Quantitative analysis of Y-Chromosome gene expression across 36 human tissues

Alexander K. Godfrey\textsuperscript{1,2}, Sahin Naqvi\textsuperscript{1,2}, Lukáš Chmátal\textsuperscript{1}, Joel M. Chick\textsuperscript{3}, Richard N. Mitchell\textsuperscript{1}, Steven P. Gygi\textsuperscript{3}, Helen Skaletsky\textsuperscript{1,5}, David C. Page\textsuperscript{1,2,5}\textsuperscript{*}

\textsuperscript{1} Whitehead Institute, Cambridge, MA, USA
\textsuperscript{2} Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
\textsuperscript{3} Department of Cell Biology, Harvard Medical School, Boston, MA, USA
\textsuperscript{4} Department of Pathology, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA
\textsuperscript{5} Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, USA

*corresponding author:
Email: dcpage@wi.mit.edu

Running title:
Human Y-Chromosome gene expression in 36 tissues

Keywords:
Y Chromosome, sex chromosomes, sex differences, \textit{EIF1AY}, \textit{EIF1AX}
ABSTRACT

Little is known about how human Y-Chromosome gene expression directly contributes to differences between XX (female) and XY (male) individuals in non-reproductive tissues. Here, we analyzed quantitative profiles of Y-Chromosome gene expression across 36 human tissues from hundreds of individuals. Although it is often said that Y-Chromosome genes are lowly expressed outside the testis, we report many instances of elevated Y-Chromosome gene expression in a non-reproductive tissue. A notable example is EIF1AY, which encodes eukaryotic translation initiation factor 1A (eIF-1A), together with its X-linked homolog EIF1AX. Evolutionary loss of a Y-linked microRNA target site enabled upregulation of EIF1AY, but not EIF1AX, in the heart. Consequently, this essential translation initiation factor is nearly twice as abundant in male as in female heart tissue at the protein level. Divergence between the X and Y Chromosomes in regulatory sequence can therefore lead to tissue-specific, Y-Chromosome-driven sex biases in expression of critical, dosage-sensitive regulatory genes.
INTRODUCTION

A wide range of diseases, collectively affecting all organ systems, manifest differentially in human males and females (Wizemann and Pardue 2001). The molecular mechanisms responsible for these differences remain poorly characterized. It was once assumed that all such differences were the products of circulating hormones (e.g., androgens, estrogens), but they are increasingly speculated to stem in part from the direct effects of sex-chromosome genes expressed in tissues throughout the body (Arnold 2012). With regard to the sex chromosomes, most attention has been paid to the X Chromosome, particularly those X-Chromosome genes that are expressed more highly in XX (female) than in XY (male) individuals because they escape X-Chromosome inactivation in XX cells (Deng et al. 2014; Tukiainen et al. 2017). Researchers often cite the Y Chromosome’s paucity of genes and those genes’ presumed specialization for reproduction as reasons to look past the Y Chromosome, if it is considered at all. But recent studies indicate that the Y Chromosome retains conserved, dosage-sensitive regulatory genes expressed in tissues throughout the body (Bellott et al. 2014), which might underlie newly found associations between the Y Chromosome and disease (Tartaglia et al. 2012; Cannon-Albright et al. 2014; Eales et al. 2019).

To better understand how Y-Chromosome genes might contribute to differences between XX and XY individuals, we sought to obtain a quantitative understanding of Y-Chromosome gene expression across the human body. We excluded Y-Chromosome genes in the two pseudoautosomal regions, where the X and Y Chromosomes are identical in sequence, and instead focused on genes in the Y Chromosome’s male-specific region (MSY) (Skaletsky et al. 2003) (Fig. 1A; Supplemental Table S1). For our purposes, it was useful to distinguish two groups of MSY genes—those that have similar but non-identical homologs on the X
Chromosome and those that do not. MSY genes without X homologs are the products of transposition or retrotransposition events that brought copies of autosomal genes to the MSY at various points during mammalian evolution (Saxena et al. 1996; Lahn and Page 1999b; Skaletsky et al. 2003). Because these MSY genes have no counterparts on the X, they could confer differences to XX and XY individuals in any tissue where they are robustly expressed. A different set of considerations pertains to the MSY genes with X homologs, most of which are remnants of the ancestral pair of autosomes from which the mammalian sex chromosomes evolved, having survived millions of years of Y-Chromosome decay (Lahn and Page 1999a; Ross et al. 2005). Previous studies suggest that the X- and Y-linked members of these homologous X–Y gene pairs encode proteins that are at least partially equivalent in function (Table 1). Nevertheless, up- or downregulated expression of the MSY gene in a particular tissue might lead to a quantitative difference between XX and XY individuals in the expression level of the X–Y gene pair overall. Because ancestral MSY genes with X homologs encode highly dosage-sensitive regulators of transcription, translation, and protein stability (Bellott et al. 2014; Naqvi et al. 2018), even small sex biases in expression could have cascading effects on genes across the genome.

The current understanding of MSY gene expression is based on limited observations from humans and other mammals. Previous studies included only a few tissue types, while employing small sample sizes or suboptimal methodologies for analyzing MSY gene expression quantitatively. These studies established that some MSY genes show testis-specific expression while others are expressed in many tissues, but they could not detect more subtle quantitative differences in MSY gene expression between tissues (Lahn and Page 1997; Skaletsky et al. 2003; Bellott et al. 2014; Cortez et al. 2014). Other studies have found that MSY genes show lower
expression levels than their corresponding X-linked homologs (Xu et al. 2002, 2008b, 2008a; Johnston et al. 2008; Trabzuni et al. 2013; Cortez et al. 2014; Johansson et al. 2016). However, most such studies focused on single or subsets of MSY genes in individual tissues or on non-human mammals. This has made it difficult to discern a consistent quantitative picture of MSY gene expression and its bearing on the difference between human XX and XY tissues. These efforts have been further complicated by complexities of the MSY’s sequence. Homology with the X Chromosome and an abundance of complex segmental duplications pose various challenges for accurately measuring the expression of MSY genes at the transcript level. Even less is known about the expression of MSY genes at the protein level due in large part to the difficulty of obtaining reagents that can distinguish X- and Y-encoded amino-acid sequences. We therefore set out to conduct a systematic and quantitative survey of MSY gene expression across a diversity of human tissues.

RESULTS

Accurately estimating MSY gene expression levels

We obtained thousands of bulk-tissue RNA-sequencing (RNA-seq) samples released by The GTEx Consortium (GTEx Consortium 2017), spanning 36 adult human tissues and hundreds of post-mortem donors. To generate a quantitative view of MSY gene expression, we sought a method that could accurately estimate the expression levels of Y-Chromosome genes using short RNA-seq reads, overcoming challenges inherent in the MSY’s sequence. Some MSY genes show ~99% identity with their corresponding X-linked homologs in nucleotide sequence (Skaletsky et al. 2003). Other MSY genes have been amplified into multi-copy gene families, with genes in these families showing upwards of 99.9% nucleotide sequence identity. In an
RNA-seq experiment, many short reads from these genes will map to multiple genomic locations. These multi-mapping reads are routinely discarded in RNA-seq analyses to avoid the uncertainty of their origins, but excluding them can lead to underestimates of gene expression (Robert and Watson 2015). We suspected that expression of MSY genes had been disproportionately underestimated in the publicly available expression-level estimates released by The GTEx Consortium, for which multi-mapping reads were discarded. In these published estimates, a much smaller fraction of MSY genes appeared to be expressed (≥1 transcript per million (TPM)) than genes from other chromosomes (MSY: 38.8%; autosomes, Chr X: 78.2 – 98.6%) (Fig. 1B; Supplemental Table S2), in line with the MSY’s deficit of uniquely mappable sequence (Supplemental Fig. S1).

To obtain accurate expression-level estimates for all MSY genes, we re-estimated expression levels genome-wide from the GTEx raw data with kallisto (Bray et al. 2016), a program that jointly infers the most likely origins of uniquely and multi-mapping reads under a statistical model. In contrast to a procedure that discards multi-mapping reads, kallisto enabled us to accurately estimate the expression levels of MSY genes in simulated RNA-seq datasets (±7.3% for the average MSY gene, when simulated at 5 TPM) (Methods), including the relative expression of Y- and X-linked homologs and the total expression of genes in multi-copy families (Supplemental Fig. S2). Kallisto’s accuracy in these tests implies that, for high levels of sequence identity (~99%), enough uniquely mapping reads are present in GTEx RNA-seq libraries to inform the correct assignment of multi-mapping reads. We then applied kallisto to the raw RNA-seq data and found that 80% of MSY genes are expressed in at least one tissue, a number more typical of other chromosomes (Fig. 1B). In some cases, our re-estimates identified expression levels more than two orders of magnitude higher than previously reported (e.g., the
HSFY gene family in testis, 32.4 TPM vs. <0.1 TPM, Fig. 1C; Supplemental Table S2). These differences were most pronounced for the MSY’s multi-copy gene families. By contrast, ancestral single-copy MSY genes produced few if any multi-mapping reads; their expression levels were therefore not systematically underestimated (Supplemental Fig. S1, S2). Nevertheless, of the approaches tested, we found kallisto to yield the most accurate estimates overall (Supplemental Fig. S2).

After performing a series of quality control steps, including outlier-sample detection and expression-level adjustment for three indicators of sample quality (Methods), we retained 6,358 RNA-seq samples spanning 36 adult tissues, collected from 337 XY donors and 178 XX donors, for our primary analysis. Overall, we detected expression of 24 of the 26 MSY genes and gene families in at least one tissue (Fig. 1D; Supplemental Fig. S3; Supplemental Table S3).

Most MSY genes without X homologs show testis-specific expression

MSY genes that lack X homologs belong to five multi-copy gene families (BPY2, CDY, DAZ, PRY, XKRY (Supplemental Table S1)). We first asked if any of these gene families is robustly expressed in a non-reproductive tissue—i.e., in a tissue found in both XX and XY donors. We identified one such instance. Genes of the DAZ (Deleted in Azoospermia) gene family, which are generally viewed as testis-specific genes involved in spermatogenesis (Vogt et al. 2008), were expressed in testis samples but also showed robust (and even 2.5-fold higher) expression in the stomach (Fig. 1D; Supplemental Fig. S4A), replicating a similar observation from a recent, smaller study (Gremel et al. 2015). (By contrast, the DAZ family’s autosomal homolog and progenitor (Saxena et al. 1996), DAZL, was not expressed in stomach samples from XY or XX donors (Supplemental Fig. S4B)). The DAZ genes’ expression in the stomach proved to be the
exception among MSY genes without X-linked homologs. One of the four remaining gene families (XKRY) was not robustly expressed in any tissue, while the others (BPY2, CDY, PRY) showed exquisitely testis-specific expression (Fig. 1D; Supplemental Fig. S3). We conclude that, overall, MSY genes without X homologs are unlikely to contribute substantially to differences between XX and XY individuals outside of the reproductive system.

Quantitative differences between X- and Y-homolog expression in XY individuals

Next, we considered the expression of MSY genes with X homologs, focusing on those X–Y gene pairs where the MSY gene is expressed predominantly in non-reproductive tissues. Because these MSY genes were typically expressed in the same tissues as their corresponding X homologs (Supplemental Fig. S5, S6; Supplemental Tables S4, S5), we specifically sought to characterize the quantitative differences in X- and Y-homolog expression.

We first asked if the MSY genes are expressed at higher or lower levels than their X-linked homologs in tissues of XY individuals. We estimated the Y-homolog-to-X-homolog expression ratio (Y/X expression ratio) in each XY tissue sample and aggregated these into tissue-level estimates (Fig. 2A, B; Supplemental Table S6). We observed differences among the X–Y pairs in their average Y/X expression ratios. Two MSY genes (TMSB4Y, TBL1Y) showed substantially lower expression than their corresponding X-linked homologs in all tissues (Fig. 2B). However, for the remaining X–Y pairs, the expression levels of the Y- and X-linked homologs were more similar. Some MSY genes (e.g., DDX3Y, USP9Y, and RPS4Y1) were typically expressed at 30–50% of the level of their X homolog, while others were often expressed at equal (e.g., KDM5D, EIF1AY) or higher (e.g., TXLNGY, NLGN4Y) levels. We
replicated these Y/X-expression-ratio estimates using independently generated RNA-seq data spanning a subset of the GTEx tissues (Supplemental Fig. S7).

Although some X–Y gene pairs had higher or lower Y/X expression ratios than others (Friedman test, $P = 1 \times 10^{-28}$), no one tissue had significantly higher or lower Y/X expression ratios overall (Friedman test, $P = 0.42$; Fig. 2C; Supplemental Fig. S8). This implies that the expression of individual, widely expressed MSY genes largely reflects gene-specific regulation rather than an MSY-wide specialization for a biological process like reproduction. Indeed, despite the absence of substantial differences between tissues, when the tissues are ranked, testis was the tissue where Y/X expression ratios are lowest on average (Fig. 2C; Supplemental Fig. S8).

For each X–Y gene pair, we next sought to determine if the expression of the X and Y homologs continues to be regulated by the same upstream factors. If so, variation in the activity of these factors from one sample to the next should yield correlated X- and Y-homolog expression. Indeed, we found that the X and Y homologs of most X–Y gene pairs showed highly correlated expression in many tissues (Fig. 2H; Supplemental Fig. S9; Supplemental Table S7). For example, the Y-linked ribosomal protein gene $RPS4Y1$ and its X-linked homolog $RPS4X$ showed tightly correlated expression in most tissues across the body (Fig. 2D, H; Supplemental Fig. S9). $RPS4Y1$’s expression levels also correlated tightly with those of ribosomal protein genes on other chromosomes, such as $RPS8$ on Chromosome 1 (Fig. 2E), but not with those of Y-linked transcription factor $ZFY$ (Fig. 2F), whose expression levels, instead, correlated with those of its X homolog $ZFX$ (Fig. 2G)). This suggests that $RPS4Y1$’s expression levels are determined in accordance with molecular function rather than chromosomal location. MSY genes that were typically expressed at only 30–50% of the levels of their X homologs (e.g.,
\textit{RPS4Y1, DDX3Y, ZFY} still showed tightly correlated expression with their X homologs in many tissues (Fig. 2B, H; Supplemental Fig. S9; Supplemental Table S7). This highly correlated expression is not an artifact of read mis-mapping between the X and Y Chromosomes, as few reads mapped to both X and Y homologs of widely expressed X–Y gene pairs, and we could independently estimate their expression levels in simulated RNA-seq datasets (Supplemental Fig. S10). Thus, even though these Y homologs show diminished expression, the ancestral regulatory elements governing their expression likely remain intact and under considerable evolutionary constraint, despite millions of years of Y-Chromosome decay in the absence of regular recombination with the X Chromosome.

\textbf{Evolutionary loss of a microRNA target site promoted elevated \textit{EIF1AY} expression in the heart}

We also found evidence of tissue-specific divergence in the regulation of X- and Y-homolog expression. Individual X–Y pairs showed Y/X expression ratios in some tissues that differed substantially from their ratios in other tissues (e.g., \textit{USP9Y}/\textit{USP9X} = 1.1 in the pituitary, compared to 0.2 – 0.6 in most tissues) (Fig. 2B), leading us to hypothesize that one member of the X–Y gene pair, but not the other, might be up- or downregulated. To explore this possibility, for each X and Y homolog separately, we identified tissues where its expression level is 30\% higher or lower than its expression level in most other tissues (Methods). All widely expressed MSY genes showed significantly higher or lower expression in at least one tissue (Fig. 3; Supplemental Table S8). We observed increased expression in a variety of tissues, including endocrine glands (e.g., pituitary, adrenal, pancreas), striated muscle (heart and skeletal), spleen, and skin.
A prominent example of elevated expression of an MSY gene, without a corresponding increase in the expression of its X-linked homolog, is that of *EIF1AY*. *EIF1AY* encodes the Y isoform of eukaryotic translation initiation factor 1A (eIF-1A). eIF-1A is one of 27 primary factors used to initiate protein synthesis in all eukaryotic lineages (Hinnebusch 2014), and the only such factor encoded on both X and Y Chromosomes in primates (Bellott et al. 2014). The X and Y isoforms of eIF-1A—encoded by *EIF1AX* and *EIF1AY*, respectively—are likely to be functionally equivalent: they differ by only a single amino acid, a conservative leucine-to-methionine substitution at a position outside of eIF-1A’s key functional domains, at which both leucine and methionine are observed in various vertebrate species (Supplemental Fig. S11).

Although *EIF1AY* and its X-linked homolog *EIF1AX* are expressed at similar levels in most tissues, we found elevated expression of *EIF1AY* in the heart, skeletal muscle, spleen, and pituitary, causing *EIF1AY* expression levels to be as much as 5.8-fold higher than those of *EIF1AX* (Fig. 4A). We replicated this tissue-specific pattern of higher *EIF1AY* expression in human RNA-seq data from an independently generated dataset (Supplemental Fig. S12).

We searched for factors that might explain *EIF1AY*’s elevated expression relative to *EIF1AX* in these tissues. Motivated by our previous studies (Naqvi et al. 2018), we wondered if these two genes might be differentially regulated by microRNAs (miRNAs), small regulatory RNAs that act as sequence-specific repressors of gene expression (Bartel 2018). A miRNA might specifically target *EIF1AX*, limiting its expression level in these tissues. When we searched the 3′-untranslated region (3′-UTR) of *EIF1AX* (Methods), the miRNA target site with the highest predicted efficacy was a match to miR-1 (Fig. 4B; Supplemental Table S9), a miRNA expressed abundantly and specifically in heart and skeletal muscle (Lim et al. 2005; Ludwig et al. 2016) (Fig. 4C). At the homologous position in the 3′-UTR of *EIF1AY*, however, this miR-1 target site
is disrupted by two nucleotide substitutions at positions critical for effective miRNA-mediated repression (Fig. 4B).

Two observations indicate that disruption of the miR-1 site in EIF1AY contributed to EIF1AY’s higher expression in the heart and skeletal muscle. First, we found that the 3′-UTR of EIF1AX, but not of EIF1AY, mediated approximately twofold repression of the reporter upon miR-1 transfection, but not upon transfection with another miRNA (Fig. 4D). miR-1’s repression of the EIF1AX-reporter construct required the target site to be intact, and repairing the two target-site substitutions within the EIF1AY-reporter construct was sufficient to confer miR-1–mediated repression. Second, the status of the miR-1 site predicts the expression pattern of EIF1AX and EIF1AY orthologs across species (Supplemental Table S10). In other primates, which both retain an intact EIF1AY gene and possess the disrupted miR-1 site, EIF1AY showed approximately twofold higher expression than EIF1AX specifically in heart and skeletal muscle (Fig. 4E, F). This expression pattern was likely acquired by primate EIF1AY, as EIF1AX orthologs in and outside of mammals do not show elevated heart expression (Supplemental Fig. S13). Together, these observations suggest that two nucleotide substitutions within an EIF1AY regulatory element contributed to tissue-specific upregulation of EIF1AY.

Male-biased expression of X–Y gene pairs at the transcriptional level

We next asked if the divergent expression we observed within XY individuals leads to differences in expression between XX and XY individuals. We found that the X-linked members of the eight most widely expressed X–Y gene pairs typically showed XX-biased expression—i.e., higher expression in tissue samples from XX individuals than in the same tissue type from XY individuals (Fig. 5). This XX-biased expression is expected because the X homologs of
widely expressed X–Y gene pairs are not subject to X-Chromosome inactivation in XX cells and thus are expressed biallelically (Carrel and Willard 2005; Tukiainen et al. 2017). In all cases, the magnitude of XX bias was 2.0-fold and typically less than 1.5-fold (Fig. 5). This is consistent with past observations that the X-linked allele on the otherwise inactivated X Chromosome shows lower expression than the X-linked allele on the fully active X (Cotton et al. 2013; Berletch et al. 2015; Tukiainen et al. 2017). Next, for each X–Y gene pair, we compared the summed expression level of the X and Y homologs in XY samples to the expression level of the X homolog in XX samples. When accounting for Y-homolog expression, the X–Y gene pairs typically showed slightly XY-biased expression, with differences in expression less than 2.0-fold. However, in tissues where the X and Y homologs of a given pair showed uncorrelated expression (Supplemental Fig. S14), and in tissues with elevated Y-homolog expression, the XY-biased expression was more prominent. For example, KDM5D showed elevated expression in the adrenal gland (Fig. 3), leading to 2.1-fold XY-biased expression of KDM5C/D (Fig. 5; Supplemental Table S11). In the pituitary gland, elevated expression of USP9Y, together with depleted expression of USP9X, yields 2.0-fold XY-biased expression of USP9X/Y (Fig. 3; Fig. 5). Upregulated EIF1AY expression in the heart leads to 5.2-fold higher expression of EIF1AX/Y in XY heart (left ventricle) tissue. Thus, at the transcriptional level, the Y-linked members of human X–Y gene pairs typically show higher expression in XY cells than the second copy of their X-linked homologs in XX cells, causing the X–Y gene pairs to show at least subtly, and sometimes substantially, male-biased expression.
Male-biased expression of EIF1A in the heart at the protein level

We sought to assess whether the male-biased expression of X–Y gene pairs at the transcript level further manifests as male-biased expression at the protein level. We generated proteome-wide measurements of protein abundance in 21 XY and 12 XX heart (left ventricle) tissue samples by multiplexed, tandem mass tag (TMT)-based mass spectrometry (Supplemental Fig. S15; Methods). These samples, which we obtained from the GTEx tissue biobank, were selected through rigorous histological review to ensure that XX and XY samples showed minimal pathology and similar cell-type composition (Methods). At a 0.22% false-discovery rate (FDR), we detected peptides that specifically match the X or Y protein isoform encoded by widely expressed X–Y gene pairs (RPS4X, RPS4Y1, EIF1AX, EIF1AY, DDX3X, DDX3Y, USP9X; Fig. 6A, B; Supplemental Fig. S16; Supplemental Table S12). Each of these proteins (except RPS4Y1) was supported by multiple, independent observations of isoform-specific peptides. Moreover, Y-specific peptides from all Y isoforms showed only background levels of signal in XX samples (Fig. 6A). Together, these observations provide strong evidence that these seven proteins are present in heart tissue. The absence of peptides from the remaining 11 proteins was consistent with their lower expression levels at the transcript level and the overall rate at which we recovered peptides from expressed genes across the genome (7/18 X–Y pair genes vs. 4,788/11,936 expressed genes: one-tailed Fisher’s exact test, \( P \approx 1.0 \); Supplemental Fig. S17).

Thus, whether these 11 remaining proteins are present in human heart tissue remains an open question.

For the three X–Y gene pairs from which both X- and Y-specific peptides were detected (DDX3X/Y, EIF1AX/Y, RPS4X/Y1), we asked if their expression is sex biased at the protein level. For each X–Y pair, we first used signal from X-specific peptides to estimate the sex bias of the X
isoform; we next used signal from peptides that match both X and Y isoforms (X–Y-shared peptides) to estimate the sex bias of the X–Y pair overall, accounting for the contribution of the Y isoform (Supplemental Fig. S15; Supplemental Table S13). These two expression ratios then allowed us to infer the relative abundances of X and Y isoforms within XY tissue. This approach contrasts with the common practice in mass-spectrometric analysis of assigning non-unique peptides to the apparently most abundant protein (e.g., (Cox and Mann 2008)), which would conflate the expression of X and Y isoforms in these samples.

We found that the X isoforms of all three X–Y pairs showed XX-biased protein abundance ($P < 5 \times 10^{-3}$, by permutation (Methods)), consistent with their escape from X-Chromosome inactivation (Fig. 6C). (In contrast, proteins encoded by X-Chromosome genes that are subject to X-Chromosome inactivation showed no or only modest sex biases in protein abundance (Supplemental Fig. S18; Supplemental Table S14).) For RPS4X/Y and DDX3X/Y, the combined expression levels of X and Y isoforms in XY tissues were slightly below the levels of the X isoforms in XX tissues on average (RPS4X/Y: mean XY/XX ratio = 0.91, $P = 0.05$ by permutation; DDX3X/Y: mean XY/XX ratio = 0.90, $P = 0.03$ by permutation) (Fig. 6C), albeit at only nominally statistically significant levels, suggesting RPS4Y1 and DDX3Y mostly if not entirely compensate for the XX-biased expression of RPS4X and DDX3X. The combined expression of EIF1AX and EIF1AY, however, showed a 1.7-fold XY bias ($P < 10^{-6}$, permutation), indicating that EIF1AY overcompensates for the XX-biased expression of EIF1AX. These estimates further imply that EIF1AY protein is 2.1-fold more abundant than EIF1AX in XY heart tissue. Using an eIF-1A antibody that recognizes both EIF1AX and EIF1AY (Supplemental Fig. S19), we corroborated EIF1AX/Y’s (i.e., eIF-1A’s) XY-biased expression in these same heart tissue samples by Western blot (Fig. 6E, F). Although $EIF1AY$...
transcripts were 5.8-fold more abundant than *EIF1AX* transcripts in heart (left ventricle) tissue, EIF1AY protein was only 2.1-fold more abundant than EIF1AX (Fig. 6D). Nevertheless, *EIF1AY*’s upregulated expression in the heart—a result of its non-coding divergence from *EIF1AX*—is sufficient to lead to a male-biased abundance of this essential translation initiation factor.

**DISCUSSION**

How do human Y-Chromosome genes contribute to differences between XX and XY individuals beyond the reproductive system? It has been tempting to speculate that MSY genes encode proteins with “male-specific” effects (Arnold 2012), as the result of protein-coding sequence divergence between MSY genes and their corresponding X homologs. Such instances might yet be uncovered. However, given past evidence attesting to the functional interchangeability of X and Y protein isoforms (Table 1) and our observations of divergent X–Y expression herein, we propose that divergence of MSY genes from their X homologs in regulatory (i.e., non-coding) sequence is an important means by which the Y Chromosome could directly give rise to differences between XX and XY individuals. Because the X–Y gene pairs encode regulators of transcription, translation, and protein stability that are highly dosage sensitive (Bellott et al. 2014; Naqvi et al. 2018), small differences in their expression levels could contribute significantly to the widespread sex differences in gene expression observed across tissues (Naqvi et al. 2019) and ultimately to phenotypic differences between the sexes.

This focus on regulatory-sequence divergence, rather than protein-coding divergence, accords with prevailing views from complex trait genetics and evolutionary developmental biology. In these contexts, phenotypic variation within and across species is thought to flow in
large part from non-coding substitutions that alter the expression of pleiotropic regulatory genes (Albert and Kruglyak 2015; Carroll 2008), genes very much like those encoded by ancestral X–Y pairs. In a similar manner, quantitative differences between males and females in disease susceptibility or morphometric traits might reflect regulatory-sequence divergence between the X and Y Chromosomes that yields sex-biased expression of the X–Y gene pairs. It is likely that many types of regulatory factors beyond microRNAs are involved in establishing these expression patterns. Factors other than miR-1, such as a heart-specific transcription factor, might additionally contribute to $\textit{EIF1AY}$’s ~5-fold higher expression than $\textit{EIF1AX}$ in the heart, as miRNAs typically repress their targets by less than twofold (Baek et al. 2008) (Fig. 4D).

One speculation is that the XY-biased expression of eIF-1A contributes to sex differences in diseases of the heart, many of which manifest with greater incidence or severity in one sex (Regitz-Zagrosek et al. 2010). As a core component of the 43S preinitiation complex in eukaryotes (Hinnebusch 2014), eIF-1A impacts the translation of many if not all mRNA transcripts in the cell (Sehrawat et al. 2018). Changes in translational regulation are a prominent molecular feature of human heart tissue from individuals with dilated cardiomyopathy (van Heesch et al. 2019), a disease with a 1.5-fold higher incidence in males than in females (Towbin et al. 2006). Although it is currently unknown whether elevated levels of eIF-1A are beneficial, harmful, or neutral in consequence, $\textit{EIF1AY}$’s expression pattern and those of other MSY genes provide new motivation to examine the Y Chromosome’s contribution to various quantitative traits.

Beyond these cases of divergent X- and Y-homolog regulation, our observations accord with the view that MSY genes encode proteins that function similarly to their X-encoded homologs, and that these shared functions are dosage sensitive across a multitude of tissues. The
tightly correlated expression of X and Y homologs we observe is typical of genes whose proteins must together be synthesized in precise quantities (Taggart and Li 2018). It is unlikely that the regulatory elements that enable the MSY genes to be expressed in this manner would survive by chance, after tens of millions of years of Y-Chromosome decay.

We have provided direct evidence that the proteins encoded by MSY genes are present in human heart tissue. Our detection of DDX3Y protein in the heart conflicts with earlier claims that DDX3Y is widely transcribed but only translated in the testis (Ditton et al. 2004). Using a DDX3Y-specific antibody, Ditton et al. detected DDX3Y protein in testis but not in brain or kidney. In our analysis, we find that DDX3Y shows lower transcript abundance in brain and kidney than in most other tissues (Supplemental Fig. S3; Supplemental Table S3), suggesting DDX3Y protein might have been present only at low levels (Gueler et al. 2012). DDX3Y transcripts might also be translated inefficiently (see below). Beyond our study, DDX3Y has been detected by Western blot in a neuronal cell line (Vakilian et al. 2015), and DDX3Y was identified as an essential gene in a leukemia cell line through a genome-wide, unbiased screen (Wang et al. 2015). Although the only known phenotype for individuals with DDX3Y deletions is spermatogenic failure (Vogt et al. 2008), milder non-reproductive phenotypes have not been excluded. Recognizing that DDX3Y protein is present in non-reproductive tissues has important implications for studies of DDX3X—an intellectual disability gene (Snijders Blok et al. 2015) and therapeutic target (Bol et al. 2015; Valiente-Echeverría et al. 2015)—which have typically disregarded the impact of its Y homolog.

We found that the protein expression levels of DDX3Y, EIF1AY, and RPS4Y1 were 1.8–4.7-fold lower than their transcript expression levels, when measured against their X homologs. It is possible that these transcripts are translated less efficiently, or that their proteins are less
stable, than those encoded by their X homologs. If true for other MSY genes, this could explain
why X–Y gene pairs often show slightly male-biased expression at the transcript level (Fig. 5):
over-expression of the Y homolog at the transcript level might be needed to achieve the requisite
level of protein abundance. However, we caution against extrapolating these results to other
MSY genes and other tissues until many more protein-level measurements are made.

Ultimately, our analyses establish that mass spectrometry can be used, in an unbiased
manner, to detect the expression of MSY proteins in a non-reproductive tissue and quantify the
levels of X–Y pair proteins across individuals, even when the X and Y isoforms differ by only a
single amino acid. This will remain a challenge for the Y (and X) protein isoforms expressed at
lower levels (Meyfour et al. 2017). However, as is the case with analyses of the Y Chromosome
in DNA and RNA sequence, a distinct picture of the Y Chromosome emerges with appropriate
analytical approaches. Deploying methods that can resolve subtle differences between genes as
standard practice, whether for RNA-seq (Bray et al. 2016; Li and Dewey 2011; Patro et al. 2017)
or mass spectrometry (Malioutov et al. 2018), promises a more complete understanding not only
of sex-chromosome genes, but also of all groups of homologous genes genome-wide.

Going forward, we anticipate that additional examples of upregulated MSY gene
expression will be revealed through expression profiling in other contexts. Particularly promising
will be the application of single-cell approaches to observe MSY gene expression in rare cell
types, whose contributions to the bulk-tissue estimates here are diluted. Indeed, a recent study
found elevated expression of TBL1Y—a gene we found to show lower expression than its X
homolog in all instances—in cells of the inner ear, with implications for syndromic hearing loss
(Di Stazio et al. 2018). Given the differences in expression between MSY genes and their X
homologs, it will be especially important to characterize how increases or decreases in the
expression of proteins encoded in X–Y pairs leads to changes across the genome, in specific cell
types, tissues, and developmental stages.

METHODS

Code used in analysis

Unless stated otherwise, all analyses were conducted in Python (v3.6.9), drawing upon software
packages numpy (v1.17.2) (Oliphant 2007; van der Walt et al. 2011), scipy (v1.3.1) (Virtanen et
al. 2020), pandas (v0.25.1) (McKinney 2010), scikit-learn (v0.21.3) (Pedregosa et al. 2011),
statsmodels (v0.10.1) (Seabold and Perktold 2010, matplotlib (v3.1.1) (Hunter 2007), and
seaborn (v0.9.0) (seaborn.pydata.org). Code and Jupyter notebooks (jupyter.org) for recreating
these analyses are available on GitHub (https://github.com/akg8/MSY-expression) and as
Supplemental Code.

Abbreviated tissue names

Tissues with long names are abbreviated in figures as follows: Adipose – Subcutaneous (Subc),
Visceral (Visc); Artery – Aorta (Aort), Coronary (Coro), Tibial (Tib); Brain – Amygdala
(Amyg), Cerebellum (Cblm), Cortex (Cort), Hippocampus (Hipp), Hypothalamus (Hypo),
Striatum (Stri), Substantia nigra (Subn); Colon – Sigmoid (Sigm), Transverse (Trns); Esophagus
– Mucosa (Muco), Muscularis (Musc); Heart – Atrial Appendage (AtrA), Left Ventricle (LVen);
Skeletal Muscle (Sk Muscle); Small Intestine (Sm Intestine).
Human transcriptome annotation and MSY genes

All human analyses use transcript/gene models defined in a custom subset of the comprehensive GENCODE version 24 transcript annotation, based on our annotation of the male-specific region of the human Y Chromosome (Skaletsky et al. 2003) (Supplemental Table S1). See Supplemental Methods for further details.

Comparison of RNA-seq analysis methods

Simulated RNA-seq libraries were generated using RSEM (v1.2.22) (Li and Dewey 2011), using a GTEx testis sample as a template. The expression levels of MSY genes and their X-linked homologs were set to predetermined levels in each simulation. Three methods were then used to estimate the expression levels of Y-Chromosome genes and their X-linked homologs: (1) reads were aligned to the genome using TopHat2 (Kim et al. 2013), and the number of uniquely mapping reads overlapping each gene was counted with featureCounts (Liao et al. 2014) (this “unique reads” approach is based on The GTEx Consortium’s procedure); (2) reads were aligned with TopHat2 and expression levels were estimated with Cufflinks (Trapnell et al. 2010) in “multi-read-correct” mode; (3) reads were input to kallisto (Bray et al. 2016), which estimated expression levels using the transcriptome annotation. See Supplemental Methods for details.

Estimating transcript expression levels from GTEx RNA-seq samples

GTEx (v7) raw data were obtained from dbGaP (dbGaP accession: phs000424.v7.p2). Transcript expression levels were then estimated in TPM units using kallisto with sequence-bias correction (--bias); transcript expression levels were summed to obtain gene expression levels. The expression levels of genes in multi-copy gene families (Supplemental Table S1) were summed to
obtain family-level estimates, which were used in place of estimates at the gene level. Within each tissue, samples that appeared to be outliers based on their genome-wide expression profile were identified and removed (Supplemental Methods). Samples from male and female donors were verified to have likely XY and XX sex-chromosome constitutions through the expression of MSY genes and \textit{XIST}. The final set of samples used for analysis is given Supplemental File S1. Samples from some tissue subsites defined by The GTEx Consortium (e.g., Brain – Cerebellum and Brain – Cerebellar Hemisphere) could not be easily distinguished by hierarchical clustering. In these cases, we merged the tissue labels, treating them as single tissue types (Supplemental Methods).

To reduce technical variation in expression levels and increase tissue-to-tissue comparability, linear regression was used to adjust expression levels for the effects of ischemic time, RNA integrity number (RIN), and the sample intronic read mapping rate (see Supplemental Methods). These adjusted expression levels were used in all analyses, except when comparing our estimated expression levels from kallisto to those released by The GTEx Consortium in Figure 1B–C and Supplemental Fig. S1.

We estimated a gene’s expression level in a tissue as its median expression level among samples from that tissue unless otherwise noted. To obtain the clustering of genes and tissues shown in Fig. 1C, the estimated expression levels of MSY genes were subject to hierarchical clustering by average linkage using correlation distances (scipy.cluster.hierarchy.linkage, with method=“average”, metric=“correlation”).
Comparison with The GTEx Consortium’s analysis based on uniquely mapped reads

The GTEx Consortium’s gene expression level estimates (v7) were downloaded from the GTEx Portal (gtexportal.org: GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct). Genes in GENCODE version 19 were matched to genes in our version-24-based annotation by Ensembl gene ID. The fraction of uniquely mapping reads per gene was estimated by aligning all possible 76-nucleotide reads from its longest transcript isoform to the transcriptome exhaustively (see Supplemental Methods).

Expression-level normalization across samples and tissues

For analyses where the expression level of a gene was compared across samples, we applied a modified version of the between-sample, size-factor normalization used in DESeq (Anders and Huber 2010). For a set of \( n \) samples, the normalization factor, \( s_i \), for sample \( i \) was calculated as

\[
s_i = \text{median}_{g \in G_C} \left( \frac{y_{gi}}{\prod_{j=1}^{n} y_{gj}} \right)^{1/n}
\]

where \( y_{gi} \) is the expression level (in TPM units) of gene \( g \) in sample \( i \) and \( G_C \) is a set of control genes. Rather than using all genes in the genome, we base our normalization factor on a set of 50 control genes that are expressed like “housekeeping” genes. These control genes were identified as the 50 genes, among all genes with mean expression levels between 10 and 100 TPM, with the most conserved expression-level ranks (i.e., whose expression-level ranks showed the lowest coefficient of variation across the samples). This approach helps to ensure that the genes driving the normalization have known properties even when comparing samples from two or more tissues in which the expression levels of many genes would be expected to differ.
Y/X expression ratios

For a given X–Y gene pair and tissue, we estimated the Y/X expression ratio in each sample as

\[
\frac{\text{(Y-homolog TPM + 0.5)}}{\text{(X-homolog TPM + 0.5)}},
\]

excluding samples where both genes were expressed below 1 TPM; the median sample-level Y/X ratio was then used as the tissue-level estimate. A tissue-level Y/X ratio was not reported where both genes were expressed below 1 TPM. For a given X–Y pair and tissue, the difference in the X and Y homolog’s expression levels was assessed with a two-sided Wilcoxon signed-rank test (Python function: scipy.stats.wilcoxon). After obtaining \(p\)-values for all X–Y pairs in all tissues, these \(p\)-values were adjusted for multiple hypotheses using the Benjamini-Hochberg (Python function: statsmodels.stats.multitest.multipletests, method = ‘fdr_bh’). To test for differences between Y/X expression ratios among X–Y pairs and among tissues, the Friedman test was applied (Python function: scipy.stats.friedmanchisquare), using the Y/X expression ratios from the 28 tissues where a ratio was estimated for the 10 most widely expressed X–Y pairs (listed in Supplemental Fig. S8).

Replication of gene expression patterns

To assess expression patterns in an independent dataset, we analyzed raw RNA-seq data from the Human Protein Atlas (HPA) Project (Uhlén et al. 2015) with kallisto using sequence-bias correction. For replication of Y/X expression ratios in Supplemental Fig. S7, we used the HPA tissues matching a GTEx tissue where at least four HPA samples from male donors were present (colon, prostate, testis). For more detailed replication of \(EIF1AY\)’s expression pattern (Supplemental Fig. S12), we used samples from all HPA tissues matching a GTEx tissue. When an HPA tissue potentially matched multiple GTEx tissues (e.g., Colon – Transverse, Colon –
Sigmoid), the best matching tissue was selected by calculating correlation coefficients between samples from the two datasets using genome-wide gene expression levels.

**Correlated expression of X and Y homologs**

Analyses of pairwise gene co-expression were performed in each tissue with at least 30 samples from male donors. Each tissue was analyzed separately, considering only those genes with expression levels \( \geq 5 \) TPM. The expression levels from each sample were first normalized by the housekeeping method described above and transformed to \( \log_2(\text{TPM} + 0.5) \) units. To control for unmodeled technical factors (e.g., batch effects) that might lead to spuriously correlated expression between the X and Y homologs of X–Y pairs, the principal components (PCs) of the \( N \) genes \( \times M \) samples matrix were calculated (Python function: sklearn.decomposition.PCA): after mean-centering the expression levels of each gene, each sample’s loading on the top principal component were extracted. For each gene, variation in expression associated with this principal component was removed by linear regression. The degree of co-expression between gene \( i \) and gene \( j \) was measured with Spearman’s correlation coefficient, \( \rho_{ij} \), of their PC-adjusted expression levels. The procedure used to obtain the significance of X–Y co-expression is described in Supplemental Methods.

**Differential expression across tissues**

The housekeeping normalization was first applied to all XY samples from all tissues. Then, for each gene of interest, its \( \log_2(\text{TPM} + 0.5) \) expression levels were compared in each pair of tissues (excluding tissues with fewer than 30 samples) with a Welch’s \( t \)-test (Python function: scipy.stats.ttest_ind with equal_var=False). A gene was considered to be significantly
differentially expressed between two tissues if the $p$-value was less than $10^{-3}$ and its average expression levels in the two tissues differed by at least 30% (1.3-fold). A gene was considered to be up-regulated (or down-regulated) in a tissue if its expression in that tissue was significantly higher (or lower) than its expression in at least 75% of the other tissues (to allow for the possibility of up-/down-regulation in multiple tissues). This analysis was limited to the 9 X–Y gene pairs where the Y homolog was robustly expressed in many tissues. TXLNG/TXLNGY was excluded because the regulation of TXLNGY expression appears to have diverged almost completely from the regulation of TXLNG.

**microRNA analyses**

Scripts from TargetScan 6.0 (Friedman et al. 2009) were used to identify and evaluate miRNA target sites in the 3′ UTRs of the X- and Y-linked homologs of each widely expressed X–Y gene pair. Sites identified in X homologs were validated using the latest TargetScan predictions (release 7.2 (Agarwal et al. 2015)) (Supplemental Table S9). miRNA expression patterns were evaluated using quantile-normalized expression values from Ludwig et al. (2016) (Ludwig et al. 2016). Among target sites for tissue-specific, highly expressed miRNAs, the miR-1 target site in EIF1AX is the target site with the greatest predicted efficacy that is preserved in one homolog of an X–Y pair but not the other. For luciferase assays, EIF1AX’s miR-1 site was changed to shuffled sequence, and EIF1AY’s disrupted miR-1 site was changed to match that of EIF1AX, using the QuikChange II kit (Agilent). Further details on the computational identification of miRNA sites and experimental validation with luciferase assays are provided in Supplemental Methods and Supplemental File S2.
Cross-species analyses of sequence and expression

Multiple sequence alignments of $E I F 1 A X / Y$ 3'-UTR and amino-acid sequences were generated with PRANK (Löytynoja and Goldman 2005) using fixed species trees (with separate clades for mammalian X- and Y-linked genes). Expression levels of $E I F 1 A X / Y$ homologs in male chimpanzee ($P a n$ $t r o g l o d y t e s$), rhesus macaque ($M a c a c a$ $m u l a t t a$), mouse ($M u s$ $m u s c u l u s$), and chicken ($G a l l u s$ $g a l l u s$) tissues were estimated with kallisto, using RNA-seq data from Brawand et al. (Brawand et al. 2011) and Merkin et al. (Merkin et al. 2012). See Supplemental Methods for further details.

Quantitative proteomic analysis of human heart tissue

Heart (left ventricle) samples from 21 male donors and 12 female donors were obtained from the GTEx tissue biobank for quantitative proteomic analysis after thoroughly screening all left ventricle samples by donor medical history and histopathological analysis (Supplemental Methods; Supplemental File S3). Quantitative proteomic analysis was performed in three 11-plex TMT experiments as previously described (Chick et al. 2016) and as detailed in Supplemental Methods. The protein encoded by a Y-linked homolog of an X–Y gene pair (Y isoform) was determined to be present in heart tissue if at least one peptide with the following two properties was detected: (1) its sequence specifically matched the Y isoform and no other protein; (2) it showed signal above background only in male samples. For proteins not encoded by X–Y gene pairs, protein abundance was estimated as previously described (Chick et al. 2016) (Supplemental Methods) (Supplemental Table S12). Protein abundances of the X and Y isoforms were estimated separately using X-isoform-specific and X- and Y-isoform-shared peptides, as detailed in Supplemental Methods.
Immunoblot experiments

Human heart-tissue lysates (from tissue obtained for the mass spectrometry analysis) were pooled by sex for immunoblotting. EIF1AX and EIF1AY protein was detected with an eIF-1A primary antibody (Abcam Ab177939, anti-rabbit), with GAPDH (Ambion AM4300, anti-mouse) as a loading control. eIF-1A levels were quantified using the Odyssey CLx Imaging System (LI-COR). Four technical replicates were performed per sex. To verify that the eIF-1A antibody recognizes both EIF1AX and EIF1AY, immunoblot experiments were performed with protein lysates from human lymphoblastoid cell lines with varying numbers of sex chromosomes (45,X; 46,XX; 46,XY; 47,XXY; 48,XXXX; 49,XXYY; 49,YYYY) and, correspondingly, varying levels of EIF1AX and EIF1AY. See Supplemental Methods for further experimental details.

DATA ACCESS

The proteomic data generated in this study have been submitted to the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al. 2018; https://www.ebi.ac.uk/pride/archive/) under accession number PXD017055. Processed data (re-estimated TPM matrices) have been submitted to Zenodo (zenodo.org) under DOI 10.5281/zenodo.3627233.

ACKNOWLEDGEMENTS

We are grateful to A.K. San Roman for sharing sex-chromosome-aneuploidy cell lines, to S.W. Eichhorn and S.E. McGeary for discussions about miRNA targeting, and to current and former members of the Page lab for valuable input over the course of this project. We thank D.W.
Author contributions: A.K.G., H.S., and D.C.P. designed the study. A.K.G. performed computational analyses. S.N. performed computational and experimental analyses of miRNA target sites. R.N.M. performed histological evaluations on human heart tissue samples. J.M.C. and S.P.G. contributed proteomic analyses, with assistance from L.C. A.K.G. performed mass-spectrometric data analysis of X and Y isoform abundance. L.C. performed immunoblotting experiments. A.K.G. and D.C.P. wrote the paper.

DISCLOSURE DECLARATION

The authors declare no competing interests.

REFERENCES


Table 1. Published evidence for functional equivalence or difference of proteins encoded by widely expressed, ancestral X–Y gene pairs.

| X–Y Pair          | a.a. % id. | Evidence supporting at least partial equivalence | Evidence supporting difference
|-------------------|-----------|--------------------------------------------------|---------------------------------
| KDM5C/KDM5D       | 87%       | KDM5C and KDM5D demethylate di- and trimethylated histone 3 lysine 4 *in vitro* (Iwase et al. 2007). *Kdm5d rescues inviability of Kdm5c-knockout mice* (Kosugi et al. 2020). | Compared to KDM5C, KDM5D shows reduced demethylase activity *in vitro* (Iwase et al. 2007). |
| USP9X/USP9Y       | 91%       | -                                                | -                               |
| DDX3X/DDX3Y       | 92%       | Human DDX3X and DDX3Y rescue cell-proliferation defect conferred by *Ddx3x* mutation in hamster cell line (Sekiguchi et al. 2004). *DDX3Y is essential for cell proliferation in a lymphoma cell line with a truncating mutation in DDX3X* (Wang et al. 2015). | -                               |
| PRKX/PRKY         | 92%       | -                                                | -                               |
| RPS4X/RPS4Y1      | 93%       | Human RPS4X and RPS4Y1 rescue cell-proliferation defect conferred by *Rps4x* mutation in hamster cell line (Watanabe et al. 1993). | -                               |
| ZFX/ZFY           | 93%       | -                                                | -                               |
| EIF1AX/EIF1AY     | 99%       | -                                                | -                               |
| NLGN4X/NLGN4Y     | 99%       | -                                                | -                               |

*Percent amino-acid sequence identity (Skaletsky et al. 2003)

b Dashes (“-”) indicate an absence of published evidence, to our knowledge.

c A functional “difference” could include quantitative differences in the same protein function (e.g., differences in enzymatic activity) or qualitatively distinct protein functions.
FIGURE LEGENDS

Fig. 1. Estimates of MSY gene expression across 36 human tissues. (A) Outside of the two pseudoautosomal regions (PAR1, PAR2), the X and Y Chromosomes have diverged in sequence. The locations of protein-coding genes and multi-copy gene families in the male-specific region of the human Y Chromosome (MSY; blue) are shown at right. The X-linked homologs of MSY genes are annotated in the non-pseudoautosomal region of the X (orange); numbers (ancestral X–Y pairs) and letters (acquired X–Y pairs) match MSY genes to their X-linked homologs. (B) Fraction genes on autosomal chromosomes (1–22), the X Chromosome, or the MSY expressed above 1 TPM in at least one tissue when multi-mapping RNA-seq reads are discarded (gray) or included (red). Error bars: minimum and maximum values among individual autosomes. (C) Each point shows estimated expression level of the HSFY gene family in a single sample when multi-mapping reads are included (red) and discarded (gray). Lines show median expression levels. The 15 tissues shown are those with the highest median expression level after discarding multi-mapping reads, in descending order. (D) Median expression levels of MSY genes and gene families (*) in each tissue, with row and column order determined by hierarchical clustering. Stars denote the tissue with the highest expression for a given gene.

Fig. 2. Quantitative comparison of X- and Y-homolog expression in XY individuals. (A) Expression of RPS4Y1 (blue) and its X-linked homolog RPS4X (tan) in individual skeletal muscle samples (50 of 255 total samples are shown). (B – C) Each point shows Y/X expression ratio for one widely expressed X–Y gene pair in one tissue; points are grouped by gene pair (B) or tissue (C). In B, colors denote higher Y-homolog expression (blue), higher X-homolog expression (tan), or no significant difference (gray) (Wilcoxon signed-rank test, FDR < 0.05). In
C, Y/X expression ratios in testis are highlighted. Highlighted point (yellow) shows the summary
data in A. (D – G) Each point shows the co-expression of an MSY gene with another gene in a
single skeletal muscle sample: \textit{RPS4Y1} vs. \textit{RPS4X} (D), \textit{RPS8} (E), \textit{ZFY} (F); \textit{ZFY} vs. \textit{ZFX} (G).

(H) Each cell shows the correlation coefficient (Spearman’s \( \rho \)) of expression for X- and Y-linked
members of an X–Y gene pair in one tissue. Stars indicate that a member of the X–Y pair shows
more correlated expression with its homolog than with 95\% of other genes in the genome. The
highlighted cell (yellow) summarizes data in C and A.

\textbf{Fig. 3. Tissue-specific up- and downregulated of X and Y homologs.} Each cell shows a gene’s
expression level in a tissue relative to its median expression level across all tissues. Stars denote
tissues where a gene’s expression level is significantly higher or lower (±30\%; Welch’s \( t \)-test, \( P < 0.001 \)) than in 75\% of other tissues.

\textbf{Fig. 4. Y-specific loss of miR-1 target site led to elevated \textit{EIF1AY} expression in XY heart
and tissue.} (A) Each point shows expression level of \textit{EIF1AY} (blue) or \textit{EIF1AX} (gold) in a
single tissue sample from an XY individual. Lines show median expression level. (B) Alignment
of 3′-UTRs of \textit{EIF1AY}, \textit{EIF1AX}, and their orthologs; miR-1 target site in pink. Key branch
points annotated with estimated divergence times in millions of years ago (mya). Fully conserved
sites annotated with “*”; sites consistent with a single evolutionary substitution event annotated
with “.”. (C) Quantile-normalized expression levels of miR-1 across human tissues. (D) Activity
of luciferase reporter fused to 3′-UTR sequences of \textit{EIF1AX} or \textit{EIF1AY} with intact (+) or
disrupted (−) miR-1 site in HEK293 cells, upon transfection with miR-1 or miR-124. Luciferase
activity of each reporter with a disrupted miR-1 site is normalized to activity of corresponding
reporter with intact site. \( P \)-values from two-sided Welch’s \( t \)-test. (E – F) Each point shows Log2(Y/X expression ratio) for EIF1AY/EIF1AX orthologs in macaque (E) and chimpanzee (F).

**Fig. 5. Tissue-specific, male-biased expression of X–Y gene pairs at the transcriptional level.** Each pair of bars shows the median expression level of an X–Y gene pair in XX (left) and XY (right) samples from one tissue. Expression is normalized to the level in XX samples. In XY samples, the sum of X-homolog (tan) and Y-homolog (blue) expression is shown. Error bars: 25th and 75th percentiles. “*”: X-homolog expression in XX samples is significantly different from the summed X- and Y-homolog expression in XY samples: FDR < 0.05.

**Fig. 6. Male-biased expression of EIF1A protein in the heart.** (A) Signal/noise values for Y-specific peptides in XX (gray) and XY (green) samples. “Set” refers to the 11-plex experiment (out of three total) in which the peptide was detected. The dotted line shows average signal/noise value in XX samples. (B) Amino-acid sequence of EIF1AX/Y: X- and Y-specific amino acids are superscripted and subscripted, respectively. X-specific (gold), Y-specific (blue), and X–Y shared (purple) peptides detected by mass spectrometry are shown, along with the number of 11-plex experiments in which each peptide was detected. (C) Relative abundance of X and Y protein isoforms in XX (\( n = 12 \)) and XY (\( n = 21 \)) heart tissue samples by mass spectrometry. For each X–Y pair, points show the levels of the X isoform (gold) or the total level of the X and Y isoform (purple) in XX samples compared to XY samples, from which the relative proportion of X and Y isoform expression in XY samples can be inferred (dotted white line). \( P \)-values by estimated by permutation. (D) Comparison of estimated Y/X expression ratios and sex-biased expression from RNA-seq and mass-spectrometry. Asterisks indicate statistical significance in
the corresponding analysis. (E) Abundance of EIF1A and GAPDH by Western blot in pooled
XX and XY protein lysates. (F) Quantification of EIF1A levels in pooled XX and XY samples
by Western blot; \( P \)-value by Welch’s \( t \)-test.
Figure 3

Log$_2$(TPM in tissue / median TPM across tissues)

30% higher or lower expression than in 75% of other tissues
Figure 6

A

Y-specific Peptide

<table>
<thead>
<tr>
<th>Set</th>
<th>Y-specific Peptide</th>
<th>Signal/noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>K.SVEDEDESK.H</td>
<td>2.57</td>
</tr>
<tr>
<td>1</td>
<td>R.GLIDSNVR.H</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>R.SDYGIGR.I</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>R.VSGALSTATSFNEK.N</td>
<td>1</td>
</tr>
</tbody>
</table>

DDX3

EF1AY

RPS4X/RPS4Y1

Normalized intensity

XX (n=12)  YY (n=21)

DDX3X/DDX3Y

Normalized intensity

XX (n=12)  YY (n=21)

EIF1AX/EIF1AY

Normalized intensity

XX (n=12)  YY (n=21)

B

EIF1AX/EIF1AY amino-acid sequence

MPKNGKGKGNRGGKQNESEKRELKRE

C

P=0.05

P=0.03

P=0.004

D

<table>
<thead>
<tr>
<th></th>
<th>RPS4X/RPS4Y1</th>
<th>DDX3X/DDX3Y</th>
<th>EIF1AX/EIF1AY</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-seq</td>
<td>1.14</td>
<td>0.78</td>
<td>4.26</td>
</tr>
<tr>
<td>mass-spec</td>
<td>0.91</td>
<td>0.30*</td>
<td>1.67*</td>
</tr>
</tbody>
</table>

E

Pool heart lysates

XX (n=12)  YY (n=21)

eIF-1A

F

Relative pixel intensity

XX (n=8 lanes each)  YY
Quantitative analysis of Y-Chromosome gene expression across 36 human tissues

Alexander K Godfrey, Sahin Naqvi, Lukas Chmatal, et al.

*Genome Res.* published online May 27, 2020
Access the most recent version at doi:10.1101/gr.261248.120

Published online May 27, 2020 in advance of the print journal.

Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see [http://genome.cshlp.org/site/misc/terms.xhtml](http://genome.cshlp.org/site/misc/terms.xhtml)). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at [http://creativecommons.org/licenses/by-nc/4.0/](http://creativecommons.org/licenses/by-nc/4.0/).

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://genome.cshlp.org/subscriptions).

To subscribe to *Genome Research* go to: [http://genome.cshlp.org/subscriptions](http://genome.cshlp.org/subscriptions)

Published by Cold Spring Harbor Laboratory Press