forces underlying these exceptions.

The evolution of mammalian sex chromosomes

To understand dosage compensation, it is important to look in detail into the evolution of sex chromosomes in mammals. Current mammalian sex chromosomes originated from a pair of autosomes in the ancestor of both marsupials and placental mammals [Cortez et al., 2014]. The formation of sex chromosomes started with a single event: the acquisition of a new promoter region upstream of the gene *Sox3* via fusion with a portion of the first exon of the gene *Dgcr8* (including the promoter region that contains a binding motif for the transcription factor CP2, TFCP2), which resulted in the creation of the male determining gene *Sry* [Sato et al. 2010]. This event occurred approximately 180 million years ago [Cortez et al., 2014]. Upon its creation, *Sry* gained the role of sex determining trigger by gaining a position at the top of the sex determination genetic cascade that already existed. As it emerged in the ancestor of all therians, *Sry* is therian-specific. Because of *Sry*'s position as a trigger of male sex determination, the chromosome carrying *Sry* became a *proto-Y chromosome* and consequently, its homologous partner became a *proto-X chromosome*. In some species including anurans (European common frog, green toad), ratites (ostrich, emu), and dipterans *(Megaselia scalaris*), the presence of the sex determination trigger gene seems to be the only main difference that the sex chromosomes harbour, and have harboured sometimes for tens of thousands of generations, suggesting that proto sex chromosomes can remain homomorphic for a very long time [Perrin 2009]. But in therian mammals, the emergence of *Sry* was followed by a chain of genomic modifications of the proto-sex chromosomes that ultimately lead to the heteromorphism between the X and Y that we observe today.

The first event of this chain of genomic modifications was the spontaneous occurrence of sexually antagonistic mutations. These mutations can appear anywhere in the genome, and are detrimental for one sex while beneficial for the other. Usually these mutations are only fixed in a population if they are much more beneficial to one sex than they are detrimental to the other, because otherwise they

would be removed by purifying selection. On the proto-Y chromosome, the existence of a male determining gene offers a new genomic environment for these mutations: if they are in close proximity to *Sry*, they will be transmitted more often to males than to females because of the short recombination distance, thus allowing for their transmission despite being potentially detrimental for females. This physical linkage allows for the accumulation of male beneficial and female detrimental mutations around the *Sry* locus. This process is self-reinforcing: the accumulation of sexually antagonistic mutations around the sex determining region selects for tighter linkage, thus allowing for an even higher accumulation of sexually antagonist mutations. The complete linkage that we observe in genomes is usually the consequence of total recombination arrest.

Evidence for such a chain of modifications can be observed in mammals. It was discovered by Lahn and Page in 1999 [Lahn & Page 1999] through the identification of evolutionary strata. They identified four different regions along the X chromosome that could be distinguished from one another based on their level of sequence divergence to their homologous regions on the Y chromosome. These strata are characterised by an almost absolute absence of recombination between the sex chromosomes, which explains the different degrees of divergence: if no recombination occurs, mutations can accumulate, and so, the older the recombination arrest, the more genetic differences have accumulated.

An important paper from the Kaessmann group [Cortez et al., 2014] dated precisely the emergence of these strata and helped paint the history of the evolution of mammalian sex chromosomes, Stratum one emerged shortly after the emergence of therian sex chromosomes, 180 million years ago, via a segmental chromosomal inversion. Shortly after the split between the marsupial and placental lineages, stratum two was created, and this same region was selected independently in the two lineages. On the placental lineage, the sex chromosomes fused to another pair of chromosomes (which remain autosomal in marsupials) [Graves 1995]. The newly formed placental neo-sex chromosomes were therefore composed of a new X/Y added region combined with the ancestral X/Y conserved region. Before the radiation of placentals, this new added region stopped recombining, thus forming the third stratum. The fourth stratum found in humans is ape specific, and a fifth stratum was identified by Ross and colleagues in 2005 [Ross et al. 2005]. On both ends of the human X and Y chromosomes there remain regions that still recombine between sex chromosomes. Because they behave similarly to autosomes, they are called pseudo autosomal regions (PARs). Humans have two PARs, one at the end of each arm of the sex chromosomes, while only one PAR is present in chimpanzee and rhesus monkey, the one corresponding to the short-arm PAR in humans [Hughes et al. 2012]. One recent small PAR was found in mouse [Ellis & Goodfellow 1989]. Marsupials do not have any PAR; their sex chromosomes are fully achiasmate (as are the ones of Drosophila) [De la Fuente et al. 2007, Karpen et al. 1996].

It is well accepted that segmental inversions are likely responsible for the recombination arrest, as suggested by the observation that the order of strata is well preserved on the X chromosome across lineages. What is still subject to debate is whether recombination arrest via segmental inversions needs the selective pressure of sexually antagonistic mutations to evolve. There are alternative, neutral models for the ultimate causes for sex chromosomes recombination arrest. On autosomes, inversions are either fixed or eliminated, while on sex chromosomes, inversions involving sex-determining regions will be kept in the heterogametic sex in the descendants. It is still unclear which among antagonistic mutations and chromosomal inversions is the cause and which is the consequence.

Consequences of recombination arrest

Following recombination arrest, the newly formed sex chromosomes face a reduced population size. Instead of being present in four copies per mating pair, Y chromosomes are only found in one copy and X chromosomes are only found in three copies. This reduced population size increases the power of genetic drift, thus reducing the strength of purifying selection on new mutations. In addition, interference within the same non-recombining region can further lower the population size. Selection at one locus increases drift at the proximal non-recombining region, since without recombination, positive and negative mutations cannot be disentangled. As a consequence, the fixation probabilities of alleles at different loci are dependent, a phenomenon known as Hill-Robertson interference [Charlesworth & Charlesworth 2000].

Another kind of Hill-Robertson interferences is also known as *genetic hitchhiking,* where an allele changes frequency because it is linked to a selected site. This means that a slightly detrimental genetic background could be selected because of the presence of a single highly beneficial mutation, thus adding to the perceived smaller effective population size of this region of the genome. The combination of the effects of linkage, selection, and genetic drift accelerate the rate of fixation of deleterious mutations and retard that of beneficial ones. This is known as Muller's ratchet: after a certain number of mutations in a non-recombining population, selection alone will not be able to reduce the total number of mutations in the population any further. Unless back mutations occur, evolution can only go towards the accumulation of deleterious mutations [Muller 1918, Charlesworth & Charlesworth 2000].

Moreover, the Y chromosome is strictly sex linked, which makes sexually antagonistic mutations more likely to be kept. In the heterogametic sex, transcription from the sex chromosomes is repressed during meiosis, a phenomenon known as meiotic sex chromosome inactivation (MSCI). During MSCI

all unpaired chromatin is silenced, and MSCI has been hypothesised to be a defence mechanism against transposon invasion and meiotic drive [reviewed in Turner 2007]. A competing hypothesis proposes that since female beneficial sexually antagonistic alleles accumulate on the X, MSCI would allow for their silencing during male meiosis. As a consequence, any mutation that could negatively impact male meiosis would be hidden from selection and could be passed on, effectively contributing to the decay of sex chromosomes. In general, because of the permanent heterozygosity of the Y chromosome, deleterious recessive mutations are free to accumulate overtime [Muller 1918]. Even loss of function mutations can be transmitted without excessively impacting fitness. Finally, Y chromosomes are also expected to decay rapidly because males usually have higher mutation rates than females, due to differences in meiosis, and also because they can have lower effective population sizes due to the behavioural aspects of partner selection.

It is important to note that the X chromosome escapes from most of these phenomena, thanks to the continued recombination along its entire length in females. However, it is still sensitive to specific evolutionary pressures, as it is always present in a single copy in males. For example, it is under stronger purifying selection against male-detrimental mutations because of the hemizygosity in males, and can undergo an accumulation of ampliconic genes in an arms race between the sex chromosomes [Soh et al. 2014].

The consequences of these effects on the Y chromosome are clear in mammals. Mammalian Y chromosomes have a lower diversity than that expected given their 4-fold drop in population size [Hellborg & Ellegren 2004]. They have accumulated non-coding DNA, mostly retrotransposons [Soh et al. 2014] and some of the genes present on modern Y chromosomes (*e.g.*, *TSPY* and *EIF1AY* in humans, *Ube1y* and *Zfy* in mouse) show a "masculinisation" of their expression pattern compared to their X homologs and their orthologs in species in which these genes are still autosomal [Cortez et al. 2014]. Finally, a major consequence (and the one most relevant to this project) is the decay of coding regions. It can be extremely severe, as there are only 17 remaining functional protein-coding genes on the human Y and 9 on the mouse Y [Cortez et al. 2014]. At later stages of decay, some broadly expressed genes have also moved from the sex chromosomes to the autosomes (*e.g.*, *Eif1a* is now on chromosome 18 in mouse and rat [Skaletsky et al, 2003]).

The genes that remained on the Y chromosomes have very specific characteristics: they are either broadly expressed dosage-sensitive regulators, or they are male-specific genes expressed exclusively in the adult testes [Cortez et al. 2014, Bellott et al. 2014]. The rate of gene loss on the Y, as inferred by species comparison and ancestral state reconstruction, suggests that this chromosome is not, however, about to disappear: most genes were lost rapidly after recombination arrest, and the remaining genes have been kept for extensive periods [Hughes et al. 2012, Cortez et al. 2014]. This strong

conservation, despite the dramatic decay of the Y chromosome, suggests that Y-linked genes are essential for male fitness, and that they are under strong purifying selection.

Overall, these dramatic changes to the Y chromosome composition come with a fitness cost to the heterogametic sex. To counter the almost total monosomy of the X chromosome in males, species with extensive Y decay have resorted to different solutions. In some lineages (mostly in fishes [Ezaz et al. 2006], but also in amphibians and reptiles [Marin et al. 2017]), there has been a turnover of the sex chromosomes, where a pair of autosomes has effectively replaced the old pair of sex chromosomes. In therians, and less extensively in monotremes and birds, the strategy has been to accommodate the Y decay via dosage compensation.

Dosage compensation

Because of the loss of function of Y-linked genes, there must have been a pressure in males to increase the activity of the X partners. The imbalance occurs at two levels: between males and females, and between sex chromosomes and autosomes [Marin et al. 2000, Heard & Carrel 2009].

This increased activity is observed at various levels in either XY or ZW sex chromosome systems, but the molecular mechanisms vary greatly across lineages. In mammals, Marie Lyon demonstrated in 1961 that sex imbalances are solved by randomly inactivating one of the X chromosomes [Lyon, 1961]. It is important to note that the imbalances between the sexes are not directly under selection. It is the imbalance between the products of sex chromosomes and of the autosomes that must be regulated by the genome, independently in males and females, to maintain fitness in both sexes.

Previous studies have compared expression levels of X-linked genes to those of orthologous genes in species with different sex chromosomes, *i.e.*, where they are autosomal [Julien et al. 2012]. They found that the messenger-RNA (mRNA) level output per active chromosome copy is, on average, the same, effectively reducing the total expression output to 0.6 times the ancestral level. It is still controversial how this global output reduction is accommodated. The old hypothesis of a global up-regulation of X-linked genes has been recently ruled out in favour of a local up-regulation of dosage sensitive genes [Pessia et al. 2012], an up-regulation of translational efficiency [Wang et al. in writing] and a downregulation of some autosomal gene partners [Julien et al. 2012].

Mechanism of X-silencing in mammals

In eutherians, a transient phase of paternally imprinted X chromosome inactivation during the 4-8 cells stage is replaced by random X inactivation at the blastocyst stage [Pinheiro & Heard 2017]. The inactivation status is then fixed for each cell, and transmitted through cell duplication resulting in adult females that are a mosaic of cells with either the paternal or the maternal X inactivated.

The mechanism by which the inactivation is triggered varies between human and mouse: mice inactivate any X chromosome beyond the first one, while humans protect one X chromosome from inactivation, and turn on the inactivation machinery on the other. This distinction is of particular importance when one considers the case of XXX or XXY trisomic cells [Migeon 2017].

In mouse, the X chromosome is inactivated when the balance between two antisense long non-coding RNAs (lncRNAs) at the X inactivation centre (XIC) (see Figure 1), in Yang et al. 2011], *Xist* and *Tsix*, is resolved [Migeon 2017]. As they are mutual transcriptional repressors, *Xist* is expressed on the inactive X (Xi) while *Tsix* is expressed on the active X (Xa). In humans, *Tsix* is truncated and is inactive, so the repression of *Xist* on the Xa relies on another unknown mechanism. For the next steps, the mechanism is similar for both lineages. On the Xi, *Xist* spreads along the chromosome, coating it, and then recruits two polycomb complexes (PRC1 and PRC2), which hypoacetylate histones, stably condensing the chromatin into heterochromatin, and thus forming the Barr body [Barr & Bertram 1949, Marin et al. 2000, Avner & Heard 2001]. The early imprinting of the paternal X is maintained in the extra-embryonic tissues in mice and cattle and in marsupials it is maintained in all cells throughout life [Wang et al. 2014, Huynh & Lee 2003, Heard & Disteche 2006].

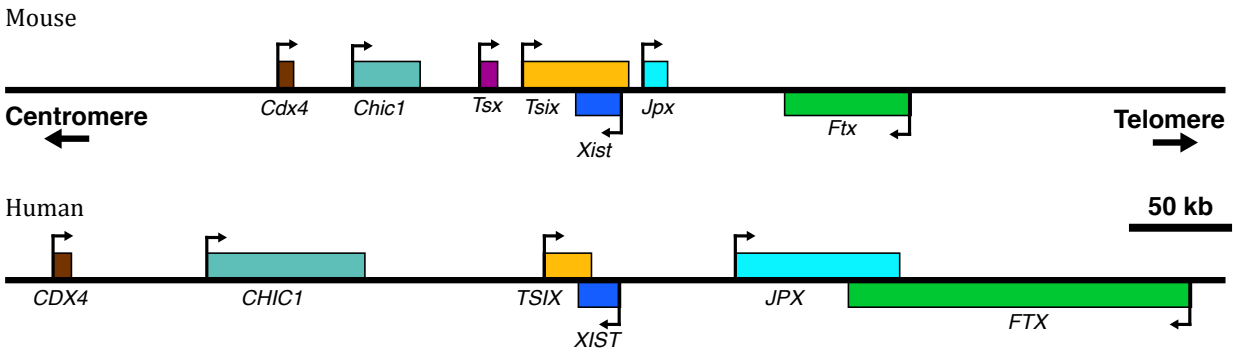

Figure 1: **X inactivation center in mouse and human**. Based on [Yang et al. 2011]

Genes escaping X chromosome inactivation (XCI)

Despite being an efficient mechanism, the silencing of one X chromosome is not absolute. It has been estimated that at least 23% of the 683 human X-linked genes escape XCI in adults in at least one tissue, cell line, or developmental stage, as do at least 15% of mouse X-linked genes [Tukiainen et al. 2017, Carrel et Willard 2005, Berletch et al. 2015]. The majority of the escapers are expressed from both female X chromosomes but inconsistently. These escapers are referred to as facultative escapers, as opposed to the constitutive escapers that are expressed from both Xa and Xi consistently [Peeters et al. 2014]. Giorgetti and colleagues attributed the difference between constitutive and facultative escapers to the structural organisation of the Xi [Giorgetti et al. 2016, more below].

Tukiainen and colleagues have generated the most recent catalogue of XCI escapers for human [Tukiainen et al. 2017]. In this study, the authors explored data from the GTEx project from 5500 transcriptomes from 29 tissues and showed that the escape from XCI results in differences in gene expression between males and females (sex-biased gene expression). For genes escaping XCI, the output from the two copies in females is expected to exceed the output from the single copy in males. They compared expression levels between males and females of all genes along the X chromosome, and observed that the non-PAR of the chromosome is dominated by female-biased escapers. These escapers are also more numerous in the more recent strata, suggesting a more selective escape status with time [Marin et al. 2000]. The PAR, on the other hand, is dominated by male-biased genes (*e.g.*, *SHOX* gene). The partial spreading of XCI beyond the PAR/non-PAR boundaries in females likely explains why the Xi PAR expression only reaches 80% of the Xa PAR expression.

In mouse, the escape status also varies between tissues, as demonstrated by a single-cell study in brain, spleen and ovaries [Berletch et al. 2015]. The escapers can also be divided into two groups depending on how consistent the escape is across organs. However, in these studies, being an escaper does not mean that there is an important contribution from the Xi, as genes often show some expression from the Xi amounting to less than 10% of the total expression from both chromosomes.

Present study

Our understanding of dosage compensation is still shadowed by our lack of a clear reason for why it has evolved. If dosage compensation was selected for in order to circumvent the deleterious effect of hemizygosity in the heterogametic sex, then one would expect dosage compensation to lead to increased levels of expression of X-linked genes in males, thus matching the ancestral expression levels. In contrast, what is observed in mammals is a decrease in the total expression output from the X in females, leading to gene-product imbalances across the genome.

Even if we assume that the balance between gene products from the X and from their autosomal partners is re-established via a combination of autosomal gene-by-gene downregulation [Julien et al. 2012], and by an increase of translational efficiency of X-linked genes [Wang et al., in writing], it is still not clear how an intermediate state including only one of either the global X-silencing or the aforementioned gene-by-gene regulations could have been maintained in the population without major deleterious effects on fitness. Because the decay of the Y chromosome is the cause for the dosage decrease, one would expect that the first response would be driven by males. Selection could act first in males and reduce the expression levels of autosomal gene partners of newly hemizygous X genes and/or increase their translational efficiency. Then, because these changes would affect females, XCI could evolve in response to changes in the stoichiometry of protein complexes and/or to translational changes.

A second open question is why some genes escape XCI. If we assume that the global silencing of the X was strongly selected for (as seems to be necessary to overcome the implications for genes that are part of protein complexes), why is it then avoided by a fraction of the genes? The silencing of the X being a global mechanism [Graves 2015], its avoidance necessitates a specific and new counter-mechanism, that thus is unlikely to arise neutrally. Because the escape from XCI is often facultative, it not only leads to differences in expression between males and females, but also between organs and between individuals (*i.e.*, females). Since some of the escapers have been associated with human disease (e.g. Turner syndrome, reviewed in [Hughes and Page, 2015]), this could lead to differences in disease predispositions between males and females, between tissues and also between individual females.

A hypothesis for the existence of escapers is based on the persistence of some genes on the Y chromosome. As genes from both the X and Y arise from the same ancestral gene, the two genes (hereafter "gametologs") retain a certain degree of homology. Therefore, Y-linked genes could assist the function of X-linked genes in males, thus alleviating the effects of hemizygosity for the X, and erasing the need for dosage compensation altogether. However, this hypothesis only applies to a potential small number of escapers. Although X escapers are enriched in gametologs [Slavney et al. 2015], not all X gametologs escape XCI. The presence of Y gametologs is not sufficient to explain some genes escaping XCI.

Finally, the difference in the frequency of escape from XCI between human and mouse (23% vs. 15%), as well as the variability in the degree of escape shown by genes, raises the question of to which extent escape from XCI is conserved across species. The lists of human and mouse escapers show very

few commonalities, which indicates a potential fast turnover of escapers. If this were the case, then we would gain novel insights by comparing more closely related species.

During my doctoral work, a clearer image has emerged regarding the differences between constitutive and facultative escapers. In 2016, Giorgetti and colleagues published a study of the structural organisation of the X chromosome [Giorgetti et al. 2016]. They compared Xa and Xi using Hi-C, ATAC-seq and RNA sequencing. They showed that in Neural Progenitor Cells (NPCs), the Xi shows an absence of active/inactive compartments and topologically associated domains (TADs), except around genes that escape XCI. They found that some genes were facultative escapers and showed variability between different NPC clones, whereas other genes were constitutive escapers and remained constant. They suggested that the escape status is acquired via the formation of topologically associated domains, which are necessary for chromatin accessibility and escape. They also noted that the particular conformation of the Xi in two megadomains is evolutionarily conserved, and could be responsible for XCI. Facultative escapers are silenced and re-expressed during XCI and would be more sensitive to stochastic changes in the chromosome 3D conformation than constitutive escapers.

Also recently, the link between escape status, presence of a Y gametolog and dosage sensitivity was studied by Naqvi and colleagues [Naqvi et al 2018]. They observed that ancestral dosage sensitivities (as measured by the conservation of miRNA target sites) were different for three groups of X-linked genes: highest for X gametologs that escape XCI, second highest for genes subject to XCI, and lowest for escapers that are not gametologs. These observations support the hypothesis that escapers that are gametologs and escapers without a Y partner have opposite reasons to escape XCI: either high dosage constraints that favour keeping the ancestral dosage, or low dosage constraints that allow for variability in escaping XCI.

When taken together, the conservation of Xi topographic domains can be perceived as being in contradiction with the extreme variability in the number and identity of escapers between human and mouse.

My aim is to use mammalian non-model species (rat, rabbit, opossum), in addition to human and mouse, to detect XCI escapers during development , understand to which degree the escape from XCI is conserved, and to identify the features that characterize developmental XCI escapers.

RESULTS

Upon joining the group, my shared interest with Prof. Henrik Kaessmann in sex chromosome evolution led us to design a project in that field. As most studies on sex chromosomes focused so far uniquely on adults, the idea was to use a novel evo-devo dataset that was nearing completion in the group to explore the developmental dimension of sex chromosome evolution.

The dataset consists of RNA sequencing (RNA-seq) data for 6 therians (human, rhesus macaque, mouse, rat, rabbit, opossum) and chicken, used as an evolutionary outgroup, and consists of 1,893 libraries, covering the development of 7 organs, 9-23 developmental stages (depending on the species) and 2-4 replicates per stage (Table 1).

Common name	**Species**	**Clade**	**Additional information**
Human	*Homo sapiens*	Eutheria	Elective abortions with normal karyotypes
Rhesus macaque	*Macaca mulatta*	Eutheria	
Mouse	*Mus musculus*	Eutheria	Outbred strain CD-1 (RjOrl:SWISS)
Rat	*Rattus novegicus*	Eutheria	Outbred strain Holtzman SD
Rabbit	*Oryctolagus cuniculus*	Eutheria	Outbred New Zealand breed
Grey short-tailed opossum	*Monodelphis domestica*	Metatheria	
Red junglefowl (chicken)	*Gallus gallus*	Aves	

Table 1: **List of species included in the dataset.** More information in Cardoso-Moreira et al., in review.

For each species, the time-series start when the organs can be identified and dissected separately from nearby tissues. That means, for example, embryonic (e) day 10.5 for the mouse and 4 weeks post conception (wpc) for human. The organs dissected were: brain (forebrain), cerebellum (hindbrain), heart, kidney, liver, and ovary or testis. Most timepoints were supported by 4 biological replicates (2 males and 2 females), except for the gonads, for which there were 2 replicates per sex. Due to the difficulty in obtaining samples, there are on average only two replicates per timepoint in primates. Due to the nature of sampling of human organs, and to a lesser extent macaque, samples were grouped together as biological replicates over broader developmental periods than in non-primates. As a consequence, samples that have passed different developmental milestones may be grouped together, thus introducing more variability in the data, compared to the other species.

When I joined the group, the available data was a collection of ~2,000 RNA-seq libraries already aligned to the genomic sequence. Because the dataset was not yet complete, some analyses had to be repeated several times following the addition or removal of individual libraries.

My doctoral projects:

- For the first months of my doctoral work, I work on a project on the evolution of sex-biased micro-RNAs (miRNAs) under the supervision of Dr. Maria Warnefors. My participation in this project, and its resulting publication in Genome Research, is discussed in its own chapter.

- My next project concerned the chromosome undergoing most changes during sex chromosome evolution: the Y-chromosome. Dr. Diego Cortez, a previous postdoctoral researcher in the group, produced transcriptome assemblies of Y chromosome genes for 15 mammals using RNA-seq data. This work was published in Nature in 2014, and was particularly novel in that it bypassed the need for assembled genomic sequence of the Y chromosome to discover Y-linked genes (which is a very resource-consuming endeavour). However, similarly to other studies on sex chromosomes, only adult samples were used. As gene annotation from RNA-seq data is dependent on genes being expressed in a given organ, the question was raised whether some potential Y-linked genes with expression restricted to development would escape annotation via this method.

  My project was, therefore, to apply the Y transcriptome assembly pipeline to the new evo-devo dataset. Given that the objective was to find Y-linked genes not previously known, it was necessary for my analyses to be thorough. Therefore, I implemented several modifications to the original pipeline that included, for example, getting rid of extensive random read drops that were originally implemented to reduce the library size and speed up computation (see Supplementary material S1). Unfortunately, the computational demands required to find the potentially missing genes forced me to abort the project. Using a dataset that included 2 males for 9 to 23 stages increased the amount of data more than 20-fold from the previous work, and it was deemed impossible to process at once through the exponential process that is transcript assembly. Adding a workaround similar to the original study in order to randomly select transcripts to create more manageable data sets, but at later stages of the pipeline, was contemplated, but finally deemed to cancel out the benefits of the developmental dimension. The possibility of finding unknown Y genes via transcriptome assembly from RNA-seq data remains, but will require further developments

in large datasets transcript assembly software. As a consequence, after a few months working on this project, my doctoral work was redirected towards other questions.

- The new direction of my research, which would end up being the main project of my doctoral work, concerns the evolution of genes that escape X chromosome inactivation (XCI). As stated in the introduction, the main goal was to study the conservation of escapers across mammals. In adults, the escape status of genes was shown to be very inconsistent both between tissues and in the amplitude of the sex bias.

Study of XCI escapers across mammals

As I was starting this project, I learned about the study of Tukiainen and colleagues, who established that differences in gene expression between males and females (sex biased expression) could be used as a proxy for escape status in human. This study was limited to adults and a large number of replicates were used. My first step was therefore to investigate how consistent is sex-biased expression in adults in our dataset, and how it compares to the escape status of X-linked genes as described in the literature.

Starting from the genome alignments, I created expression tables in both reads per kilobase of exon model per million mapped reads (RPKMs) and counts per million (CPMs) using the package EdgeR (version 3.16.5, [Robinson et al. 2010]), which also normalizes the data using the method TMM. The gene annotations used included both the known set of protein coding genes and an annotation of novel long non-coding RNAs created by Dr. Ray Marin in the group. The tables were created using only reads mapping to a unique locus. Because sex chromosomes are known to contain a large number of repetitive elements, and because the homology between the X and Y chromosomes could potentially lead to an underestimation of gene expression if reads mapping to both gametologs were discarded, most analyses were also performed on expression tables that also took into account reads mapping to multiple locations. However, the inclusion of multi-mapped reads did not change any of the conclusions described below.

Due to an almost complete absence of data for females in rhesus macaque, this species was excluded from this work.

Sex-biased expression in adults

My first step was to study how consistent is sex bias in adults in our dataset, and how it compares to the escape status of X-linked genes in the literature.

To do so, I compared male and female expression levels using a Student's t-test on mouse data (I pooled together the last 2 developmental stages that correspond to sexual maturity) for 862 X-linked genes in all somatic organs.

I found that 373 genes (43%) were biased in at least one organ. There were 67 genes known in the literature to escape XCI that were present in our annotation, and 45 of them (67%) were biased in adults. 30% of the biased genes were overexpressed in males, 49% were overexpressed in females, and 20% were expressed at significantly higher level in both male and female samples, depending on the organ. Because only 45 of the 373 sex-biased genes in adults were known to be escapers in the literature, the sex-bias cannot be explained by XCI escape in the majority (88%) of cases for adult genes.

Sex-biased expression during development

My next step was to apply Tukiainen's method for detecting XCI escapers via sex-bias in expression. In my case, instead of many adult biological replicates, the development dimension could provide the power needed to uncover XCI escape. I started by testing the difference in expression between male and female samples, grouped across development. I used a Mann-Whitney U-test (as it does not require normality) combining all stages for each gene, but it resulted in an excessive number of false positives, even after False-Discovery-Rate multiple test correction (which in turn, made some clearly interesting genes barely significant, if at all). At the same time in the group, Dr. Margarida Cardoso Moreira was studying genes with significant changes in temporal expression during organ development, termed Developmentally Dynamic Genes (DDGs). She found that the majority (79-91%) of protein-coding genes are classified as DDGs, and thus I realized that my analyses needed to take into account the developmental trajectory of the genes, and that the average over development would not be sufficient.

I therefore undertook an exploratory approach, where I visually classified genes as either sex-biased or not. I produced 3336 plots representing the median expression level per sex of 861 mouse protein-coding genes in the 5 somatic organs (only genes expressed at more than 1RPKM in the organ were studied), as well as 4117 plots representing 1116 human genes. I divided development in 3 periods: pre-natal, post-natal, and sexually mature, which were scored independently (Figure 2). To

avoid confirmation bias, I was unaware of the names of the genes I was scoring (i.e., used Ensembl identifiers). The scoring was: strongly female-biased, female-biased, same between sexes, male-biased, and strongly male-biased.

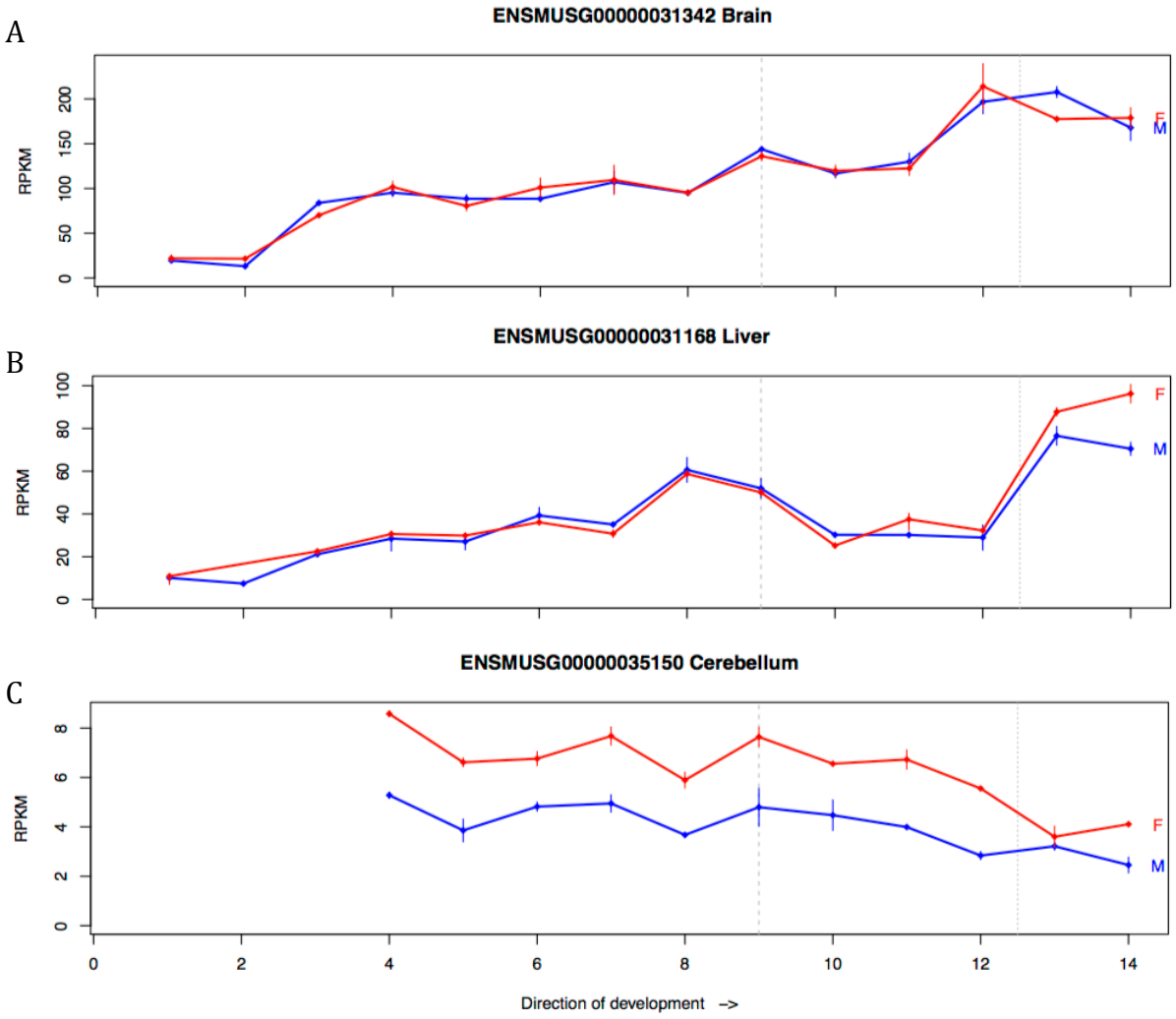

Figure 2: **Example of the original plots used for sex-biased expression scoring in mouse**. The x-axis shows stages from early (left) to late (right) development. The y-axis represents the median expression level across replicates. The vertical bars mark birth (left) and sexual maturity (right). Red: female samples, blue: male samples. Vertical dashed lines correspond to the neonate stage (left) and sexual maturity (right).

I found that most genes (83% in mouse and 70% in human) do not show sexually dimorphic expression (e.g., *Gpm6b,* Fig. 2A); that among the 146 remaining genes showing dimorphism in mouse, 94 showed clear dimorphism at the adult stage with no trend in bias at any other stage (39% towards male over expression, 61% towards female overexpression) (e.g., *Ebp* in liver, Fig. 2B); and finally, that only a minority of genes (5 in mouse) show continuous dimorphism throughout development (e.g., *Eif2s3x* in cerebellum, Fig. 2C) [Berletch et al. 2015, Marks et al. 2015]. Most notably, all dimorphic protein-coding genes with female overexpression across the 3 developmental

periods and organs in mouse were X gametologs (i.e., *Ddx3x, Kdm5c, Eif2s3x, Kdm6a*), except for *Pbdc1*.

This exploratory approach yielded promising results, so I proceed to select a statistical method that would take advantage of the strength of the data, that is, that the developmental points are not independent but are instead a time series. Statistical analyses were therefore conducted via linear model with the Limma package in R [Ritchie et al. 2015, Law et al. 2014]. To keep consistency across projects within the group, I used the same expression tables as those used in Cardoso-Moreira et al., where expression levels were estimated in CPM. The model was fitted separately for each organ via a design matrix taking into account the development stage and sex of the samples. Scale normalization across libraries was done via the TMM normalization as implemented in Limma. An additional filter removed genes that had consistently less than 3 CPM (implemented by the filterByExpr function). Voom transformation, taking into account the design of the model, was applied to the normalized and filtered DGEList object, as it is recommended for samples with variable library sizes. Afterwards, the model was fitted vial lmFit, and empirical Bayes smoothing to the standard errors via eBayes. In an attempt to normalize the data, and to not be over sensitive to X linked genes, all statistics were performed genome-wide. The significant p-values for the sex parameter of the model were recovered via TopTable, using *Benjamin–Hochberg* multiple test correction over only the X chromosome. The complete list of p-values for each X-linked expressed gene, organ and species, will be provided upon request.

Mouse developmental female-biased gene expression

Gene	Significance	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank
Xist	***	14.58	16.89	16.79	18.08	16.22	16.79	**1**
Kdm6a/Utx	***	7.00	1.33	5.56	10.47	7.14	7.00	**2**
Kdm5c	***	8.85	2.41	5.56	8.09	6.63	6.63	**3**
Eif2s3x	***	2.31	6.51	3.45	5.96	4.48	4.48	**4**
Pbdc1/Cxorf26	**	5.04	0.00	2.67	9.36	2.62	2.67	**5**
5530601H04Rik	**	5.89	1.81	0.65	8.09	2.62	2.62	**6**
Jpx	**	5.62	1.22	2.15	3.06	1.94	2.15	**7**
Firre	**	2.36	1.22	1.78	5.54	2.02	2.02	**8**
Ddx3x	N.S.	0.37	0.00	1.56	3.01	0.11	0.37	**42**
Tspyl2	N.S.	0.00	0.00	0.19	0.92	0.47	0.19	**87**
Sox3	N.S.	0.55	0.00	0.17	0.01	0.20	0.17	**92**
Rbmx	N.S.	0.06	0.00	0.02	2.68	0.13	0.06	**218**
Uba1	N.S.	0.00	0.00	0.03	0.08	0.11	0.03	**313**
Usp9x	N.S.	0.00	0.00	0.84	0.01	0.00	0.00	**517**
Zfx	N.S.	0.00	0.00	0.48	0.10	0.00	0.00	**517**
Total								**988**

Table 2: **Top 8 most female-biased X-linked genes across development in mouse (top) and remaining X gametologs (bottom).** Gene names in bold indicate X gametologs, underlined genes are lncRNAs involved in XCI. Significance: *** = p-value <0.001, ** = p-value <0.01, * = p-value <0.05, N.S. = p-value > 0.05. The columns brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance.

The most strongly continuously female-biased gene during development is the lncRNA *Xist* (p-value = $1.6e^{-17}$). This is not surprising because *Xist* is a female-specific gene in all eutherians. It can therefore be used as a positive control for this approach when exploring datasets with more variability between biological replicates. The following three highest female-biased genes, *Kdm6a*, *Kdm5d*, *Eif2s3x* (p-values of $1e^{-7}$, $2.3e^{-7}$, and $3.3e^{-5}$, respectively) are all X gametologs. The next two genes, *Pbdc1* and *5530601H04Rik*, are a pair of head-to-head protein-coding and lncRNAs genes with unknown function, but already known to escape XCI by locally remodeling the chromatin [Lopes et al. 2011]. Finally, the last 2 significant genes, *Jpx* and *Firre*, are two lncRNAs actively involved in XCI. Jpx controls the initiation of XCI by activating *Xist* on Xi [Tian et al. 2010] and *Firre* is in charge of preserving the long term H3K27me3 on the Xi [Yang et al. 2015].

Surprisingly, despite being involved in XCI, *Tsix* does not show sex-biased expression in mouse (rank: 246). The gene *Ftx* (which is also known to be involved in mouse XCI [Chureau et al. 2010, Furlan et al. 2018]) is not part of the gene annotations used in this work.

Rat developmental female-biased gene expression

Gene	Significance	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank
Xist	***	27.12	12.60	26.47	16.12	18.90	18.90	**1**
Pbdc1/Cxorf26	***	12.22	2.30	7.19	3.25	9.54	7.19	**2**
Eif2s3x	***	11.21	0.59	3.87	1.73	10.22	3.87	**3**
Jpx	***	8.56	2.30	3.68	3.69	7.28	3.69	**4**
5530601H04Rik (by homology with mouse)	**	8.61	2.11	2.92	2.55	6.33	2.92	**5**
Med14	**	4.44	1.36	0.01	2.16	4.87	2.16	**6**
5530601H04Rik (by homology with mouse)	N.S.	2.98	1.22	0.45	0.55	1.61	1.22	**7**
Dach2	N.S.	1.41	0.01	0.84	1.15	1.18	1.15	**8**
Kdm6a/Utx	N.S.	1.57	0.00	1.04	0.44	4.45	1.04	**12**
Kdm5c	N.S.	4.44	0.55	0.12	1.06	0.51	0.55	**30**
Zfx	N.S.	0.36	0.00	0.00	0.53	0.33	0.33	**84**
Uba1	N.S.	0.22	0.00	0.00	0.12	1.57	0.12	**307**
Usp9x	N.S.	1.19	0.00	0.00	0.11	0.38	0.11	**360**
Rbmx	N.S.	0.46	0.00	0.00	0.07	0.35	0.07	**473**
Ddx3x	N.S.	1.75	0.00	0.06	0.06	2.02	0.06	**475**
Total								**884**

Table 3: **Top 8 most female-biased X-linked genes across development in rat (top) and remaining X gametologs (bottom).** Gene names in bold indicate X gametologs, underlined genes are lncRNAs involved in XCI. Significance: *** = p-value <0.001, ** = p-value <0.01, * = p-value <0.05, N.S. = p-value > 0.05. The columns Brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance. Two newly annotated lncRNAs are homologous to mouse *5530601H04Rik,* and are named as such.

Similarly to mouse, *Xist* is the most female-biased genes (p-value= $1.2e^{-19}$). *Pbdc1* and *5530601H04Rik* are also in the top 5 most female-biased genes in rat, which indicates that their sex-biased expression is likely conserved among rodents. In rat, the gametologs *Eif2s3x* and *Med14* are also significantly female-biased, as is *Jpx*.

Ftx and *Firre* are not annotated in rat, and the expression of the gametolog *Sox3* was below our cutoff of 1 RPKM.

Rabbit developmental female-biased gene expression

Gene	Significance	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank
Xist (5' Part)	***	21.97	16.13	16.67	21.86	14.86	16.67	**1**
Xist (3' Part)	***	11.23	14.72	12.25	14.79	10.08	12.25	**2**
Ddx3x	N.S.	0.89	0.23	0.96	2.30	1.09	0.96	**3**
Eif2s3x	N.S.	0.89	1.51	1.49	0.53	0.91	0.91	**4**
Kdm6a/Utx	N.S.	0.40	0.14	0.51	0.70	4.40	0.51	**5**
Mrpl32	N.S.	0.49	0.45	0.96	0.19	0.52	0.49	**6**
Slc25a5	N.S.	0.24	0.45	0.33	0.00	1.90	0.33	**7**
Novel lncRNA *XLOC_042844*	N.S.	0.22	0.03	0.38	0.49	0.29	0.29	**8**
Tspyl2	N.S.	0.01	0.00	0.35	0.70	0.04	0.04	**72**
Usp9y	N.S.	0.40	0.00	0.03	0.00	0.14	0.03	**79**
Zfx	N.S.	0.32	0.00	0.03	0.00	0.18	0.03	**106**
Rps4x	N.S.	0.00	0.00	0.09	0.00	0.03	0.00	**355**
Kdm5c	N.S.	0.00	0.00	0.11	0.00	0.08	0.00	**355**
Rbmx	N.S.	0.00	0.00	0.00	0.00	0.00	0.00	**355**
Uba1	N.S.	0.00	0.00	0.03	0.00	0.05	0.00	**355**
Pbdc1/Cxorf26	N.S.	0.00	0.00	0.14	0.00	0.28	0.00	**355**
Total								**621**

Table 4: **Top 8 most female-biased X-linked genes across development in rabbit (top) and remaining X gametologs (bottom).** Gene names in bold indicate X gametologs, underlined genes are lncRNAs involved in XCI. Significance: *** = p-value <0.001, ** = p-value <0.01, * = p-value <0.05, N.S. = p-value > 0.05. The columns Brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance. The *Xist* coordinates are split between 2 transcript annotations.

In rabbit only *Xist* shows significant female-biased expression during development (the *Xist* annotation is divided into 2 transcripts, and both are significant). Although none of the other genes passed the significance threshold, we can still observe a clear divide between *Ddx3x, Eif2s3x, Kdm6a* and the rest of the gametologs. Interestingly, *Pbdc1* is not biased at all, which makes its sex bias a rodent-specific trait (*5530601H04Rik* doesn't have a known homolog outside of rodents). The homologs of other lncRNAs involved in inactivation machinery (*Jpx, Ftx*) are not known in rabbit. As in rat, *Sox3* shows too low expression in rabbit to be analysed.

Human developmental female-biased gene expression

Gene	Significance	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank
XIST	***	6.82	15.22	5.14	1.21	7.22	6.82	**1**
TSIX	*	1.43	1.88	2.69	0.62	2.51	1.88	**2**
GABRA3 (exons 5 and 6)	N.S.	0.74	1.13	3.26	0.00	2.51	1.13	**3**
KDM5C	N.S.	1.43	1.67	1.10	0.20	0.70	1.10	**4**
RPS4X	N.S.	0.97	1.59	0.63	0.05	1.84	0.97	**5**
FATE1	N.S.	1.43	1.21	0.97	0.04	0.15	0.97	**6**
ARMCX5	N.S.	0.38	0.78	0.76	0.20	1.84	0.76	**7**
ZFX	N.S.	1.06	1.55	0.68	0.36	0.00	0.68	**8**
DDX3X	N.S.	0.93	0.76	0.53	0.00	0.31	0.53	**14**
SOX3	N.S.	0.88	0.38	0.37	0.00	0.18	0.37	**55**
KDM6A/UTX	N.S.	0.33	0.83	0.85	0.34	0.07	0.34	**60**
TBL1X	N.S.	0.60	0.32	0.85	0.03	0.15	0.32	**72**
PRKX	N.S.	0.30	0.68	0.22	0.00	0.00	0.22	**149**
TMSB4X	N.S.	0.54	0.63	0.01	0.00	0.15	0.15	**266**
NLGN4X	N.S.	0.22	0.52	0.07	0.13	0.00	0.13	**313**
TXLNG	N.S.	0.48	0.10	0.85	0.00	0.00	0.10	**426**
RBMX	N.S.	0.09	0.33	0.42	0.07	0.00	0.09	**498**
PCDH11X	N.S.	0.09	0.07	0.30	0.04	0.00	0.07	**563**
EIF1AX	N.S.	0.22	0.10	0.06	0.00	0.00	0.06	**587**
USP9X	N.S.	0.04	0.04	0.07	0.00	0.00	0.04	**703**
TSPYL2	N.S.	0.00	0.08	0.06	0.00	0.00	0.00	**972**
PBDC1	N.S.	0.05	0.10	0.10	0.00	0.00	0.05	**631**
Total								**1002**

Table 5: **Top 8 most female-biased X-linked genes across development in human (top) and remaining X gametologs (bottom).** Gene names in bold indicate X gametologs, underlined genes are lncRNAs involved in XCI. Significance: *** = p-value <0.001, ** = p-value <0.01, * = p-value <0.05, N.S. = p-value > 0.05. The columns Brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance. The genes coordinates used in this analysis for *GABRA3* were retired since, and are now annotated as corresponding to exons 5 and 6 of *GABRA3*. *FTX*, *JPX* and *Firre* (involved in mouse XCI) are part of the gene annotations used.

In human, only *XIST* (p-value = $1.5e^{-7}$), and surprisingly, *TSIX* (p-value = 0.013) show significant female-biased gene expression during development. The two genes are antisense to each other, but *TSIX* is not known to be involved in XCI in humans (and as described, *Tsix* not show sex-biased expression in mouse).

The gametologs *KDM5C*, *RPS4X* and *ZFX* are in the top 10 most female-biased genes, despite not crossing the significance threshold. Three genes show more female-biased expression than the most

biased gametologs: *GABRA3* (gamma-aminobutyric acid (GABA) A receptor subunit alpha 3), *FATE1*, and *ARMCX5*. Neither is known to escape XCI.

Across all eutherians, the conservation of female-bias is strongest for *Xist*, which can be explained by the fact that it is the most fundamental contributor to XCI.

The comparison between the eutherians (mouse, rat, rabbit and human) identifies 3 gametologs that show female-biased expression (though not always passing the significance threshold) in at least 2 of the 4 species: *Kdm6a*, *Eif2s3x*, *Kdm5c*. In addition, *Ddx3x* is third most female-biased gene in rabbit, and is close to significance in both human and mouse.

Opossum developmental female-biased gene expression

Gene	Significance	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank
Novel lncRNAs, XLOC_045717	***	8.97	9.16	23.58	5.75	11.04	9.16	**1**
Frmd7	***	5.87	9.16	9.31	2.06	9.58	9.16	**1**
Novel lncRNAs, XLOC_045517	***	8.97	8.98	19.74	1.52	10.04	8.98	**3**
Novel lncRNAs, XLOC_044938	***	8.97	9.05	13.03	1.58	8.53	8.97	**4**
Rsx	***	8.83	7.41	21.12	5.67	10.16	8.83	**5**
Hmgb3	***	6.35	9.16	13.53	3.55	8.64	8.64	**6**
Rragb	N.S.	0.93	1.61	1.92	0.55	1.71	1.61	**7**
Klf8	N.S.	2.02	2.02	1.06	0.09	0.22	1.06	**8**
Thoc2	N.S.	0.00	0.22	0.03	0.21	0.32	0.21	**30**
Rpl10	N.S.	0.00	0.17	0.34	0.09	2.36	0.17	**42**
Phf6	N.S.	0.00	1.83	0.03	0.21	0.14	0.14	**47**
Rps4	N.S.	0.00	0.22	0.03	0.04	0.46	0.04	**150**
Mecp2	N.S.	0.00	0.01	0.25	0.03	0.09	0.03	**205**
Hsfx	N.S.	0.00	0.04	0.03	0.03	0.00	0.03	**224**
Rbmx	N.S.	0.00	0.44	0.03	0.11	0.02	0.03	**234**
Atrx	N.S.	0.00	0.02	0.02	0.02	0.11	0.02	**314**
Rmb10	N.S.	0.00	0.74	0.02	0.03	0.00	0.02	**341**
Hcfc1	N.S.	0.00	0.17	0.01	0.03	0.02	0.02	**407**
Uba1	N.S.	0.00	0.83	0.02	0.02	0.00	0.02	**416**
Tfe3	N.S.	0.00	0.72	0.02	0.03	0.00	0.02	**429**
Kdm5d	N.S.	0.00	0.58	0.01	0.03	0.00	0.01	**452**
Sox3	N.S.	0.00	0.90	0.01	0.05	0.00	0.01	**468**
Total								**534**

Table 6: **Top 8 most female-biased X-linked genes across development in opossum (top) and remaining X gametologs (bottom).** Gene names in bold indicate X gametologs, underlined genes are lncRNAs involved in XCI. Significance: *** = p-value <0.001, ** = p-value <0.01, * = p-value <0.05, N.S. = p-value > 0.05. The columns Brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance.

In opossum, 6 genes show significant levels of female-biased gene expression (p-value between 6.9e-10 and 2.3e-9). *Rsx*, the lncRNA analogous to the eutherian's *Xist*, is unsurprisingly among them [Grant et al. 2012]. The protein-coding gene *Frdm7* has already been described as sex-biased with 60% of its expression coming from the Xi [Wang et al. 2014], but is not known to be associated with the XCI machinery. It is known to cause nystagmus (involuntary eye movement) in humans when mutated with a penetrance of 100% in males and 53% in females [Richards et al. 2015]. The gametolog *Hmgb3* shows strong sex-bias, and is the only gametolog to show female-biased expression in this species. Remarkably, 3 newly annotated lncRNAs are strongly significant: XLOC_045517,

which is antisense to the only gametolog that is sex-biased (*Hmgb3*), XLOC_44398, which is lowly expressed, and XLOC_45717, which is expressed at moderately high levels throughout the dataset (between 10 and 20 RPKM on average, see Figure 3)

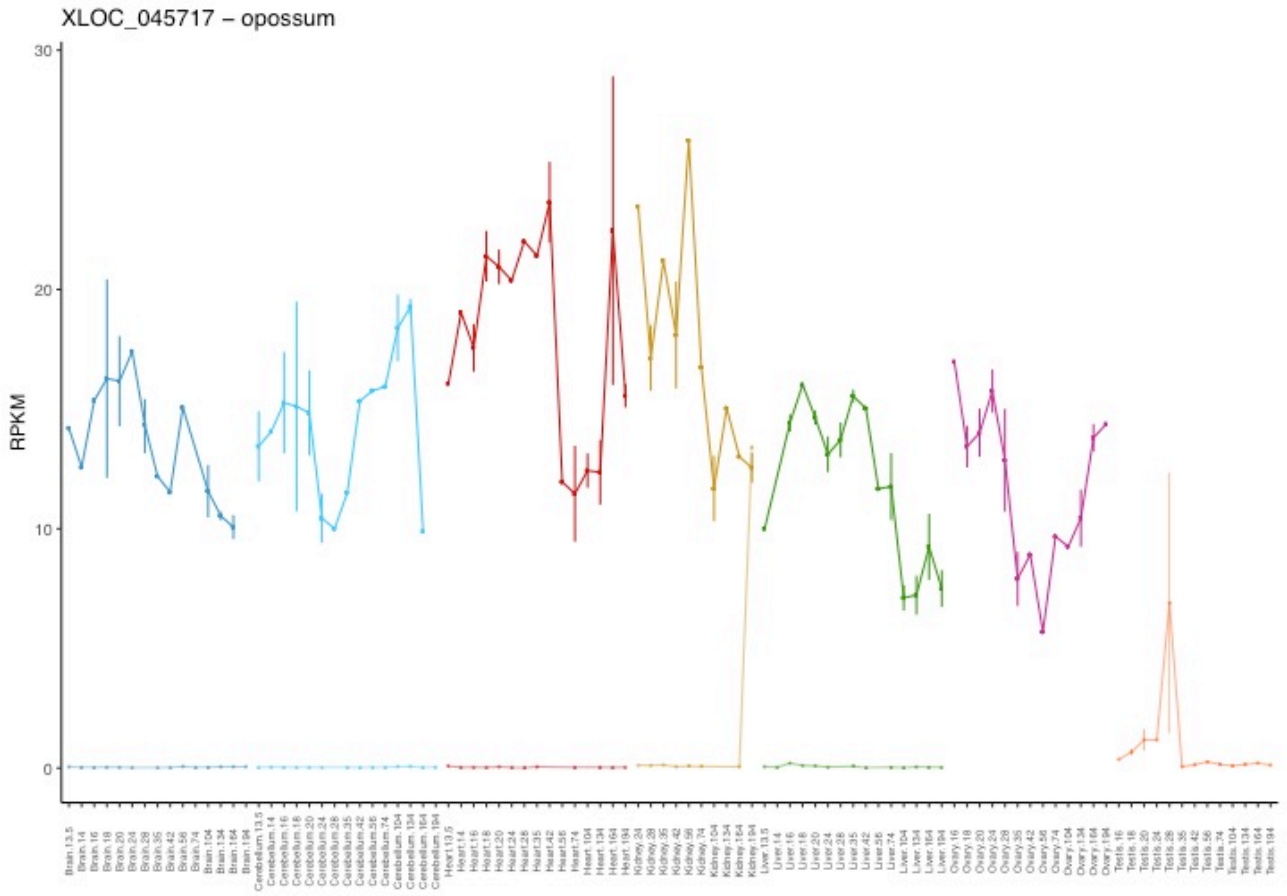

Figure 3: **Expression of the newly annotated lncRNA XLOC_45717 throughout opossum organ development.** The x-axis represents developmental stages from early (e13.5, left) to late (194 days post birth, right) development for forebrain/brain (dark blue), hindbrain/cerebellum (light blue), heart (red), kidney (yellow), liver (green), ovary (pink) and testis (orange). The y-axis shows expression levels. The vertical bars show the range of expression between replicates. Expression in males is represented by thin lines (near 0, bottom) and in females by bold lines. All following plots were created with ggplot2 [Wickham 2016]

The discovery of these 3 strongly female-biased lncRNAs raised the possibility that these genes could be involved in XCI in opossum. A new project has therefore been started in the group, which aims to characterize the spatial distribution of XLOC_45717's RNA in the cell via Fluorescent In-Situ Hybridization (FISH), and test the hypothesis that its RNA coats the X chromosome. Dr. Mari Sepp is leading this project.

Across all 5 species I found that only a small fraction of genes show significant female-biased gene expression during development. These mostly comprise *Xist/Rsx*, lncRNAs actively involved in XCI, and a subset of X gametologs. My next step consisted in understanding the reason why some

gametologs are consistent escapers (the ones on the top most strongly female-biased genes) and others are not.

Escape from XCI and Y gametologs

Given that the current protein-coding Y-linked genes were most likely initially retained because of their dosage sensitivity as broadly expressed regulators [Bellott et al. 2014], their X homologs were probably haploinsufficient. Therefore, initially, all gametologs had to avoid XCI in females in order to have both alleles expressed in the two sexes. However, several Y-linked genes have evolved male-specific functions [Cortez et al. 2014]. We therefore hypothesized that because Y genes with testis-specific expression will not complement the function of their X gametologs in male somatic tissues, there is no need for these X gametologs to escape XCI and therefore to show a difference in expression between males and females during development. Only Y gametologs that are still ubiquitously expressed have the possibility of complementing their X gametologs in every organ and so only this set of gametologs should escape XCI.

To test this hypothesis, I created expression tables for chicken genes that are orthologous to current Y genes. As the chicken orthologs are situated on a pair of chromosomes that have not been selected as sex chromosomes (chromosome pairs 1 and 4), they were not subject to the expression changes pressures that are typical of sex chromosomes, and are expected to have maintained, on average, the ancestral expression levels.

I investigated expression levels for chicken and mouse, and for chicken and human at corresponding stages of development, as reported by Dr. Cardoso-Moreira (see Table 7). By comparing the ratio of ancestral (chicken) expression levels with the current expression levels, I could categorize Y genes as broadly expressed or male-specific.

Species	**Corresponding developmental stages**							
Mouse	e16.5	e17.5	e18.5	P0	P3	P14	P28	P63
Human	12wpc	13wpc	19wpc	19wpc	20wpc	Toddler (2-4 years)	youngAdult (25-32 years)	youngMidAge (39-41 years)
Chicken	e10	e12	e12	e12	e14	P0	P70	P155

Table 7: **Corresponding developmental stages across species as used in this study**. The vertical line marks sexual maturity.

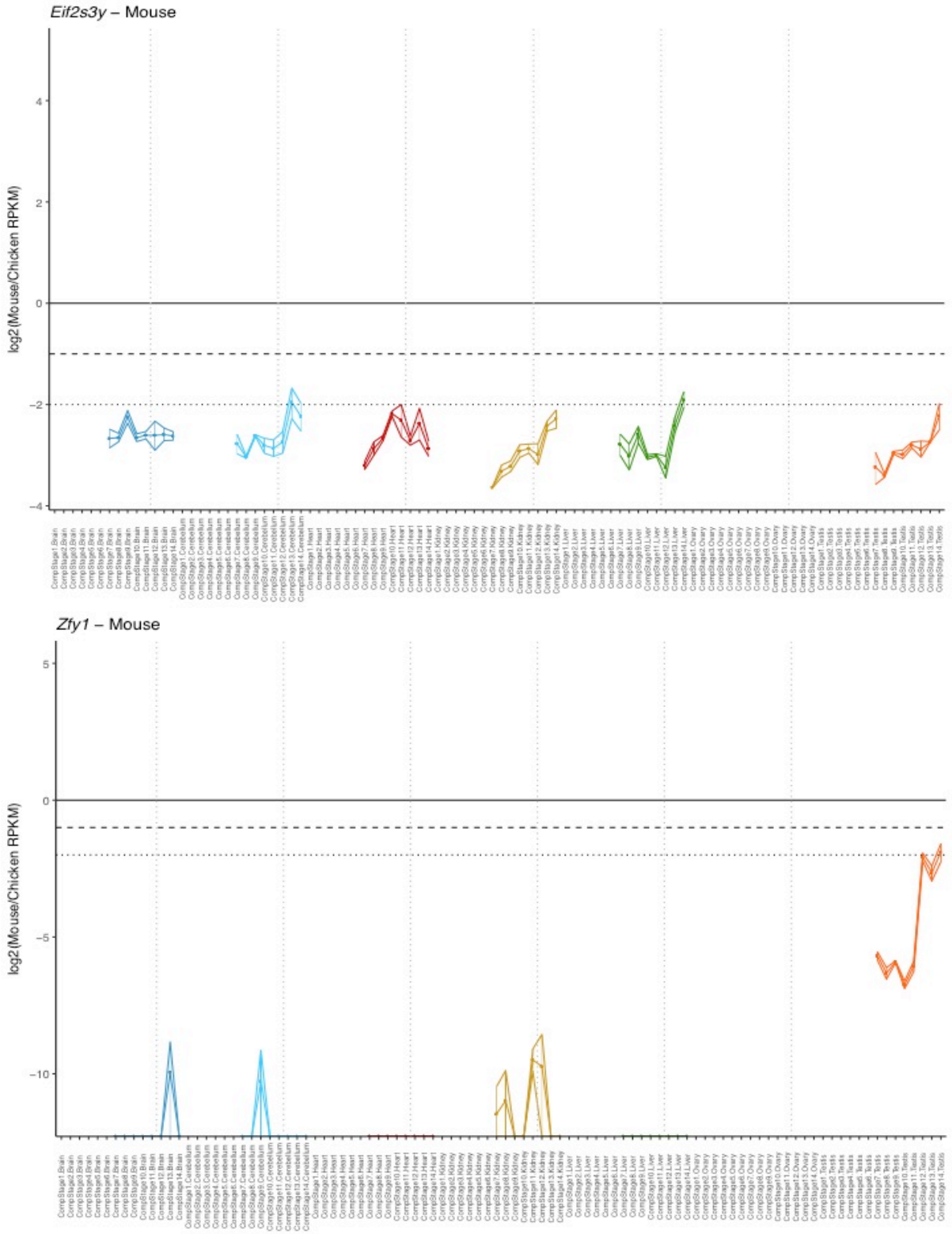

Figure 4**: Loss of ancestral expression for *Eif2s3y*, classified as broadly expressed, and for *Zfy1*, classified as male-specific, in mouse**. The x-axis shows stages from early (corresponding to an e16.5 mouse, left) to late (corresponding to a 63 days post birth mouse, right) for forebrain/brain (dark blue), hindbrain/cerebellum (light blue), heart (red), kidney (yellow), liver (green), ovary (pink) and testis (orange). The y-axis represents the $\log_2$ ratio of expression difference between mouse and chicken, per chromosome copy. A $\log_2$ (ratio) of -2 corresponds to a reduction of expression levels by a factor of 4 in the eutherian lineage. The width of the interval represents the minimum and maximum difference in expression ratio in our data (minimum mouse expression divided by maximum chicken expression, and vice-versa.) The vertical lines represent the data points.

When we combine these results with the previous analysis of constant gene expression differences between the sexes during mouse development we obtain the following table:

Gene	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank	Y homolog expression
Kdm6a	7.00	1.33	5.56	10.47	7.14	7.00	**2**	Ubiquitous
Kdm5c	8.85	2.41	5.56	8.09	6.63	6.63	**3**	Ubiquitous
Eif2s3x	2.31	6.51	3.45	5.96	4.48	4.48	**4**	Ubiquitous
Ddx3x	0.37	0.00	1.56	3.01	0.11	0.37	**42**	Ubiquitous
Tspyl2	0.00	0.00	0.19	0.92	0.47	0.19	**87**	N.a.
Sox3	0.55	0.00	0.17	0.01	0.20	0.17	**92**	Testis-specific
Rbmx	0.06	0.00	0.02	2.68	0.13	0.06	**218**	Testis-specific
Uba1	0.00	0.00	0.03	0.08	0.11	0.03	**313**	Testis-specific
Usp9x	0.00	0.00	0.84	0.01	0.00	0.00	**517**	Testis-specific
Zfx	0.00	0.00	0.48	0.10	0.00	0.00	**517**	Testis-specific

Table 8: **Comparison of levels of female-biased developmental expression for X gametologs and the expression patterns of their Y homologs in mouse.** The columns Brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance.

From this table it is clear that the X gametologs that most consistently show female biased expression have ubiquitously expressed Y homologs. *Ddx3x* appears to be the exception. However, the developmental expression of *Ddx3x* in males and females (Figure 5) suggests that this gene is also female biased and that it was our statistical approach that failed to identify it as such (i.e. false negative). Therefore, in mouse the pattern is clear: X gametologs of Y genes with ubiquitous expression escape XCI (i.e., show female biased expression), whereas X gametologs of Y genes with testis-specific expression do not.

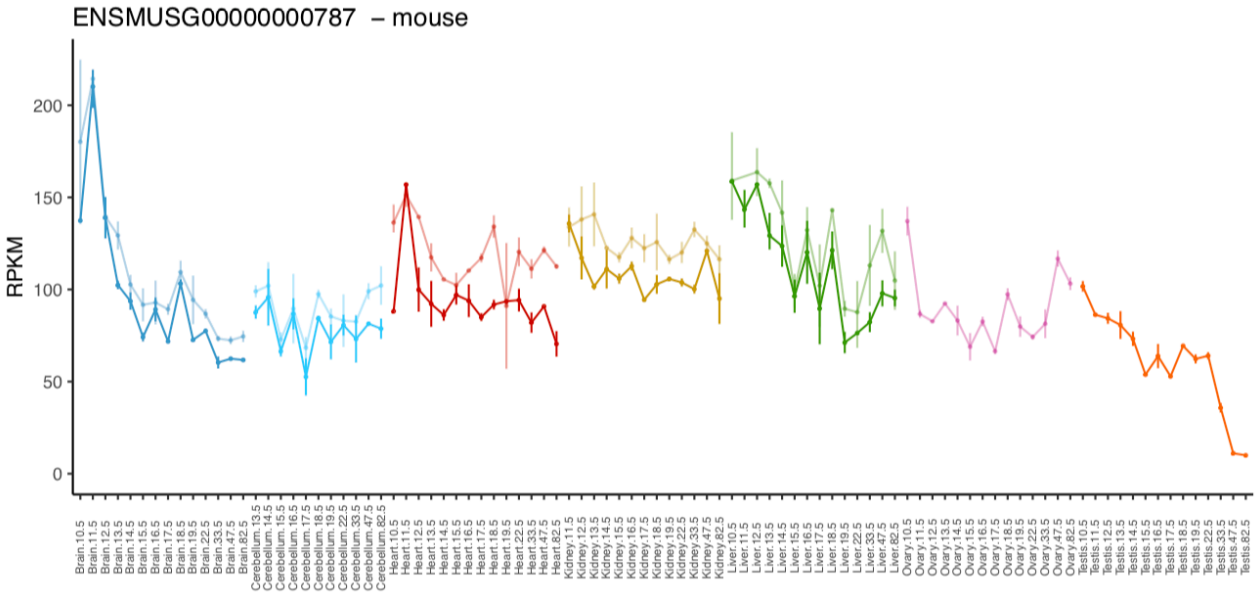

Figure 5: **Developmental expression of *Ddx3x* in mouse males and females**. The x-axis represents stages from early (e10.5, left) to late (63 days post birth, right) development for brain/forebrain (dark blue), hindbrain/cerebellum (light blue), heart (red), kidney (yellow), liver (green), ovary (pink) and testis (orange). The last 2 time points in each organ correspond to sexual maturity. The y-axis shows expression levels. Plotted is the median expression in each stage with the bars showing the range of expression among replicates. Expression for males in bold lines (bottom) and for females in thin lines (top).

The results for human are less clear (Table 9). As in mouse, the X gametologs that were among the top 10 genes with the most significant sex-biased expression all have ubiquitously expressed Y gametologs. However, 2 X gametologs that do not show female-biased expression also have ubiquitously expressed Y gametologs.

Gene	Brain	Cerebellum	Heart	Kidney	Liver	Median	Rank	Y homolog expression
KDM5C	1.43	1.67	1.10	0.20	0.70	1.10	**4**	Ubiquitous
RPS4X	0.97	1.59	0.63	0.05	1.84	0.97	**5**	Ubiquitous
ZFX	1.06	1.55	0.68	0.36	0.00	0.68	**8**	Ubiquitous
DDX3X	0.93	0.76	0.53	0.00	0.31	0.53	**14**	Ubiquitous
SOX3	0.88	0.38	0.37	0.00	0.18	0.37	**55**	Variable
KDM6A/Utx	0.33	0.83	0.85	0.34	0.07	0.34	**60**	Ubiquitous
TBL1X	0.60	0.32	0.85	0.03	0.15	0.32	**72**	No expr.
PRKX	0.30	0.68	0.22	0.00	0.00	0.22	**149**	No expr.
TMSB4X	0.54	0.63	0.01	0.00	0.15	0.15	**266**	Very low & Ubiquitous
NLGN4X	0.22	0.52	0.07	0.13	0.00	0.13	**313**	Variable
TXLNG	0.48	0.10	0.85	0.00	0.00	0.10	**426**	No expr.
RBMX	0.09	0.33	0.42	0.07	0.00	0.09	**498**	Low expr. & Testis-specific
PCDH11X	0.09	0.07	0.30	0.04	0.00	0.07	**563**	Variable
EIF1AX	0.22	0.10	0.06	0.00	0.00	0.06	**587**	Ubiquitous
USP9X	0.04	0.04	0.07	0.00	0.00	0.04	**703**	Ubiquitous
TSPYL2	0.00	0.08	0.06	0.00	0.00	0.00	**972**	**N.a.**

Table 9: **Comparison of the levels of female-biased developmental expression for X gametologs and the expression patterns of their Y homologs in human**. The columns Brain, cerebellum, heart, kidney, liver show the $-\log_{10}$ of the p-value in the respective organ. The median is calculated across all 5 organs, and is used to rank expressed X-linked genes in decreasing order of significance.

In the case of *EIF1AX*, it is possible that similarly to what happened with mouse *Ddx3x*, we simply failed to identify female-biased expression even though it exists (because of a combination of the statistical method used, the low number of replicates, and the small effect size of female-male differences). As show in the figure below, we can observe a trend of female overexpression in kidney, liver, and early heart.

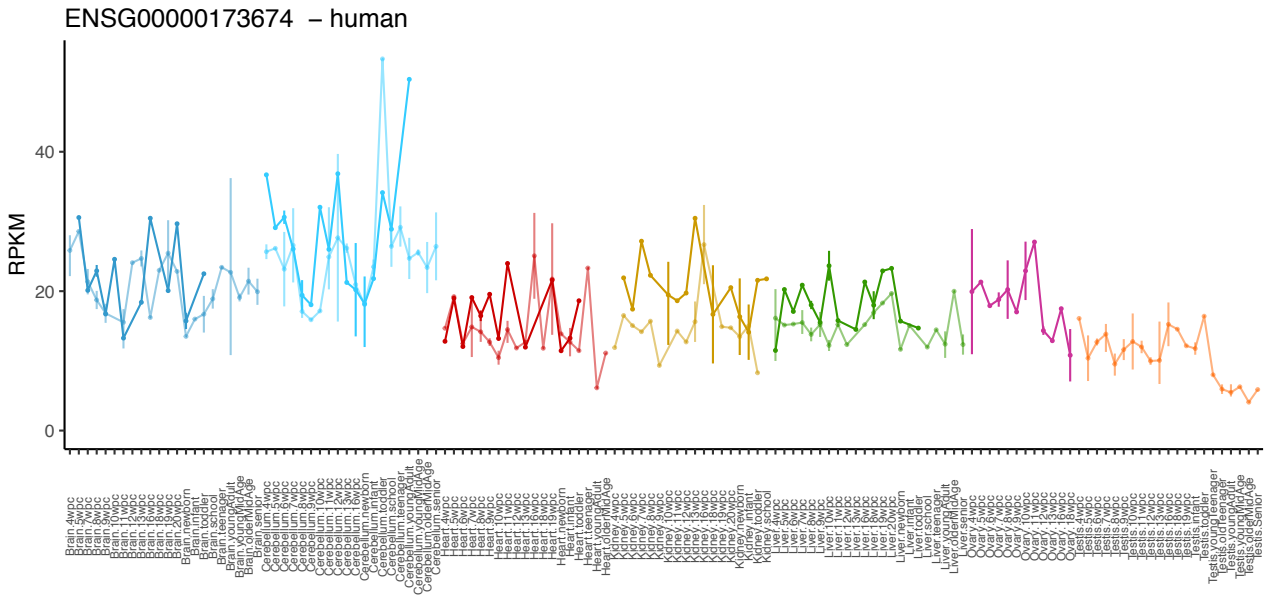

Figure 6: **Developmental expression of *EIF1AX* in human males and females**. The x-axis represents stages from early (4wpc, left) to late (senior, right) development for brain/forebrain (dark blue), hindbrain/cerebellum (light blue), heart (red), kidney (yellow), liver (green), ovary (pink) and testis (orange). Plotted is the median expression in each stage with the bars showing the range of expression among replicates. Expression for males in thin lines (bottom) and for females in bold lines (top).

However, *USP9X* does not show any evidence for female-biased expression. It is unclear why for this single X gametolog, its Y gametolog ubiquitous expression in somatic organs is not correlated with a consistent female-bias.

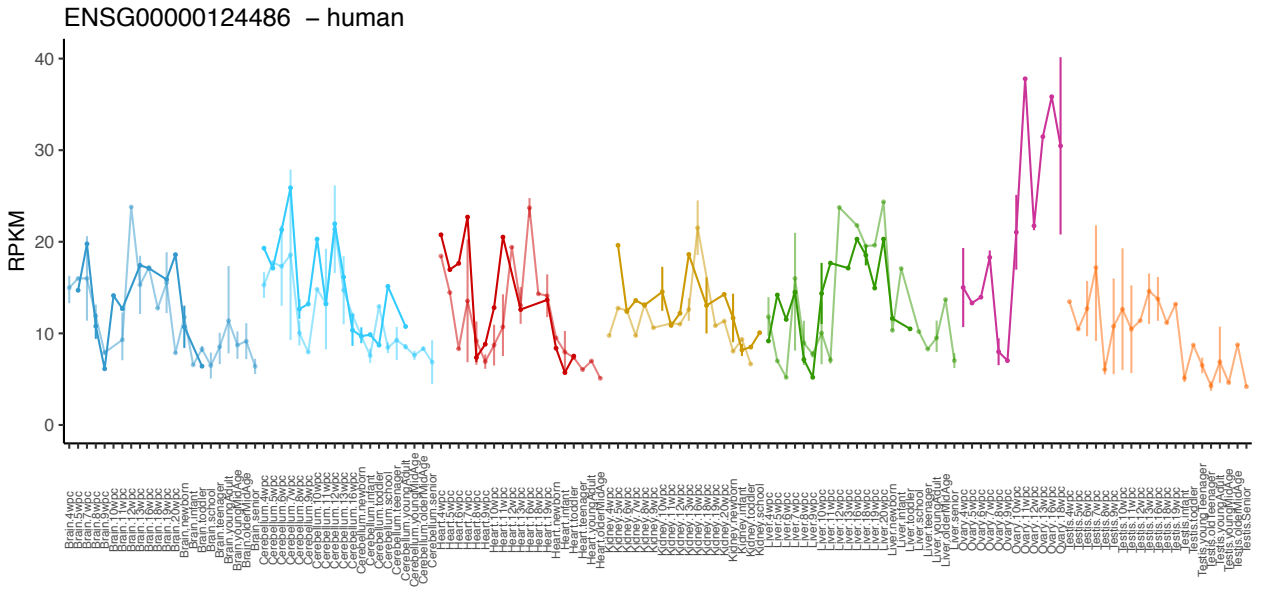

Figure 7: **Developmental expression of *USP9X* in human males and females**. The x-axis represents stages from early (4wpc, left) to late (senior, right) development for brain/forebrain (dark blue), hindbrain/cerebellum (light blue), heart (red), kidney (yellow), liver (green), ovary (pink) and testis (orange). Plotted is the median expression in each stage with the bars showing the range of expression among replicates. Expression for males is depicted in thin lines and for females in bold lines.

In conclusion, our data suggest that there are only two reasons for a gene to be systematically differentially over-expressed during development in females: either it is a gametolog of a broadly expressed Y gene, or is a lncRNA directly involved in XCI. These are discussed in page 42.

Allelic contributions from Xa and Xi

Given that most X-linked genes do not show consistent differences in expression between the sexes, and that it was shown in 1949 with the discovery of the Barr body, that the Xi is mostly inactive and condensed, we expect the physiology and chromatin state of the female Xa to be similar to those of the single male X. Therefore, one can expect that the contribution of the Xa to the female expression would correspond to the total expression observed in males, and that the additional expression in females (when present) would be due to the contribution of the Xi (Figure 8).

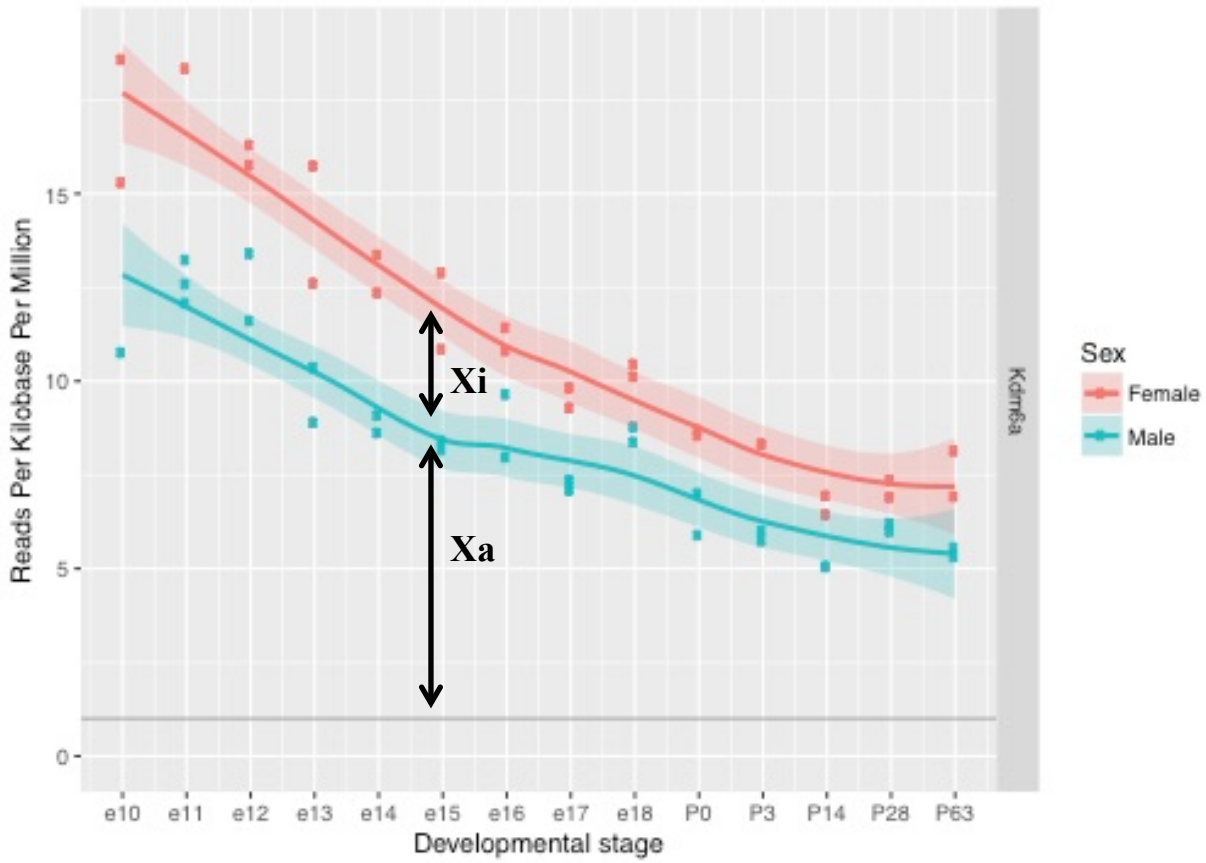

Figure 8: **Model of the allelic contribution of Xa and Xi to the total female expression, using the mouse *Kdm6a* in brain as an example**. The x-axis represent development from left to right, and the y-axis represents the expression level. Coloured lines represent the expectation for the mean and the range correspond to 90% confidence interval with loess smoothing.

So far, the relative contributions of the Xa and Xi in females and the sex-bias in expression levels have not been jointly analysed in the same study. So for my final project in this doctoral work I compared the allelic contribution of each X to the total expression output in females and to the expression output in males.

To do so, one needs to be able to attribute reads to either the active or to the inactive X. However, in eutherians one X chromosome is randomly inactivated in each cell [Migeon 2017]. Any RNA-seq dataset from bulk tissue contains cells with the maternal X inactivated (mXi), as well as cells with the paternal X inactivated (pXi). Most genes will therefore show expression coming from 2 alleles in the same proportion as that of the cell populations of mXi and pXi. Previous studies have therefore used single-cell RNA-seq in mouse and human to attribute expression to either the maternal or paternal X in adults [Berletch et al. 2015].

As developmental single-cell RNA-seq data was not available at the time of my doctoral work, I used another approach. In marsupials, XCI operates differently from eutherians in that XCI is restricted to the paternal X in females. This means that in female opossums, only genes that escape XCI will show expression coming from 2 alleles in bulk tissue data.

The majority of opossum samples in the dataset are from individuals for whom we do not know the genomic sequence, and obtaining genomic data is no longer possible for many of the samples given that they were fully used for the RNA extractions. Therefore, in order to attribute expression to either the paternal or maternal alleles, I had to find the alleles present in each sample directly from the RNA-seq data. For that, I used the mpileup tool from samtools [Li et al. 2009] which counts how many times each of the 4 nucleotides are encountered at each position of aligned reads (the input are the alignment bam files and the output are vcf tables with allele counts).

However, given the nature of RNA sequencing and alignment, sequencing errors, genomic variations from the reference, and the allowance of alignment mismatches during the attribution of a read to a locus, virtually all positions within genes show more than one nucleotide. Therefore, given that opossums are diploid, I first discarded the least common nucleotides at each locus, if more than 2 were represented. I then differentiated false SNPs from real alleles by removing X-linked variable positions that were also present in male samples (suggestive of misalignment), and only allowed for those variable positions that were supported by at least 50 reads, with at least 10% of them coming from a second allele.

The results of this approach are illustrated using the gene *Dkc1* (Figure 9). *Dkc1* is known to escape XCI in marsupials. It was shown to have 37.5% of its contribution coming from the pXi in opossum

brain at e13 [Wang et al. 2014]. Despite not being significantly female-biased in our analysis (p-value = 0.2, 10th most consistently biased gene during development), the trend of female overexpression is clearly visible in our data. Out of 128 female samples, 21 had alleles that passed all filters. On average, the contribution of the second allele was 31.6% over all informative organs, stages, and replicates. The sample in our dataset for which we detected 2 alleles that is closest in time to the e13 brain examined by Wang and colleagues is that of a 2 days post-birth brain, and it shows 38.7% of contribution from the second allele, which shows a good agreement.

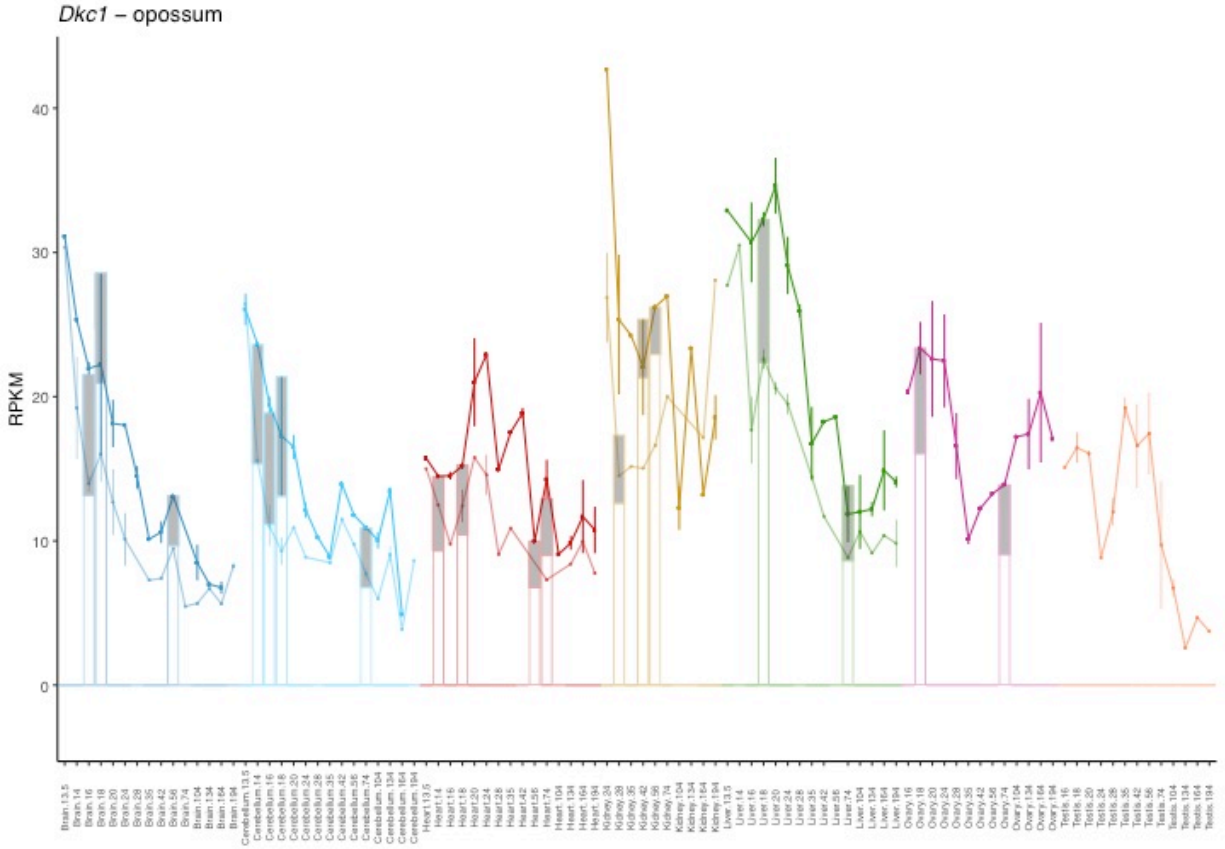

Figure 9: **Expression in opossum of *Dkc1* in males and females, with estimated contribution from each allele**. X-axis shows stages from early (e13.5, left) to late (180 days post birth, right) development for forebrain/brain (dark blue), hindbrain/cerebellum (light blue), heart (red), kidney (yellow), liver (green), ovary (pink) and testis (orange). The second timepoint corresponds to a neonate. The y-axis represents the median expression level between replicates. Expression for males is depicted by the thin lines and in females by the bold lines. The vertical lines represent the maximum and minimum expression among replicates. The barplots represent the relative contribution of the major allele (empty) and the minor allele (grey) on average in heterozygous samples.

Unfortunately, *Dkc1* was the only known sex-biased gene for which we detected a high number of heterozygous samples. This method relies completely on the natural frequency of heterozygous individuals for a gene. Given that our samples come from relatively inbred individuals, the probability of finding heterozygous genes is low. Moreover, the absence of genomic sequence data for the individuals sampled requires the use of several filters to separate technical variants from real

heterozygosity, which in turn, prevent us from identifying any minor allele representing less than 10% of the expression output.

Despite promising, the optimal application of this method requires opossums created by crosses from 2 parents from divergent lineages, in order to ensure the highest possible levels of heterozygosity, and for the existing alleles to be known via DNA sequencing of the sampled individuals.

SEX-BIASED MICRO-RNAS

Upon my arrival in the group, a pilot project was proposed to me regarding sex-biased expression of micro-RNAs. The available data consisted of adult samples for mouse gonads (2 females, 3 males), and opossum gonads (2 females, 3 males) and liver (3 females, 2 males).

I performed the first exploratory analyses for this project. I aligned fastq files containing the short reads against the genome using Bowtie 1.1.2 [Langmead et al. 2009]. In Opossum, chromosome 1 is longer than what the Bowtie script will accommodate, so I had to transform the coordinates to artificially create an additional shorter chromosome. I used bedtools 2.19.1 [Quinlan 2014] to create read count tables. I then used DESeq2 [Love et al. 2014] in R to find significantly sex-biased miRNAs. As expected, the somatic tissues showed a much higher correlation in expression levels between the sexes than the gonads. Interestingly, an in accordance with previous studies, numerous X-linked miRNAs in mouse and opossum showed a higher expression in testes than ovaries (Figure 10) [Song et al. 2009; Meunier et al. 2013]

As no miRNAs were yet annotated on the mouse Y chromosome, I searched for unknown Y-linked miRNAs using the same transcriptome subtraction approach as used by Cortez and colleagues in 2014, and that I later modified to detecting potentially missing Y genes (Supplementary material S1). I removed from male libraries all reads that were also present in females. However, when I mapped the remaining reads to the Y chromosome, most reads mapped to locations known to produce piRNAs, and no satisfying new miRNA was found.

These analyses were repeated by Dr. Maria Warnefors, who was at the time a postdoctoral fellow in the group, and were included in our publication in Genome Research [Warnefors, M., Mössinger, K., Halbert, J., **Studer, T**., VandeBerg, J. L., Lindgren, I., ... & Kaessmann, H. (2017). Sex-biased microRNA expression in mammals and birds reveals underlying regulatory mechanisms and a role in dosage compensation. *Genome research*.]

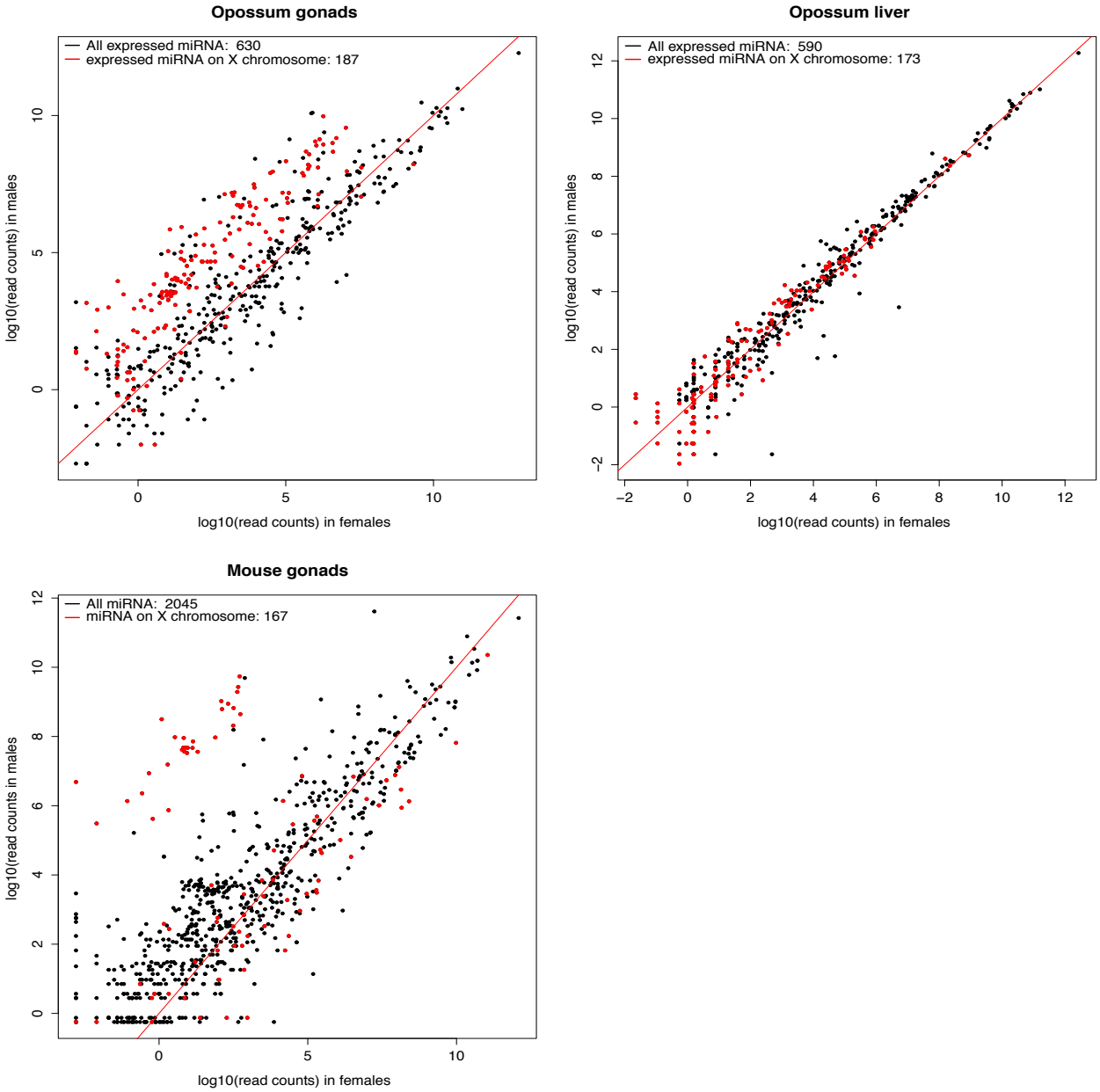

Figure 10: **Comparison of expression levels between the sexes in opossum gonads (top left) and liver (top right) and mouse gonads (bottom left).** Each point represents a miRNA that is expressed in at least one sample in the organ. The x-axis represents the median $\log_{10}$ of read counts between female replicates, and the y-axis represents the median log10 of read counts in males. The total number of expressed miRNAs and X-linked miRNAs (red points) is indicated.

DISCUSSION

Across all 5 therian species, we observe a strong correlation between consistent female biased expression across organs and development in X gametologs and the ubiquitous expression of their Y chromosome partners.

Caveats in the model used to identify sex-biased expression

As shown for a few individual genes (e.g., *Ddx3x* in mouse, *EIF1AX* in human), the statistical analysis used to identify consistent sex-biased expression during development (based on a linear model) may have sometimes failed to identify some genes as significantly biased (false negatives), which can explain some of the differences in the number of significant sex-biased genes between species.

The false negatives can be due to several factors, some of which can have variable effects between species, including the amount of genetic diversity in the species, and the grouping of replicates of slightly different stages to increase the number of biological replicates per timepoint. These factors will increase the variance in gene expression among replicates and therefore decrease the power to call differential expression. Other factors have a similar impact across species, and are intrinsic to the statistical model. Indeed, Limma fits linear models to the data to reject or not the hypothesis that two groups are equal. However, as shown by the work of Dr. Margarida Cardoso-Moreira, most genes are developmentally dynamic in at least one organ, and the temporal profiles may not be linear.

In order to overcome these technical problems, a collaboration was initiated between M.Sc. Svetlana Ovchinnikova, a PhD student in Dr. Simon Anders group, in order to find more relevant tests that can be applied to developmentally dynamically regulated genes in a non-linear manner. Her initial results were consistent with this work for the genes discussed here.

Consistent sex-bias across development, organs, and species

Across all eutherians species studied here, some genes are consistently strongly sex-biased.

X gametologs

Although already known to be over-represented among XCI escapers in adult tissues and cell cultures [Park et al. 2010], X gametologs show additional specific characteristics when their expression levels are compared between the sexes across development for multiple organs.

Their expression patterns fall into 2 categories: *Kdm6a, Eif2s3x, Kdm5c* and to a lesser extent *Ddx3x* are among the most strongly, consistently female-biased genes, across all 4 eutherian species, while the rest of the gametologs are not sex-biased during development. When the breadth of expression of their Y homologs is considered, the correlation between consistent sex-bias and broad expression of the Y homolog is strong.

The likely explanation for this correlation is that broadly expressed Y gametologs can compensate the expression of their X gametologs in males, as they retain a certain level of redundancy in function. This functional redundancy was shown, for example, by Yamauchi and colleagues in 2016 when they effectively replaced the Y-linked *Eif2s3y* by overexpressing their gametolog *Eif2s3x*. The potential complementation of the expression levels of X gametologs by their Y partners has been described in previous studies, and I observed it as well (Supplementary Figure S2). It is notable that, the hemizygosity of X gametologs is blamed for the phenotypes observed in the Turner syndrome (XO females), suggesting that for some X-linked genes the presence of genes expressed from a second X chromosome in females or from a Y chromosome in males is required. The candidate genes for Turner syndrome symptoms (webbed neck, puffy hands and feet, sterility) are mostly XCI escapers. (Reviewed in Hughes & Page 2015)

On the other hand, Y-linked protein-coding genes that developed a male-specific function through the evolution of male-specific tissue expression, are no longer expressed in somatic tissues. Therefore, their X gametologs are susceptible to the same evolutionary forces that act on any other X-linked genes which have lost their Y homolog, and which, in therians, lead to dosage compensation via X chromosome silencing.

Correlation between consistent sex-bias and dosage sensitivity

The observation that only X gametologs of broadly expressed Y-linked genes consistently escape XCI during development (unlike the rest of the genes escaping XCI) correlates well with our knowledge on the dosage sensitivity of these genes [Park et al, 2010]. The recent discovery by [Naqvi et al. 2018] that dosage sensitivity differs between general escapers of XCI, genes not escaping XCI, and escapers that are gametologs of Y genes, from low to high sensitivity, respectively, reinforces the

idea that there are two kinds of escape from XCI: high dosage sensitivity forces some genes to maintain expression on two chromosomes (either XX or XY, depending on the sex) and low dosage sensitivity allows expression levels to be dictated by either 1 or 2 chromosomes with little impact on fitness, resulting in strong variability in escape between tissues, cell lines, individuals, species, and developmental stages.

Overall, the conserved ancestral expression from 2 X chromosomes in females and from X and Y chromosomes in males is likely the "default" mode of sex chromosome expression, which will necessarily result in a consistent bias between the sexes regarding the expression of the X-linked genes. Therefore, the global XCI in females is possibly the secondary consequence of dosage compensation in males, either by the increase in transcriptional efficiency of X-linked genes and/or by the downregulation of their autosomal partners. A recent study conducted on the European common frog *Rana temporaria*'s homomorphic sex chromosomes supports this global model [Ma et al. 2018]. This study compared expression levels between males and females at 5 different developmental stages and for 3 adult organs (brain, liver and gonads). Among the genes studied, no gene showed consistent sex bias across development, and only 2% of genes showed sex bias across all three adult stages. They also didn't find any differences in the male/female expression ratio between the autosomes and the sex chromosomes. Very interestingly, two differences are observed between the results in anurans and therians: 1) because of the homomorphism of their sex chromosomes, the notion of "gametology" doesn't apply to this frog, and 2) no sex-biased gene was consistently significantly sex biased across developmental stages, and no loss of expression was observed between sex chromosomes and autosomes. This suggests that in absence of Y chromosome degradation, no consistent gene expression bias evolves. Most likely, in anurans there is no imbalance between the autosomes and the sex chromosomes in either sex, and therefore there is no need for dosage compensation systems, which in turn does not create the need for consistent sex biased expression.

Female-biased lncRNAs

Unsurprisingly, *Xist* is the most female-biased gene due to its female-specificity. LncRNAs involved in XCI upstream of *Xist,* like *Jpx* and *Firre*, show strong sex-biased expression (when expressed above 1 RPKM). However, unlike *Xist*, they are not female-specific.

Their sex-bias is directly linked to their female-specific function, as they are required to have an efficient level of activity in females, but not in males.

Other genes that are sex-biased

A third and final category of genes that show consistent sex bias are protein-coding genes and lncRNAs, some already known to escape XCI, that are not gametologs and that do not have any known direct involvement in XCI. Prior to this work, their characteristics as escapers were not particularly remarkable when compared to other escapers, and went mostly unnoticed among the large number of genes escaping XCI. However, their expression pattern throughout development groups them together with genes associated with one unique process: XCI.

Given that understanding the molecular mechanisms responsible for the establishment of XCI is a hot topic in the field of sex chromosome dosage compensation, these genes are prime candidates for having a function in XCI. The functional study of these genes is already underway, as Dr. Mari Sepp, a postdoctoral fellow in the group, is currently exploring the spatial localization in the cell via Fluorescent In Situ Hibridization (FISH) of one of these candidate genes in Opossum: XLOC_045717.

Allelic contribution in genes escaping XCI

Because of the very condensed state of the inactive X (Xi) in mammals, and because the vast majority of X-linked genes show equivalent expression level between both sexes, it is reasonable to think that the general physiology of the active X (Xa) in females is equivalent to the one of the only male X. Therefore, we assumed that the female overexpression observed in a handful of genes is the result of the Xi contribution to expression.

Testing this hypothesis has, however, proven difficult.

Opossum *Dkc1* gene promisingly fits the model of allelic contribution of Xa and Xi. However, the properties of the data available made it impossible to draw any more conclusions. The crossing of two well-diverged parental lines is crucial to obtain samples with heterozygous genes. Having genomic sequence for the individuals would also help lower the threshold for calling informative SNPs (which in the present approach requires at least 10% of contribution from the minor allele). Finally, advances in single-cell sequencing technologies will allow to perform this analysis on eutherians, as well as on any lineage showing random XCI.

CONCLUSION

In this work, I compared the expression levels of genes on the X chromosome between males and females during development, for multiple organs and for multiple species. I discovered that only lncRNAs directly involved in X chromosome silencing, X gametologs of broadly expressed Y-linked genes, and a few protein-coding genes and lncRNAs with unknown function were significantly consistently biased across development and organs. Although some genes did not always pass the significance threshold, these trends were strongly conserved across species, in particular for lncRNAs and X gametologs of broadly expressed Y-linked genes. Due to their similarities in expression patterns with genes directly involved in XCI, the remaining genes with unknown function are of particular interest as candidates for having a function in XCI. Refinement of the statistical method used to call sex-biased expression will allow the reduction of false negative results.

I proposed a model stating that the contribution of the inactive X to the total gene expression in females is responsible for the difference in total expression level between the sexes. The only gene informative to test this hypothesis is in conformity with the model. Further testing of this model will require more suitable data.

Finally, I presented my contribution to a peer-reviewed publication regarding sex-bias in micro-RNAs expression levels in eutherians. I highlighted the presence of a group of X-linked micro-RNAs that are overexpressed in male gonads, and searched for new micro-RNAs located on the mouse Y chromosome but without success.

REFERENCES

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

Bachtrog, D., Hom, E., Wong, K. M., Maside, X., & de Jong, P. (2008). Genomic degradation of a young Y chromosome in Drosophila miranda. *Genome biology*, *9*(2), R30.

Barr, M. L., & 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. In *Problems of Birth Defects* (pp. 101-102). Springer, Dordrecht.

Bellott, D. W., Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Cho, T. J., ... & Kremitzki, C. (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*, *508*(7497), 494.

Berletch, J. B., Ma, W., Yang, F., Shendure, J., Noble, W. S., Disteche, C. M., & Deng, X. (2015). Escape from X inactivation varies in mouse tissues. *PLoS genetics*, *11*(3), e1005079.

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

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., ... & Huang, X. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome biology*, *16*(1), 30.

Charlesworth, B., & Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *355*(1403), 1563-1572.

Chen, J., Wildhardt, G., Zhong, Z., Roeth, R., Weiss, B., Steinberger, D., ... & Rappold, G. (2009). Enhancer deletions of the SHOX gene as a frequent cause of short stature: the essential role of a 250 kb downstream regulatory domain. *Journal of medical genetics*, *46*(12), 834-839.

Chureau, C., Chantalat, S., Romito, A., Galvani, A., Duret, L., Avner, P., & Rougeulle, C. (2010). Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region. *Human molecular genetics*, *20*(4), 705-718.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., ... & Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature*, *508*(7497), 488.

De la Fuente, R., Parra, M. T., Viera, A., Calvente, A., Gómez, R., Suja, J. Á., ... & Page, J. (2007). Meiotic pairing and segregation of achiasmate sex chromosomes in eutherian mammals: the role of SYCP3 protein. *PLoS genetics*, *3*(11), e198.

Ellis, N., & Goodfellow, P. N. (1989). The mammalian pseudoautosomal region. *Trends in Genetics*, *5*, 406-410.

Ezaz, T., Stiglec, R., Veyrunes, F. & Graves, J. A. M. Relationships between vertebrate ZW and XY sex chromosome systems. *Curr. Biol.* **16**, R736–R743 (2006)

Furlan, G., Hernandez, N. G., Huret, C., Galupa, R., van Bemmel, J. G., Romito, A., ... & Rougeulle, C. (2018). The Ftx noncoding locus controls X chromosome inactivation independently of its RNA products. *Molecular cell*, *70*(3), 462-472.

Giorgetti, L., Lajoie, B. R., Carter, A. C., Attia, M., Zhan, Y., Xu, J., ... & Dekker, J. (2016). Structural organization of the inactive X chromosome in the mouse. *Nature*, *535*(7613), 575.

Grant, J., Mahadevaiah, S. K., Khil, P., Sangrithi, M. N., Royo, H., Duckworth, J., ... & Elgar, G. (2012). Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *nature*, *487*(7406), 254.

Graves, J. A. M. (1995). The evolution of mammalian sex chromosomes and the origin of sex determining genes. *Phil. Trans. R. Soc. Lond. B*, *350*(1333), 305-312.

Graves, J. A. M. (2016). Evolution of vertebrate sex chromosomes and dosage compensation. *Nature Reviews Genetics*, *17*(1), 33.

Heard, E., & Disteche, C. M. (2006). Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes & development*, *20*(14), 1848-1867.

Heard, E., & Carrel, L. (2009). Foreword: Coping with sex chromosome imbalance. *Chromosome Research*, *17*(5), 579-583.

Hegyi, H., & Hassan, H. (2018). Pervasive chromatin remodeling at X-inactivation escape genes in schizophrenic males. *bioRxiv*, 300624.

Hellborg, L., & Ellegren, H. (2004). Low levels of nucleotide diversity in mammalian Y chromosomes. *Molecular Biology and Evolution*, *21*(1), 158-163.
Hughes, J. F., & Page, D. C. (2015). The biology and evolution of mammalian Y chromosomes. *Annual review of genetics*, *49*, 507-527.
Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Graves, T., Fulton, R. S., ... & Wang, Q. (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature*, *483*(7387), 82.
Huynh, K. D., & Lee, J. T. (2003). Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature*, *426*(6968), 857.
Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., ... & Kaessmann, H. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS biology*, *10*(5), e1001328.
Karpen, G. H., Le, M. H., & Le, H. (1996). Centric heterochromatin and the efficiency of achiasmate disjunction in Drosophila female meiosis. *Science*, *273*(5271), 118-122.
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, *14*(4), R36.
Lahn, B. T., & Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science*, *286*(5441), 964-967.
Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, *10*(3), R25.
Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, *15*(2), R29.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.
Lopes, A. M., Arnold-Croop, S. E., Amorim, A., & Carrel, L. (2011). Clustered transcripts that escape X inactivation at mouse XqD. *Mammalian genome*, *22*(9-10), 572.
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.
Ma, W. J., Veltsos, P., Toups, M. A., Rodrigues, N., Sermier, R., Jeffries, D. L., & Perrin, N. (2018). Tissue Specificity and Dynamics of Sex-Biased Gene Expression in a Common Frog Population with Differentiated, Yet Homomorphic, Sex Chromosomes. *Genes*, *9*(6).
Marks, H., Kerstens, H. H., Barakat, T. S., Splinter, E., Dirks, R. A., van Mierlo, G., ... & Kalkan, T. (2015). Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome biology*, *16*(1), 149.
Marín, I., Siegal, M. L., & Baker, B. S. (2000). The evolution of dosage-compensation mechanisms. *Bioessays*, *22*(12), 1106-1114.
Marin, R., Cortez, D., Lamanna, F., Pradeepa, M. M., Leushkin, E., Julien, P., ... & Trefzer, T. (2017). Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome research*.
Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., ... & Kaessmann, H. (2013). Birth and expression evolution of mammalian microRNA genes. *Genome research*, *23*(1), 34-45.
Migeon, B. R. (2017). Choosing the active X: the human version of X inactivation. *Trends in Genetics*.
Muller, H. J. (1918). Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics*, *3*(5), 422.
Naqvi, S., Bellott, D. W., Lin, K. S., & Page, D. C. (2018). Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. *Genome research*, gr-230433.
Park, C., Carrel, L., & Makova, K. D. (2010). Strong purifying selection at genes escaping X chromosome inactivation. *Molecular biology and evolution*, *27*(11), 2446-2450.
Peeters, S. B., Cotton, A. M., & Brown, C. J. (2014). Variable escape from X-chromosome inactivation: Identifying factors that tip the scales towards expression. *Bioessays*, *36*(8), 746-756.

Perrin, N. (2009). Sex reversal: a fountain of youth for sex chromosomes?. *Evolution: International Journal of Organic Evolution*, *63*(12), 3043-3049.
Pessia, E., Makino, T., Bailly-Bechet, M., McLysaght, A., & Marais, G. A. (2012). Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proceedings of the National Academy of Sciences*, 201116763.
Pinheiro, I., Dejager, L., & Libert, C. (2011). X-chromosome-located microRNAs in immunity: might they explain male/female differences? The X chromosome genomic context may affect X-located miRNAs and downstream signaling, thereby contributing to the enhanced immune response of females. *Bioessays*, *33*(11), 791-802.
Pinheiro, I., & Heard, E. (2017). X chromosome inactivation: new players in the initiation of gene silencing. *F1000Research*, *6*.
Richards, M. D., & Wong, A. (2015). Infantile nystagmus syndrome: clinical characteristics, current theories of pathogenesis, diagnosis, and management. *Canadian Journal of Ophthalmology/Journal Canadien d'Ophtalmologie*, *50*(6), 400-408
Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, *47*(1), 11-12.
Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, *43*(7), e47-e47.
Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.
Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., ... & Frankish, A. (2005). The DNA sequence of the human X chromosome. *Nature*, *434*(7031), 325.
Sato, Y., Shinka, T., Sakamoto, K., Ewis, A. A., & Nakahori, Y. (2010). The male-determining gene
Royo, H., Seitz, H., ElInati, E., Peters, A. H., Stadler, M. B., & Turner, J. M. (2015). Silencing of X-linked microRNAs by meiotic sex chromosome inactivation. *PLoS genetics*, *11*(10), e1005461.
SRY is a hybrid of DGCR8 and SOX3, and is regulated by the transcription factor CP2. *Molecular and cellular biochemistry*, *337*(1-2), 267-275.
Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., ... & Chinwalla, A. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*(6942), 825.
Slavney, A., Arbiza, L., Clark, A. G., & Keinan, A. (2015). Strong constraint on human genes escaping X-inactivation is modulated by their expression level and breadth in both sexes. *Molecular biology and evolution*, *33*(2), 384-393.
Song, R., Ro, S., Michaels, J. D., Park, C., McCarrey, J. R., & Yan, W. (2009). Many X-linked microRNAs escape meiotic sex chromosome inactivation. *Nature genetics*, *41*(4), 488.
Soh, Y. S., Alföldi, J., Pyntikova, T., Brown, L. G., Graves, T., Minx, P. J., ... & Rozen, S. (2014). Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell*, *159*(4), 800-813.
Tian, D., Sun, S., & Lee, J. T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell*, *143*(3), 390-403.
Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., ... & Cummings, B. B. (2017). Landscape of X chromosome inactivation across human tissues. *Nature*, *550*(7675), 244.
Turner, J. M. (2007). Meiotic sex chromosome inactivation. *Development*, *134*(10), 1823-1831.
Wang, X., Douglas, K. C., VandeBerg, J. L., Clark, A. G., & Samollow, P. B. (2014). Chromosome-wide profiling of X-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, Monodelphis domestica. *Genome research*, *24*(1), 70-83.
Warnefors, M., Mössinger, K., Halbert, J., Studer, T., VandeBerg, J. L., Lindgren, I., ... & Kaessmann, H. (2017). Sex-biased microRNA expression in mammals and birds reveals underlying regulatory mechanisms and a role in dosage compensation. *Genome research*.
Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873-881.

Yamauchi, Y., Riel, J. M., Ruthig, V. A., Ortega, E. A., Mitchell, M. J., & Ward, M. A. (2016). Two genes substitute for the mouse Y chromosome for spermatogenesis and reproduction. *Science*, *351*(6272), 514-516.

Yang, C., Chapman, A. G., Kelsey, A. D., Minks, J., Cotton, A. M., & Brown, C. J. (2011). X-chromosome inactivation: molecular mechanisms from the human perspective. *Human genetics*, *130*(2), 175-185.

Yang, F., Deng, X., Ma, W., Berletch, J. B., Rabaia, N., Wei, G., ... & Noble, W. S. (2015). The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome biology*, *16*(1), 52.

Zhou, Q., & Bachtrog, D. (2012). Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science*, *337*(6092), 341-345.

ACKNOWLEGEMENTS

I would like to thank immensely **Henrik Kaessmann** for the opportunity to work in his lab, for helping me direct my project towards a topic that fascinates both of us, and for the scientific guidance.

Thank you also to the other members of my thesis committee – **Gudrun Rappold, Eileen Furlong**, and **Christine Clayton** – for accepting being part of the committee and for the scientific discussion.

Also, I would like to thank particularly warmly **Margarida Cardoso-Moreira**, who has proven time and time again to be of immense support. Not only was your scientific guidance always incredibly welcomed, but your friendship and mentorship were treasured even more.

I would also like to thank greatly all the members of **the Kaessmann lab**, in Heidelberg as much as the ones who remained in Lausanne. **Maria Warnefors**, I appreciated remarkably our short work together. All of you, your inputs during lab meetings helped me progress more than you may realise.

Furthermore, I would like to thank the members of the **HBIGS** graduate school committee for offering me to join as a student and for the training in a variety of refreshing topics.

Finally, I would like to thank all the people who have contributed in making me feel at home in Heidelberg. I am especially grateful to have met **Evgeny Leushkin**, **Fanziska Gruhl**, **Mihai Petrovici** and **Cyril Mongis**, may our friendship last a lifetime! And I would like to thank very much **Cindy Dupuis Alexandre Dubuis** and **my family** for keeping me so close to their heart despite the distance.

SUPPLEMENTARY MATERIAL

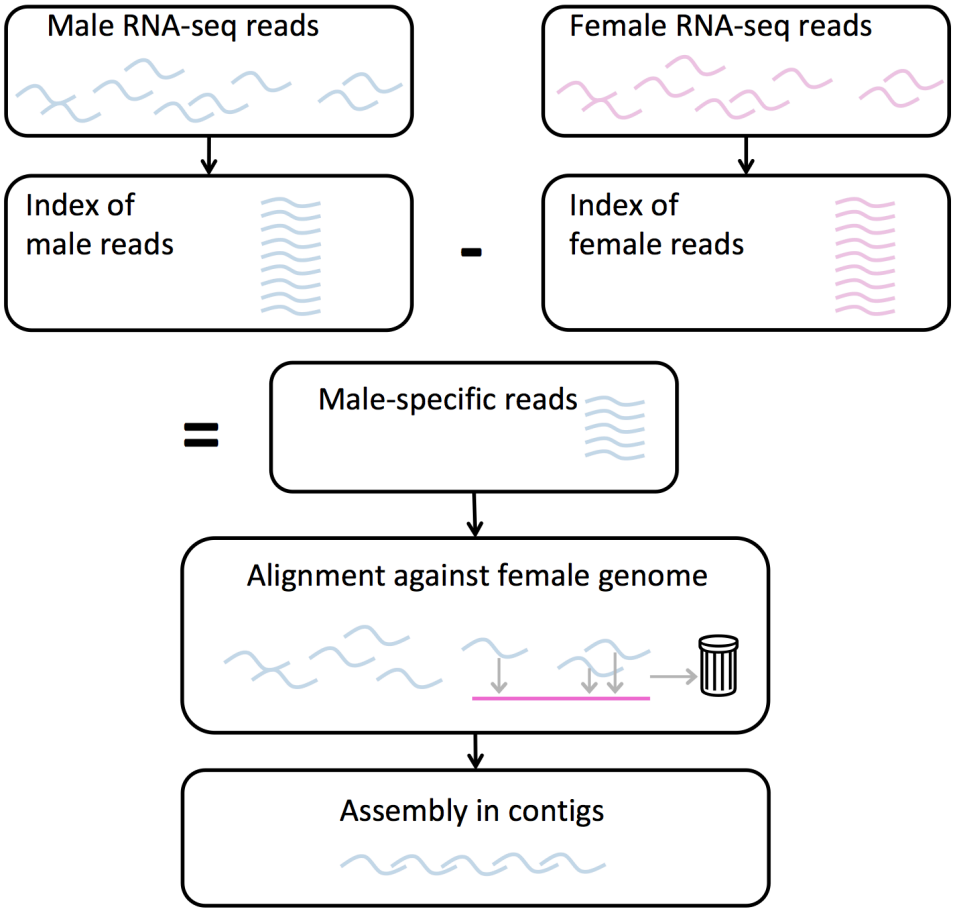

Figure S1: **Pipeline used for Y-chromosome transcriptome assembly and Y-linked miRNA detection**. The male and female indexes were created using my own Python script. The alignment against the female genome was done with TopHat version 2.1.1 [Kim et al. 2013]. The assembly into contigs was performed via Bridger-r2014-12-01 [Chang et al. 2015]. The pipeline used for Y-linked miRNA detection did not need contig assembly, so it stops before the last step.

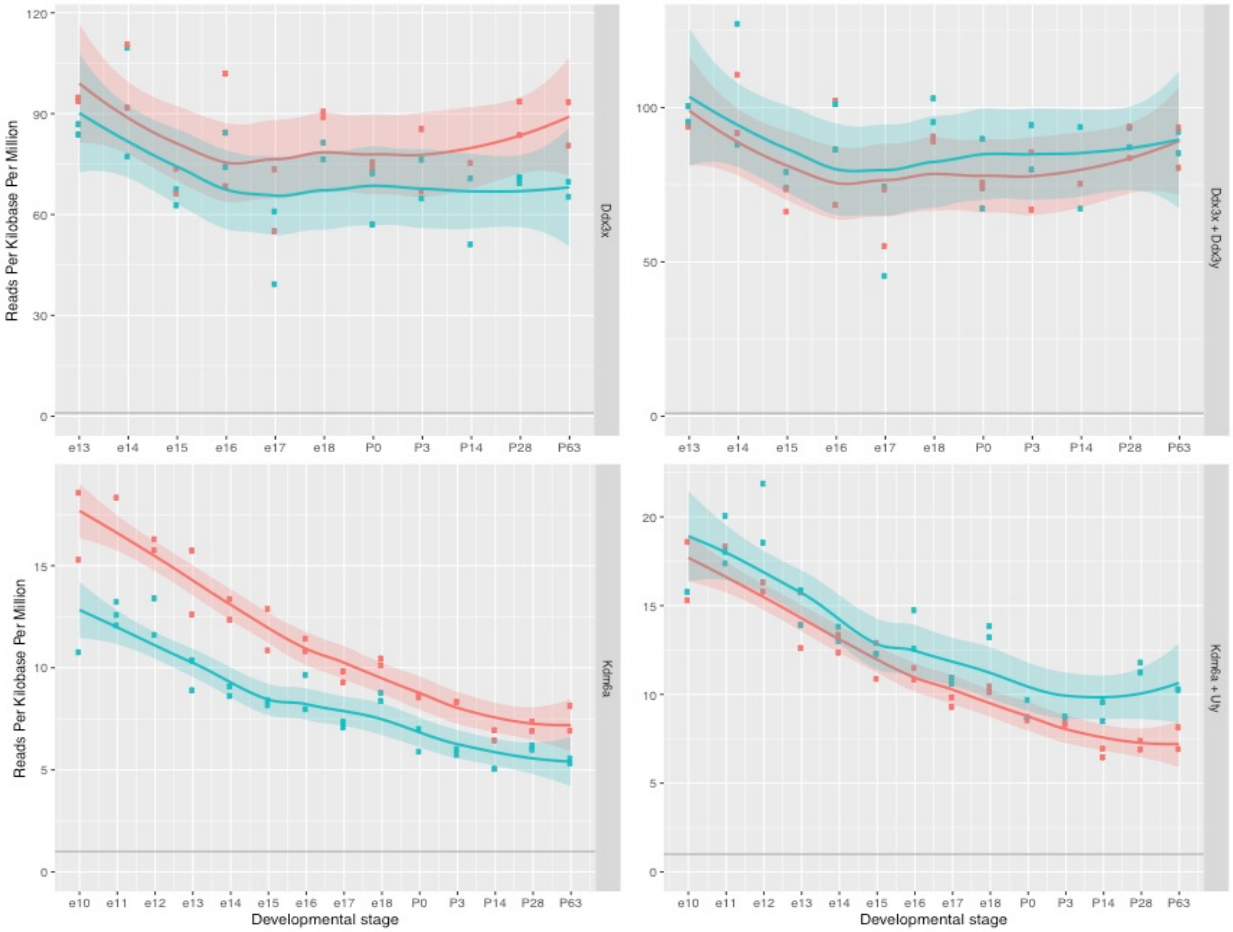

Figure S2: **Combined expression from both sex chromosomes for *Ddx3x/Ddx3y* in cerebellum (top), and *Kdm6a/Uty* in brain (bottom), in mouse**. Plots on the left show the expression of the X gametolog in males (blue) and females (red), while plots on the right show the expression of the X gametolog combined with the expression of its Y homolog. The x-axis shows stages from early (e13 for *Ddx3x/Ddx3y*, e10 for *Kdm6a/Uty*) to late (63 days post birth) development. The y-axis represent the expression level in RPKM.

If the function of X and Y gametologs is at least partially redundant, broadly expressed Y gametologs can effectively complement the under-expressed X gametolog in males and increase the correlation of expression levels between the sexes. In Figure S2, we observe that combination of expression from both X and Y gametologs in males slightly exceeds the expression of the X gametolog in females. This is could be the result of functional divergence between the gametologs, that could reduce the redundancy in function.