A Quantitative View of Y-Chromosome Gene Expression across the Human Body

by

Alexander Kamitsuka Godfrey
A.B., Molecular Biology
Princeton University, 2010

Submitted to the Department of Biology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Signature of Author ..... Alexander K. Godfrey
Department of Biology
May 15, 2020

Certified by ..... David C. Page
Professor of Biology
Director, Whitehead Institute
Investigator, Howard Hughes Medical Institute
Thesis Supervisor

Accepted by ..... Mary Gehring
Associate Professor of Biology
Member, Whitehead Institute
Co-Director, Biology Graduate Committee

A Quantitative View of Y-Chromosome Gene Expression across the Human Body

by

Alexander Kamitsuka Godfrey

Submitted to the Department of Biology on May 15, 2020
in partial fulfillment of the requirements
for Degree of Doctor of Philosophy of Biology

ABSTRACT

Human Y-chromosome genes have long been known to play pivotal roles in two biological processes—sex determination and spermatogenesis. Recent studies have uncovered evidence that Y-chromosome genes also perform important functions beyond the reproductive tract. Little is known about the roles of Y-chromosome genes in these contexts, or how their expression might directly lead to differences between XX (female) and XY (male) individuals.

This thesis presents a quantitative analysis of human Y-chromosome gene expression across 36 adult tissues collected from hundreds of individuals. Compared to past efforts, this work greatly expands the number of tissues in which Y-chromosome gene expression has been measured and offers a richly quantitative picture that could not previously be achieved. In contrast to traditional views of the Y chromosome, we show that Y-chromosome genes are abundantly expressed in a variety of tissues outside the reproductive tract. We additionally find that regulatory-sequence differences between the X and Y chromosomes can lead to Y-chromosome-driven, male-biased expression of critical regulatory genes. In one notable example, evolutionary loss of a conserved microRNA site on the Y chromosome enabled a Y-linked copy of eukaryotic initiation factor 1A, *EIF1AY*, but not its X-linked homolog *EIF1AX*, to be highly expressed in the heart. As a result, this essential translation initiation factor is nearly twice as abundant in male as in female heart tissue at the protein level. We were able to arrive at these conclusions through careful application of analytic and experimental methods suited specifically to the Y chromosome's complexity; guidelines for the selection and use of these methods are discussed. Taken as a whole, these efforts shed new light on the Y chromosome's evolution and possible roles in sex differences and suggest promising future avenues for Y-chromosome research.

Thesis Supervisor: David C. Page
Title: Professor of Biology

ACKNOWLEDGEMENTS

The work described in these pages would not have been possible without the support of numerous mentors, colleagues, friends, and family members…

I am deeply grateful to my advisor, David Page, for the opportunity to work at a true scientific frontier, especially one with such social resonance; for helping me understand the virtues of simplicity and for many valuable lessons in communication; for giving me the freedom to learn and grow; for being a role model of principled leadership; for helping me see that there is beauty, creativity, and humanity in science. I am confident that all I have learned from you will serve me well wherever my professional and personal life takes me.

I would also like to thank the members of my thesis committee, Bob Horvitz and Dave Bartel, for their rigorous perspectives on my work and their encouragement over the years, as well as for their comments on this thesis.

The members of the Page Lab have provided a wonderful scientific home during my time in graduate school. I want to particularly thank Mina Kojima for being a wonderful baymate for many years, as well as Holly Christensen and other inhabitants of Room 411 for making this a great home base. I am grateful for Jorge Adarme and Susan Tocio, for all of their seen and unseen efforts to support my graduate work, and for their support in the chaotic few days before my defense. And I'd like to thank Tatyana Pyntikova, Natalia Koutseva, and Laura Brown for maintaining the Page Lab's distinctive culture of laid-back fun. My deepest appreciation and admiration go out to Winston Bellott and Helen Skaletsky for all of their sage advice over the years, and to Sahin Naqvi, alongside whom I learned most of what is presented in this thesis. Finally, I have greatly enjoyed working with other founding members of the Sex Differences "Subgroup", especially Adrianna San Roman, Lukáš Chmátal, and Laura Blanton. I am amazed at how you have transformed this scientific initiative and can't wait to see where you take it next.

I couldn't be more thankful for my wonderful group of friends—many of the closest friends I have ever made—for countless hours of conversation, many late nights and early mornings, adventures around Cambridge, Somerville, and beyond. Thank you for helping make these the best years of my life thus far. And to my dear friend, Sean, this just wouldn't have been the same without you.

Finally, and most importantly, I am thankful for my parents: the givers of my X and Y chromosomes, tireless cheerleaders in all my pursuits. I do not thank you enough and could not thank you enough for all of your support, enthusiasm, and inspiration, for all you have given to get me to the place where I am today. These pages are dedicated to you.

CONTENTS

Chapter 1. Introduction

The diploid human genome comprises roughly six billion DNA base pairs. Substituting a base at a single position can have stark phenotypic effects, from a change in eye color, to dramatically increased risk of disease. Dwarfing these minute variations in DNA sequence is the variation in sex-chromosome complement as XX or XY. Substituting a ~160 megabase (Mb) X chromosome for the much smaller ~60 Mb Y chromosome constitutes 100 million base pairs lost or gained.

What is phenotypic consequence of such a massive genetic variation? It is known that transient expression of a single gene on the Y chromosome is sufficient to "nudge" the embryo down a path towards the development of male anatomy (Sekido and Lovell-Badge 2009). In XY individuals, Y-chromosome genes are essential for the production of sperm (Krausz and Casamonti 2017); in XX individuals, two X chromosomes appear to be essential for the survival of oocytes (Toniolo and Rizzolio 2007). Beyond the gonad, having two X chromosomes or one X and one Y chromosome is correlated with numerous other differences between females (typically XX) and males (typically XY) in anatomy, physiology, and the manifestation of disease. But little is known conclusively about how the sex-chromosome complement as XX or XY contributes (or does not contribute) directly to these non-reproductive differences. Even less is known about the Y chromosome's contribution to the equality or inequality of XX and XY.

The motivation of this thesis work has been to gain a deeper understanding of human Y-chromosome genes, how they express themselves particularly outside the reproductive tract, and what this might mean for differences between XX and XY individuals. A useful framing is to ask how Y-chromosome genes are different from genes on the X chromosome. Differences between the X and Y chromosomes are the product of millions of years of dramatic evolutionary history, during which an ordinary pair of autosomes was transformed into the differentiated X and Y chromosomes we find today. This introductory chapter therefore opens with a brief review of the mechanics of sex-chromosome evolution in general but using the mammalian X and Y chromosomes as its primary example. I then trace the intellectual history of Y chromosomes from their discovery to the present day. For much of that time, the Y chromosome was viewed as a genetic wasteland or as specialized for reproduction; only very recently has it become compelling to pursue the Y chromosome's roles outside the gonad. These older views, however, are difficult to unseat and have likely contributed to the current deficit of understanding of Y-chromosome genes. An attempt to synthesize what we currently know and do not know about Y-chromosome genes beyond the reproductive tract forms the last section of this chapter.

SEX-CHROMOSOME EVOLUTION AND Y-CHROMOSOME DEGENERATION

Differentiated X and Y chromosomes (or Z and W chromosomes[1]) evolve from ordinary pairs of autosomes. Pairs of sex chromosomes have formed independently many times in many animal and plant lineages (Bull 1983). The mammalian X and Y chromosomes are

[1] By convention, in species where females are homogametic—i.e., all of their gametes contain the same sex chromosome—the sex chromosomes are called X and Y, with females having XX and males XY. In species where males are homogametic (e.g., as in birds), the sex chromosomes are called Z and W, with males having ZZ and females ZW. For simplicity, I will use "X" and "Y" to talk about the general case of sex chromosome evolution.

derived from a pair of autosomes present ~300 million years ago (Mya) in the common ancestor of amniotes (Ross et al. 2005; Lahn and Page 1999). The avian Z and W chromosomes are derived from a distinct set of ancestral amniote autosomes (Fridolfsson et al. 1998; Ross et al. 2005; Bellott et al. 2010; Nanda et al. 1999). The first step in the formation of a pair of sex chromosomes is thought to be the acquisition of a new sex-determining gene (or genes). For differentiation of the proto-X and proto-Y chromosomes to proceed, X–Y crossing-over must then be suppressed at the sex-determining locus (e.g., through a chromosomal inversion (Lahn and Page 1999)). The absence of X–Y interchange transforms former pairs of autosomal alleles into independently segregating X- and Y-linked genes. The X- and Y-linked members of these homologous X–Y gene pairs can then acquire mutations independently and diverge. Outside of this region of suppressed recombination—in the X and Y chromosomes' "pseudoautosomal" regions—the X and Y chromosomes remain identical in sequence. It is the divergent, non-pseudoautosomal regions of the X and Y chromosomes that are the focus of this thesis.[2]

The hallmark of sex-chromosome evolution is the degeneration of the sex-specific Y (or W) chromosome (Ohno 1967). Genes originally present on the ancestral autosome pair are asymmetrically lost from the Y but retained on the X. Outside the human X and Y chromosomes' small pseudoautosomal regions, the human Y chromosome retains only ~3% of the ancestral autosomal genes (17 of ~640) compared to ~98% that are found on the X (Skaletsky et al. 2003; Bellott et al. 2014; Ross et al. 2005). Similarly, only 4% of ancestral autosome genes remain intact on the chicken W chromosome compared to 97% on the chicken Z (Bellott et al. 2017). Studies of more recently formed sex-chromosome systems support the view that decay of Y-linked sequences is rapid and progresses over time. In various species of the *Drosophila* genus, autosomes became fused to the X and Y chromosomes at various points in the past, forming "neo" sex of various ages. The neo-Y

[2] Throughout this chapter, it can be assumed that all discussion of the X and Y chromosomes refers to their non-pseudoautosomal regions.

chromosomes of *D. albomicans* (formed ~0.1Mya), *D. miranda* (~1.5 Mya), and *D. pseudoobscura* (~15 Mya) form a gradient of decay, with ~40% of ancestral genes inactivated on the neo-Y after 1.5 My and few traces of homology with the neo-X and ancestral autosome detectable by 15 My (Zhou et al. 2012; Carvalho and Clark 2005; Zhou and Bachtrog 2012; Mahajan and Bachtrog 2017). The older Y chromosome of *D. melanogaster* (~70 Mya) has lost all traces of its ancestral origin (Mahajan and Bachtrog 2017; Carvalho et al. 2000, 2001; Vicoso and Bachtrog 2015).

The loss of genes from the Y chromosome entails both a reduction in gene dosage in XY individuals and an imbalance in gene dosage between XX and XY individuals. To counteract these changes, systems of X-chromosome dosage compensation have evolved alongside Y-chromosome degeneration. Different strategies for dosage compensation, utilizing different molecular mechanisms, have evolved in different lineages. These include upregulating genes on the single X in XY animals (*D. melanogaster*), downregulating both Xs in XX animals (*C. elegans*), and randomly inactivating an X in XX individuals (while also upregulating the X in both XX and XY) (mammals) (Disteche 2012). In mammals, X-chromosome dosage compensation appears to evolve on a gene-by-gene basis in response to the loss of Y-chromosome genes: ancestral X chromosome genes with surviving Y homologs are not subject to X-chromosome inactivation (XCI) (Jegalian and Page 1998; Carrel and Willard 2005).

The Y chromosome's tendency to degenerate stems from its inability to regularly recombine with a homologous chromosome during meiosis (Charlesworth and Charlesworth 2000). On an autosome, linked beneficial and deleterious mutations that co-occur on a single haplotype can be separated by recombination, enabling natural selection to increase the frequency of the beneficial mutation while independently eliminating the deleterious mutation. (The same mechanism is available to the X chromosome in females.) By contrast, Y-linked sequences are transmitted as uninterrupted haplotype blocks from father to son, generation after generation. Interference between linked causes selection to be inefficient: deleterious mutations can

"hitchhike" to fixation alongside a strongly beneficial mutation or lead to the elimination of adaptive mutations depending on the aggregate fitness effect (Rice 1987; Orr and Kim 1998; Bachtrog 2013; Charlesworth and Charlesworth 2000). Theoretical models predict that the rate of gene loss might slow over time, as the smaller number of genes reduces the chance of interference between segregating mutations (Bachtrog 2008). Nevertheless, both empirical and theoretical observations have reinforced a fundamental intuition—omnipresent in Y-chromosome research—that Y-chromosome genes will tend towards decay in the absence of (and perhaps even in the presence of) selective forces to resist it.

THE EVOLVING UNDERSTANDING OF Y-CHROMOSOME GENES

Early debates over the nature of the Y chromosome

Sex chromosomes were first described at the turn of the 19th century, in studies of chromosome behavior during meiosis. Studying the meiotic divisions of the fire wasp *Pyrrhocoris apternis* (an insect we know today to have an XX (female)/XO (male) sex-chromosome system), Henking observed a peculiar chromatin element—"element *x*"—that entered only one half of the sperm resulting from a given set of meiotic divisions (Henking 1891). The first suggestion that this "X" chromosome functioned as the inherited basis of sex arrived shortly thereafter, in 1902 (McClung 1902). In 1905, reports by Nettie Stevens and Edmund Wilson revealed yet another type of chromosome constitution. Some species had an "unequal" pair of chromosomes, with one larger and one smaller (Stevens 1905; Wilson 1905a, 1905b)—one X and one Y.

Dueling views of Y chromosomes soon emerged, fueled by early genetic studies of the fruit fly (*Drosophila melanogaster*) on the one hand and fish on the other. Fly geneticists argued that the Y chromosome was degenerate or entirely lacking in genes. In 1914, Muller supported this position with three lines of evidence (Muller 1914). First, Y chromosomes seemed to vary frivolously in size in closely related species. Second, citing Calvin Bridges' then-unpublished studies of chromosomal non-disjunction, the Y

chromosome appeared immaterial to sex determination or viability (Bridges 1916). XX and XXY flies were female, XO and XY flies were male, and OO and OY flies were never observed. (Bridges would note that males lacking a Y chromosome showed impaired fertility, suggesting the Y might not be entirely devoid of genes, but this point received little attention (Bridges 1916).) Third, if genes existed on the Y chromosome, they were never dominant to genes on the X chromosome. In fact, the X-linked recessive mode of inheritance, a linchpin in proving the chromosomal basis of heredity, presupposed as a necessary condition the degeneracy of the Y (Morgan 1910).

Quite a different view of the Y chromosome was taking shape from studies of fish. In *Lebistes* (guppies) and medaka, external coloration traits were shown to exhibit Y-linkage (Schmidt 1920; Aida 1921; Winge 1922). In 1927, a survey of 18 *Lebistes* traits—a wide variety of colored spots and stripes and fin ornamentations—found that nine mapped to the Y, three mapped to the X, and five appeared to occasionally cross over between X and Y (and one trait was autosomal) (Winge 1927). Winge proposed that the occasional interchange between the X and Y chromosomes was a function of proximity to a male determining factor on the relatively gene-rich Y chromosome (Winge 1923, 1927). In response to skepticism from the fly geneticist Thomas Hunt Morgan, who suggested that these coloration genes might in fact be on an autosome fused to the Y chromosome, Winge notes:

> "Possibly… it is to some extent the general idea of genes on the Y chromosome which leads Morgan to seek for a more complicated explanation… but it must surely seem more remarkable, and more unexpected, that the X chromosome in *Drosophila* should contain many genes and the Y none than that *Lebistes* should be found to have genes both in X and Y. … That a dominant male-determining factor in the Y chromosome of *Lebistes* should seem remarkable to us is, then, really only due to the fact that we find essentially different conditions in the well-investigated *Drosophila*" (Winge 1927).

It is clear that debates over the degeneracy of the Y have been present since the beginning.

The degenerate Y with perhaps one sex-determining gene

XY sex chromosomes were identified cytologically in mammals in 1921 (Painter 1921). Encouraged by studies of fish, researchers embarked on a hunt for Y-linked traits in humans. Over the ensuing decades, analyses of family trees yielded a multitude of supposedly Y-linked traits, including webbed toes, hairy ears, and scaly skin. But in 1957, Curt Stern re-analyzed and debunked each of these reports (Stern 1957). Two years later, reports of 47,XXY (Klinefelter syndrome) males and 45,X (Turner syndrome) females established that the human Y chromosome encoded a sex-determining gene (Jacobs and Strong 1959; Ford et al. 1959), but this was the exception.

Susumu Ohno's classic 1967 monograph on sex chromosomes helped to cement the view of the degenerate Y chromosome as part of a unified evolutionary narrative. Through a sweeping review of cytological and genetic observations about sex chromosomes, he formally proposed that differentiated sex chromosomes evolved from autosomes, a process characterized by the exclusive degeneration of the Y (or W) and complete conservation of the X (or Z) (Ohno 1967). Although many of his claims would be shown to be correct decades later through molecular-genetic and genomic analyses, the reasoning underlying some are no longer valid. For example, a key piece of evidence for the progressive differentiation of sex chromosomes was that, both across vertebrates and within vertebrate lineages, "lower" vertebrates tended to have cytologically indistinguishable sex chromosomes, whereas "higher" and "more evolved" vertebrate species had more differentiated sex chromosomes. By this logic, the Y chromosome of humans and other mammals must indeed be in a very advanced state of decay, perhaps retaining just one sex-determining gene.

Ohno also briefly speculated about the nature of the sex-determining gene and concluded that its action must be very limited in nature. Alfred Jost's 1947 experiments had shown that castrating XY rabbits *in utero*, prior to any detectable sexual differentiation, led them to develop with female reproductive anatomy (Jost 1947). This implied that the key decision underlying sex determination was whether the bipotential

gonad developed as testis or ovary. Through further analysis of past literature, he reasoned that this decision must be made during a narrow window of development in somatic cells of the gonad. (Indeed, decades later, characterization of the mammalian sex-determining gene, *SRY*, would show that mouse SRY exerts its pivotal action through perhaps two hours of expression in a single gonadal cell lineage (Sekido and Lovell-Badge 2009). Moreover, *SRY* is Y chromosome's lone sex-determining gene. No other Y-chromosome gene is known to be directly required for the differentiation of primary or secondary sexual characteristics.)

The Y chromosome as fertility factor

By the early 2000s, the commonly held view of the human Y chromosome had expanded to include a second function—a role in spermatogenesis.

In the 1980s, efforts commenced to construct a map of the human Y chromosome. The Y chromosome's lack of recombination precluded the construction of traditional maps based on genetic linkage, so the Y chromosome was mapped instead using naturally occurring deletions and translocations. Males carrying some of these deletions or rearrangements were infertile. Through these efforts, a putative gene for spermatogenesis proposed by Tiepolo and Zuffardi (Tiepolo & Zuffardi, 1976) was localized to a segment of the Y-chromosome long arm (Andersson et al., 1988). Follow-up investigations implied that Y-chromosome deletions might be a very common cause of spermatogenic failure: in a screen of males with non-obstructive azoospermia (failure to detect sperm in semen, with physical blockage excluded), 12 of 89 individuals had overlapping Y-chromosome deletions (Reijo et al. 1995). Further studies identified Y-chromosome deletions in two other regions, implying that there might be multiple spermatogenesis genes (Vogt et al. 1996). (Today, five major classes of recurrent Y-chromosome deletions have been defined and constitute the most common, identified genetic causes of impaired spermatogenesis (e.g., 5 – 10% in non-obstructive azoospermia (Hughes and Rozen 2012; Krausz and Casamonti 2017).)

Meanwhile, advances in molecular genetics techniques gave researchers increasing success in cloning genes from the Y chromosomes of humans and other mammals. The emerging view of the Y chromosome from these reports reaffirmed the Y chromosome's connection to spermatogenesis. Many of the genes identified appeared to be expressed only in testes, typically only in spermatogenic cells (Reijo et al. 1995; Lahn and Page 1997; Ma et al. 1993; Arnemann et al. 1991; Manz et al. 1993; Mitchell et al. 1991). Some of these Y-chromosome genes were identified as ubiquitously expressed in some species but testis-specific in others (Koopman et al. 1989; Mardon et al. 1990; Schneider-Gädicke et al. 1989). Some genes had homologs on the X chromosome that were widely expressed. Thus it appeared that Y-chromosome genes displayed a tendency to evolve testis-specific expression and, possibly, a specialization for spermatogenesis.

Fertility soon became ensconced in the evolutionary narrative of the Y chromosome. *DAZ*, a gene that was recurrently deleted in men with spermatogenic failure, had actually been acquired from autosome during mammalian evolution, contravening expectations of the Y's inexorable decay (Saxena et al. 1996). What factors might favor the acquisitions of a Y-chromosome gene? In 1931, R. A. Fisher first speculated that the male-limited inheritance of Y chromosomes should cause them to accumulate genes that have opposing fitness effects in the sexes—specifically, that increase the fitness of male carriers but decrease the fitness of female carrier (Fisher 1931). He was inspired by Winge's observations of Y-linked coloration traits in guppies: these showy displays would help males attract mates (and thus increase fitness), but would only increase predation risk for females (and thus decrease fitness). These theories of "sexually antagonistic" genes would later be given more formal (Rice 1984; Charlesworth and Charlesworth 1980) and empirical (Rice 1992; Prasad et al. 2007) support.

In this manner, genes promoting male fertility or spermatogenesis might find particularly fertile grounds on the Y chromosome. (Although it is not immediately clear why genes specialized for spermatogenesis would be harmful to female carriers, some

empirical support for this idea can be found in *Drosophila* (Innocenti and Morrow 2010; Zhou and Bachtrog 2012).) These theories lent formal credence to the seemingly intuitive view that genes involved in spermatogenesis should be found on Y chromosomes. Thus, it was proposed that genes with male-specific roles in reproduction might be able survive Y-chromosome decay (Graves 1995; Lahn and Page 1997).

The survival of widely expressed, dosage-sensitive genes on the Y chromosome

In 2003, the sequence of the non-pseudoautosomal region of the human Y chromosome—also called its "male-specific" region (MSY)—was published, delineating its full complement of genes and the structure of its sequence (Skaletsky et al., 2003). Three classes of sequence were identified and termed "X-degenerate", "ampliconic", and "X-transposed". X-degenerate regions were single-copy regions containing genes that were homologous to, but diverged from, genes on the X chromosome, reflecting the ancestral origin of the sex chromosomes as a pair of autosomes. The ampliconic regions were structurally complex regions defined by lengthy (10s to 100s of kilobases) repeated segments in palindromic and tandem orientation, with intra-repeat sequence identity typically greater than 99.9%. The X-transposed regions showed 99% identity with X-chromosomal sequence and resulted from a X-to-Y transposition event that occurred 3–4 million years ago in the human lineage. Genes in the ampliconic regions all showed testis-specific expression, and were the genes commonly deleted in men with spermatogenic failure. By contrast, nearly all genes in the X-degenerate region (with the notable exception of *SRY*) showed apparently ubiquitous expression (i.e., in many tissues). (Only two genes were found in the X-transposed region, one showing testis-specific expression and the other showing expression mostly in the brain.)

What to make of the Y chromosome's widely expressed genes—i.e., genes without obvious male-specific roles? Lahn and Page argued that not one, but two factors could account for survival of genes on the Y chromosome (Lahn and Page 1997). They argued

that, aside from the Y chromosome's testis-specific genes, these widely expressed genes, too, were "functionally coherent" (i.e., non-random). In addition to their ubiquitous expression, they appeared to encode proteins with cellular "housekeeping" functions (e.g., one was a ribosomal protein gene, another was the eukaryotic initiation factor 1A (eIF1A), etc.). They had homologs on the X chromosome that were not subject to X-chromosome inactivation (XCI), and the proteins encoded by the X and Y homologs appeared based on sequence and (limited) experimental data (Watanabe et al. 1993) to be functionally redundant. As a result, XX cells expressed two X-linked copies of these genes, while XY cells expressed one X-linked and one Y-linked copy. They therefore proposed that these Y-chromosome genes survived to maintain the dosage of critical housekeeping functions beyond the reproductive tract.

However, this argument was not satisfying to all. Some human Y-chromosome genes appeared to be missing or present as pseudogenes on other mammalian Y chromosomes, raising the possibility that different Y chromosomes contained random, decaying subsets of the ancestral autosomes (Graves 2006). Observations in mice suggested that these widely expressed Y-chromosome genes were expressed at lower levels than their homologs on the X chromosome, consistent with partial decay (Xu et al., 2002). Particularly troubling was the fact that some mammalian species had lost their Y chromosome altogether (Arakawa et al. 2002), which seemed to reinforce the Y chromosome's ultimate dispensability, and led to predictions of its imminent demise (Aitken and Graves 2002).

A fuller understanding of these other Y-chromosome genes would only be revealed through sequencing additional mammalian Y chromosomes. Sequencing the Y chromosomes of chimpanzee and rhesus macaque showed that widely expressed human Y-chromosome genes were conserved on primate Y chromosomes (Hughes et al., 2005, 2012). Notably, each ancestral gene on the human Y chromosome was found to be intact on the macaque Y chromosome. This indicated no gene loss had occurred on the human Y chromosome for the past ~25 million years, implying its gene content had stabilized.

In 2014, Bellott et al. sequenced and analyzed the ancestral portions of eight mammalian Y chromosomes (from seven placental mammals and one marsupial), permitting a more detailed investigation of the genes that had survived Y-chromosome decay (Bellott et al. 2014). Of ~600 genes inferred to be present on the ancestral pair of autosomes, 36 remain intact on the Y chromosome of one or more species as part of a homologous X–Y gene pair. (At the time of writing, I am unaware of any study reporting an ancestral gene that survived on the Y but was lost from the X.) The pattern of gene survival across these lineages was highly non-random. For example, some genes remained intact on each of the seven Y chromosomes from placental mammals or on all eight mammalian Ys. In a separate study, simulations of sex-chromosome evolution showed that this configuration of gene survival was inconsistent with a model of indiscriminate gene loss, in which no genes are more likely to survive than others (Cortez et al. 2014).

The features of ancestral genes with intact Y homologs (X–Y gene pairs) affirmed earlier suspicions that dosage sensitivity is an important property contributing to the survival of Y-chromosome genes (Lahn and Page 1997; Kaiser et al. 2011). Compared to ancestral genes that lost their Y homologs, surviving X–Y gene pairs were more broadly expressed, subject to stronger purifying selection, and tended to encode important regulators of gene expression, including histone demethylases, transcription factors, and translation initiation factors (Bellott et al. 2014). Ancestral X–Y gene pairs also showed greater evidence of dosage sensitivity: they showed a higher computationally predicted probability of haploinsufficiency, and the X homologs of some pairs were known from clinical genetics studies to be haploinsufficient (Lederer et al. 2011). Finally, they noted that these ancestral Y-chromosome genes likely rescue the high degree of embryonic lethality found among 45,X human embryos.

With these and other observations, Bellott et al. concluded that selection enabled ancestral genes with either of two features to survive genetic decay on the Y chromosome: (1) an ancestral or acquired function in male reproduction; (2) a highly dosage-sensitive

functions that needed throughout the body. Subsequent studies have found that widely expressed, dosage-sensitive genes are also preferentially retained on the stickleback Y chromosome (White et al. 2015) and on avian W chromosomes (Bellott et al. 2017), suggesting this is likely to be a common feature of sex-chromosome evolution.

THE NEXT FRONTIER: Y-CHROMOSOME GENES BEYOND THE REPRODUCTIVE TRACT

Whatever the fate of the human Y chromosome in 10, 50, or 100 million years, comparative genomics studies have yielded evidence that Y-chromosome genes are important outside of the gonad. A defining challenge now in Y-chromosome research is to describe and understand the functions of individual Y-chromosome genes in specific, non-reproductive biological processes.

An important motivation for this endeavor is the growing recognition of differences between females and males in health and disease (Clayton, 2018; Institute of Medicine, 2001). Many autoimmune disorders, cardiovascular diseases, and neurological disorders manifest with different frequencies, severities, or qualities in males and females (Ngo et al., 2014; Regitz-Zagrosek et al., 2015; Werling & Geschwind, 2013). In some cases, as in some autoimmune disorders, one's sex is the greatest known risk factor for developing the disease (Voskuhl, 2011). Thus, understanding molecular and cellular differences between the sexes could yield insight into the etiology of disease and uncover potential avenues for treatment.

Three major factors could account for these differences: (1) the direct effects of the sex-chromosome complement as XX or XY in cells and tissues throughout the body; (2) differing profiles of circulating hormones in females and males, resulting from the hormonal output of ovaries or testes and their interactions with the rest of the endocrine system; (3) differences in the environment of females and males (a term encompassing the health implications of one's gender). Hormonal and the environmental effects were

once assumed to account for all differences between males and females, but the sex-chromosome complement as XX or XY (and thus possibly the Y chromosome itself) is increasingly recognized to be significant (San Roman and Page 2019; Arnold 2012).

To the extent that sex differences stem from the sex-chromosome complement as XX or XY, the widely expressed ancestral X–Y gene pairs are prime suspects for the mediators of these effects. The X homologs of these pairs are not subject to XCI, meaning they are biallelically expressed in XX cells, and thus are more highly expressed in XX than in XY cells (Carrel and Willard 2005; Tukiainen et al. 2017). (By contrast, X-chromosome genes that are subject to XCI typically do not show sex-biased expression.) They are also highly dosage sensitive (Bellott et al. 2014; Naqvi et al. 2018), meaning small differences in their expression levels could have large phenotypic effects. Moreover, they encode potent regulators of gene expression. The DNA-binding motif of the transcription factor *ZFX*, the X homolog of one X–Y gene pair, was found to be enriched at genes showing sex-biased expression in multiple mammalian species (Naqvi et al. 2019), suggesting it might make a sizable contribution to sex-biased gene expression in mammals.

However, if the X- and Y-linked homologs of an X–Y gene pair are equivalent—that is, if they encode identical proteins, and if they are expressed identically, such that X copies in an XX cell is always equivalent to the summed expression of the single X and Y copies in an XY cell—then there is no material difference between XX and XY. Thus, understanding the functions of ancestral Y-chromosome genes becomes a question of understanding the differences between Y-chromosome genes and their corresponding X-linked homologs.

At the onset of sex-chromosome differentiation, the X- and Y-linked members of each X–Y gene pair were identical, as similar as two alleles of any autosomal gene. But over millions of years of evolution, accumulated X- or Y-specific nucleotide substitutions might have enabled the two homologs to diverge. Below I review what we know about the differences between the X- and Y homologs of individual X–Y gene pairs along two dimensions—gene expression and protein function. I will focus on the X–Y gene pairs

where both X and Y homologs are expressed in many tissues. (For a number of X–Y pairs, divergence between the X and Y homologs is obvious. Most notably, the sex-determining gene *SRY* is the Y-linked member of an X–Y pair, whose X-linked homolog is the transcription factor *SOX3* (Foster and Graves 1994). The Y homologs of other X–Y pairs have evolved testis-specific expression, though their X homologs remain widely expressed (Bellott et al. 2014; Cortez et al. 2014).) These are the genes that are most likely to contribute to XX–XY differences in non-reproductive tissues.

Differences in X- and Y-homolog expression

Our understanding of Y-chromosome gene expression, and differences between X- and Y-homolog expression, is derived from three types of studies. First are surveys of whole Y chromosomes or large groups of Y-chromosome genes. The earliest studies of this type only considered expression in a binary ("on"/"off") manner (Lahn and Page 1997; Skaletsky et al. 2003). These studies established that Y-chromosome genes could generally be described as showing either testis-specific or ubiquitous (i.e., in many tissues) expression, while the X homologs of Y-chromosome genes showed ubiquitous expression. More recent studies of Y-chromosome gene expression based on RNA-sequencing (RNA-seq), a highly quantitative method, surveyed expression in handfuls of tissues (~5–8) with only small numbers of samples (~2–4) per tissue (Cortez et al. 2014; Bellott et al. 2014). They therefore lacked the power to detect subtle quantitative variation in Y-chromosome gene expression from one tissue to the next or to compare these patterns to those of their X homologs. As a result, these studies were only able to affirm the testis-specific/broadly expressed dichotomy from earlier work. Cortez et al. compared the expression levels of Y-chromosome genes to their X homologs in an aggregated manner, by first averaging the expression of all Y-chromosome genes from multiple tissues and species (Cortez et al. 2014). This multi-gene, multi-tissue, multi-species analysis suggested that, on average, Y-chromosome genes show lower expression

than the X homologs of Y-chromosome genes, especially in non-testis tissue. However, the patterns of individual genes (in individual species) could not be assessed.

A second set of studies have focused on the expression of individual X–Y gene pairs in single tissues or small sets of tissues. In brain tissue from mice, the Y homologs of X–Y pairs have generally been found to show lower expression than their X counterparts in XY tissue (Xu et al. 2008a, 2002). Moreover, the sum of X- and Y-homolog expression in XY tissue was less than the biallelic X-homolog expression in XX tissue. Two studies also observed some differences in the spatial patterns of X- and Y-homolog expression, for *Kdm6a*/*Uty* in mice (Xu et al. 2008b) and for *NLGN4X*/*NLGN4Y* and *PCDH11X*/*PCDH11Y* (an X–Y gene pair in the X-transposed region) in humans (Johansson et al. 2016). Outside of the brain, one recent study reported that Y-linked *TBL1Y* shows higher expression than X-linked *TBL1X* in cells of the human inner ear (Di Stazio et al. 2018).

A third group of studies set out to study dosage-compensation of the X chromosome or sex-differences in expression and came upon the X–Y pairs as a result. Trabzuni et al. and Johnston et al. compared the (biallelic) expression of X homologs in XX samples to the sum of X- and Y-homolog expression in XY samples in human brain tissue (Trabzuni et al. 2013) and human lymphoblastoid cell lines (LCLs) (Johnston et al. 2008). Observations from the human brain were similar to those from mice, with lower Y expression that also did not compensate for the higher X expression in XX samples. In human LCLs, however, some Y homologs showed higher expression than their X counterparts, leading to higher overall expression in the XY samples. A caveat of both studies is that both measured expression with microarrays, a technology that is not ideally suited for comparing the expression levels of different genes and might not fully distinguish the sequences of X and Y homologs.

Taken as a whole, it appears that Y-chromosome genes often show lower expression than their X homologs, but the disparate nature of these studies makes it hard to discern a clear and reliable view of any one gene in any one species. These studies have

also repeatedly examined a small number of well-studied tissues—e.g., the brain, testis, muscle, etc. As is evident from the case of *TBL1X*/*TBL1Y* in the inner ear, surveying a broader array of tissues will likely reveal unexpected patterns. One aim of Chapter 2 is to remedy these problems by constructing a more comprehensive, quantitative picture of Y-chromosome gene expression across many tissues and from a single species (humans).

Finally, virtually nothing is known about the expression of Y-chromosome genes at the protein level. Reliable antibodies that can distinguish X and Y protein isoforms (which show ~85–99% identity at the amino-acid sequence level) are not widely available. In a rare study to generate a custom antibody for a Y-encoded protein, Ditton et al. detected DDX3Y protein in testis but not brain or kidney, despite *DDX3Y* transcript expression in many tissues (Ditton et al. 2004). From this observation, they claimed that DDX3Y protein is only translated in the testis. However, there are reasons to doubt the reproducibility of this now well-cited study (see Gueler et al. (Gueler et al. 2012); and see below and Chapter 2). Further efforts to generate Y-specific antibodies are welcomed (Rastegar et al. 2015), but the specificity of these reagents must be rigorously established. Quantitative mass-spectrometry-based approaches provide a potential way forward, but methods to distinguish protein isoforms encoded by homologous genes are still not widely used (Malioutov et al. 2018). Standard analytic approaches are not able to distinguish Y-encoded proteins from their corresponding X-encoded isoforms or possibly other proteins. As evidence of this, publicly available analyses of human proteomic data (e.g., the Human Proteome Map, humanproteome.org, (Kim et al. 2014)) list Y-encoded proteins like DDX3Y as abundantly expressed in female-specific tissues. A proof-of-concept effort to quantify Y-encoded proteins by mass spectrometry is presented in Chapter 2.

Differences in the functions of X and Y protein isoforms

The X and Y protein isoforms of widely expressed human X–Y gene pairs show ~85–99% amino-acid sequence identity. Comparing the amino-acid sequences of the X and Y

isoforms does not yield conspicuous evidence of functional divergence, such as X–Y differences at key positions within substrate-recognition domains (A.K.G., unpublished). Both cross-species and within-human studies of their protein-coding sequences suggest that both X and Y homologs have evolved and remain under purifying selection, albeit with Y-linked sequences showing evidence of relaxed constraint (Rozen et al. 2009; Poznik et al. 2016; Wilson and Makova 2009). Thus, sequence-based analyses do not point to marked divergence in protein function. This accords with the evolutionary argument that these ancestral Y-chromosome genes survived due to selection to maintain the dosage-sensitive functions they share with their X homologs (Bellott et al. 2014).

A number of studies provide empirical evidence that the X and Y protein isoforms of some pairs are least partially redundant in function. In hamster cell lines showing arrested growth due to point mutations in the X homologs *Rps4x* or *Ddx3x*, transfection of the corresponding human Y homolog, *RPS4Y1* or *DDX3Y*, (or the corresponding human X homolog) rescues growth (Watanabe et al. 1993; Sekiguchi et al. 2004). Similarly, an unbiased, genome-wide screen for cell-type-specific essential genes identified *DDX3Y* as essential for cell proliferation in a cancer cell line with a mutation in *DDX3X* (Wang et al. 2015). (This latter observation is one strand of evidence rebutting the previously mentioned claim that DDX3Y protein is not produced outside the testis.)

Other studies point to the redundancy of X and Y homologs *in vivo*. In mice, XX animals homozygous for loss-of-function (LOF) mutations in X-linked H3K27me3 demethylase *Kdm6a* show completely penetrant embryonic lethality, which is partially rescued by its Y homolog *Uty* (i.e., some, but not all, $X^{Kdm6a-}Y^{Uty+}$ mice live to be fertile adults) (Lee et al. 2012; Welstead et al. 2012; Shpargel et al. 2012). Compound hemizygous XY mice, lacking functional copies of both *Kdm6a* and *Uty*, show completely embryonic lethality that phenocopies that of homozygous XX mice (Shpargel et al. 2012). This indicates that the partial viability of $X^{Kdm6a-}Y^{Uty+}$ mice is attributable to functional *Uty*, rather than some other factor associated with an XY sex-chromosome constitution. Moreover, KDM6A and UTY have been shown to associate with each other and some of

the same transcription-activating complexes (Shpargel et al. 2012; Gozdecka et al. 2018). A recent study reported analogous X–Y redundancy for the H3K4me3 demethylases *Kdm5c* (X-linked) and *Kdm5d* (Y-linked) (Kosugi et al. 2020).

Nevertheless, X and Y protein isoforms might only be partially equivalent in function. The clearest example of this comes from X-encoded KDM6A and Y-encoded UTY. Although the two proteins appear to be redundant with respect to some gene-regulatory functions (as described above), KDM6A functions as a H3K27me3 demethylase, whereas UTY lacks or shows severely reduced demethylase activity *in vitro* and in cellular assays (Hong et al. 2007; Lan et al. 2007; Shpargel et al. 2012; Walport et al. 2014). *In vivo* evidence that UTY retains some, but not all, of KDM6A's function is manifest in human cancers. *KDM6A* has been identified as one of the most recurrently mutated genes in human cancers (van Haaften et al. 2009; Kandoth et al. 2013), but shows a distinct mutational profile in some cancers compared to others. In hematopoietic cancers of the lymphoid lineage, such as T-cell acute lymphoblastoid leukemia (T-ALL), *KDM6A* mutations cluster in the catalytic demethylase domain, suggesting its demethylase activity is required for tumor suppression; these mutations are also found more frequently in males, consistent with no compensation (or incomplete compensation) by *UTY* (Arcipowski et al. 2016; Meulen et al. 2015). By contrast, in myeloid-lineage cancers like acute myeloid leukemia (AML), as well as in other cancers, *KDM6A* mutations occur throughout the gene and co-occur with loss of *UTY*, implying the two genes function redundantly as tumor suppressors (Gozdecka et al. 2018; van Haaften et al. 2009; Dunford et al. 2017). Mechanistic studies have confirmed that KDM6A's demethylase activity is required for tumor suppression in T-ALL but dispensable in AML, and UTY serves tumor suppressor in AML (Ntziachristos et al. 2014; Gozdecka et al. 2018).

Making inferences about the relative functions of X and Y homologs from germline mutations in humans is more challenging. LOF mutations in the X homologs of four human X–Y gene pairs (*DDX3X*, *KDM5C*, *KDM6A*, *USP9X*) have been identified as

causative of distinct congenital intellectual disability syndromes (Snijders Blok et al. 2015; Reijnders et al. 2016; Deciphering Developmental Disorders 2017; Jensen et al. 2005; Lederer et al. 2011), but the implications for the functions of their Y homologs are somewhat ambiguous. For example, *KDM6A* mutations cause Kabuki syndrome, an intellectual disability syndrome associated various skeletal and developmental abnormalities (Lederer et al. 2011). *KDM6A* mutations have been found in both XX and XY individuals, and XY individuals appear to show more severe intellectual disability, but all individuals have distinct mutations with potentially varying effects on KDM6A function and expression, confounding the comparison of XX and XY individuals (Banka et al. 2014; Bögershausen et al. 2016). Moreover, a moderating effect of an XY sex-chromosome constitution also cannot formally be excluded. Retrospective screening of Kabuki syndrome patients lacking *KDM6A* mutations has not uncovered mutations in *UTY* (Bögershausen et al. 2016), but a preferable design would include Y-chromosome genes in the initial screen. *De novo DDX3X* LOF mutations are associated with a distinct intellectual disability syndrome and are predominantly found in XX individuals (Snijders Blok et al. 2015). The few mutations identified in XY individuals appear to be hypomorphic and have been inherited from unaffected mothers (Snijders Blok et al. 2015; Nicola et al. 2019; Kellaris et al. 2018). This suggests that complete LOF mutations in *DDX3X* cause embryonic XY lethality, due to lack of compensation by *DDX3Y*. However, *DDX3Y*'s inability to compensate could be entirely explained by a lower expression level than *DDX3X* in the relevant cell type(s), partial loss of protein function, or some combination of both.

On the Y-chromosome, rare but recurrent deletions encompassing two genes, *DDX3Y* and *USP9Y*, cause azoospermia but are not associated with non-reproductive phenotypes (Vogt et al. 2008). Both genes appear to contribute to this phenotype (Tyler-Smith and Krausz 2009; Krausz and Casamonti 2017). The absence of obvious non-reproductive phenotypes has been taken as evidence that these genes function only in reproduction (Vogt et al. 2008), but this conclusion seems premature without carefully

screening *DDX3Y*-/*USP9Y*-deleted individuals for more subtle abnormalities, which could reveal quantitative phenotypic differences. If indeed DDX3Y functions throughout the body but is expressed at a lower level than DDX3X, we might not expect *DDX3Y* mutations to confer obvious phenotypes (similar to hypomorphic *DDX3X* mutations in XX individuals). The contribution of *DDX3X* to spermatogenesis is not known, because few XY individuals with *DDX3X* mutations have been identified and fertility was not reported. The overall picture is consistent with *DDX3Y* (compared to *DDX3X*) functioning in a manner that is more restricted to spermatogenesis; whether this has resulted from divergence at the expression or protein level (or both) cannot yet be determined.

Thus, the X and Y protein isoforms of X–Y gene pairs can be redundant with respect to some protein functions but not others, with these shared and divergent functions operative in closely related cell types. This partial divergence might have evolved if some aspects of their functions are dosage-sensitive, while others are not. Alternatively, natural selection might have favored different dosages of these molecular activities in females and males. The intriguing observation that UTY apparently lost its demethylase activity independently in the ancestors of humans and mice could be construed as evidence for either hypothesis. Ultimately, a richer understanding of the differences between X and Y protein isoforms would come from unbiased comparisons of their genome- or proteome-wide activities—e.g., comparing the transcriptome-wide effects of knocking down a Y-chromosome gene or its X homolog; comparing genome-wide occupancies by ChIP-seq; or, comparing their protein interaction partners by immunoprecipitation and mass spectrometry. These experiments could also reveal if Y protein isoforms have any functions not demonstrated by their X counterparts. To my knowledge, among widely expressed X–Y gene pairs, no example of a Y-specific protein function has been reported.

Challenges and opportunities

Arguably the greatest challenge to progress in understanding the functions of Y-chromosome genes is the routine exclusion of the Y chromosome from biological studies of all types. As is clear from the discussion above, much of what we know about the function of Y-chromosome genes was inferred from studies focused primarily on their X homologs. Exome- and genome-sequencing studies searching for mutations underlying disease typically do not analyze Y-chromosomal sequences. The reason for the Y's exclusion in any given study is often not knowable. However, one factor is likely to be the persistent view of the Y chromosome as degenerate and specialized for reproduction. Technical reasons are also cited, including the reduced sequencing depth of the Y chromosome compared to autosomes (a factor which, unfortunately, often leads to exclusion of the X as well (Wise et al. 2013)) and fears about the complexity of the Y chromosome's sequence. Indeed, some portions of the Y chromosome, such as the palindrome-rich ampliconic regions, require fully customized analyses (Teitz et al. 2018). But the Y-chromosome genes of greatest interest beyond the reproductive tract lie within single-copy regions that can be readily analyzed with appropriate off-the-shelf tools (Chapter 2), opening these sequences up to all researchers. The exclusion of the Y chromosome from mainstream research efforts makes it difficult to discern when the absence of "hits" on the Y chromosome represents a true negative or a false negative result. The upside of these challenges is that many are easy to overcome. Reams of new biological data are being generated every day; these data will have much to say about Y-chromosome genes, if only someone decides to analyze them.

SUMMARY

The Y chromosome's gene content (or lack thereof) has long been a matter of speculation and generalization. The human Y chromosome was once thought to be a genetic wasteland with perhaps one sex-determining gene. Molecular-genetic studies of the

1990s and early 2000s established the Y chromosome's involvement in spermatogenesis. Recently, sequencing the Y chromosomes of various species has uncovered evidence for the importance of Y-chromosome genes beyond the reproductive tract, possibly contributing to differences between XX and XY individuals in health and disease. Knowledge of the functions of individual Y-chromosome genes remains in its infancy, largely due to the ongoing exclusion of the Y chromosome from biological studies, but much could be learned by harnessing the wealth of new data being generated every day. Chapter 2 presents a case study of this very point.

REFERENCES

Aida T. 1921. On the Inheritance of Color in a Fresh-Water Fish, APLOCHEILUS LATIPES Temmick and Schlegel, with Special Reference to Sex-Linked Inheritance. *Genetics* **6**: 554–573.

Aitken RJ, Graves JAM. 2002. The future of sex. *Nature* **415**: 963.

Arakawa Y, Nishida-Umehara C, Matsuda Y, Sutou S, Suzuki H. 2002. X-chromosomal localization of mammalian Y-linked genes in two XO species of the Ryukyu spiny rat. *Cytogenet Genome Res* **99**: 303–309.

Arcipowski KM, Martinez CA, Ntziachristos P. 2016. Histone demethylases in physiology and cancer: a tale of two enzymes, JMJD3 and UTX. *Curr Opin Genet Dev* **36**: 59–67.

Arnemann J, Jakubiczka S, Thüring S, Schmidtke J. 1991. Cloning and sequence analysis of a human Y-chromosome-derived, testicular cDNA, TSPY. *Genomics* **11**: 108–114.

Arnold AP. 2012. The end of gonad-centric sex determination in mammals. *Trends Genet* **28**: 55–61.

Bachtrog D. 2008. The temporal dynamics of processes underlying Y chromosome degeneration. *Genetics* **179**: 1513–1525.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* **14**: 113–124.

Banka S, Lederer D, Benoit V, Jenkins E, Howard E, Bunstone S, Kerr B, McKee S, Lloyd IC, Shears D, et al. 2014. Novel KDM6A (UTX) mutations and a clinical and molecular review of the X-linked Kabuki syndrome (KS2). *Clin Genet* **87**: 252–258.

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, Koutseva N, Zaghlul S, Graves T, Rock S, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**: 494–499.

Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T, et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat Genet* **49**: 387–394.

Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, Brown LG, Rozen S, Warren WC, Wilson RK, et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**: 612–616.

Bögershausen N, Gatinois V, Riehmer V, Kayserili H, Becker J, Thoenes M, Simsek-Kiper PÖ, Barat-Houari M, Elcioglu NH, Wieczorek D, et al. 2016. Mutation Update for Kabuki Syndrome Genes KMT2D and KDM6A and Further Delineation of X-Linked Kabuki Syndrome Subtype 2. *Hum Mutat* **37**: 847–864.

Bridges CB. 1916. Non-Disjunction as Proof of the Chromosome Theory of Heredity. *Genetics* **1**: 1–52.

Bull JJ. 1983. *Evolution of Sex Determining Mechanisms*. The Benjamin/Cummings Publishing Company, Menlo Park, CA.

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

Carvalho AB, Clark AG. 2005. Y Chromosome of D. pseudoobscura Is Not Homologous to the Ancestral Drosophila Y. *Science* **307**: 108–110.

Carvalho AB, Dobo BA, Vibranovski MD, Clark AG. 2001. Identification of five new genes on the Y chromosome of Drosophila melanogaster. *Proc Natl Acad Sci* **98**: 13225–13230.

Carvalho AB, Lazzaro BP, Clark AG. 2000. Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. *Proc Natl Acad Sci* **97**: 13239–13244.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Phil Trans R Soc B* **355**: 1563–1572.

Charlesworth D, Charlesworth B. 1980. Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genet Res* **35**: 205–214.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488–493.

Deciphering Developmental Disorders Study. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**: 433–438.

Di Stazio MD, Collesi C, Vozzi D, Liu W, Myers M, Morgan A, Adamo PAD, Girotto G, Rubinato E, Giacca M, et al. 2018. TBL1Y: a new gene involved in syndromic hearing loss. *Eur J Hum Genet* **354**: 466–474.

Disteche CM. 2012. Dosage compensation of the sex chromosomes. *Annu Rev Genet* **46**: 537–560.

Ditton HJ, Zimmer J, Kamp C, Meyts ER-D, Vogt PH. 2004. The AZFa gene DBY (DDX3Y) is widely transcribed but the protein is limited to the male germ cells by translation control. *Hum Mol Genet* **13**: 2333–2341.

Dunford A, Weinstock DM, Savova V, Schumacher SE, Cleary JP, Yoda A, Sullivan TJ, Hess JM, Gimelbrant AA, Beroukhim R, et al. 2017. Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nat Genet* **49**: 10–16.

Fisher RA. 1931. The evolution of dominance. *Biol Rev* **6**: 345–368.

Ford CE, Jones KW, Polani PE, Almeida JC de, Briggs JH. 1959. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* **1**: 711–713.

Foster JW, Graves JA. 1994. An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. *Proc Natl Acad Sci* **91**: 1927–1931.

Fridolfsson AK, Cheng H, Copeland NG, Jenkins NA, Liu HC, Raudsepp T, Woodage T, Chowdhary B, Halverson J, Ellegren H. 1998. Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proc Natl Acad Sci* **95**: 8147–8152.

Gozdecka M, Meduri E, Mazan M, Tzelepis K, Dudek M, Knights AJ, Pardo M, Yu L, Choudhary JS, Metzakopian E, et al. 2018. UTX-mediated enhancer and chromatin remodeling suppresses myeloid leukemogenesis through noncatalytic inverse regulation of ETS and GATA programs. *Nat Genet* **50**: 883–894.

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

Graves JAM. 2006. Sex chromosome specialization and degeneration in mammals. *Cell* **124**: 901–914.

Gueler B, Sonne SB, Zimmer J, Hilscher B, Hilscher W, Græm N, Meyts ER-D, Vogt PH. 2012. AZFa protein DDX3Y is differentially expressed in human male germ cells during development and in testicular tumours: new evidence for phenotypic plasticity of germ cells. *Hum Reprod* **27**: 1547–1555.

Henking H. 1891. Untersuchungen über die ersten Entwicklungsv-organge in den Eiern der Insekten. H. Über Spermatogense und deren Beziehung zur Eientwicklung bei Pyrrhocoris apertus. *Zeitschrift für wissenschaftliche Zoologie* **51**: 685–786.

Hong S, Cho Y-W, Yu L-R, Yu H, Veenstra TD, Ge K. 2007. Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. *Proc Natl Acad Sci* **104**: 18439–18444.

Hughes JF, Rozen S. 2012. Genomics and Genetics of Human and Primate Y Chromosomes. *Annu Rev Genomics Hum Genet* **13**: 83–108.

Innocenti P, Morrow EH. 2010. The sexually antagonistic genes of Drosophila melanogaster. *PLoS Biol* **8**: e1000335. doi:10.1371/journal.pbio.1000335

Jacobs PA, Strong JA. 1959. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* **183**: 302–303.

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

Jensen LR, Amende M, Gurok U, Moser B, Gimmel V, Tzschach A, Janecke AR, Tariverdian G, Chelly J, Fryns J-P, et al. 2005. Mutations in the JARID1C Gene, Which Is Involved in Transcriptional Regulation and Chromatin Remodeling, Cause X-Linked Mental Retardation. *Am J Hum Genet* **76**: 227–236.

Johansson MM, Lundin E, Qian X, Mirzazadeh M, Halvardson J, Darj E, Feuk L, Nilsson M, Jazin E. 2016. Spatial sexual dimorphism of X and Y homolog gene expression in the human central nervous system during early male development. *Biol Sex Differ* **7**: 5. doi:10.1186/s13293-015-0056-4

Johnston CM, Lovell FL, Leongamornlert DA, Stranger BE, Dermitzakis ET, Ross MT. 2008. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet* **4**: e9. doi:10.1371/journal.pgen.0040009

Jost A. 1947. Sur les effets de la castration précoce de l'embryon male de lapin. *C R Seances Soc Biol Fil* **141**: 126–129.

Kaiser VB, Zhou Q, Bachtrog D. 2011. Nonrandom gene loss from the Drosophila miranda neo-Y chromosome. *Genome Biol Evol* **3**: 1329–1337.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502**: 333–339.

Kellaris G, Khan K, Baig SM, Tsai I-C, Zamora FM, Ruggieri P, Natowicz MR, Katsanis N. 2018. A hypomorphic inherited pathogenic variant in DDX3X causes male intellectual disability with additional neurodevelopmental and neurodegenerative features. *Hum Genomics* **12**: 11. doi:10.1186/s40246-018-0141-y

Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. 2014. A draft map of the human proteome. *Nature* **509**: 575–581.

Koopman P, Gubbay J, Collignon J, Lovell-Badge R. 1989. Zfy gene expression patterns are not compatible with a primary role in mouse sex determination. *Nature* **342**: 940–942.

Kosugi M, Otani M, Kikkawa Y, Itakura Y, Sakai K, Ito T, Toyoda M, Sekita Y, Kimura T. 2020. Mutations of histone demethylase genes encoded by X and Y chromosomes, Kdm5c and Kdm5d, lead to noncompaction cardiomyopathy in mice. *Biochem Biophys Res Commun* **525**: 100–106.

Krausz C, Casamonti E. 2017. Spermatogenic failure and the Y chromosome. *Hum Genet* **136**: 637–655.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286**: 964–967.

Lahn BT, Page DC. 1997. Functional coherence of the human Y chromosome. *Science* **278**: 675–680.

Lan F, Bayliss PE, Rinn JL, Whetstine JR, Wang JK, Chen S, Iwase S, Alpatov R, Issaeva I, Canaani E, et al. 2007. A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature* **449**: 689–694.

Lederer D, Grisart B, Digilio MC, Benoit V, Crespin M, Ghariani SC, Maystadt I, Dallapiccola B, Verellen-Dumoulin C. 2011. Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *Am J Hum Genet* **90**: 119–24.

Lee S, Lee JW, Lee S-K. 2012. UTX, a Histone H3-Lysine 27 Demethylase, Acts as a Critical Switch to Activate the Cardiac Developmental Program. *Dev Cell* **22**: 25–37.

Ma K, Inglis JD, Sharkey A, Bickmore WA, Hill RE, Prosser EJ, Speed RM, Thomson EJ, Jobling M, Taylor K. 1993. A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor AZF controlling human spermatogenesis. *Cell* **75**: 1287–1295.

Mahajan S, Bachtrog D. 2017. Convergent evolution of Y chromosome gene content in flies. *Nature communications* **8**: 785.

Malioutov D, Chen T, Airoldi E, Jaffe JD, Budnik B, Slavov N. 2018. Quantifying homologous proteins and proteoforms. *Mol Cell Proteomics* **18**: 162–168.

Manz E, Schnieders F, Brechlin AM, Schmidtke J. 1993. TSPY-Related Sequences Represent a Microheterogeneous Gene Family Organized as Constitutive Elements in DYZ5 Tandem Repeat Units on the Human Y Chromosome. *Genomics* **17**: 726–731.

Mardon G, Luoh SW, Simpson EM, Gill G, Brown LG, Page DC. 1990. Mouse Zfx protein is similar to Zfy-2: each contains an acidic activating domain and 13 zinc fingers. *Mol Cell Biol* **10**: 681–688.

McClung CE. 1902. The accessory chromosome—sex determinant? *Biol Bull* **3**: 43–84.

Meulen JV der, Sanghvi V, Mavrakis K, Durinck K, Fang F, Matthijssens F, Rondou P, Rosen M, Pieters T, Vandenberghe P, et al. 2015. The H3K27me3 demethylase UTX is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. *Blood* **125**: 13–21.

Mitchell MJ, Woods DR, Tucker PK, Opp JS, Bishop CE. 1991. Homology of a candidate spermatogenic gene from the mouse Y chromosome to the ubiquitin-activating enzyme El. *Nature* **354**: 483–486.

Morgan TH. 1910. Sex-limited inheritance in Drosophila. *Science* **32**: 120–122.

Muller HJ. 1914. A gene for the fourth chromosome of Drosophila. *J Exp Zool* **17**: 325–336.

Nanda I, Shan Z, Schartl M, Burt DW, Koehler M, Nothwang H, Grützner F, Paton IR, Windsor D, Dunn I, et al. 1999. 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat Genet* **21**: 258–259.

Naqvi S, Bellott DW, Lin KS, Page DC. 2018. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. *Genome research* **28**: 474–483.

Naqvi S, Godfrey AK, Hughes JF, Goodheart ML, Mitchell RN, Page DC. 2019. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* **365**: eaaw7317. doi:10.1126/science.aaw7317

Nicola P, Blackburn PR, Rasmussen KJ, Bertsch NL, Klee EW, Hasadsri L, Pichurin PN, Rankin J, Raymond FL, Study D, et al. 2019. De novo DDX3X missense variants in males appear viable and contribute to syndromic intellectual disability. *Am J Med Genet A* **179**: 570–578.

Ntziachristos P, Tsirigos A, Welstead GG, Trimarchi T, Bakogianni S, Xu L, Loizou E, Holmfeldt L, Strikoudis A, King B, et al. 2014. Contrasting roles of histone 3 lysine 27 demethylases in acute lymphoblastic leukaemia. *Nature* **514**: 513–517.

Ohno S. 1967. *Sex Chromosomes and Sex-Linked Genes*. Springer-Verlag, Berlin, Germany.

Orr HA, Kim Y. 1998. An adaptive hypothesis for the evolution of the Y chromosome. *Genetics* **150**: 1693–1698.

Painter TS. 1921. The Y-chromosome in mammals. *Science* **53**: 503–504.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Sayres MAW, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* **48**: 593–599.

Prasad NG, Bedhomme S, Day T, Chippindale AK. 2007. An evolutionary cost of separate genders revealed by male-limited evolution. *Am Nat* **169**: 29–37.

Rastegar DA, Tabar MS, Alikhani M, Parsamatin P, Samani FS, Sabbaghian M, Gilani MAS, Ahadi AM, Meybodi AM, Piryaei A, et al. 2015. Isoform-Level Gene Expression Profiles of Human Y

Chromosome Azoospermia Factor Genes and Their X Chromosome Paralogs in the Testicular Tissue of Non-Obstructive Azoospermia Patients. *J Proteome Res* **14**: 3595–3605.

Reijnders MRF, Zachariadis V, Latour B, Jolly L, Mancini GM, Pfundt R, Wu KM, Ravenswaaij-Arts CMA van, Veenstra-Knol HE, Anderlid B-MM, et al. 2016. De Novo Loss-of-Function Mutations in USP9X Cause a Female-Specific Recognizable Syndrome with Developmental Delay and Congenital Malformations. *Am J Hum Genetics* **98**: 373–381.

Reijo R, Lee T-Y, Salo P, Alagappan R, Brown LG, Rosenberg M, Rozen S, Jaffe T, Straus D, Hovatta O, et al. 1995. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA–binding protein gene. *Nat Genet* **10**: 383–393.

Rice WR. 1987. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116**: 161–167.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**: 735–742.

Rice WR. 1992. Sexually antagonistic genes: experimental evidence. *Science* **256**: 1436–1439.

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

Rozen S, Marszalek JD, Alagappan RK, Skaletsky H, Page DC. 2009. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. *Am J Hum Genet* **85**: 923–928.

San Roman AK, Page DC. 2019. A strategic research alliance: Turner syndrome and sex differences. *Am J Med Genet C Semin Med Genet* **181**: 59–67.

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, Rozen S, Dinulos MB, Disteche CM, et al. 1996. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat Genet* **14**: 292–299.

Schmidt J. 1920. The genetic behaviour of a secondary sexual character. *C R Trav Lab Carlsberg* **14**, 1–8.

Schneider-Gädicke A, Beer-Romero P, Brown LG, Nussbaum R, Page DC. 1989. ZFX has a gene structure similar to ZFY, the putative human sex determinant, and escapes X inactivation. *Cell* **57**: 1247–1258.

Sekido R, Lovell-Badge R. 2009. Sex determination and SRY: down to a wink and a nudge? *Trends Genet* **25**: 19–29.

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

Shpargel KB, Sengoku T, Yokoyama S, Magnuson T. 2012. UTX and UTY Demonstrate Histone Demethylase-Independent Function in Mouse Embryonic Development. *PLoS Genet* **8**: e1002964. doi:10.1371/journal.pgen.1002964

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

Snijders Blok LS, Madsen E, Juusola J, Gilissen C, Baralle D, Reijnders MRF, Venselaar H, Helsmoortel C, Cho MT, Hoischen A, et al. 2015. Mutations in DDX3X Are a Common Cause of Unexplained Intellectual Disability with Gender-Specific Effects on Wnt Signaling. *Am J Hum Genet* **97**: 343–352.

Stern C. 1957. The problem of complete Y-linkage in man. *Am J Hum Genet* **9**: 147–166.

Stevens NM. 1905. *Studies in Spermatogenesis with Especial Reference to the "Accessory Chromosome"*. Carnegie Institution, Washington, DC.

Teitz LS, Pyntikova T, Skaletsky H, Page DC. 2018. Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages. *Am J Hum Genet* **103**: 261–275.

Toniolo D, Rizzolio F. 2007. X Chromosome and Ovarian Failure. *Semin Reprod Med* **25**: 264–271.

Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME, Hardy J, Ryten M, North American Brain Expression Consortium. 2013. Widespread sex differences in gene expression and splicing in the adult human brain. *Nat Commun* **4**: 2771. doi:10.1038/ncomms3771

Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, et al. 2017. Landscape of X chromosome inactivation across human tissues. *Nature* **550**: 244–248.

Tyler-Smith C, Krausz C. 2009. The will-o'-the-wisp of genetics--hunting for the azoospermia factor gene. *N Engl J Med* **360**: 925–927.

van Haaften G van, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, Edkins S, Hardy C, O'Meara S, Teague J, et al. 2009. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet* **41**: 521–523.

Vicoso B, Bachtrog D. 2015. Numerous Transitions of Sex Chromosomes in Diptera. *PLoS Biol* **13**: e1002078. doi:10.1371/journal.pbio.1002078

Vogt PH, Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, Köhn FM, Schill WB, Farah S, Ramos C, et al. 1996. Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. *Hum Mol Genet* **5**: 933–943.

Vogt PH, Falcao CL, Hanstein R, Zimmer J. 2008. The AZF proteins. *Int J Androl* **31**: 383–394.

Walport LJ, Hopkinson RJ, Vollmar M, Madden SK, Gileadi C, Oppermann U, Schofield CJ, Johansson C. 2014. Human UTY(KDM6C) is a male-specific Nϵ-methyl lysyl demethylase. *J Biol Chem* **289**: 18302–18313.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* **350**: 1096–1101.

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

Welstead GG, Creyghton MP, Bilodeau S, Cheng AW, Markoulaki S, Young RA, Jaenisch R. 2012. X-linked H3K27me3 demethylase Utx is required for embryonic development in a sex-specific manner. *Proc Natl Acad Sci* **109**: 13004–13009.

White MA, Kitano J, Peichel CL. 2015. Purifying Selection Maintains Dosage-Sensitive Genes during Degeneration of the Threespine Stickleback Y Chromosome. *Mol Biol Evol* **32**: 1981–1995.

Wilson EB. 1905a. Studies on chromosomes. I. The behavior of the idiochromosomes in Hemiptera. *J Exp Zool* **2**: 371–405.

Wilson EB. 1905b. Studies on chromosomes. II. The paired microchromosomes, idiochromosomes and heterotropic chromosomes in hemiptera. *J Exp Zool* **2**: 507–545.

Wilson MA, Makova KD. 2009. Evolution and survival on eutherian sex chromosomes. *PLoS Genet* **5**: e1000568. doi:10.1371/journal.pgen.1000568

Winge Ö. 1923. Crossing-over between the X and the Y chromosome in Lebistes. *J Genet* **13**: 201–217.

Winge Ö. 1922. One-sided masculine and sex-linked inheritance in Lebistes reticulatus. *J Genet* **12**: 145–162.

Winge Ö. 1927. The location of eighteen genes in Lebistes reticulatus. *J Genet* **18**: 1–43.

Wise AL, Gyi L, Manolio TA. 2013. eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses. *Am J Hum Genetics* **92**: 643–647.

Xu J, Burgoyne PS, Arnold AP. 2002. Sex differences in sex chromosome gene expression in mouse brain. *Hum Mol Genet* **11**: 1409–1419.

Xu J, Deng X, Disteche CM. 2008a. Sex-specific expression of the X-linked histone demethylase gene Jarid1c in brain. *PLoS One* **3**: e2553. doi:10.1371/journal.pone.0002553

Xu J, Deng X, Watkins R, Disteche CM. 2008b. Sex-specific differences in expression of histone demethylases Utx and Uty in mouse brain and neurons. *J Neurosci* **28**: 4521–4527.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* **337**: 341–345.

Zhou Q, Zhu H, Huang Q, Zhao L, Zhang G, Roy SW, Vicoso B, Xuan Z, Ruan J, Zhang Y, et al. 2012. Deciphering neo-sex and B chromosome evolution by the draft genome of Drosophila albomicans. *BMC Genomics* **13**: 109. doi:10.1186/1471-2164-13-109

Chapter 2. A Quantitative View of Y-Chromosome Gene Expression across the Human Body

Alexander K. Godfrey, Sahin Naqvi, Lukáš Chmátal, Joel M. Chick, Richard N. Mitchell, Steven P. Gygi, Helen Skaletsky, David C. Page

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.

Acknowledgments: We thank A.K. San Roman for sharing sex-chromosome-aneuploidy cell lines, 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. Bellott, J.F. Hughes, and A.K. San Roman for critical reading of the manuscript. This work was supported by Biogen, Inc., the American Heart Association, the National Institutes of Health, the Whitehead Institute, the Howard Hughes Medical Institute, and generous gifts from Brit and Alexander d'Arbeloff and Arthur W. and Carol Tobin Brill.

Adapted from Godfrey, A.K., Navqi, S., Chmátal, L., Chick, J.M., Mitchell, R.N., Gygi, S.P., Skaletsky, H., Page, D.C. 2020. *Forthcoming.*

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 initiation factor 1A (EIF1A), 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 males and females in any tissue where they are robustly expressed. A different set of considerations pertains to the MSY

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.[a]	Evidence supporting at least partial equivalence[b]	Evidence supporting difference[b,c]
KDM6A/UTY	86%	*Uty* rescues inviability of *Utx*-knockout mice (Lee et al. 2012; Shpargel et al. 2012; Welstead et al. 2012). Concomitant loss of *UTX* and *UTY* in cancer (Gozdecka et al. 2018; van Haaften et al. 2009). UTX and UTY demethylate trimethylated histone 3 lysine 27 *in vitro* (Walport et al. 2014).	Compared to UTX, UTY shows substantially reduced or absent demethylase activity *in vitro* and in cellular assays (Hong et al. 2007; Lan et al. 2007; Shpargel et al. 201 Walport et al. 2014).
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 activi in *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%	-	-

[a] 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 enzymatic activity) or qualitatively distinct protein functions.

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, with studies examining only a few tissue types, while employing small sample sizes or suboptimal methodologies for quantitatively analyzing MSY gene expression. Previous studies have firmly established that MSY genes can be placed into two groups—with some genes showing testis-specific expression and others showing expression in many tissues—but these studies could not detect more subtle quantitative differences in the expression levels of MSY genes 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 have focused on single or subsets of MSY genes in individual tissues, or have studied 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, chrX: 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

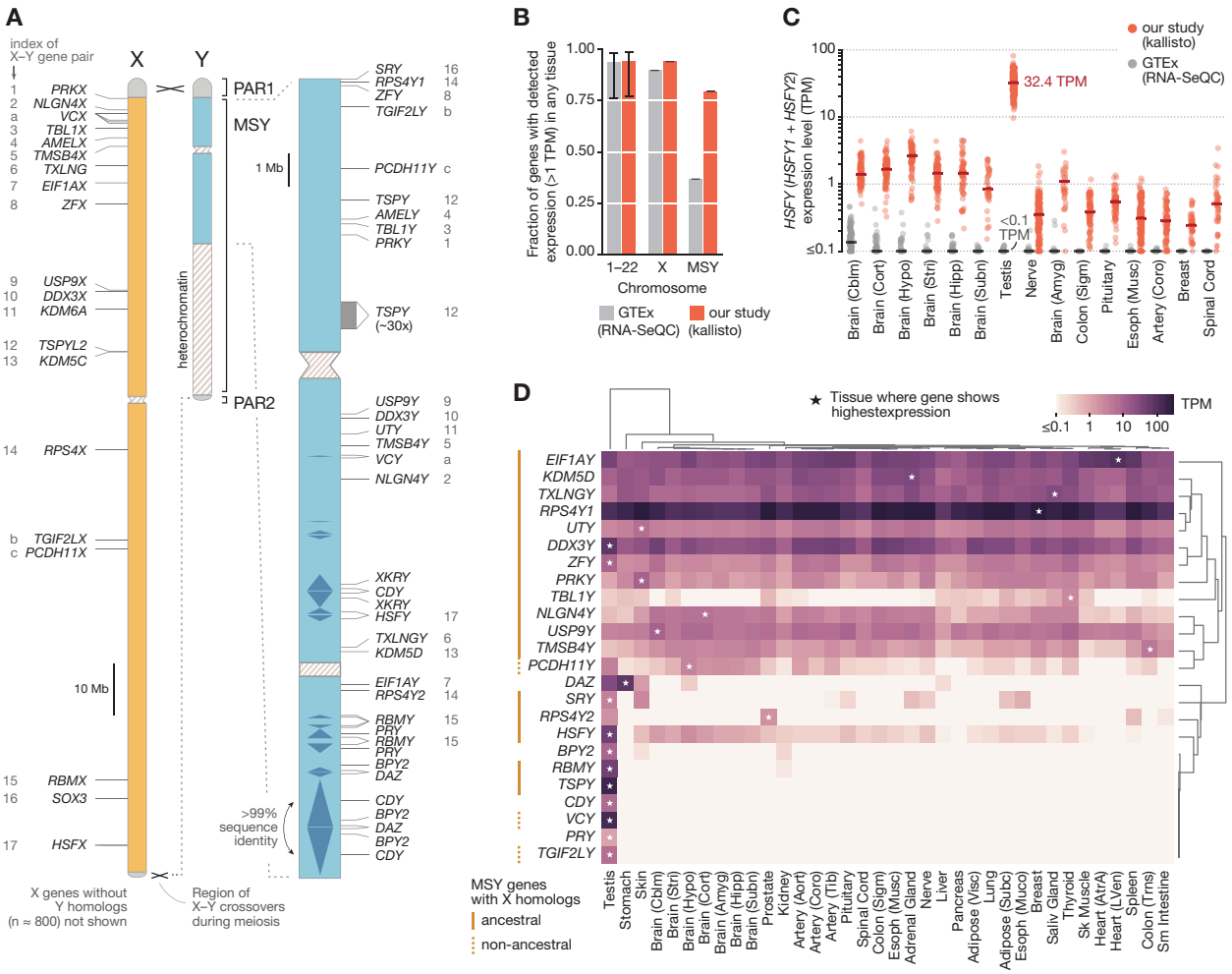


Fig. 1. Estimates of MSY gene expression across 36 human tissues. (**A**) Outside of the two pseudoautosomal regions (PAR1, PAR2), the X and Y chromosomes have diverged in sequence. The locations of protein-coding genes and multi-copy gene families in the male-specific region of the human Y chromosome (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 HSFY 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 in each tissue, with row and column order determined by hierarchical clustering. Stars denote the tissue with the highest expression for a given gene.

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., *HSFY* 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 collectively form five multi-copy gene families (*BPY2*, *CDY*, *DAZ*, *PRY*, *XKRY*). 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. *DAZ* (*Deleted in Azoospermia*), which is best known for its role in spermatogenesis and is generally thought to be expressed exclusively in testes (Vogt et al. 2008), was 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, *DAZ*'s autosomal homolog and progenitor (Saxena et al. 1996), *DAZL*, was not expressed in stomach samples from XY or XX donors (Supplemental Fig. S4B)). *DAZ*'s expression in the stomach proved to be the exception among MSY genes without X-linked homologs. Of the remaining four gene families, one (*XKRY*) was not robustly expressed in any tissue, while three (*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 substantially contribute to differences between XX and XY individuals outside of the reproductive tract.

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 (*TMSB4Y*/*TMSB4X* < 0.01 in all tissues; *TBL1Y*/*TBL1X* < 0.22 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).

Some X–Y gene pairs had higher or lower Y/X expression ratios than others (Friedman test, $P = 1 \times 10^{-28}$), but 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).

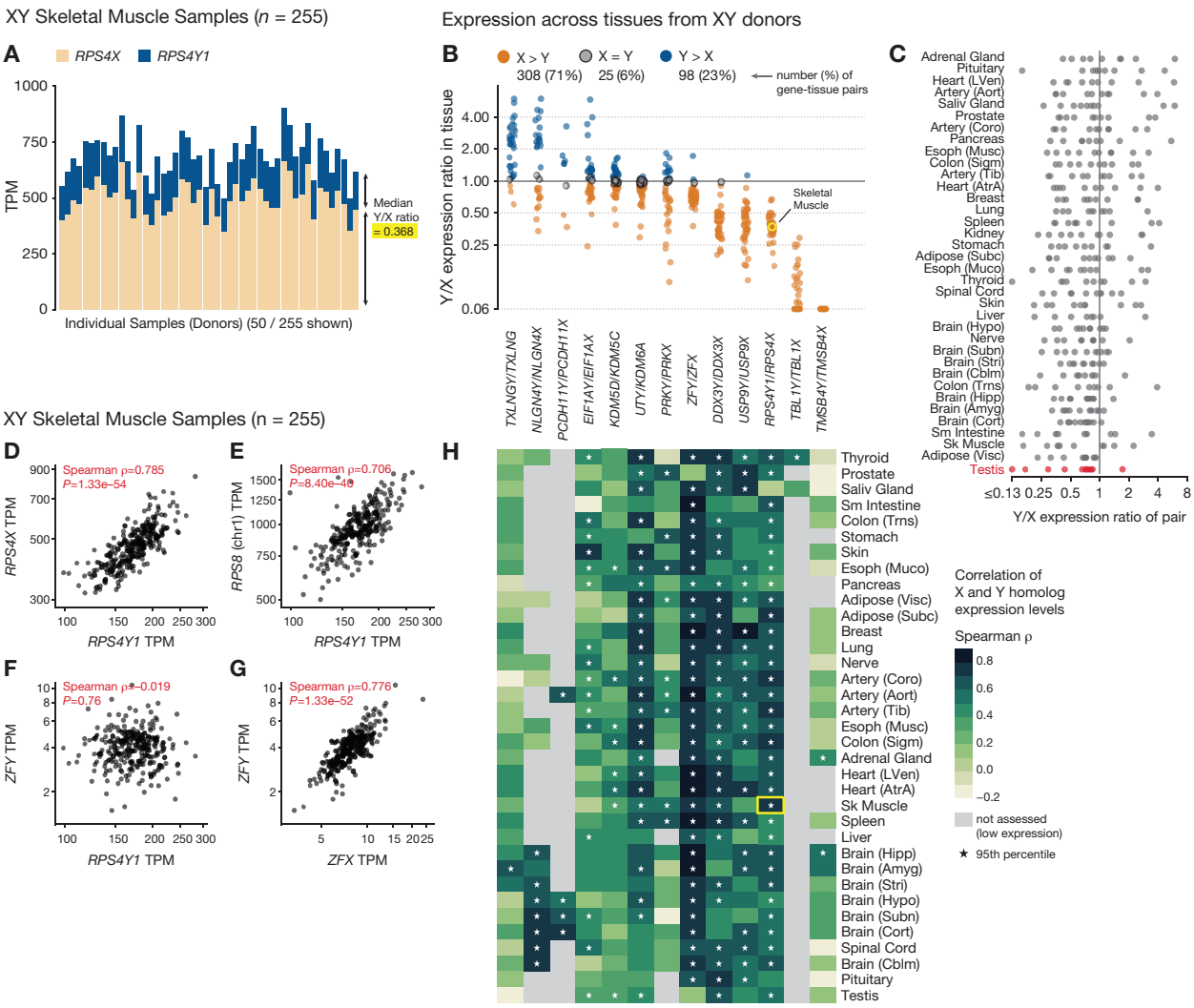


Fig. 2. Quantitative comparison of X- and Y-homolog expression in XY individuals. (**A**) Expression of *RPS4Y1* (blue) and its X-linked homolog *RPS4X* (tan) in individual skeletal muscle samples (50 of 255 total samples are shown). (**B** – **C**) Each point shows Y/X expression ratio for one widely 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 of data in **A**. (**D** – **G**) Each point shows the co-expression of an MSY gene with another gene in a single skeletal muscle sample: *RPS4Y1* vs. *RPS4X* (**D**), *RPS8* (**E**), *ZFY* (**F**); *ZFY* vs. *ZFX* (**G**). (**H**) Each cell shows the correlation coefficient (Spearman ρ) 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**.

Although MSY genes and their corresponding X homologs often differ in expression level, we wondered if their expression 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

expression between an MSY gene and its X homolog. Indeed, we found that the X- and Y-linked 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 precisely determined in accordance with molecular function rather than chromosomal location. Remarkably, MSY genes that were typically expressed at only 30–50% of the levels of their X homologs (e.g., *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.

Evolutionary loss of a microRNA target site promoted elevated *EIF1AY* expression in the heart

Although MSY genes and their corresponding X homologs often showed correlated expression, implying co-regulation, we also found evidence of tissue-specific divergence in regulation. Individual X–Y pairs showed Y/X expression ratios in some tissues that were substantially higher or lower than their ratios in other tissues (e.g., *USP9Y*/*USP9X* = 1.1 in the pituitary, compared to 0.2 – 0.6 in most tissues) (Fig. 2B). We hypothesized that one member of the X–Y gene pair, but not the other, might be up- or downregulated in these tissues. To explore this possibility, for

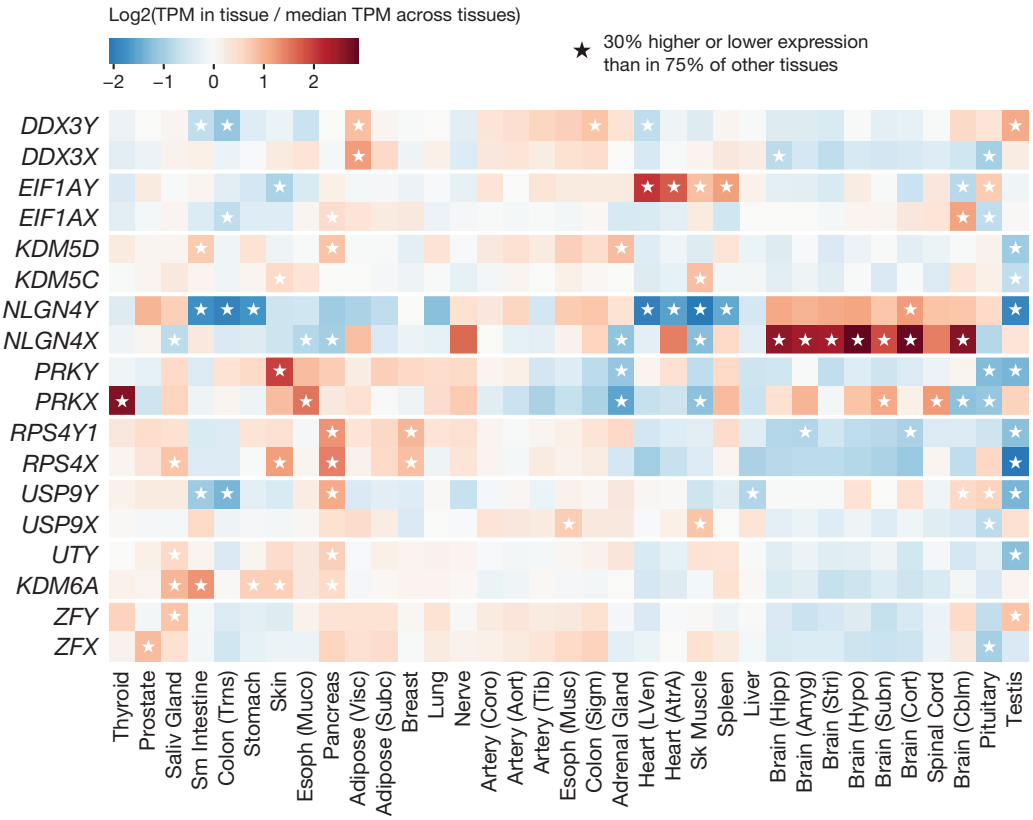


Fig. 3. Tissue-specific up- and downregulated of X and Y homologs. Each cell shows a gene's expression level in a tissue relative to its median expression level across all tissues. Stars denote tissues where a gene's expression level is significantly higher or lower (±30%; Welch's *t*-test, $P < 0.001$)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

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 up- or downregulated 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 particularly striking 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 eukaryotic translation initiation factor 1A (EIF1A), one of 27 primary factors used to initiate protein synthesis in all eukaryotic linages (Hinnebusch 2014), and the only such factor encoded on both X and Y chromosomes in primates (Bellott et al. 2014). The proteins encoded by

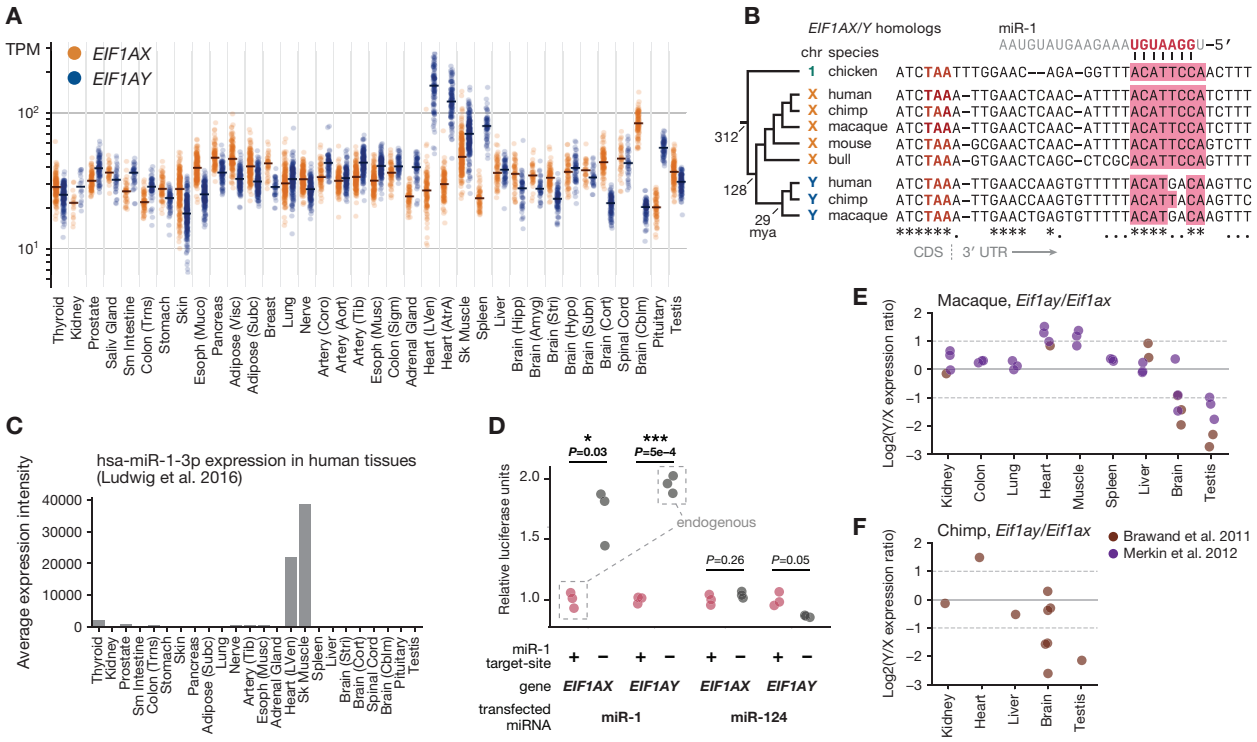


Fig. 4. Y-specific loss of miR-1 target site led to elevated *EIF1AY* expression in XY heart and tissue. (**A**) Each point shows expression level of *EIF1AY* (blue) or *EIF1AX* (gold) in a single tissue sample from an XY individual. Lines show median expression level. (**B**) Alignment of 3′-UTRs of *EIF1AY*, *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 *EIF1AX* or *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**).

EIF1AX and *EIF1AY* are likely to be functionally equivalent: they differ by only a single amino acid, a conservative leucine (EIF1AX)-to-methionine (EIF1AY) substitution at a position outside of EIF1A'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 upregulated 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 critical positions 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). However, *EIF1AX* orthologs in mammals and *EIF1AX*/*Y* orthologs on non-mammalian autosomes are not upregulated in the heart, suggesting this expression pattern was acquired specifically by primate *EIF1AY* (Supplemental Fig. S13). Together, these observations suggest that two nucleotide substitutions within an *EIF1AY* regulatory element contributed to tissue-specific upregulation of *EIF1AY*.

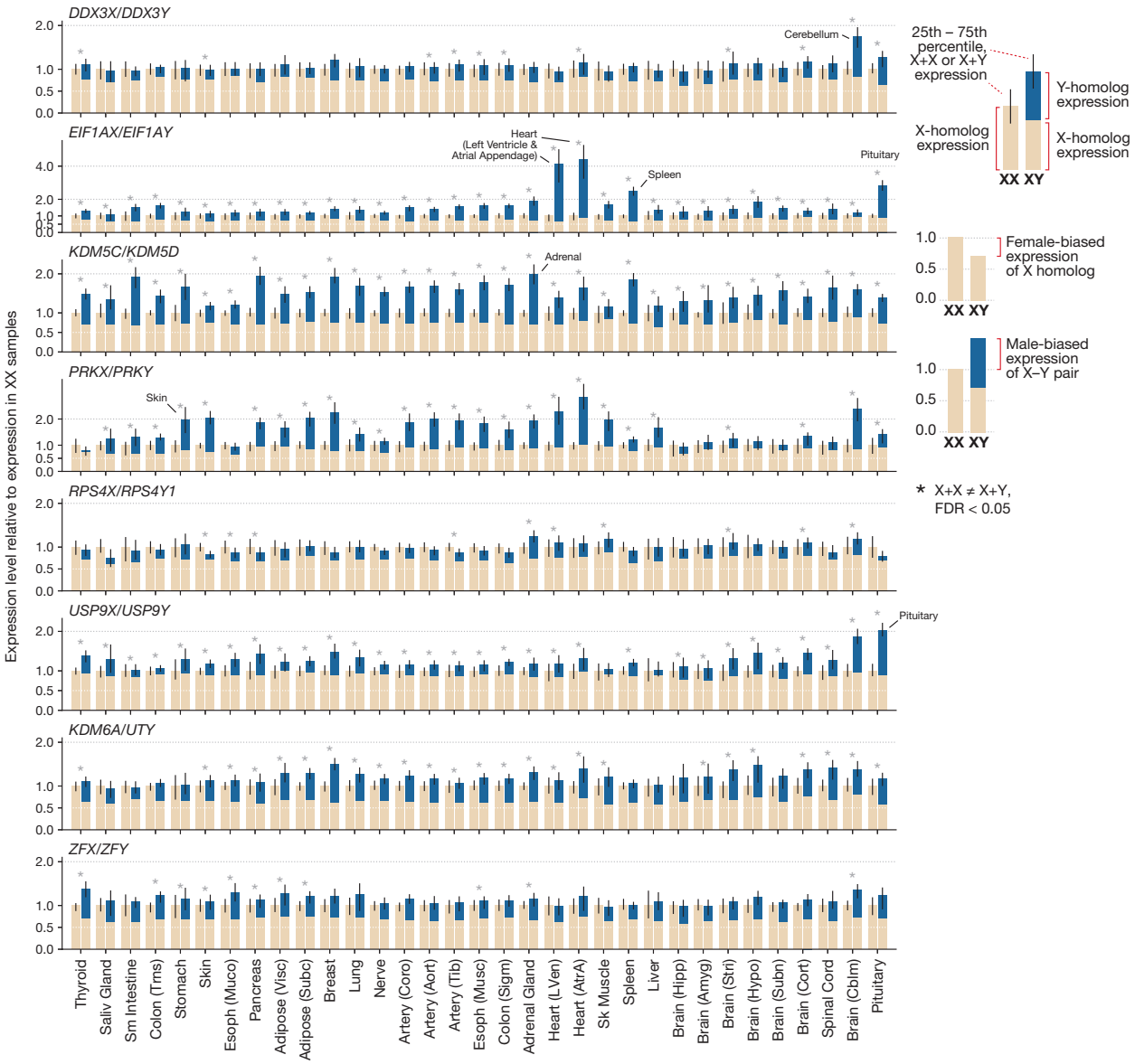


Fig. 5. Tissue-specific, male-biased expression of X–Y gene pairs at the transcriptional level. Each pair of bars shows the median expression level of an X–Y gene pair in XX (left) and XY (right) samples from one tissue. Expression is normalized to the level in XX samples. In XY samples, the sum of X-homolog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

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- (female-) biased expression—i.e.,

higher expression in tissue samples from XX individuals than in the same tissue type from XY individuals (Fig. 5). This female-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 bias towards higher XX expression was less than 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 male-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 upregulated Y-homolog expression, the male-biased expression was more prominent. For example, *KDM5D* showed upregulated expression in the adrenal gland (Fig. 3), leading to 2.1-fold male-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 male-biased expression of *USP9X/Y* (Fig. 3; Fig. 5). Most strikingly, upregulated *EIF1AY* expression in the heart leads to 5.2-fold higher expression of *EIF1AX*/*Y* in male 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;

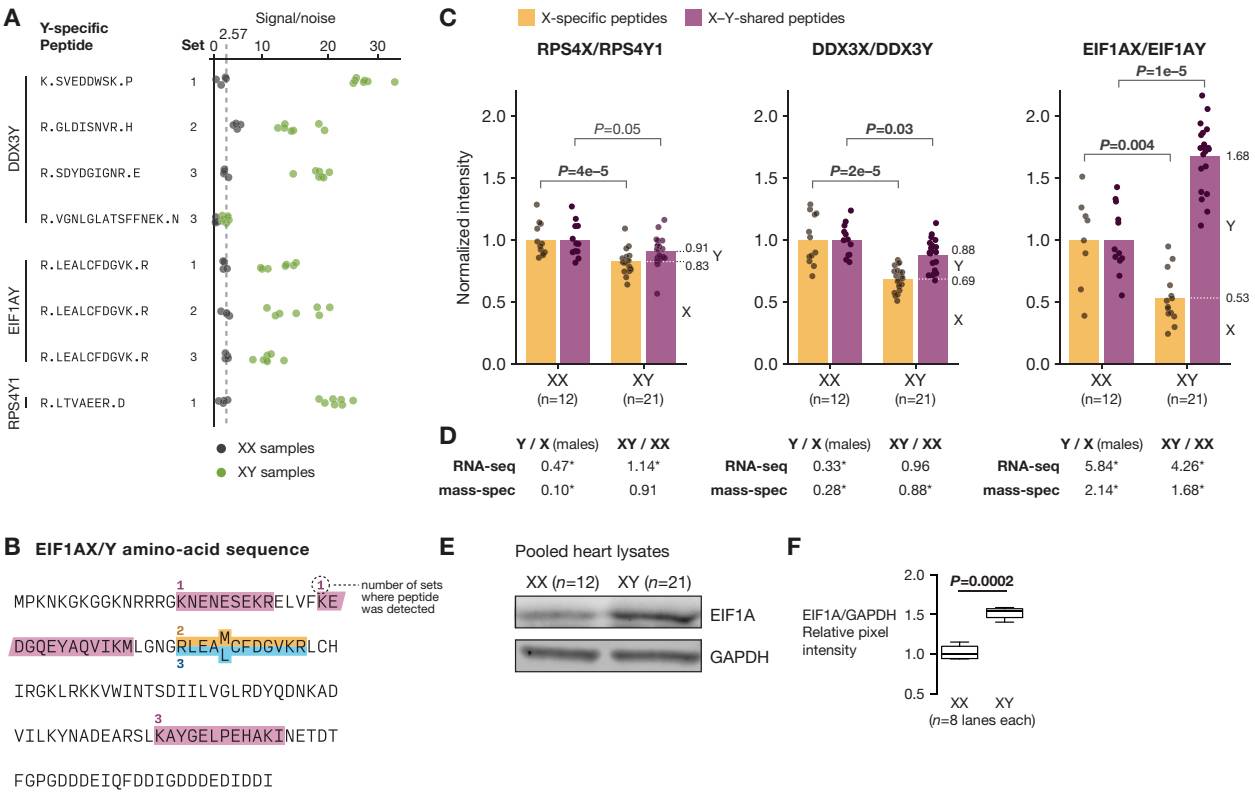


Fig. 6. Male-biased expression of EIF1A protein in the heart. (**A**) Signal/noise values for Y-specific peptides in XX (gray) and XY (green) samples. "Set" refers to the 11-plex experiment (out of three total) in which the peptide was detected. The dotted line shows average signal/noise value in XX 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.

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 protein-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 female-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 males were slightly below the levels of the X isoforms in females on average (RPS4X/Y: mean male/female ratio = 0.91, $P = 0.05$ by permutation; DDX3X/Y: mean male/female ratio = 0.88, $P = 0.03$ by permutation) (Fig. 6C), albeit at only nominally statistically significant levels, suggesting RPS4Y1 and DDX3Y mostly if not entirely compensate for the female-biased expression of RPS4X and DDX3X. The combined

expression of EIF1AX/Y, however, showed a 1.7-fold male bias ($P < 10^{-6}$, permutation), indicating that EIF1AY overcompensates for the female-biased expression of EIF1AX. These estimates further imply that EIF1AY protein is 2.1-fold more abundant than EIF1AX in male heart tissue. Using an antibody that recognizes both EIF1AX and EIF1AY (Supplemental Fig. S19), we corroborated EIF1AX/Y's male-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 tract? 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—possibly a heart-specific transcription factor—might additionally contribute to *EIF1AY*'s ~5-fold higher expression than *EIF1AX* in the heart, as miRNAs typically repress their targets by less than twofold (Baek et al. 2008) (Fig. 4D).

An intriguing speculation is that the male-biased expression of EIF1A 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), EIF1A 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 (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 EIF1A are beneficial, harmful, or neutral in consequence, *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. That such coordinated X–Y co-

expression persists even when the MSY gene is expressed at much lower levels than its X-linked homolog implies that small deviations from the optimal expression level would impair fitness.

We have provided direct evidence that the proteins encoded by MSY genes are present in human heart tissue. We found that the expression levels of EIF1AY and RPS4Y1 proteins were ~3- and 5-fold lower than their transcript expression levels, when measured against their X homologs. It is possible that *EIF1AY* and *RPS4Y1* 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 (Fig. 5): over-expression of the Y homolog at the transcript level might be needed to achieve the requisite level of protein expression. However, we caution against extrapolating these results to other MSY genes and other tissues until many more protein-level measurements are made.

Our detection of DDX3Y protein in the heart conflicts with the conclusion of a previous study 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 expression 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), or that brain and kidney might not be representative. Indeed, DDX3Y has since been detected by Western blot in a neuronal cell line (Vakilian et al. 2015). *DDX3Y* was also identified as essential gene in a leukemia cell line through a genome-wide, unbiased screen (Wang et al. 2015), implying the protein was present. 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 the MSY gene.

Our analyses further 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 undoubtedly 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 all sets of genes across the genome that retain substantial homology.

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), scipy (v1.3.1), pandas (v0.25.1), scikit-learn (v0.21.3), and statsmodels (v0.10.1). Plots were generated using functions from matplotlib (v3.1.1) and seaborn (v0.9.0). Code and Jupyter notebooks for recreating these analyses are available on GitHub (https://github.com/akg8/MSY-expression).

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, comprising the union of transcripts that (1) belong to the "GENCODE Basic" annotation and (2) are recognized by the Consensus Coding Sequence project (Pruitt et al. 2009). Filtering the comprehensive annotation in this manner enriches for full-length, manually curated transcripts defined by two distinct sources. The list of protein-coding human MSY genes analyzed in this study is 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 in these simulated libraries: (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). 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 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

$$s_i = \underset{g \in G_C}{\text{median}} \frac{y_{gi}}{\left(\prod_{j=1}^{n} y_{gj}\right)^{\frac{1}{n}}}$$

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 (Y-homolog TPM + 0.5) / (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, raw RNA-seq data was obtained from the Human Protein Atlas (HPA) Project (Uhlén et al. 2015). We estimated the expression levels of genes in these samples using kallisto with sequence bias correction and our subsetted GENCODE annotation. 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 ≥5 TPM. The expression levels from each sample were first normalized by the housekeeping method described above and transformed to log2(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 × 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 by the Spearman correlation coefficient, ρ_{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 log2(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 *EIF1AX*/*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 *EIF1AX*/*EIF1AY* homologs in male chimpanzee (*Pan troglodytes*), rhesus macaque (*Macaca mulatta*), mouse (*Mus musculus*), and

chicken (*Gallus gallus*) 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). Multiplexed quantitative proteomic analysis was performed as previously described (Chick et al. 2016) and as detailed in Supplemental Methods. Three TMT 11-plex reactions were performed. 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/EIF1AY protein was detected with an EIF1A primary antibody (Abcam Ab177939, anti-rabbit), with GAPDH (Ambion AM4300, anti-mouse) as a loading control. EIF1A levels were quantified using the Odyssey CLx Imaging System (LI-COR). Four technical replicates were performed per sex. To verify that the EIF1A 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,XYY; 48,XXXY; 49,XXXXY; 49,XYYYY) 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 deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al. 2018) with the dataset identifier PXD017055. Processed data (re-estimated TPM matrices) are available at Zenodo (DOI: 10.5281/zenodo.3627233).

REFERENCES

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**: e05005. doi:10.7554/eLife.05005.001

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197–212.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106. doi:10.1186/gb-2010-11-10-r106

Arnold AP. 2012. The end of gonad-centric sex determination in mammals. *Trends Genet* **28**: 55–61.

Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. 2008. The impact of microRNAs on protein output. *Nature* **455**: 64–71.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* **173**: 20–51.

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, Koutseva N, Zaghlul S, Graves T, Rock S, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**: 494–499.

Berletch JB, Ma W, Yang F, Shendure J, Noble WS, Disteche CM, Deng X. 2015. Escape from X Inactivation Varies in Mouse Tissues. *PLoS Genet* **11**: e1005079. doi:10.1371/journal.pgen.1005079

Bol GM, Xie M, Raman V. 2015. DDX3, a potential target for cancer treatment. *Mol Cancer* **14**: 188.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Cannon-Albright LA, Farnham JM, Bailey M, Albright FS, Teerlink CC, Agarwal N, Stephenson RA, Thomas A. 2014. Identification of specific Y chromosomes associated with increased prostate cancer risk: Y Chromosome and Prostate Cancer Risk. *Prostate* **74**: 991–998.

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

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**: 25–36.

Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. 2016. Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**: 500–505.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488–493.

Cotton AM, Ge B, Light N, Adoue V, Pastinen T, Brown CJ. 2013. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol* **14**: R122. doi: 10.1186/gb-2013-14-11-r122

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367–1372.

Deng X, Berletch JB, Nguyen DK, Disteche CM. 2014. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet* **15**: 367–378.

Di Stazio M, Collesi C, Vozzi D, Liu W, Myers M, Morgan A, D'Adamo PA, Girotto G, Rubinato E, Giacca M, et al. 2018. TBL1Y: a new gene involved in syndromic hearing loss. *Eur J Hum Genet* **27**: 466–474.

Ditton HJ, Zimmer J, Kamp C, Meyts ER-D, Vogt PH. 2004. The AZFa gene DBY (DDX3Y) is widely transcribed but the protein is limited to the male germ cells by translation control. *Hum Mol Genet* **13**: 2333–2341.

Eales JM, Maan AA, Xu X, Michoel T, Hallast P, Batini C, Zadik D, Prestes PR, Molina E, Denniff M, et al. 2019. Human Y Chromosome Exerts Pleiotropic Effects on Susceptibility to Atherosclerosis. *Arterioscler Thromb Vasc Biol* **39**: 2386–2401.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**: 92–105.

Gozdecka M, Meduri E, Mazan M, Tzelepis K, Dudek M, Knights AJ, Pardo M, Yu L, Choudhary JS, Metzakopian E, et al. 2018. UTX-mediated enhancer and chromatin remodeling suppresses myeloid leukemogenesis through noncatalytic inverse regulation of ETS and GATA programs. *Nat Genet* **50**: 883–894.

Gremel G, Wanders A, Cedernaes J, Fagerberg L, Hallström B, Edlund K, Sjöstedt E, Uhlén M, Pontén F. 2015. The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling. *J Gastroenterol* **50**: 46–57.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213.

Hinnebusch AG. 2014. The Scanning Mechanism of Eukaryotic Translation Initiation. *Annu Rev Biochem* **83**: 779–812.

Hong S, Cho Y-W, Yu L-R, Yu H, Veenstra TD, Ge K. 2007. Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. *Proc Natl Acad Sci* **104**: 18439–18444.

Iwase S, Lan F, Bayliss P, Torre-Ubieta L de la, Huarte M, Qi HH, Whetstine JR, Bonni A, Roberts TM, Shi Y. 2007. The X-Linked Mental Retardation Gene SMCX/JARID1C Defines a Family of Histone H3 Lysine 4 Demethylases. *Cell* **128**: 1077–1088.

Johansson MM, Lundin E, Qian X, Mirzazadeh M, Halvardson J, Darj E, Feuk L, Nilsson M, Jazin E. 2016. Spatial sexual dimorphism of X and Y homolog gene expression in the human central nervous system during early male development. *Biol Sex Differ* **7**: 5. doi:10.1186/s13293-015-0056-4

Johnston CM, Lovell FL, Leongamornlert DA, Stranger BE, Dermitzakis ET, Ross MT. 2008. Large-Scale Population Study of Human Cell Lines Indicates that Dosage Compensation Is Virtually Complete. *PLoS Genet* **4**: e9. doi:10.1371/journal.pgen.0040009

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36. doi:10.1186/gb-2013-14-4-r36

Kosugi M, Otani M, Kikkawa Y, Itakura Y, Sakai K, Ito T, Toyoda M, Sekita Y, Kimura T. 2020. Mutations of histone demethylase genes encoded by X and Y chromosomes, Kdm5c and Kdm5d, lead to noncompaction cardiomyopathy in mice. *Biochem Biophys Res Commun* **525**: 100–106.

Lahn BT, Page DC. 1999a. Four Evolutionary Strata on the Human X Chromosome. *Science* **286**: 964–967.

Lahn BT, Page DC. 1997. Functional Coherence of the Human Y Chromosome. *Science* **278**: 675–680.

Lahn BT, Page DC. 1999b. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat Genet* **21**: 429–433.

Lan F, Bayliss PE, Rinn JL, Whetstine JR, Wang JK, Chen S, Iwase S, Alpatov R, Issaeva I, Canaani E, et al. 2007. A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature* **449**: 689–694.

Lee S, Lee JW, Lee S-K. 2012. UTX, a Histone H3-Lysine 27 Demethylase, Acts as a Critical Switch to Activate the Cardiac Developmental Program. *Dev Cell* **22**: 25–37.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323. doi:10.1186/1471-2105-12-323

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**: 769–773.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci* **102**: 10557–10562.

Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, Rheinheimer S, Meder B, Stähler C, Meese E, et al. 2016. Distribution of miRNA expression across human tissues. *Nucleic Acids Res* **44**: 3865–3877.

Malioutov D, Chen T, Jaffe J, Airoldi E, Carr S, Budnik B, Slavov N. 2018. Quantifying Homologous Proteins and Proteoforms. *Mol Cell Proteomics* **18**: 162–168.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**: 1593–1599.

Naqvi S, Bellott DW, Lin KS, Page DC. 2018. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. *Genome Res* **28**: 474–483.

Naqvi S, Godfrey AK, Hughes JF, Goodheart ML, Mitchell RN, Page DC. 2019. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* **365**: eaaw7317. doi:10.1126/science.aaw7317

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419.

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, et al. 2018. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **47**: D442–D450.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, et al. 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**: 1316–1323.

Regitz-Zagrosek V, Oertelt-Prigione S, Seeland U, Hetzer R. 2010. Sex and Gender Differences in Myocardial Hypertrophy and Heart Failure. *Circ J* **74**: 1265–1273.

Robert C, Watson M. 2015. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol* **16**: 177. doi:10.1186/s13059-015-0734-x

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

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, Rozen S, Dinulos MB, Disteche CM, et al. 1996. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat Genet* **14**: 292–299.

Sehrawat U, Koning F, Ashkenazi S, Stelzer G, Leshkowitz D, Dikstein R. 2018. Cancer-Associated Eukaryotic Translation Initiation Factor 1A Mutants Impair Rps3 and Rps10 Binding and Enhance Scanning of Cell Cycle Genes. *Mol Cell Biol* **39**: e00441-18.

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

Shpargel KB, Sengoku T, Yokoyama S, Magnuson T. 2012. UTX and UTY Demonstrate Histone Demethylase-Independent Function in Mouse Embryonic Development. *PLoS Genet* **8**: e1002964. doi:10.1371/journal.pgen.1002964

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

Snijders Blok L, Madsen E, Juusola J, Gilissen C, Baralle D, Reijnders MRF, Venselaar H, Helsmoortel C, Cho MT, Hoischen A, et al. 2015. Mutations in DDX3X Are a Common Cause of Unexplained Intellectual Disability with Gender-Specific Effects on Wnt Signaling. *Am J Hum Genet* **97**: 343–352.

Taggart JC, Li G-W. 2018. Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes. *Cell Syst* **7**: 580-589.e4. doi:10.1016/j.cels.2018.11.003

Tartaglia NR, Ayari N, Hutaff-Lee C, Boada R. 2012. Attention-Deficit Hyperactivity Disorder Symptoms in Children and Adolescents with Sex Chromosome Aneuploidy: XXY, XXX, XYY, and XXYY. *J Dev Behav Pediatr* **33**: 309–318.

Towbin JA, Lowe AM, Colan SD, Sleeper LA, Orav EJ, Clunie S, Messere J, Cox GF, Lurie PR, Hsu D, et al. 2006. Incidence, Causes, and Outcomes of Dilated Cardiomyopathy in Children. *JAMA* **296**: 1867–1876.

Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME, Hardy J, Ryten M, Consortium NABE. 2013. Widespread sex differences in gene expression and splicing in the adult human brain. *Nat Commun* **4**: 2771. doi:10.1038/ncomms3771

Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, et al. 2017. Landscape of X chromosome inactivation across human tissues. *Nature* **550**: 244–248.

Uhlén M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. *Science* **347**: 1260419. doi:10.1126/science.1260419

Valiente-Echeverría F, Hermoso MA, Soto-Rifo R. 2015. RNA helicase DDX3: at the crossroad of viral replication and antiviral immunity: DDX3 in viral replication and immunity. *Rev Med Virol* **25**: 286–299.

van Haaften G, Dalgliesh GL, Davies H, Chen L, Bignell G, Greenman C, Edkins S, Hardy C, O'Meara S, Teague J, et al. 2009. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet* **41**: 521–523.

van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, Kirchner M, Maatz H, Blachut S, Sandmann C-L, et al. 2019. The Translational Landscape of the Human Heart. *Cell* **178**: 242-260.

Vogt PH, Falcao CL, Hanstein R, Zimmer J. 2008. The AZF proteins. *Int J Androl* **31**: 383–394.

Walport LJ, Hopkinson RJ, Vollmar M, Madden SK, Gileadi C, Oppermann U, Schofield CJ, Johansson C. 2014. Human UTY(KDM6C) Is a Male-specific Nϵ-Methyl Lysyl Demethylase. *J Biol Chem* **289**: 18302–18313.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* **350**: 1096–1101.

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

Welstead GG, Creyghton MP, Bilodeau S, Cheng AW, Markoulaki S, Young RA, Jaenisch R. 2012. X-linked H3K27me3 demethylase Utx is required for embryonic development in a sex-specific manner. *Proc Natl Acad Sci* **109**: 13004–13009.

Wizemann TM and Pardue M. 2001. *Exploring the Biological Contributions to Human Health: Does Sex Matter?* National Academies Press, Washington, DC.

Xu J, Burgoyne PS, Arnold AP. 2002. Sex differences in sex chromosome gene expression in mouse brain. *Hum Mol Genet* **11**: 1409–1419.

Xu J, Deng X, Disteche CM. 2008a. Sex-Specific Expression of the X-Linked Histone Demethylase Gene Jarid1c in Brain. *PLoS One* **3**: e2553. doi:10.1371/journal.pone.0002553

Xu J, Deng X, Watkins R, Disteche CM. 2008b. Sex-Specific Differences in Expression of Histone Demethylases Utx and Uty in Mouse Brain and Neurons. *J Neurosci* **28**: 4521–4527.

SUPPLEMENTAL METHODS

Human transcriptome annotation and MSY genes

The list of protein-coding human MSY genes analyzed in this study is based on our annotation of the male-specific region of the human Y chromosome (Skaletsky et al. 2003) and is delineated in Supplementary file 1. We note the following differences between the set of protein-coding genes analyzed here from the set of protein-coding described in the GENCODE v24 annotation. First, we excluded eight clone-based Ensembl genes (nomenclature: AC######.#), which do not have official HGNC symbols and were either removed or re-classified as pseudogenes in subsequent versions of the GENCODE annotation. Second, we excluded *PRORY* from our analyses, which was not part of our original annotation (Skaletsky et al. 2003). Its cDNA sequence returns no matches by BLAST in the NCBI EST database, and we could not find compelling evidence of its transcription (e.g., RNA-seq reads that span exon-exon junctions) in the GTEx samples. Third, we included *PRKY* and *TXLNGY* (*CYorf15A/B*) among the list of protein-coding genes. Both are currently listed as pseudogenes in the public annotation, as the result of significant structural differences with their X-linked homologs: *PRKY* lost an exon near the 3′ end of its coding region, creating a premature termination codon and making it a candidate for nonsense mediated decay; *TXLNGY* comprises two transcription units (*CYorf15A* and *CYorf15B*), homologous to the 5′ and 3′ ends of X-linked *TXLNG*. However, both Y-linked genes retain significant open reading frames, and comparisons of their sequences on the human and rhesus macaque Y chromosomes suggest that the coding sequences of *PRKY* and *CYorf15A* remain under purifying selection (dN/dS < 1) (Hughes et al. 2012).

RNA-seq simulations

Simulated RNA-Seq libraries were generated using RSEM (v1.2.22) (Li and Dewey 2011). Sequencing parameters for the simulation were obtained by running rsem-calculate-expression with option --estimate-rpsd on GTEx testis sample GTEX-P4QS-2126-SM-3NMCF, supplying our modified transcriptome annotation as a reference. The output file of expression-level estimates ("isoforms.results") was then modified to set the expression levels of Y-chromosome genes and their X-linked homologs to predetermined levels, following one of four scenarios:

(1) MSY genes/gene families = 0 TPM; X-linked homologs unmodified (i.e., kept at levels estimated in GTEX-P4QS-2126-SM-3NMCF
(2) MSY genes/gene families = 1 TPM, X-linked homologs = 2 TPM
(3) MSY genes/gene families = 5 TPM, X-linked homologs = 10 TPM
(4) MSY genes/gene families = 5 TPM, X-linked homologs set to a random value between 0 and 10 TPM

For genes with multiple transcript isoforms, the relative abundance of each isoform was assigned in proportion to a random number drawn from a heavy-tailed distribution (Pareto with tail index $\alpha = 0.5$). The relative abundance of individual members of multi-copy gene families were determined similarly, such that the summed expression level of genes in the family equaled 1, 2,

5, or 10 TPM as indicated. These relative isoform abundances were drawn anew in each simulation, to sample different configurations of alternative-isoform expression. 50 simulated RNA-seq libraries were generated for each of the four expression-level scenarios using rsem-simulate-reads, with 50 million 76bp paired-end reads in each library (median depth of samples in GTEx is ~78 million reads). Because of the random read-sampling process used in the simulations, the observed expression level for gene *g* in simulated library *j* (as given by the output of rsem-simulate-reads) deviated slightly from its idealized value (e.g., 1, 2, 5, or 10 TPM). To correct for this source of error when plotting results in Supplemental Fig. S2, the estimated expression levels obtained by various methods for gene *g* in simulated library *j* were multiplied by the ratio of idealized-to-observed values for that gene in that library.

Three methods were used to estimate the expression levels of Y-chromosome genes and their X-linked homologs in these simulated libraries. First, simulated reads were aligned to the genome using tophat2 (Kim et al. 2013) (v2.1.1; using parameters --no-mixed --no-discordant) and the number of uniquely mapping reads overlapping the exons of each gene was counted with featureCounts (Liao et al. 2014) (v1.6.2), requiring both reads from each fragment to be mapped (-B) and each read to entirely overlap annotated exons (--fracOverlap 1). Read-count values for each gene were converted to TPM units, with the length of each gene given by the total length of the union of its exons. This procedure ("unique reads" in Supplemental Fig. S2) is similar to that followed by the GTEx Consortium. Second, after aligning simulated reads with tophat, Cufflinks (Trapnell et al. 2010) (v2.2.1) was used to estimate the expression levels of annotated transcripts (-G) in "multi-read-correct" mode (-u), and estimates in FPKM units were converted to TPM. Third, simulated reads were input to kallisto (Bray et al. 2016) (v0.42.5), and the expression levels of annotated transcripts were estimated with sequence-bias correction (--bias).

On average, the "unique reads" method over-estimated MSY gene expression levels and produced less precise (i.e., more variable) estimates than kallisto or Cufflinks in simulated datasets. The over-estimated expression levels are likely the result of discarding multi-mapping reads—with fewer mapped reads in each library, each gene receives a larger proportion of reads in the library overall, thus inflating the TPM value (TPM units convey a gene's expression level as a fraction of the total expression). The imprecision of the unique-reads method is likely due to the presence of alternative transcript isoforms. The "unique reads" method calculates the density of reads mapping to each gene using a fixed length for that gene, defined by the concatenation of all constitutive and alternative exons. When only a short alternative isoform is expressed, read density for the gene (and, correspondingly, its expression level) will be underestimated. The variability in the estimates produced by the unique-reads method thus likely reflects the random mixture of alternative isoforms present in any simulated library.

Quality control of GTEx RNA-seq analysis

Initial screening of samples using sample- and donor-level metadata

All RNA-seq samples meeting the following criteria were downloaded for initial consideration: (1) RIN (SMRIN) ≥ 6.0; (2) not annotated as "severely" autolyzed (SMATSSCR != 3); (3) donor deemed eligible by GTEx (INCEXC == True); (4) not flagged for removal by GTEx (SMTORMVE !=

'FLAGGED'); (5) was generated from a primary, solid tissue (i.e., whole-blood samples and cell-line samples were excluded); (6) was generated from a tissue where at least 10 samples from male donors were available.

Identification and removal of gene-expression outlier samples

Gene-expression outlier samples were detected and removed following a procedure similar to that used in Wright et al. (Wright et al. 2014). Separately for each tissue, the pairwise Pearson correlation coefficient, r_{ij}, was calculated between the log2(TPM + 0.1) expression levels of samples i and j, using only genes with a median expression level ≥3 TPM among the samples of that tissue. The median similarity between sample i and other samples from that tissue was calculated as $\overline{r_i} = \text{median}_j(r_{ij})$ and re-expressed in median absolute deviations, $D_i = |\overline{r_i} - \bar{\bar{r}}| / \text{median}_j(|\overline{r_j} - \bar{\bar{r}}|)$, where $\bar{\bar{r}} = \text{median}_j(\overline{r_j})$. Samples with $D_i > 6$ were marked as outliers and removed from subsequent analyses. This process was repeated iteratively until no such samples remained. After outlier removal, hierarchical clustering was performed on the remaining samples to confirm the efficacy of this approach. Even after outlier removal, we found that the breast samples clustered into two highly dissimilar clusters. Breast samples from one of these clusters were found to be indistinguishable from adipose samples and were also excluded as outliers. A list of samples passing all filtering and outlier detection steps is given in Supplemental File S1.

Merging similar tissues

The filtered set of samples spanned 28 organs/body sites and 49 tissue types, meaning multiple tissue types were sometimes collected from the same organ/body site (e.g., three tissue types were taken from the esophagus: "gastroesophageal junction", "mucosa", and "muscularis"). In cases where we could not clearly separate samples of two or more tissue types by hierarchical clustering, we merged these tissue labels, treating them as single tissue types: "Brain – Cerebellum" ← (Brain – Cerebellum, Brain – Cerebellar Hemisphere); "Brain – Cortex" ← (Brain – Cortex, Brain – Anterior cingulate cortex (BA24), Brain – Frontal Cortex (BA9)); "Brain – Striatum" ← (Brain – Caudate (basal ganglia), Brain – Nucleus accumbens (basal ganglia), Brain – Putamen (basal ganglia)); "Esophagus – Muscularis" ← (Esophagus – Gastroesophageal Junction, Esophagus – Muscularis). Multiple samples from the same tissue type of a single donor were treated as technical replicates. The expression levels of each gene across these replicates were averaged to obtain a single sample representing this donor–tissue combination.

Adjusting expression levels for the effects of covariates

Expression levels were corrected for the effects of three covariates: the intronic read mapping rate (SMNTRNRT), sample ischemic time (SMTSISCH), and RNA integrity number (RIN) (SMRIN). The effects of ischemic time and RIN on gene expression have been previously noted (Gallego Romero et al. 2014; Ferreira et al. 2018), and all three variables were significantly correlated with the expression levels of many Y- and X-chromosome genes across multiple tissues. For samples collected from the brain, donor ischemic time (TRISCHD) was used in place of sample ischemic time, because the latter information was not available. To perform this

correction, for each tissue separately, the linear model $\boldsymbol{y_g} = b_0 + \boldsymbol{Xb} + \varepsilon$ was fit, where $\boldsymbol{y_g}$ is a vector of gene *g*'s normalized log2(TPM + 0.1) expression levels across *n* samples, b_0 is an intercept term, $\boldsymbol{X}$ is the $n \times 4$ matrix of covariates (the three sample-quality variables, plus a fourth variable for sex), $\boldsymbol{b}$ is a 4 × n matrix of fixed-effect coefficients for the covariates, and ε is a vector of n residuals. To increase the comparability of expression levels across tissues, the covariates were centered on common values in all tissues: 8.0 for RIN, 0.12 for intronic read mapping, 100 minutes for ischemic time (or 450 minutes for brain tissues, which were processed separately from other tissues and had uniformly longer ischemic time). Gene *g*'s adjusted expression levels were calculated as $\boldsymbol{y_g^*} = \boldsymbol{y_g} - \boldsymbol{Xb}$.

Estimating gene mappability

To estimate the short-read mappability of each gene in GENCODE v19, its longest transcript isoform was selected, and all possible 76-nt reads were generated ($n - 76 + 1$ reads in total, where *n* is the length of the transcript in nucleotides) using a sliding window. These reads were aligned to the full transcriptome annotation with bowtie (v1.2) (Langmead et al. 2009), allowing 0 mismatches (-v 0) and reporting up to 200 alignments (-k 200). If a read aligned to the transcript isoforms of more than one gene or to more than one position within a single transcript isoform, it was classified as multi-mapping (and otherwise as uniquely mapping). The gene's mappability was then calculated as the fraction of reads from that gene that are uniquely mapping. A chromosome's mappability was calculated as the fraction of all reads generated from genes on that chromosome that mapped uniquely.

Correlated expression of X and Y homologs

The significance of the correlation between gene *i* and gene *j* was assessed by calculating the proportion of genes in the genome as or more correlated in expression with gene *i* than gene *j*, and vice versa. Specifically, the correlation coefficients between gene *i* and all *N* expressed genes were calculated and ordered from largest to smallest; let r_{ij} be the rank of gene *j* in this list. (For example, $r_{ij} = 2$ if only one gene in the genome shows a higher correlation with gene *i* than gene *j*.) The procedure was repeated for gene *j*, and the rank of gene *i*, r_{ji}, was obtained. The significance of the correlation between gene *i* and *j* was estimated by the average rank, normalized by the number of expressed genes in that tissue: $\left(r_{ij} + r_{ji}\right)/2N$. For example, in skeletal muscle, where 9,445 genes are expressed above 5 TPM, eight genes show higher correlation with *DDX3X* than *DDX3Y*, and two genes show higher correlation with *DDX3Y* than *DDX3X*; therefore, the average, normalized rank is $(3 + 9)/(2 * 9445) = 0.0006$.

microRNA analyses

For each X–Y gene pair, the 3′ UTRs of the X homolog, the Y homolog, and their autosomal chicken ortholog were aligned using PRANK (Löytynoja and Goldman 2005) with default parameters. Scripts from TargetScan 6.0 (Friedman et al. 2009) were then used to identify all potential miRNA target sites in the aligned sequences (targetscan_60.pl) and calculate their

context+ scores and percentiles (targetscan_60_context_scores.pl). Context+ scores, rather than the more recent context++ scores (Agarwal et al. 2015), were used because 3P-seq data needed to calculate context++ scores were not available for the human Y chromosome or chicken. Sites identified in X homologs were validated in the context++ model (TargetScan 7.2 (Agarwal et al. 2015)) (Supplemental Table S9). miRNA-target-site presence/absence in X- and Y-homolog 3′-UTRs was then compared to miRNA expression patterns across human tissues to generate predictions about their differential effects on X- and Y-homolog expression. miRNA expression patterns were assessed using quantile normalized expression values from Ludwig et al. (2016) (Ludwig et al. 2016) (https://ccb-web.cs.uni-saarland.de/tissueatlas/). Expression levels from the two donors were averaged, and only tissues matching a tissue in the GTEx dataset were analyzed. Among target sites for tissue-specific, highly expressed miRNAs, the miR-1 target site in EIF1AX is the target site with the highest context+ score–percentile (i.e., greatest predicted efficacy) preserved in one homolog of an X–Y pair but not the other.

For luciferase assays, the entire *EIF1AY* 3′-UTR and the first 1015bp of the *EIF1AX* 3′-UTR were amplified from human genomic DNA and cloned into the psiCheck-2 vector backbone (Promega) by restriction digest (Pme1, Not1). Using the QuikChange II kit (Agilent), *EIF1AX*'s miR-1 site was changed to shuffled sequence; *EIF1AY*'s disrupted miR-1 site was changed to match that of *EIF1AX*. Each psiCheck plasmid, along miR-1 or miR-124 duplexes, was transfected into HEK293 cells with Lipofectamine 2000 (Thermo Fisher Scientific). *Renilla* and firefly absorbance were quantified 24h post-transfection using the Dual-Luciferase Reporter Assay System (Promega), and the ratio of *Renilla*-to-firefly absorbance was calculated. Primers for cloning and mutagenesis, and miRNA oligonucleotide sequences, are listed in Supplemental File S2.

Analyses of EIF1AX/Y sequence and expression across species

Non-human RNA-seq data are from Brawand et al. and Merkin et al. (Brawand et al. 2011; Merkin et al. 2012). Kallisto was used to estimate transcript abundances (with options --bias and, for Brawand et al. data, --single -l 275 -s 15), supplying Ensembl version 98 transcript annotations for chimpanzee (Pan_tro_3.0), rhesus macaque (Mmul_10), and chicken (GRCg6a) and the GENCODE vM23 Basic annotation for mouse. For species with an intact, Y-linked *EIF1AY* ortholog (chimpanzee, rhesus), the cDNA sequences of *EIF1AX* and *EIF1AY* orthologs were aligned, and the well-aligned portion of each sequence was inserted into the annotation in place of the annotated sequence(s) listed by Ensembl. This was done to prevent differences in the completeness or correctness of the annotated X- and Y-linked sequences from skewing estimated Y/X expression ratios. After transcript abundance estimation, Y/X expression ratios were calculated in each tissue sample. For the expression patterns shown in Supplemental Figure S12, the expression levels from the samples of a given species were adjusted using the housekeeping normalization method described in Methods; the expression level of *EIF1AX* in each sample was then divided by *EIF1AX*'s median expression level observed among all samples from that species.

Analysis of protein abundance in human heart tissue

Selection of human heart tissue samples for protein quantification

GTEx heart (left ventricle) samples from 21 male donors and 12 female donors were selected for quantitative proteomic analysis after thoroughly screening all left ventricle samples by donor medical history and histopathological analysis. The two goals of this screening process were to identify samples with minimal pathology and to minimize differences between XX and XY samples (e.g., adiposity, fibrosis, hypertrophy) that might introduce biases. In the first round of screening, the pathology notes released by the GTEx Consortium (SMPTHNTS) were reviewed, and samples were excluded if the notes indicated >5% adipose tissue, more than "minimal" fibrosis or hypertrophy, or evidence of infarction, ischemia, or myocarditis. Next, samples were excluded based on donor medical history and circumstances of death. Specifically, a sample was excluded if the first underlying cause (DTHFUCOD) or immediate cause (DTHCOD) of death primarily affected the heart or cardiovascular system (e.g., cardiac arrest, myocardial infarction, cardiovascular disease), or if the donor had a recorded history of myocardial infarction (MHHRTATT), heart disease (MHHRTDIS, MHHRTDISB). Finally, the 56 remaining samples underwent a further round of expert histological review using the histology images available on the GTEx portal (https://gtexportal.org/). Each sample was scored for the content of adipose tissue and interstitial fibrosis, and for the degree of myocyte hypertrophy. Samples were excluded if they showed >3% adipose tissue; >2% fibrosis; or moderate myocyte hypertrophy in combination with borderline adipose and/or borderline fibrosis. Tissue from the remaining 33 remaining samples (21 male, 12 female) was obtained from the GTEx biobank (Supplemental File S3).

Human heart tissue proteomics: data generation

A total of 33 human heart samples (left ventricle sampled 1 cm above apex), 21 males and 12 females, stored in PAXgene at −80 °C, were obtained from Gene Expression Tissue (GTEx) Biobank. Samples were rinsed from the PAXgene buffer by ice-cold PBS, pulverized in 1.5 ml RIPA lysis buffer with Roche complete protease inhibitors, and sonicated for 2 min using 0.5 pulses.

Following the procedure outlined in Chick et al. (Chick et al. 2016), heart samples (~40 mg) were reduced with 5 mM dithiothreitol (DTT) for 30 min at 54°C followed by alkylation with 20mM iodoacetamide for 30 min at room temperature in the dark. The alkylation reaction was quenched by adding 15 mM DTT for 15 min at room temperature in the dark. A 200µl sample aliquot was then methanol/chloroform precipitated. The samples were allowed to air dry before being resuspended in 300 µl of 8 M urea buffer supplemented with 50 mM Tris at pH 8.2. The urea concentration was diluted down to ~1.5 M urea with 50 mM Tris. Proteins were quantified using a BCA assay. Protein was then digested using a combination of Lys-C/trypsin at an enzyme-to-protein ratio of 1:100. First, protein was digested overnight with Lys-C followed by 6-h digestion with trypsin all at 37 °C. Samples were then acidified using formic acid to approximately pH 3. Samples were desalted using a SepPak column, and eluents were dried using a vacuum centrifuge. Peptide pellets were resuspended in 110 µl of 200 mM HEPES buffer, pH 8, and peptides were quantified by a BCA assay. Approximately 70 µg of peptides (100 µl of sample

+ 30 μl of 100% acetonitrile) were then labeled with 15 μl of 20 μg μl^{-1} of the corresponding TMT 11-plex reagent for 2 h at room temperature. The reaction was quenched using 8 μl of 5% hydroxylamine for 15 min. Peptides were then acidified using 150 μl of 1% formic acid, each set of 11 samples was mixed and desalted using a SepPak column. In total, 3 TMT 11-plex reactions were performed to analyze all 33 samples. The full labeling scheme for the heart samples is provided in Figure 6—source data.

Each of the 3 TMT experiments was separated by basic, reversed-phase chromatography. Samples were loaded onto an Agilent 300 Extend C18 column (5 μm particles, 4.6 mm ID and 220 mm in length). Using an Agilent 1100 quaternary pump equipped with a degasser and a photodiode array detector (set at 220- and 280-nm wavelength), peptides were separated using a 50 min linear gradient from 18% to 40% acetonitrile in 10 mM ammonium bicarbonate, pH 8, at a flow rate of 0.8 ml min−1. Peptides were separated into a total of 96 fractions that were consolidated into 24. Samples were subsequently acidified with 1% formic acid and vacuum centrifuged to near dryness. Each fraction was desalted via StageTip, dried via vacuum centrifugation, and reconstituted in 1% formic acid for liquid chromatography tandem mass spectrometry (LC–MS/MS) processing.

Peptides from every odd fraction (12 fractions total) from basic reverse-phase fractionation were analysed using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific) equipped with a Proxeon ultra high pressure liquid chromatography unit. Peptide mixtures were separated on a 100 μm ID microcapillary column packed first with ~0.5 cm of 5 μm Magic C18 resin followed by 40 cm of 1.8 μm GP-C18 resin. Peptides were separated using a 3-h gradient of 6–30% acetonitrile gradient in 0.125% formic acid with a flow rate of ~400 nl min−1. In each data collection cycle, one full MS scan (400–1,400 m/z) was acquired in the Orbitrap (1.2 × 105 resolution setting and an automatic gain control (AGC) setting of 2 × 105). The subsequent MS2–MS3 analysis was conducted with a top 10 setting or a top speed approach using a 2-s duration. The most abundant ions were selected for fragmentation by collision induced dissociation (CID). CID was performed with a collision energy of 35%, an AGC setting of 4 × 103, an isolation window of 0.5 Da, a maximum ion accumulation time of 150 ms and the rapid ion trap setting. Previously analyzed precursor ions were dynamically excluded for 40 s.

During the MS3 analyses for TMT quantification, precursors were isolated using a 2.5-Da m/z window and fragmented by 35% CID in the ion trap. Multiple fragment ions (SPS ions) were co-selected and further fragmented by HCD. Precursor ion selection was based on the previous MS2 scan and the MS2–MS3 was conducted using sequential precursor selection (SPS) methodology. HCD used for the MS3 was performed using 55% collision energy and reporter ions were detected using the Orbitrap with a resolution setting of 60,000, an AGC setting of 50,000 and a maximum ion accumulation time of 150 ms.

Human heart tissue proteomics: peptide quantification and protein-abundance estimation

Software tools were used to convert mass spectrometric data from raw file to the mzxml format (Huttlin et al. 2015). Erroneous charge state and monoisotopic m/z values were corrected as per previous publication (Huttlin et al. 2015). MS/MS spectra assignments were made with the Sequest algorithm (Eng et al. 1994) using an indexed Ensembl database (Ensembl version

GRCh37.61). Databases were prepared with forward and reversed sequences concatenated according to the target-decoy strategy (Elias and Gygi 2007). All searches were performed using a static modification for cysteine alkylation (57.0215 Da) and TMT on the peptide N termini and lysines. Methionine oxidation (15.9949 Da) was considered a dynamic modification. Mass spectra were searched with trypsin specificity using a precursor ion tolerance of 10 p.p.m. and a fragment ion tolerance of 0.8 Da. Sequest matches were filtered by linear discriminant analysis as described previously, first to a data set level error of 1% at the peptide level based on matches to reversed sequences (Elias and Gygi 2007). Peptide probabilities were then multiplied to create protein rankings and the data set was again filtered to a final data set level error of 1% false discovery rate (FDR) at the protein level. The final protein-level FDR fell well below 1% (~0.22% peptide level).

Peptide quantitation using TMT reporter ions was accomplished as previously published (Ting et al. 2011; McAlister et al. 2014). In brief, a 0.003 Da m/z window centered on the theoretical m/z value of each reporter ion was monitored for each of the 11 reporter ions, and the intensity of the signal closest to the theoretical m/z value was recorded. TMT signals were also corrected for isotope impurities based on the manufacturer's instructions. Peptides were only considered quantifiable if the total signal-to-noise for all channels was >200 with an isolation specificity of >0.75.

For proteins not encoded by X–Y gene pairs, peptides were assigned to protein matches using a reductionist model, where all peptides were explained using the least number of proteins. The signal-to-noise values in each channel were then divided by the sum of all signal-to-noise values in that channel, such that each channel had the same summed value. Protein quantitation was then performed by summing the signal-to-noise values for all peptides for a given protein. Within each 11-plex TMT experiment, protein quantitative measurements were then scaled to 100, such that equal expression across all channels would be equal to 100/11 ≈ 9.1.

Protein abundances of the X and Y isoforms were estimated separately as follows. Within each 11-plex experiment, raw signal-to-noise values in each of the 11 channels were first normalized (divided) by the summed signal/noise value for all peptides in that channel: $\tilde{y}_{ij} = y_{ij} / \sum_{p \in P} y_{pj}$, where y_{ij} is the raw signal/noise value for peptide i in channel j, $\tilde{y}_{ij}$ is the channel-normalized signal/noise value, and P is the set of all detected peptides. Among all detected peptides, we then identified those that specifically matched the amino-acid sequence of an X homolog of an X–Y pair, of a Y homolog of an X–Y pair, or both X and Y homologs of an X–Y pair but not the sequence of any other protein. We detected all three classes (X-specific, Y-specific, X-Y-shared) of peptides for RPS4Y1/RPS4X, EIF1AY/EIF1AX, and DDX3Y/DDX3X, the three most highly expressed X–Y pair genes. (We further detected X-specific peptides for USP9X/USP9Y but no Y-specific or X-Y-shared peptides.) Y-specific peptides showed low but roughly constant signal in female channels, with mean signal/noise = 2.57: we used this value as an estimate of the non-specific background for all peptides and subtracted this value from all peptides in all channels, setting any negative values to 0. For each of the three X–Y pairs separately, we then estimated the relative abundance of the X isoform in channel j, $a_j^{(X)}$, as the percentage of signal from all X-specific peptides in channel j out of the total signal across all channels,

$$a_j^{(\mathrm{X})} = \frac{\sum_{p \in P_X} \tilde{y}_{pj}}{\sum_{k=1}^{11} \sum_{p \in P_X} \tilde{y}_{pk}} * 100,$$

where P_X is the set of all X-specific peptides for the given X–Y pair. We repeated this calculation using X–Y-shared peptides to obtain $a_j^{(\mathrm{XY})}$, the relative abundance of the sum of X and Y isoforms in channel j. To obtain the male-to-female expression ratio for the X homolog specifically, $\phi_{\mathrm{MF}}^{(\mathrm{X})}$, we pooled the abundance estimates across all three 11-plex experiments and divided its average abundance in male channels by its average abundance in female channels, i.e.,

$$\phi_{\mathrm{MF}}^{(\mathrm{X})} = \frac{\frac{1}{21} \sum_{j \in C_M} a_j^{(\mathrm{X})}}{\frac{1}{12} \sum_{j \in C_F} a_j^{(\mathrm{X})}},$$

where C_M and C_F are the sets of channels from male and female donors, respectively. The male-to-female expression ratio for the sum of X and Y homolog expression, $\phi_{\mathrm{MF}}^{(\mathrm{XY})}$, was obtained similarly. p-values for the sex bias in expression were estimated by permuting the sample labels within each 11-plex experiment one million times and calculating the proportion of permutations that yielded more extreme male-to-female expression ratios. (This procedure was also used to estimate p-values for the sex bias of non-X–Y-pair proteins, shown in Supplemental Fig. S18.) Finally, the Y-to-X expression ratio within males was estimated as $(\phi_{\mathrm{MF}}^{(\mathrm{XY})} - \phi_{\mathrm{MF}}^{(\mathrm{X})})/\phi_{\mathrm{MF}}^{(\mathrm{X})}$.

Immunoblotting

To prepare human heart lysates, 40 mg of human heart tissue (obtained originally for the mass spectrometry analysis) was rinsed twice with 2 ml of ice-cold PBS and pulverized in 1500 µl of RIPA buffer with protease inhibitor cocktail (Roche, Catalog number: 11836170001). Lysates were incubated 30 min on ice and spun at 10,000 g for 20 min at 4 °C, and the supernatant was collected. Aliquots of the heart lysates were pooled by sex (21 male samples, 12 female samples), mixed with NuPAGE Sample Reducing Agent 10x (Invitrogen, #NP0009) and NuPAGE LDS Sample Buffer 4x (Invitrogen, #NP0007), incubated for 10 min at 90 °C, and then chilled on ice for 2 min. Proteins were separated for 3.5 hr at 80 V on a NuPAGE 4-12% Bis-Tris gel (Invitrogen, #NP0322BOX) and transferred to a nitrocellulose membrane. The membrane was blocked for 1 hr in Tris-buffered saline containing 0.1% Tween-20 (TBST) and 5% non-fat milk at room temperature, and then incubated with primary antibodies in TBST overnight at 4 °C on a shaking platform (GAPDH: Ambion AM4300, anti-mouse, 1:106 dilution; EIF1A: Abcam Ab177939, anti-rabbit, 1:5000 dilution). The monoclonal EIF1A antibody was generated using a proprietary synthetic peptide within amino-acids 50 – 144 of human EIF1AX. After three washes with TBST, the membrane was incubated at room temperature for 1 hr with TBST and 1% milk containing anti-mouse and anti-rabbit secondary antibodies labeled with fluorescent dyes detectable at different wavelengths (LI-COR IRDye 680RD Goat anti-Mouse 925-68070, 1:20000 dilution; LI-COR IRDye 800CW Goat anti-Rabbit 926-32211, 1:20000 dilution). Following three further washes in Tris-buffered saline lacking Tween-20, fluorescent signal was recorded using

an Odyssey CLx imager (LI-COR) with Image Studio software (version 5.2.5). Fluorescent signal corresponding to EIF1A was normalized to signal for GAPDH in each lane, with four technical replicates (i.e., lanes) per sex.

Lymphoblastoid cell lines (LCLs) from individuals with sex chromosome aneuploidies were derived in our lab, except the following that were previously reported: 47,XYY (Repping et al. 2003), 49,XYYYY (GM11419, from Coriell) and Sirota et al. (Sirota et al. 1981). Cells were pelleted, washed in ice-cold PBS, and snap frozen in liquid nitrogen. Approximately 5 million cells per line were lysed in 250 µl M-PER lysis buffer (Thermo Scientific #78503) supplemented with protease inhibitor. Lysates were incubated on ice for 15 min and spun for at 14,000 rpm for 15 min at 4 °C, and the supernatant was collected. LCL lysates were mixed with sample reducing agent and sample buffer, incubated for 10 min at 90 °C, and chilled on ice for 2 min. Proteins were separated for 3 hr at 80 V on a NuPAGE 4-12% Bis-Tris gel and transferred to a nitrocellulose membrane. Primary-antibody incubation was performed as described above. Membranes were subsequently incubated with TBST and 1% milk containing peroxidase-conjugated secondary antibodies at 1:5000 dilution (Jackson ImmunoResearch Peroxidase AffiniPure Donkey Anti-Mouse 715-035-151 & Anti-Rabbit 711-035-152) for 1 hr at room temperature. Following three washes with TBST, proteins on the membranes were detected by addition of Lumi-Light Western Blotting Substrate (Roche 12015200001).

SUPPLEMENT REFERENCES

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**: e05005. doi:10.7554/eLife.05005

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. 2016. Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**: 500–505.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**: 207–214.

Eng JK, McCormack AL, Yates JR. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**: 976–989.

Ferreira PG, Muñoz-Aguirre M, Reverter F, Godinho CPS, Sousa A, Amadoz A, Sodaei R, Hidalgo MR, Pervouchine D, Carbonell-Caballero J, et al. 2018. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat Commun* **9**: 490. doi:10.1038/s41467-017-02772-x

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**: 92–105.

Gallego Romero I, Pai AA, Tung J, Gilad Y. 2014. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol* **12**: 42. doi:10.1186/1741-7007-12-42

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82–86.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, et al. 2015. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**: 425–440.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36. doi:10.1186/gb-2013-14-4-r36

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25. doi:10.1186/gb-2009-10-3-r25

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323. doi:10.1186/1471-2105-12-323

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci* **102**: 10557–10562.

Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, Rheinheimer S, Meder B, Stähler C, Meese E, et al. 2016. Distribution of miRNA expression across human tissues. *Nucleic Acids Res* **44**: 3865–3877.

McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP. 2014. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**: 7150–7158.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**: 1593–1599.

Repping S, Skaletsky H, Brown L, Daalen SKM van, Korver CM, Pyntikova T, Kuroda-Kawaguchi T, Vries JWA de, Oates RD, Silber S, et al. 2003. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nat Genet* **35**: 247–251.

Sirota L, Zlotogora Y, Shabtai F, Halbrecht I, Elian E. 1981. 49, XYYYY. A case report. *Clin Genet* **19**: 87–93.

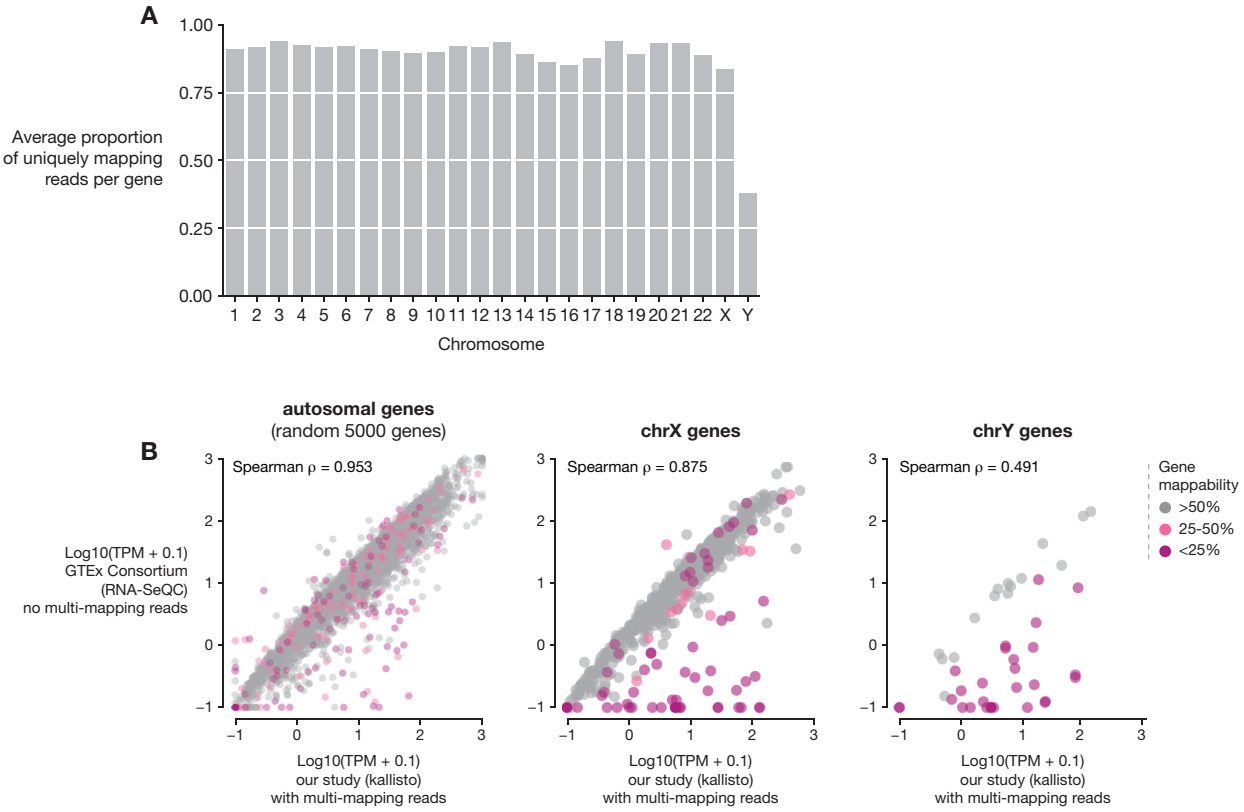
Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825–837.

Ting L, Rad R, Gygi SP, Haas W. 2011. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**: 937–940.

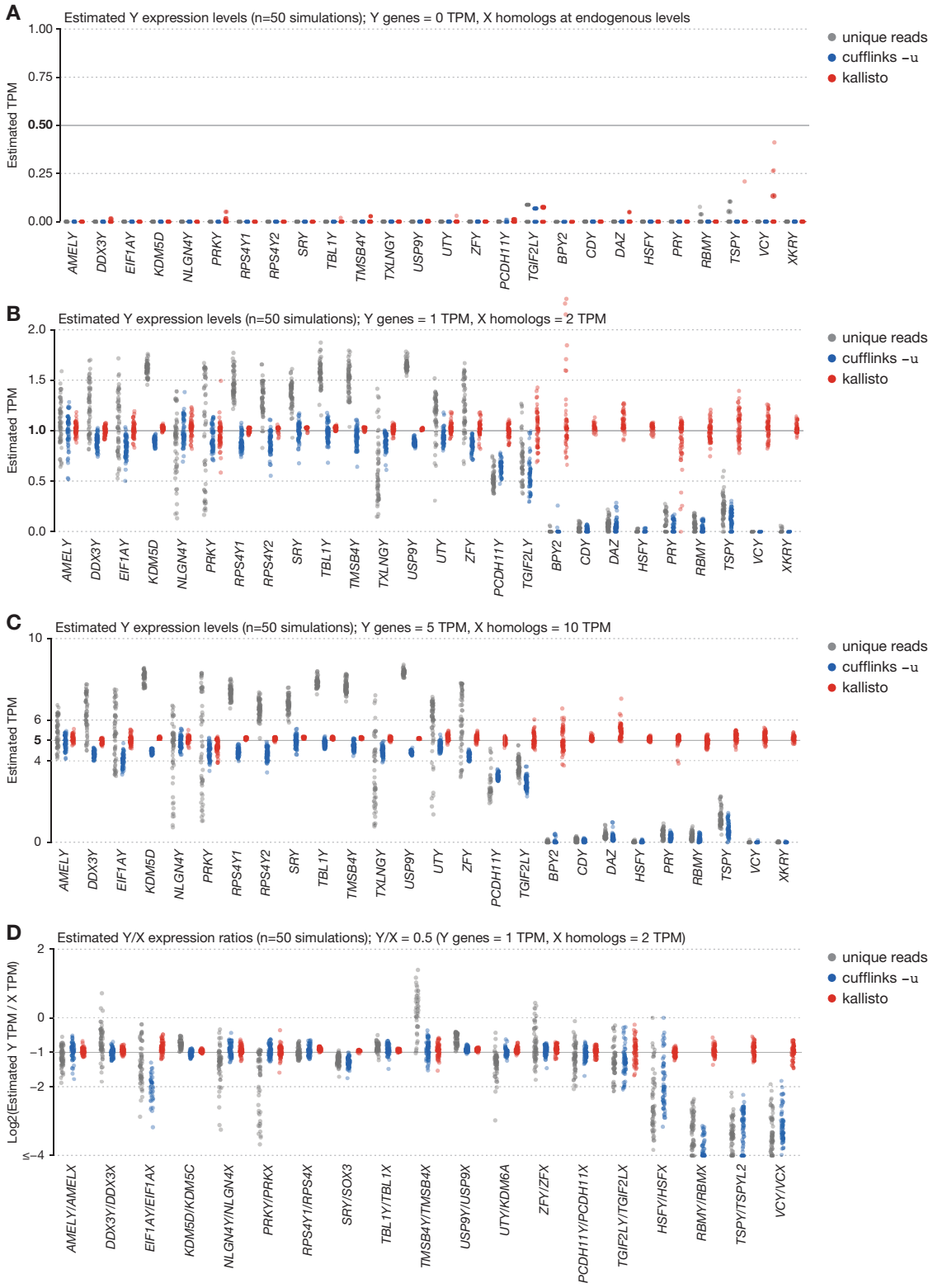
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJ van, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou Y-H, et al. 2014. Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46**: 430–437.

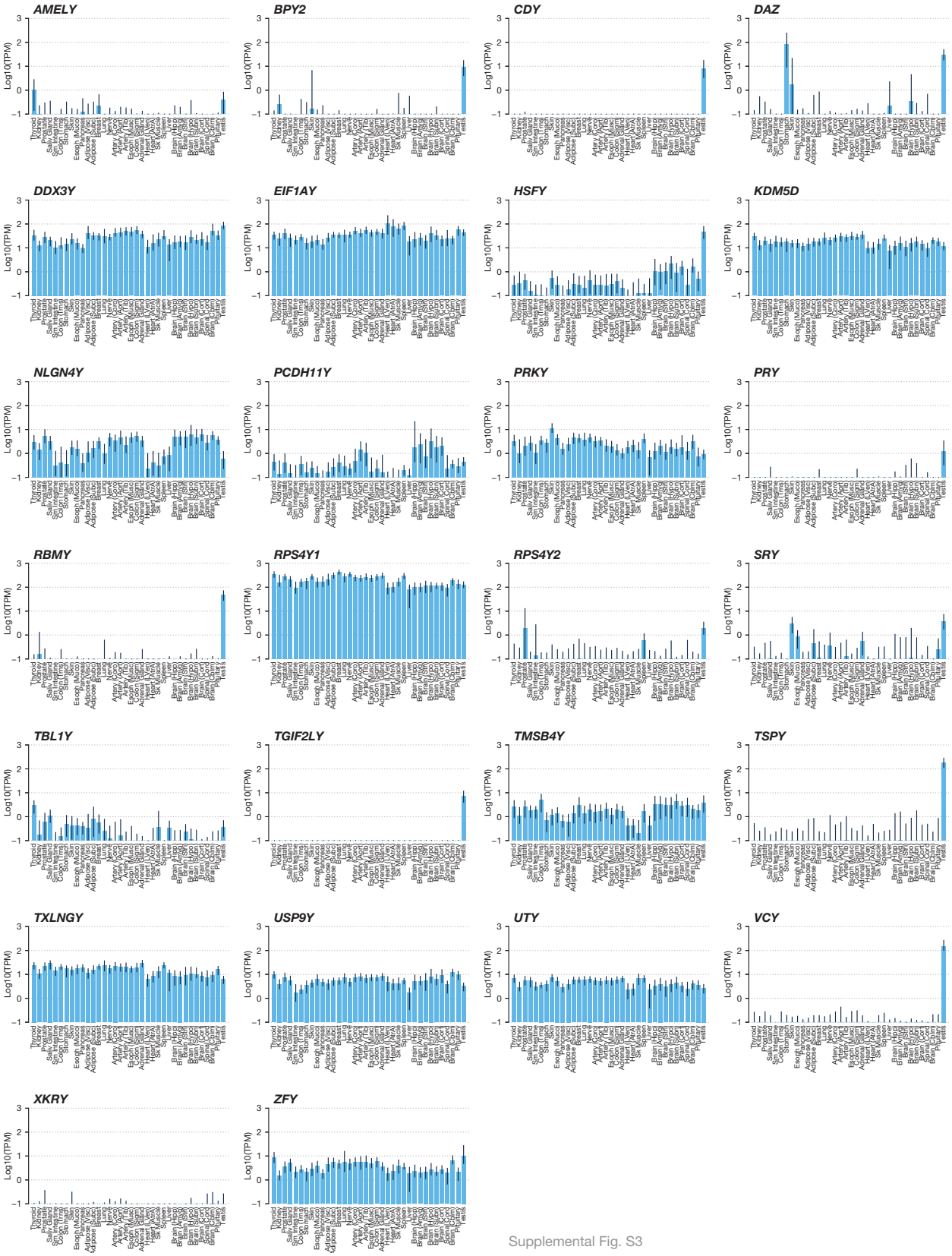
SUPPLEMENTAL FIGURES



Supplemental Fig. S1. Discarding multi-mapping reads disproportionately underestimates MSY gene expression. (**A**) Bars show the average proportion of reads from genes on each chromosome that can be aligned uniquely. Chromosome "Y" refers to genes in the MSY; chromosome "X" similarly excludes genes in the pseudoautosomal region. (**B**) Each point shows the expression level of one gene in the testis as estimated without multi-mapping reads (GTEx Consortium, via RNA-SeQC) and with multi-mapping reads (our study, via kallisto). Each point is colored according to the proportion of reads from that gene that can be aligned uniquely (gene mappability).

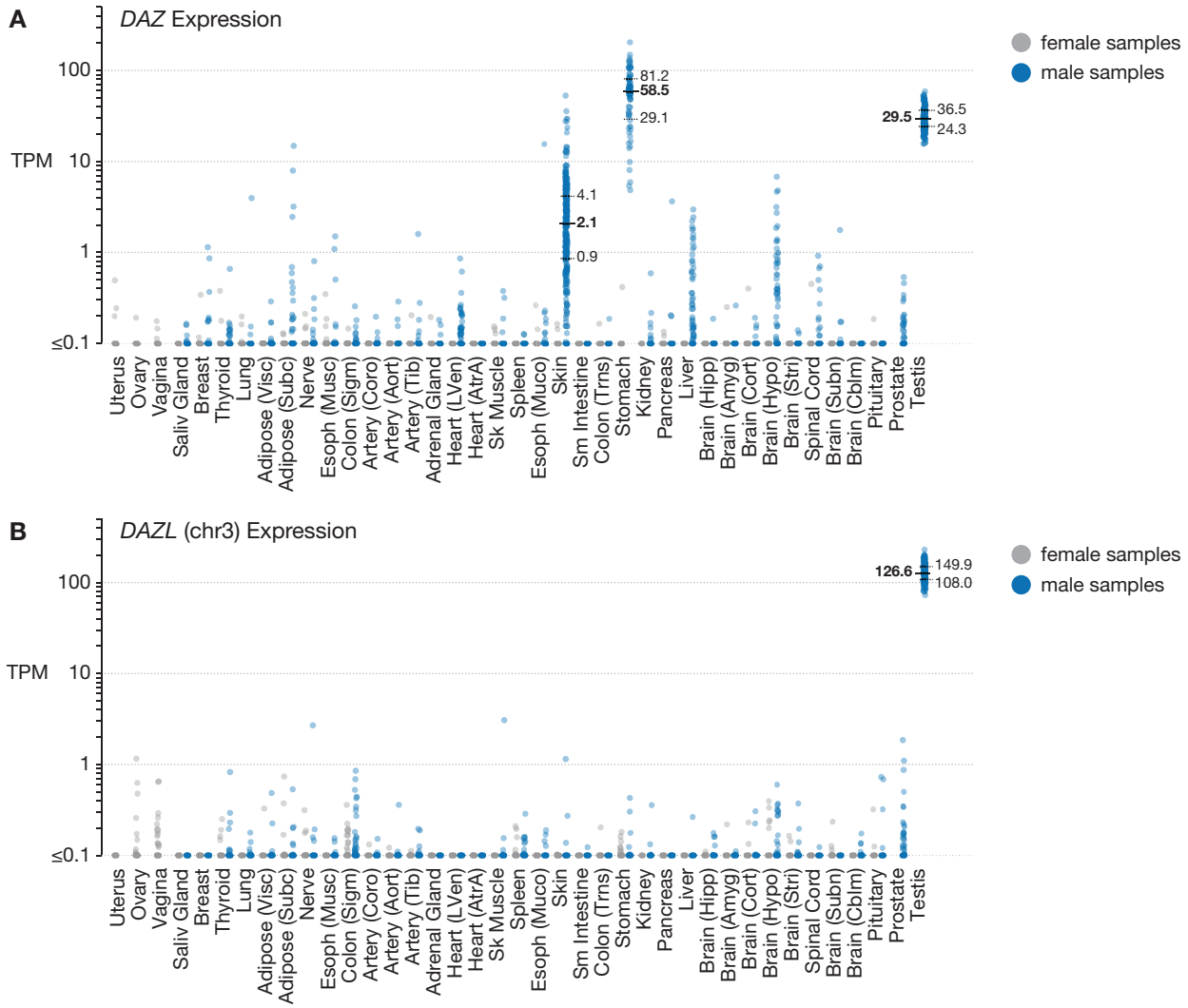

A
Estimated Y expression levels (n=50 simulations); Y genes = 0 TPM, X homologs at endogenous levels
B
Estimated Y expression levels (n=50 simulations); Y genes = 1 TPM, X homologs = 2 TPM
C
Estimated Y expression levels (n=50 simulations); Y genes = 5 TPM, X homologs = 10 TPM
D
Estimated Y/X expression ratios (n=50 simulations); Y/X = 0.5 (Y genes = 1 TPM, X homologs = 2 TPM)
unique reads
cufflinks –u
kallisto
Estimated TPM
Log2(Estimated Y TPM / X TPM)
AMELY
DDX3Y
EIF1AY
KDM5D
NLGN4Y
PRKY
RPS4Y1
RPS4Y2
SRY
TBL1Y
TMSB4Y
TXLNGY
USP9Y
UTY
ZFY
PCDH11Y
TGIF2LY
BPY2
CDY
DAZ
HSFY
PRY
RBMY
TSPY
VCY
XKRY
AMELY/AMELX
DDX3Y/DDX3X
EIF1AY/EIF1AX
KDM5D/KDM5C
NLGN4Y/NLGN4X
PRKY/PRKX
RPS4Y1/RPS4X
SRY/SOX3
TBL1Y/TBL1X
TMSB4Y/TMSB4X
USP9Y/USP9X
UTY/KDM6A
ZFY/ZFX
PCDH11Y/PCDH11X
TGIF2LY/TGIF2LX
HSFY/HSFX
RBMY/RBMX
TSPY/TSPYL2
VCY/VCX

< Supplemental Fig. S2. Kallisto accurately estimates MSY gene expression levels in simulated RNA-seq datasets. Each point shows the expression level of an MSY gene (**A** – **C**) or the Y/X expression ratio (**D**) in a simulated RNA-seq dataset as estimated by a procedure that discards multi-mapping reads (gray), Cufflinks in multi-correct "-u" mode (blue), or kallisto (red). 50 simulated RNA-seq datasets were generated for each of three scenarios: (**A**) MSY genes not expressed (0 TPM) and X homologs of MSY genes kept at levels in sample GTEX-P4QS-2126-SM-3NMCF; (**B**, **D**) MSY genes set to 1 TPM and X homologs set to 2 TPM; (**C**) MSY genes set to 5 TPM and X homologs set to 10 TPM.

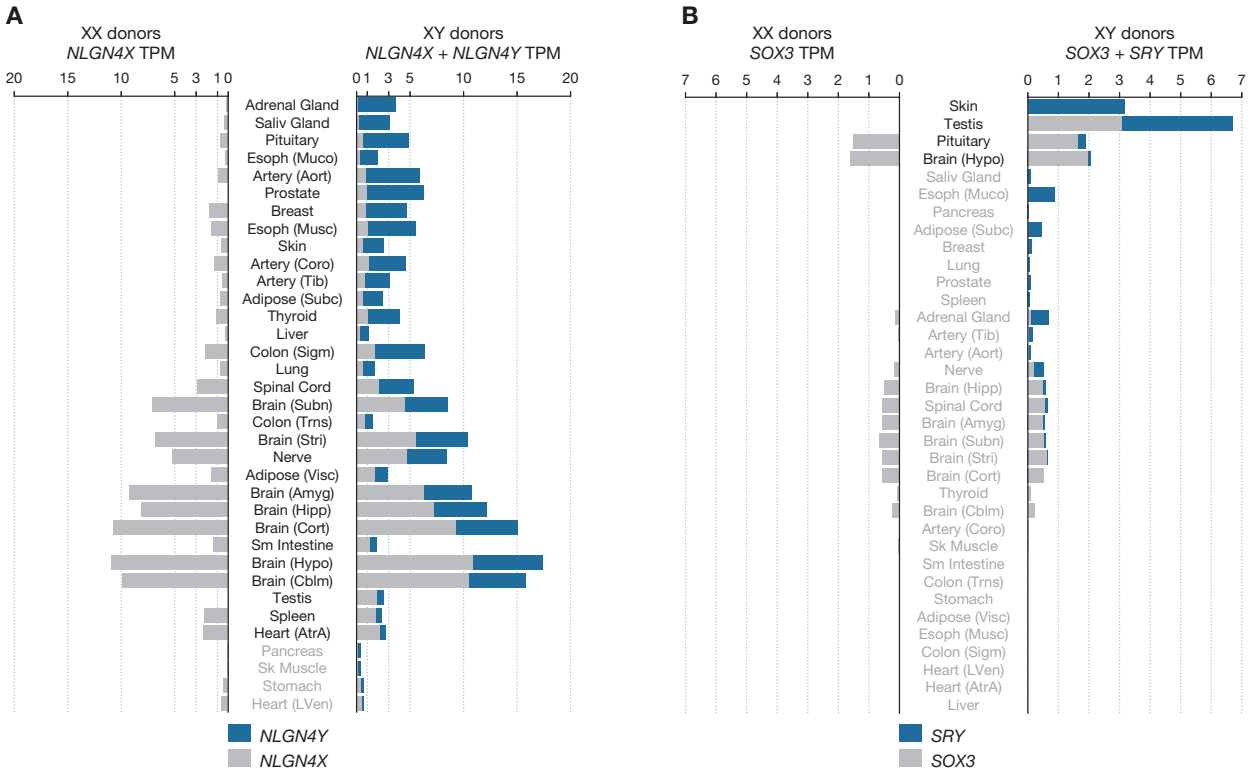


Supplemental Fig. S3

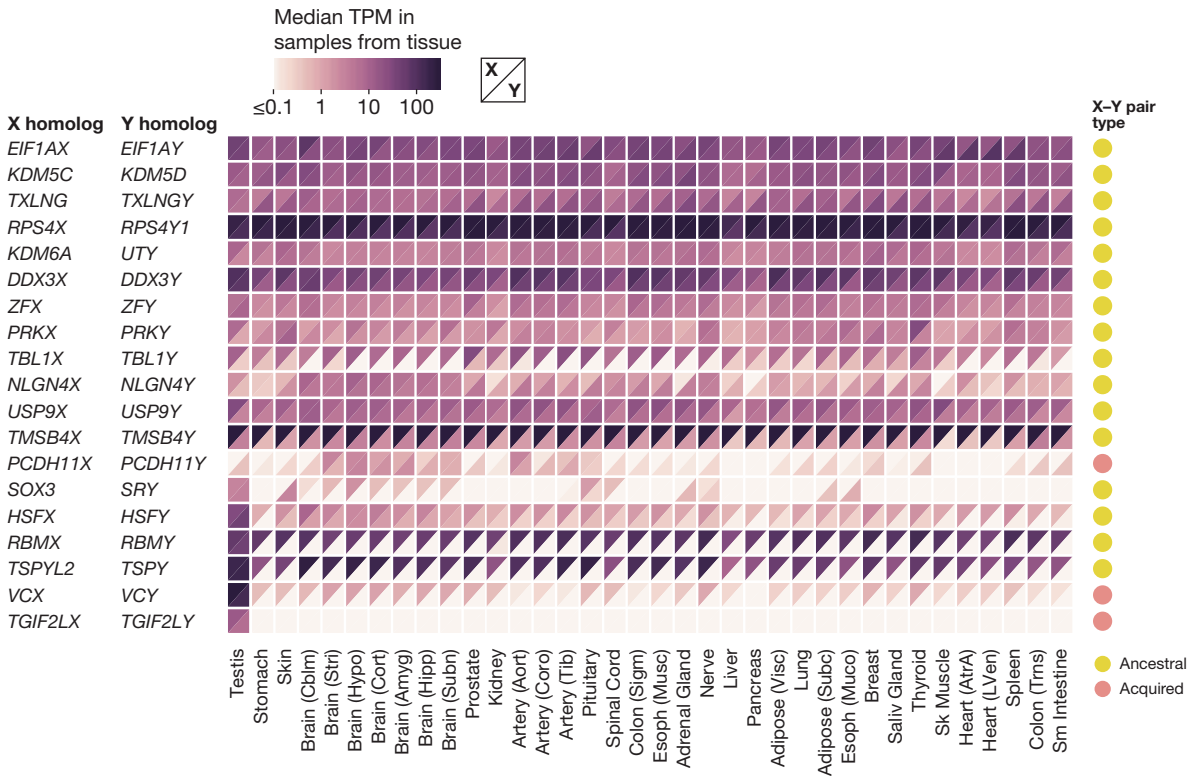
< Supplemental Fig. S3. Expression patterns of individual MSY genes. For each gene, blue bars show the gene's median expression level across the samples of each tissue; error bars: 5th and 95th percentiles.



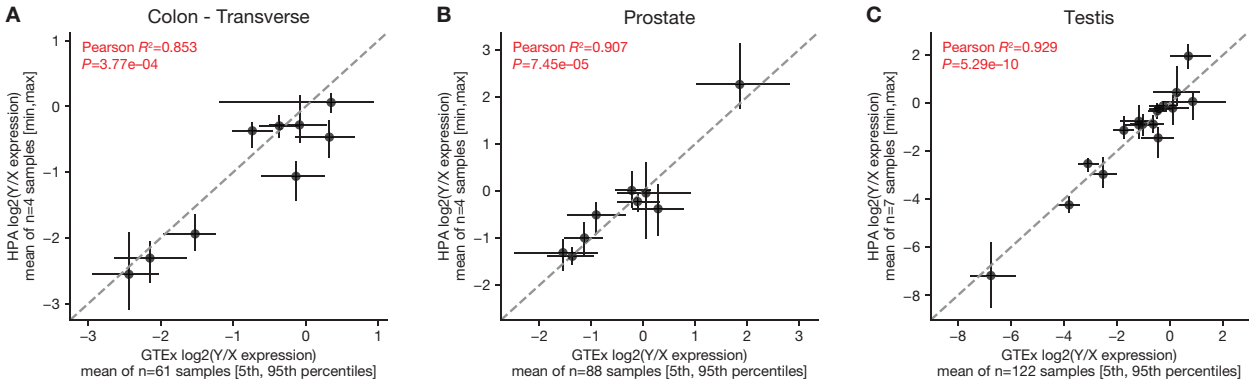
Supplemental Fig. S4. Expression of the *DAZ* gene family in the stomach. (**A**) Each point shows the expression level of *DAZ* in a single tissue sample from an XY (blue) or XX (gray) donor. For skin, stomach, and testis, lines show the 25th, 50th (median), and 75th percentiles. The absence of *DAZ* expression in XX donors suggests that *DAZ*'s stomach expression is not the result of mis-mapped reads from a gene on another chromosome. (**B**) The *DAZ* gene family arose from transposition of *DAZL* on chr3 to the Y chromosome. Each point shows *DAZL*'s expression in a single sample from XY or XX donors. The absence of *DAZL* expression in stomach samples suggests that *DAZ* acquired stomach expression after transposition to the Y chromosome.



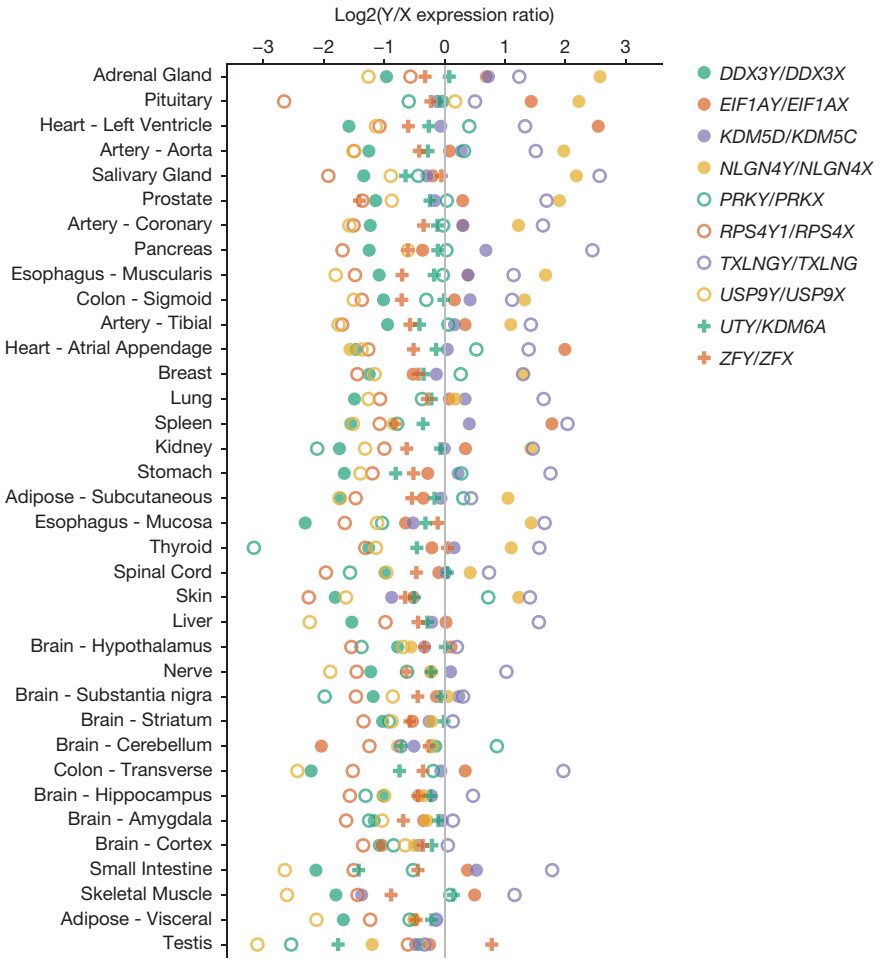
Supplemental Fig. S5. *NLGN4Y* and *SRY*, unlike other MSY genes, show broader expression than their X-linked homologs, leading to male-specific expression of *NLGN4X/Y* and *SOX3/SRY* in some tissues. (**A**) Expression levels of *NLGN4X* in XX donors (left) or the summed expression of *NLGN4X* and *NLGN4Y* in XY donors (right). (**B**) Expression levels of *SOX3* in XX donors (left) or the summed expression of *SOX3* and *SRY* in XY donors (right). In both panels, tissues are ordered by the fraction of expression coming from the Y-linked homolog. Tissues in gray text are those where the total expression of the X–Y pair is less than 1 TPM in both sexes.



Supplemental Fig. S6. Comparison of X- and Y homolog expression breadth in tissues from XY donors. Each cell shows the expression levels of the X-linked (upper left triangle) and Y-linked (lower right triangle) homologs of one X–Y gene pair in one tissue in XY donors.



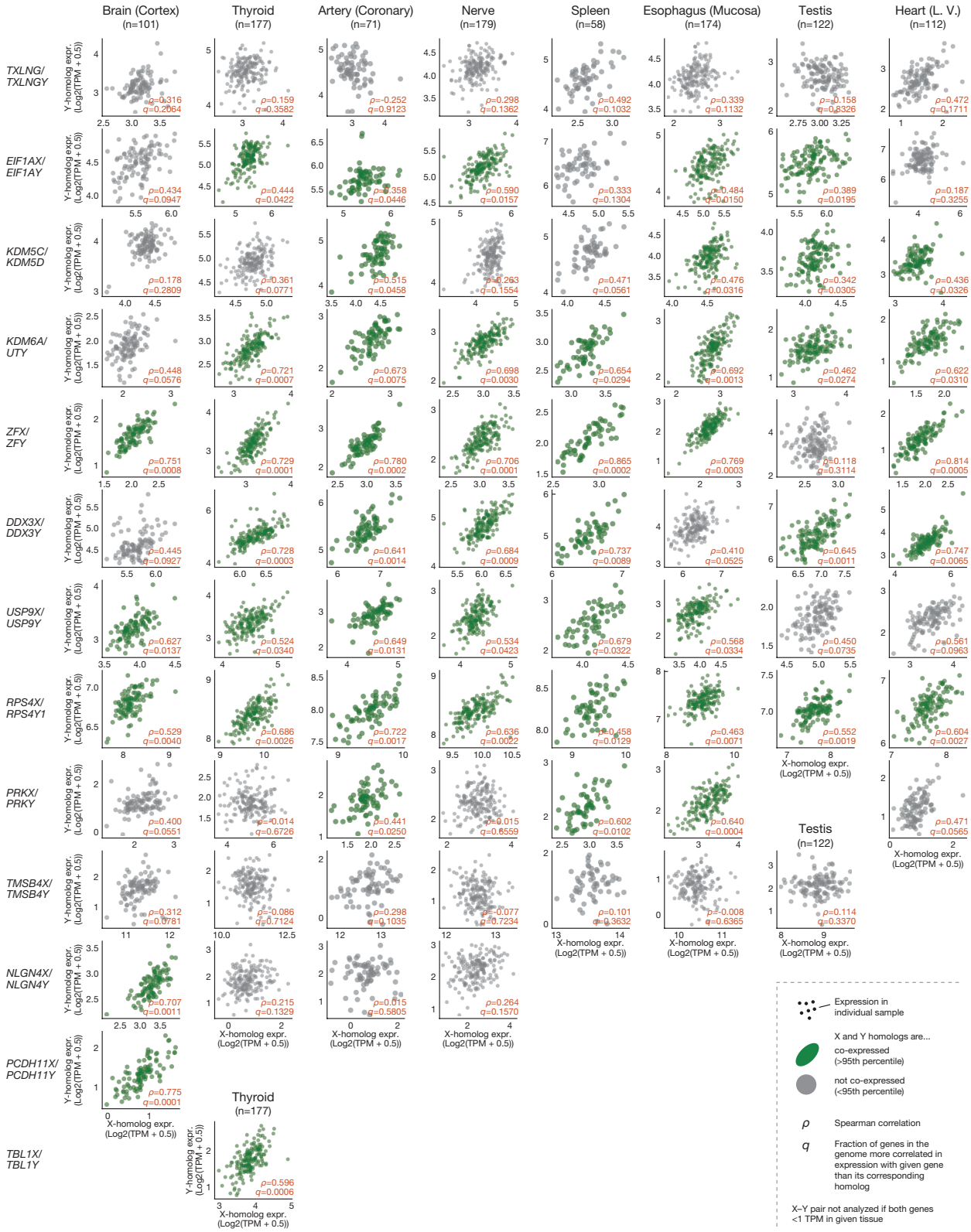
Supplemental Fig. S7. Replication of estimated Y/X expression ratios in RNA-seq data from the Human Protein Atlas (HPA). (**A** – **B**) Mean Y/X ratios for 9 widely expressed X–Y pairs in GTEx (*x*-axis; error bars: 5th – 95th percentiles) and HPA samples (*y*-axis; error bars: min, max among samples) from (**A**) colon or (**B**) prostate. (**C**) Mean Y/X ratios for all homologous X–Y pairs in GTEx and HPA testis samples.

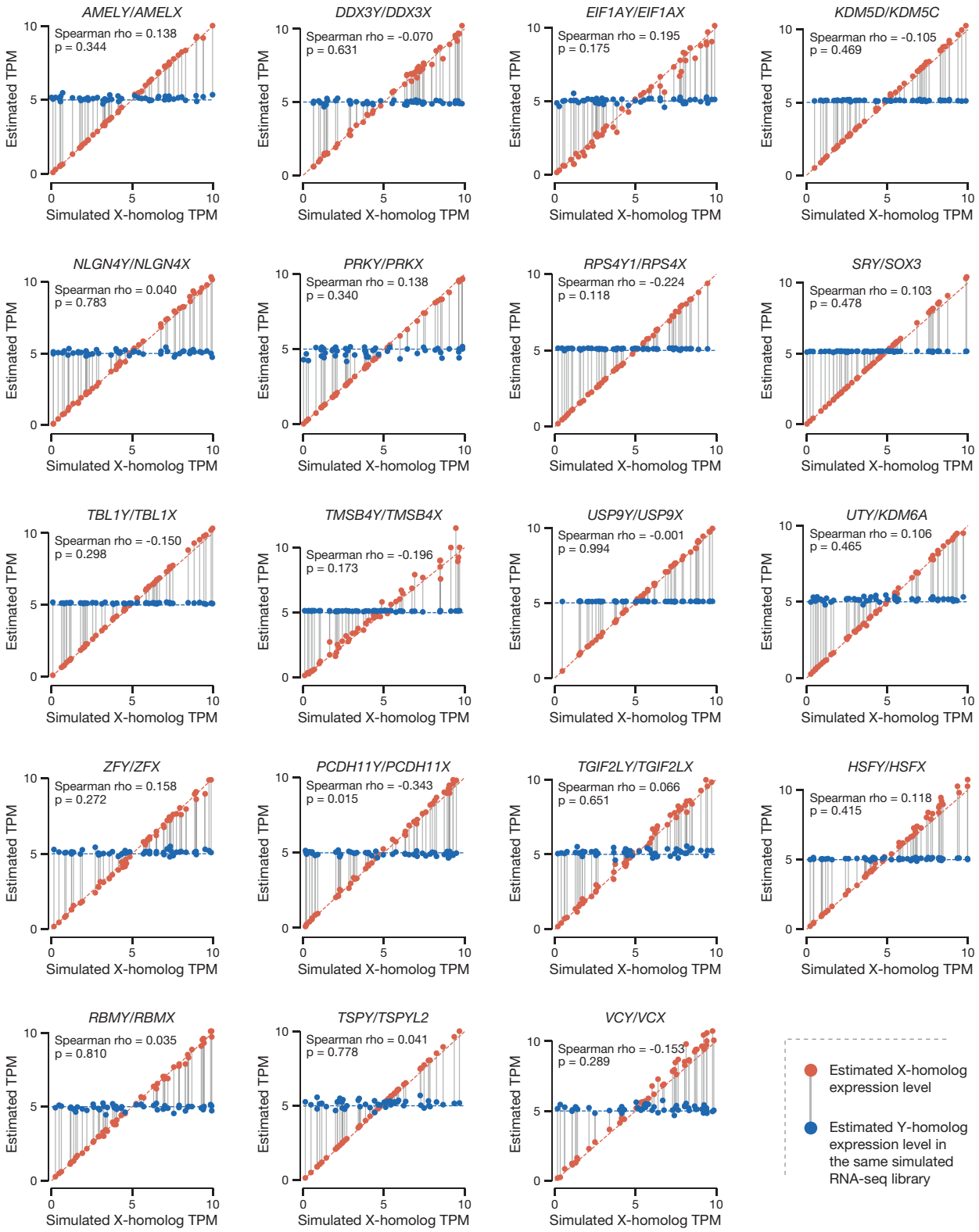


Supplemental Fig. S8. Y/X expression ratios do not differ substantially between tissues. Each point shows the estimated Y/X expression ratio of one X–Y pair in one tissue. Tissues are ordered by the average Y/X expression ratio across all pairs.

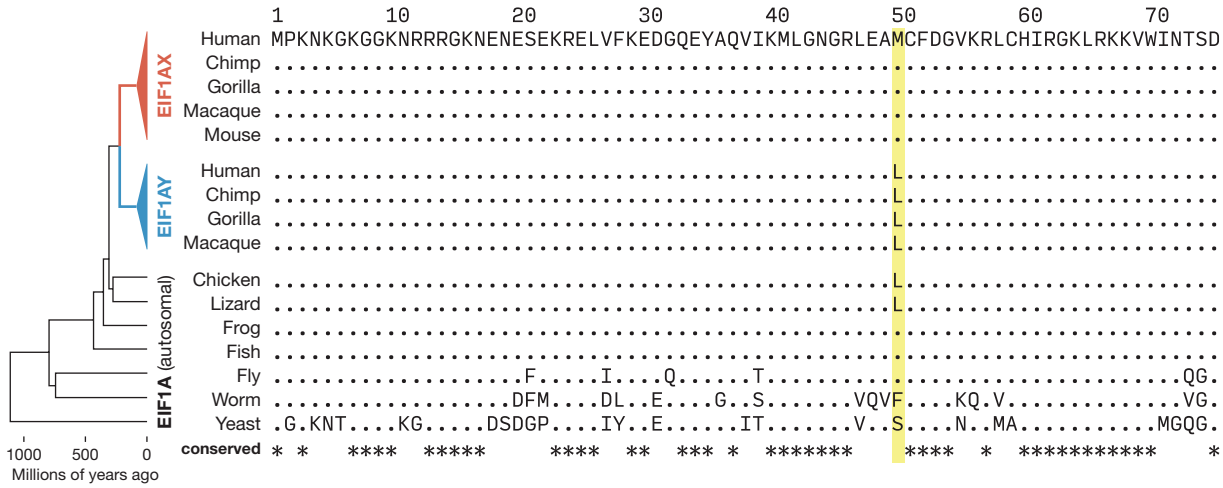
> Supplemental Fig. S9 (next page). Co-expression of the X- and Y-linked members of X–Y gene pairs in individual tissues. Each point shows the expression levels of the X- and Y-linked members of an X–Y gene pair in a single tissue sample. Each plot corresponds to a cell of the heatmap shown in Figure 3C. This subset of tissues (8/36) was selected to showcase a diversity of tissue types and cases where X–Y pairs are correlated and uncorrelated in expression.

Supplemental Fig. S9

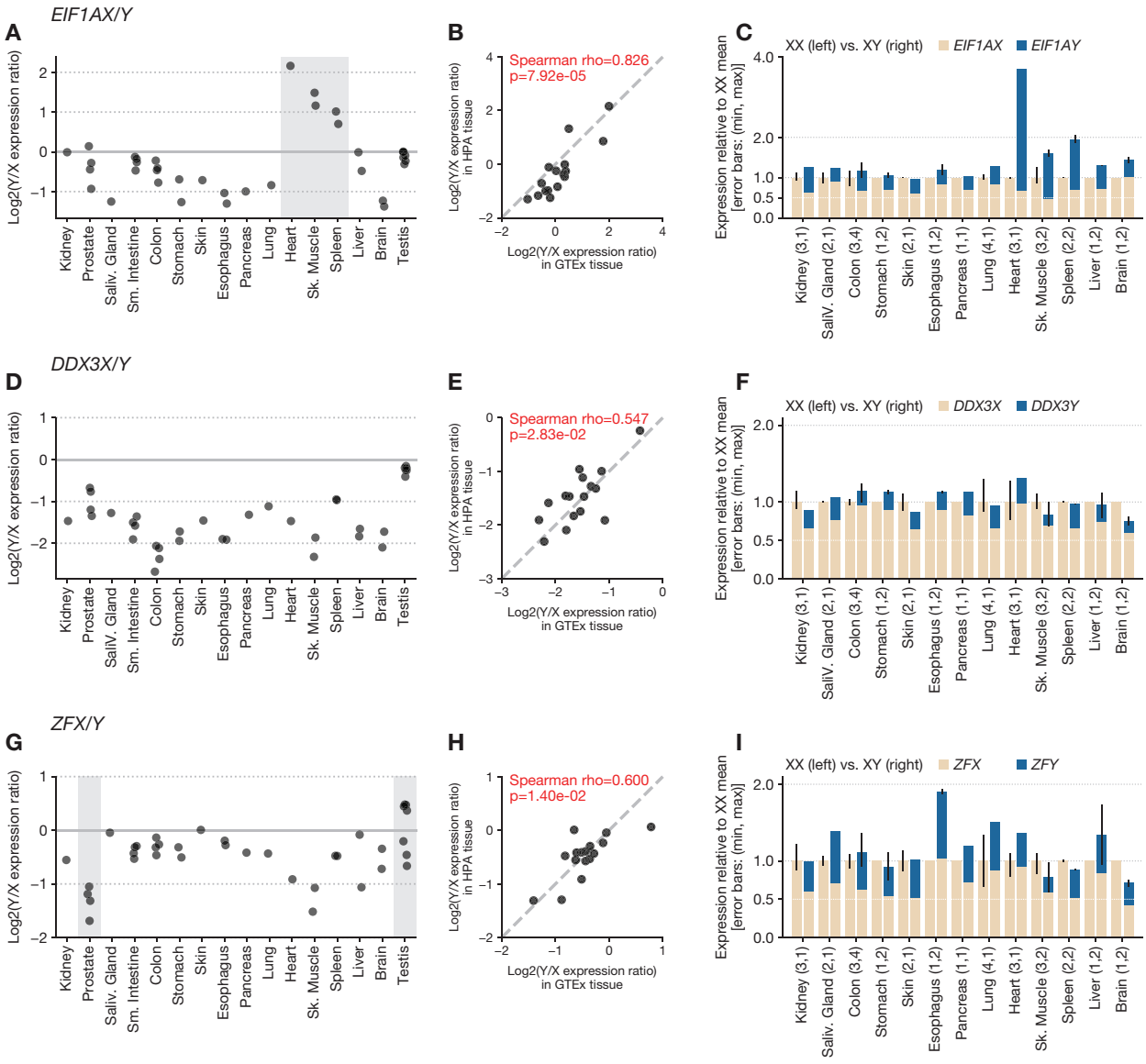



AMELY/AMELX
Spearman rho = 0.138
p = 0.344
DDX3Y/DDX3X
Spearman rho = -0.070
p = 0.631
EIF1AY/EIF1AX
Spearman rho = 0.195
p = 0.175
KDM5D/KDM5C
Spearman rho = -0.105
p = 0.469
NLGN4Y/NLGN4X
Spearman rho = 0.040
p = 0.783
PRKY/PRKX
Spearman rho = 0.138
p = 0.340
RPS4Y1/RPS4X
Spearman rho = -0.224
p = 0.118
SRY/SOX3
Spearman rho = 0.103
p = 0.478
TBL1Y/TBL1X
Spearman rho = -0.150
p = 0.298
TMSB4Y/TMSB4X
Spearman rho = -0.196
p = 0.173
USP9Y/USP9X
Spearman rho = -0.001
p = 0.994
UTY/KDM6A
Spearman rho = 0.106
p = 0.465
ZFY/ZFX
Spearman rho = 0.158
p = 0.272
PCDH11Y/PCDH11X
Spearman rho = -0.343
p = 0.015
TGIF2LY/TGIF2LX
Spearman rho = 0.066
p = 0.651
HSFY/HSFX
Spearman rho = 0.118
p = 0.415
RBMY/RBMX
Spearman rho = 0.035
p = 0.810
TSPY/TSPYL2
Spearman rho = 0.041
p = 0.778
VCY/VCX
Spearman rho = -0.153
p = 0.289
Estimated TPM
Simulated X-homolog TPM
0
5
10
Estimated X-homolog expression level
Estimated Y-homolog expression level in the same simulated RNA-seq library

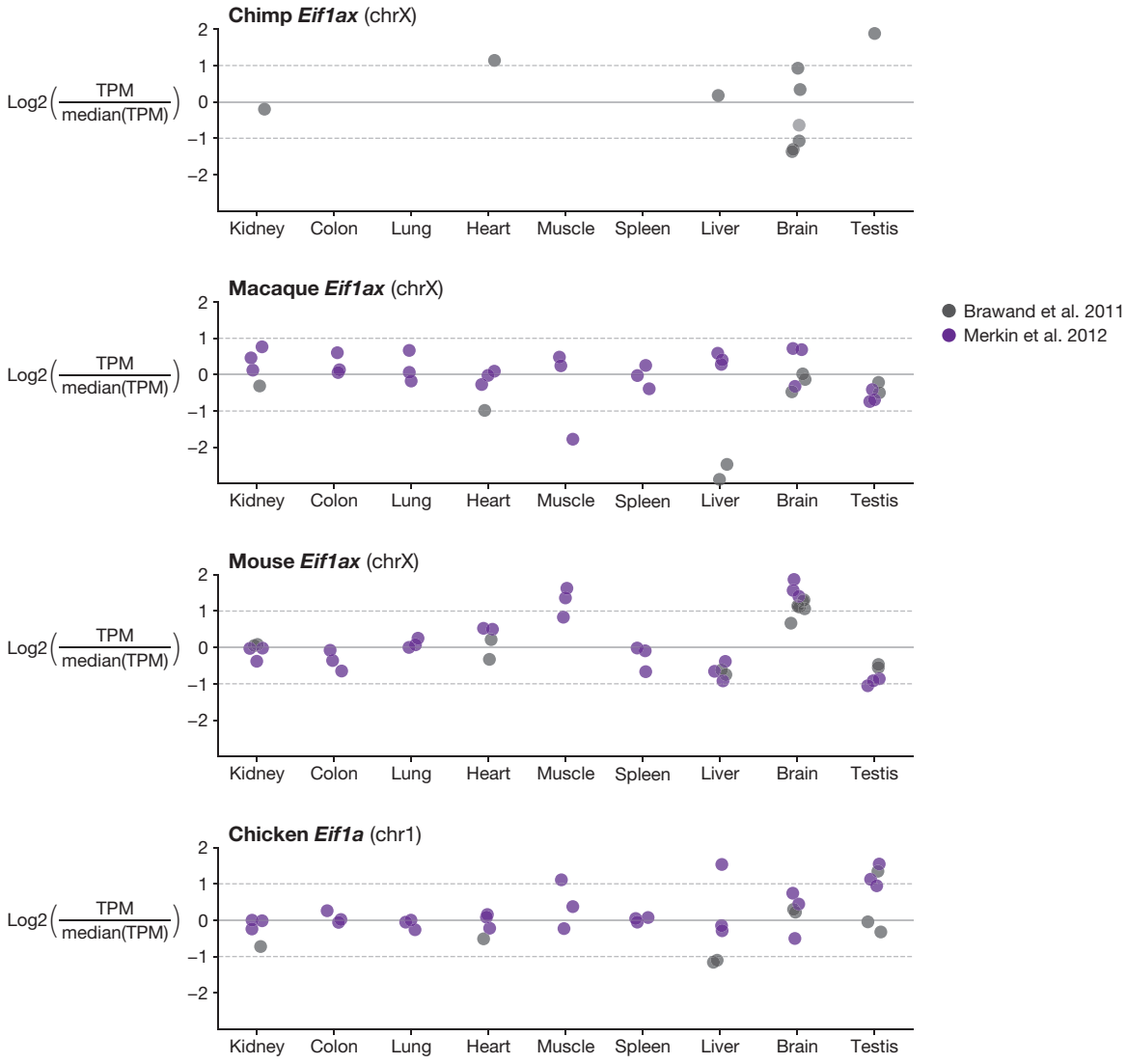
< Supplemental Fig. S10. The expression levels of MSY genes and their corresponding X-linked homologs are independently estimated. 50 simulated RNA-seq libraries were generated. In each, the expression level of the MSY gene was set to 5 TPM and its X-linked homolog was set to a random value between 0 and 10 TPM. The estimated expression levels of the X and Y homologs in each library are shown as a pair of red (X) and blue (Y) points. The correct expression level of 5 TPM was estimated for MSY genes irrespective of the expression levels of their X-linked homologs.



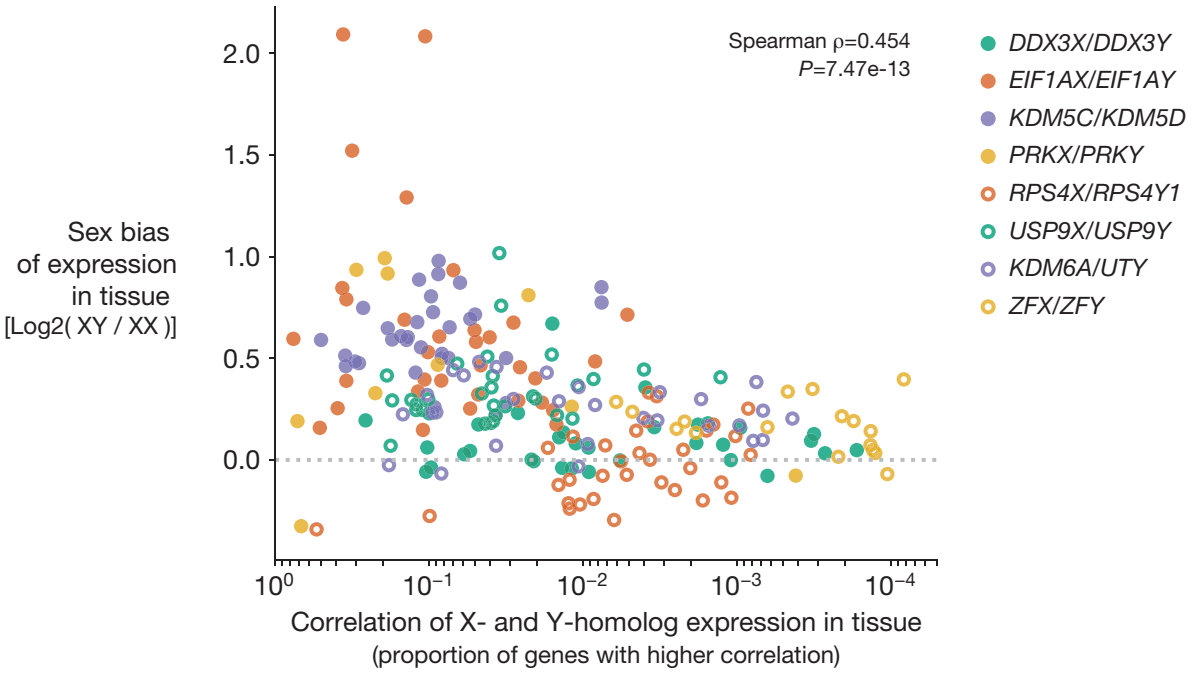
Supplemental Fig. S11. Similarity of EIF1AX and EIF1AY proteins. Amino-acid sequence of human EIF1AX (positions 1 – 75 of 144) aligned with sequences of huamn EIF1AY and EIF1AX/EIF1AY homologs around the single position at which human EIF1AX and EIF1AY differ (position 50, yellow). "." indicates identity to human EIF1AX, shown at top. Position 50 is the only position in this region at which amino-acid substitutions are found among vertebrates.



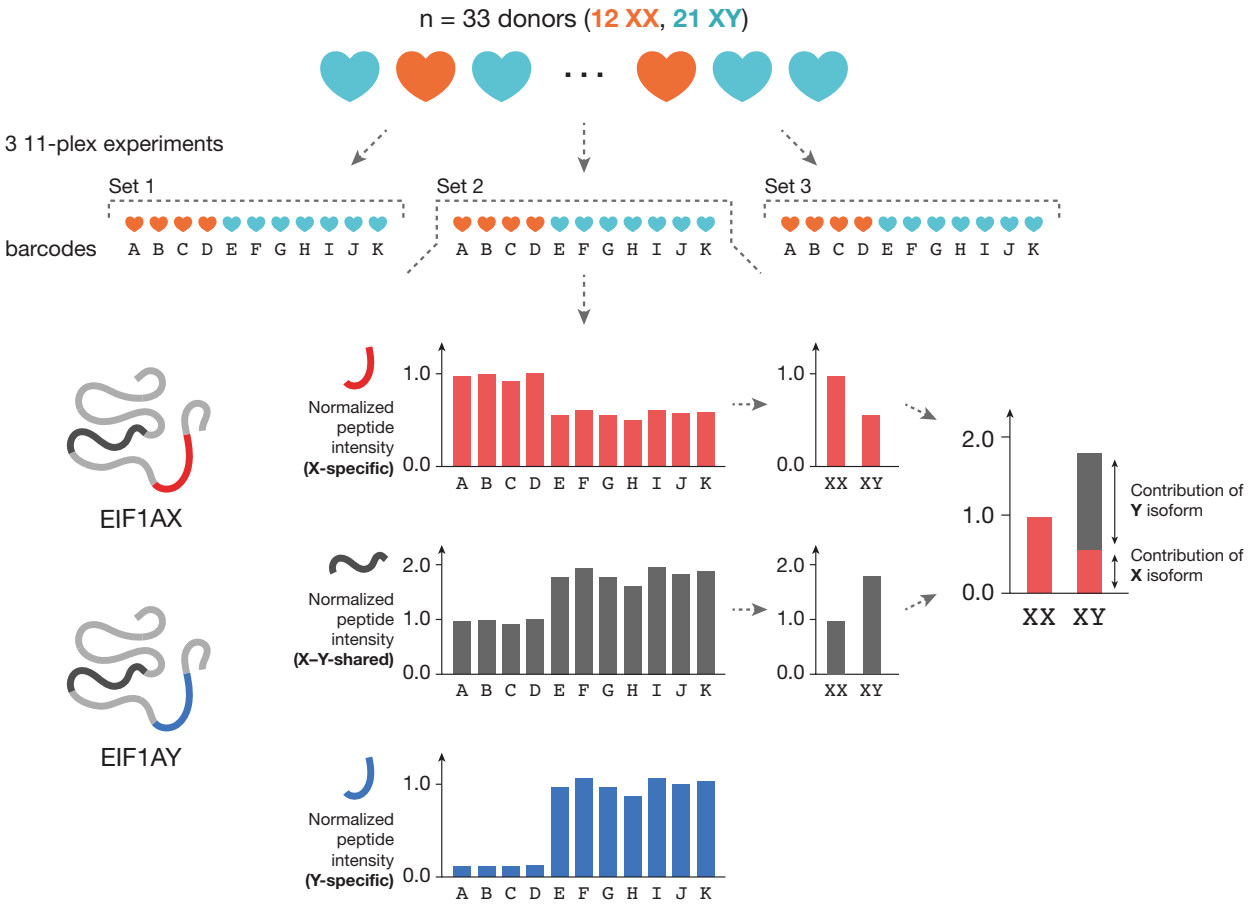
Supplemental Fig. S12. Replication of EIF1AX/EIF1AY expression pattern across human tissues using Human Protein Atlas (HPA) RNA-seq data. (**A**) Y/X expression ratio for *EIF1AX/Y* in individual XY tissue samples from the HPA dataset. Highlighted box (gray) shows tissues with notable X–Y expression divergence (e.g., as shown in Fig. 3). (**B**) Y/X expression ratio for *EIF1AX/Y* in each of 16 tissues estimated using data from GTEx (*x*-axis) and HPA (*y*-axis). (**C**) Expression of *EIF1AX* (tan) in XX samples (left) vs. summed expression of *EIF1AX* (tan) and *EIF1AY* (blue) in XY samples (right) in HPA tissue samples collected from both sexes. Expression is normalized to mean expression level in XX samples from each tissue; error bars show min and max observed values. For each tissue, the number of samples from female and male samples, respectively, are given in parentheses. For comparison with *EIF1AX/Y*, similar plots are shown for *DDX3X/Y* (**D** – **F**) and *ZFX/Y* (**G** – **I**); figure legends as in **A** – **C**.



Supplemental Fig. S13. Expression patterns of *EIF1AX* orthologs, which retain intact miR-1 target sites. Each point shows the expression level of an *EIF1AX* ortholog in a single tissue sample collected from a male donor. Within each species, expression levels are normalized to the median expression level observed across the samples. Fewer points are shown for chimpanzee because this species was not analyzed in Merkin et al. 2012.



Supplemental Fig. S14. Highly co-expressed X–Y gene pairs show little sex-biased expression. Each point compares the degree to which the X and Y homologs of an X–Y pair show correlated expression in one tissue (*x*-axis) vs. the sex-biased expression of the X–Y pair in the same tissue (*y*-axis). The degree of correlated expression between gene A and gene B is measured as the proportion of genes in the genome more correlated in expression with gene A than gene B (Methods). Highly co-expressed X–Y pairs – in which the X and Y homologs are likely tightly co-regulated – show weak sex-biased expression. By contrast, when the expression of a pair of X and Y homologs is uncorrelated – indicating their regulation has likely diverged – more prominent sex-biased expression is observed.



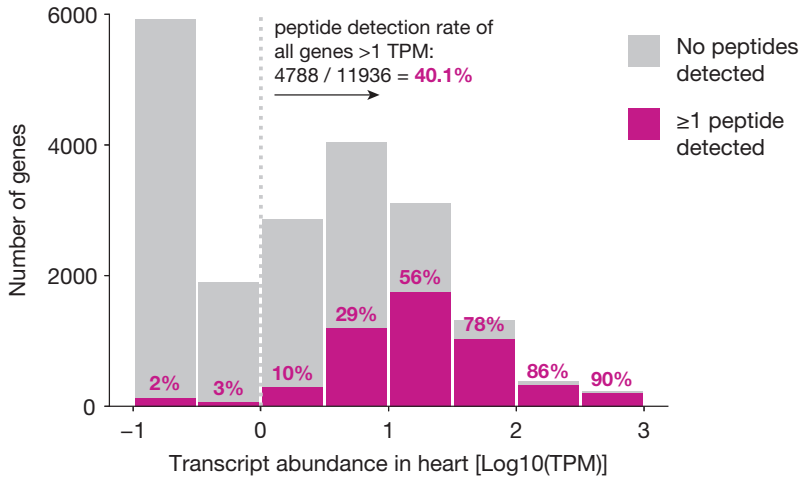
Supplemental Fig. S15. Strategy for estimating X and Y isoform expression from multiplexed proteomics data. 12 XX and 21 XY heart (left ventricle) tissue samples were analyzed by mass spectrometry in three 11-plex experiments, consisting of 4 XX and 7 XY samples each. Within each 11-plex experiment, isobaric tandem mass tags (TMTs) are used as barcodes to quantify the relative abundance of a given peptide in each sample. Peptides from the X and Y protein isoforms of a given X–Y pair can match the sequences of both proteins (dark gray; X–Y-shared) or can be specific to the X (red) or Y (blue) isoform. Y-specific peptides are used to confirm the presence of the Y isoform in the sample. X-specific and X–Y-shared peptides are used to assess sex biases in expression; information from these two peptide classes is then integrated to infer the relative contribution of X and Y isoforms to overall expression in XY individuals.

Supplemental Fig. S16. Recovery of peptides from DDX3X, DDX3Y, RPS4X, and RPS4Y1. Amino-acid sequence of DDX3X/Y (**A**) and RPS4X/Y1 (**B**): 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.

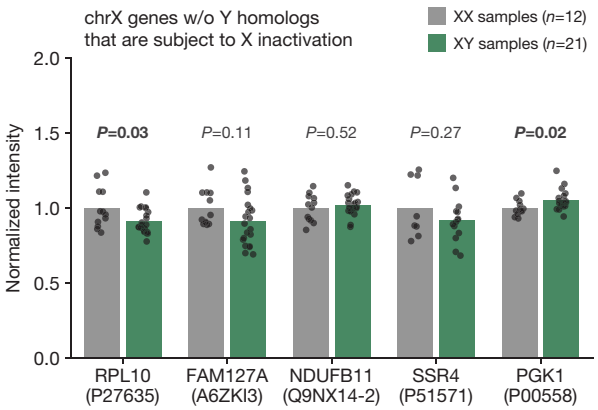
A

	Gene	Heart TPM
≥1 peptide detected	*RPS4X*	204.4
	EIF1AY	101.6
	RPS4Y1	95.6
	DDX3X	35.6
	EIF1AX	18.6
	DDX3Y	11.7
	USP9X	11.2
no peptides detected	*KDM5D*	10.1
	KDM5C	9.8
	USP9Y	4.6
	ZFX	3.2
	KDM6A	2.8
	UTY	2.3
	PRKY	1.9
	ZFY	1.9
	PRKX	1.3
	NLGN4X	0.5
	NLGN4Y	0.2

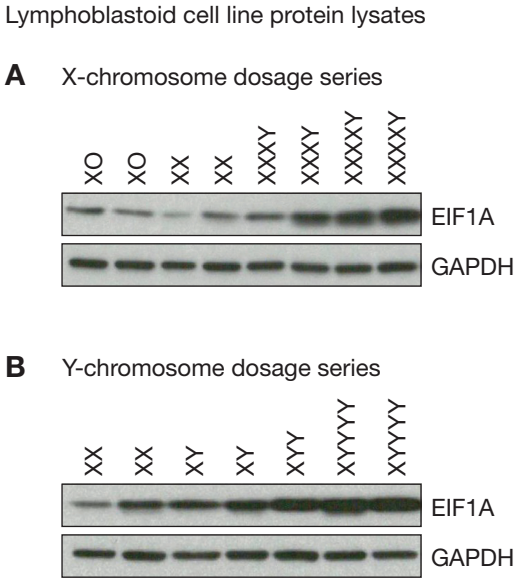
B



Supplemental Fig. S17. Peptides from highly transcribed genes were more often detected by mass spectrometry. (**A**) The median transcript expression level (TPM) in heart (left ventricle) samples from XY donors is given for the X- and Y-linked homologs of the 9 most widely expressed X–Y gene pairs. Peptides were detected for the seven genes that had the highest transcript expression levels. (**B**) All genes genome-wide were grouped into eight bins based on their transcript expression level in the heart. The percentage of genes in each bin with ≥1 peptide detected by mass spectrometry is shown in pink.



Supplemental Fig. S18. XX and XY expression of non-X–Y pair proteins. Relative expression of five non-X–Y-pair proteins in XX (gray) and XY (green) samples, used as negative controls. Each protein corresponds to an X-chromosome gene that is subject to X inactivation in XX cells and is not sex-biased at the transcript level.



Supplemental Fig. S19. EIF1A antibody recognizes both EIF1AX and EIF1AY proteins. Western blots showing signal detected by an EIF1A antibody in protein lysates prepared from human lymphoblastoid cell lines with various numbers of X (**A**) or Y (**B**) chromosomes. GAPDH was used as a loading control. Because *EIF1AX* escapes X-chromosome inactivation (XCI), EIF1AX expression should increase in cells with additional numbers of X chromosomes. Because the Y chromosome is not subject to a program of epigenetic silencing analogous to XCI, expression of EIF1AY should also increase in cells with additional Y chromosomes. As shown, signal detected by the EIF1A antibody increases in cells with additional X chromosomes (**A**) and in cells with additional Y chromosomes (**B**), implying the EIF1A antibody recognizes both EIF1AX and EIF1AY.

SUPPLEMENTAL TABLES and FILES

Supplemental Tables and Files, described below, are available upon request and at pagelab.wi.mit.edu/publications.

Supplemental Table S1. MSY genes and their X-linked homologs analyzed in this study.
Accession numbers for genes analyzed in this study, correspondences between X and Y homologs, and membership of genes in multi-copy gene families.

Supplemental Table S2. Gene mappability vs. expression level when multi-mapping reads are included or discarded.
Estimated mappability (fraction of reads that are uniquely mappable) for genes across the genome, together with expression levels in GTEx with/without multi-mapping reads. These values are used in Fig. 1B–C and Supplemental Fig. 1.

Supplemental Table S3. Median expression level (TPM) of MSY genes among samples from each tissue.
Estimated expression level of each MSY gene in each tissue, as plotted in Fig. 1D.

Supplemental Table S4. Median expression level (TPM) of X homologs of MSY genes among XY samples from each tissue.
Estimated expression level of each X homolog in each tissue in XY individuals, as plotted in Supplemental Fig. S6.

Supplemental Table S5. Median expression level (TPM) of X homologs of MSY genes among XX samples from each tissue.
Estimated expression level of each X homolog in each tissue in XX individuals.

Supplemental Table S6. Median Y/X expression ratios for X–Y gene pairs among samples from each tissue.
Y/X expression ratio estimated for each X–Y gene pair in each tissue, as plotted in Fig. 2B–C.

Supplemental Table S7. Co-expression of the X and Y homologs of each widely expressed X–Y gene pair in each tissue.
Statistics describing degree of X–Y co-expression for each pair and tissue, as presented in Fig. 2H.

Supplemental Table S8. Results of tissue-vs-tissue differential expression analysis for X and Y homologs.
Statistics describing differential expression of each gene across tissues, as presented in Fig. 3.

Supplemental Table S9. miRNA target sites in X–Y pair genes.
Candidate miRNA sites identified in X–Y pair genes along with context+ scores and (for X homologs only) context++ scores.

Supplemental Table S10. Accession numbers for EIF1AX/EIF1AY homolog sequences used in alignments.

Supplemental Table S11. Sex-biased expression of X–Y pairs, using the sum of X- and Y-homolog expression in XY samples.
Statistics describing degree of sex-biased expression for each X–Y gene pair in each tissue.

Supplemental Table S12. Quantification of peptides in each heart tissue sample.
Signal/noise values for each peptide detected in each sample across all three 11-plex experiments.

Supplemental Table S13. Relative protein abundances of X and Y isoforms in heart tissue samples.
Estimated protein abundances for X and Y isoforms in each sample based on analysis of X-specific and X–Y-shared peptides.

Supplemental Table S14. Initial estimates of protein abundance in 33 human heart tissue samples, using reductionist approach to assign peptides to proteins (refer to Supplemental Table S13 for levels of X and Y isoforms).
Estimated protein abundances proteome-wide using razor-peptide method.

Supplemental File S1. GTEx RNA-seq samples included in analyses after all filtering steps applied.

Supplemental File S2. Primer and duplex sequences for miRNA transfection experiment.

Supplemental File S3. GTEx heart samples analyzed by quantitative proteomics.

Chapter 3. Conclusion

CONCLUDING REMARKS

The work described in Chapter 2 advances our understanding of the non-reproductive activities of human Y-chromosome genes in a number of important ways. First, setting aside the Y-chromosome genes that are indeed predominantly expressed in the testis, the Y chromosome's widely expressed genes do not show any stereotyped pattern of expression, such as biased expression towards reproductive tissues. Although two genes (*DDX3Y* and *ZFY*) do show approximately twofold higher expression in the testis than in other tissues, others are lowly expressed in the testis. Overall, their expression patterns are more different than they are similar, with each gene's expression pattern likely following its molecular function. Thus, the Y-linked ribosomal protein gene *RPS4Y1* is co-expressed with other ribosomal protein genes across samples and tissues, rather than with other Y-chromosome genes. Many other Y-chromosome genes are tightly co-expressed with their X-linked homologs. This strongly implies that their expression is regulated in an orderly manner and remains under strong, gene-specific constraints. This view of the Y chromosome is not that of a "male" chromosome, but of a chromosome that encodes a variety of dosage-sensitive genes.

Second, this expanded survey of Y-chromosome gene expression provides a clearer answer to the question of whether Y-chromosome genes show lower expression

than their X homologs. To some extent, each X–Y pair shows a characteristic Y-to-X expression ratio (e.g., *RPS4Y1* appears to almost always show lower expression than *RPS4X*; *KDM5D* and *KDM5C* typically show similar expression levels). The most common observation is that a Y-chromosome gene shows lower expression than its X homolog, but this is not always the case. Sometimes the Y homolog shows much higher expression, due to differential X- and Y-homolog regulation by tissue-specific regulatory factors (e.g., microRNAs (miRNAs)). Higher Y-homolog at the transcript expression level can then be translated to the protein level. Although we found *EIF1AY*'s expression pattern—specifically, its upregulation in the heart—to be conserved in other primates, it remains to be seen whether the human Y-chromosome gene expression patterns are conserved more broadly across mammals.

Third, the observation that Y-chromosome genes sometimes show up-regulated expression in non-reproductive tissues (at both transcript and protein levels, as the result of defined primary-sequence changes) is itself a surprise. The possibility that individual Y-chromosome genes will be found to be highly expressed in other tissues or cell types should not be discounted.

Finally, with respect to differences in expression between XX and XY individuals, we find that—in the typical case—the combined expression of X and Y homologs (at the transcript level) in XY individuals is slightly higher than the biallelic expression of the X homolog in XX individuals. It is possible that the over-expression of the Y homolog compensates for some deficiency. For example, the Y-encoded transcript might be less efficiently translated, or the Y protein isoform might show weakened activity, necessitating a higher transcript expression level. This would suggest that expression of the Y homolog equalizes the effective gene dose between XX and XY individuals in most cases. However, there are two scenarios in which an inequality between XX and XY would manifest. The first is the case where the X and Y protein isoforms have diverged (partially) in function. The second, which has most clearly been documented in Chapter 2, is when the X and Y homologs have diverged starkly in expression, possibly in a tissue-

specific manner, leading to higher overall expression of the X–Y gene pair in either XX or XY individuals (e.g., higher EIF1AX/EIF1AY in the XY heart). These exceptional cases are particularly worthy for follow-up studies of XX–XY differences.

FUTURE DIRECTIONS

Expanding our quantitative understanding of human Y chromosome gene expression

The findings presented in Chapter 2 show that individual Y-chromosome genes show varied and distinct expression patterns that do not necessarily fit stereotyped expectations for a "male-specific" chromosome. For example, the starkly elevated expression of *EIF1AY* in heart tissue could not have been predicted from evolutionary theory. It seems reasonable to suspect then that other such patterns of expression will be uncovered as additional tissues, specific cell types, and developmental stages are surveyed. A good example of this is provided by Y-linked *TBL1Y*. We found it to be very lowly expressed compared to its X homolog *TBL1X* in all tissues (see Chapter 2), but it was recently found to be more highly expressed than *TBL1X* in cells of the inner ear (Di Stazio et al. 2018). The existence of rare cell types where particular Y-chromosome genes are highly expressed could explain the somewhat mysterious survival of genes like Y-linked *TMSB4Y*. *TMSB4Y* remains intact on the Y-chromosomes of humans and macaques (Hughes et al. 2012; Skaletsky et al. 2003; Bellott et al. 2014) despite its extremely low expression compared to its X homolog *TMSB4X* (~1000-fold) in adult human tissues (see Chapter 2). One possibility is that *TMSB4Y* is nearly "dead", but another is that robust *TMSB4Y* expression is restricted to a specific group of cells, and selection to preserve its expression in those cells is what maintains the survival of the gene. Our analyses of the GTEx dataset represent a major expansion in the number and diversity of biological contexts in which Y-chromosome gene expression has been measured, but with projects like the Human Cell Atlas (Regev et al. 2017), it should be possible to construct an even more complete and unbiased picture of Y-chromosome gene expression.

As differences between X- and Y-homolog expression are documented, it will be interesting to search for the primary-sequence differences that underlie them. Differential microRNA (miRNA) targeting might underlie other cases of X–Y expression divergence beyond *EIF1AX*/*EIF1AY* in the heart, though we have yet to identify an example as clear cut (i.e., where a broadly conserved, tissue-specific miRNA has an efficacious target site in one homolog but not the other, and the X and Y homologs diverge in expression in this tissue). Transcription factors (TFs) and some RNA-binding proteins (RBPs) (i.e., those that increase transcript stability) might also contribute to differences in X and Y homolog expression. Computationally predicting the targets of TFs and RBPs is a challenging task, but large-scale resources that empirically document the targets of these and other regulatory factors continue to grow and will likely improve with advances in machine learning (Dominguez et al. 2018; McGeary et al. 2019).

An important complement to these efforts will be measuring expression at the protein level. Developing antibodies specific for X and Y protein isoforms continues to be a worthwhile goal, as antibodies are multi-purpose reagents that can be used for semi-quantitative protein detection as well as immunoprecipitation. However, mass spectrometry holds more promise for obtaining unbiased quantitative measurements for all X and Y isoforms, assuming appropriate analytic approaches are used to distinguish X- and Y-derived peptides. The *ad hoc* approach taken in Chapter 2 represents one solution, but, in future studies, it would be preferable to use a method that attempts to assign all "multi-mapping peptides" to their most likely parent proteins under a unified statistical framework (Malioutov et al. 2018).

In Chapter 2, we found that the protein expression levels of DDX3Y, EIF1AY, and RPS4Y1 (relative to the levels of the corresponding X isoforms) were lower than expected based on our estimates at the transcript level. Experienced gained while conducting this research cautions against making broad generalizations about Y protein expression, until the levels of additional proteins can be obtained in this and other tissues. Nevertheless, there are four (or more) possible explanations for the lower-than-expected protein

expression of DDX3Y, EIF1AY, and RPS4Y1 in the heart. First, their transcript expression levels might, in effect, be overestimated in our analyses of the RNA-seq data. Compared to their X homologs, Y-chromosome genes might produce a greater fraction of alternatively or incorrectly spliced transcripts that are not translated or whose protein products are very unstable. Some reads from these abortive transcripts will be counted and increase the estimates of transcript expression, even though they are not indicative of protein production. (This "messy splicing" hypothesis is motivated by my own anecdotal observations.) Second, Y protein levels might be *under*estimated, because the protein products encoded by alternatively spliced Y transcripts are not annotated in the peptide databases against which mass spectrometry data are searched. Third, Y-derived transcripts encoding the canonical protein isoforms might be inefficiently translated, a possibility that could be tested with ribosome footprint profiling data. Finally, the Y-encoded protein isoforms might themselves be less stable than their corresponding X-encoded counterparts. Assuming the lower transcript levels seen, on average, for Y-chromosome genes are the result of Y-specific mutations that partially degraded motifs needed for transcription[3], then it seems likely that Y-specific mutations will also have interfered with sequence motifs governing splicing, translational efficiency, and protein stability.

Head-to-head comparisons of X- and Y-homolog function

As described in Chapter 1, analyses of the cellular and organismal phenotypes conferred by mutations in X homologs provide useful clues about the ways in which X and Y homologs might differ. However, a fuller understanding would be provided by directly comparing their functions through functional genomic experiments. For a given X–Y

[3] Formally, the lower expression of Y-chromosome genes relative to their X homologs could be an epigenetic phenomenon. However, compared to their X homologs, Y-chromosome genes appear to evolve under relaxed purifying selection (Wilson and Makova 2009), suggesting their reduced expression is at least partially encoded in the Y chromosome's primary sequence.

pair, the X and Y homologs could be separately knocked down in cells; the genome-wide transcriptional responses from perturbing X or Y would then be measured and compared. Studying the effects of *increasing* X- or Y-homolog expression are also of interest, given the expression patterns observed in Chapter 2. Care must be taken to ensure that the perturbations to X and Y homologs are both highly specific and comparable in degree. Additionally, epitope-tagged versions of the X and Y homologs could be generated in cells or mice—ideally by inserting the tag at the endogenous locus—and used to compare their genome-wide occupancy (e.g., by ChIP-seq) or protein binding partners. At the time of writing, I have yet to come across any published study of this kind or indeed any study that has specifically measured the function of widely expressed Y-chromosome gene on a genome-wide scale. These studies would help to clarify why Y-chromosome genes sometimes fail to fully compensate for loss of their X-linked homologs and, furthermore, if Y protein isoforms have acquired new functions.

Forward genetics for the Y chromosome

The experiments described above will help to elucidate the nature of Y-chromosome genes and their encoded proteins at the molecular level, but, ultimately, connecting this understanding to phenotype will require phenotype-driven approaches. A promising development is the use of high-throughput, cellular phenotype screens, in which CRISPR-based approaches are used to perturb virtually all genes in the genome in single cells in parallel (Dixit et al. 2016). As demonstrated by Wang et al., these screens are capable of retrieving hits to Y-chromosome genes (*DDX3Y* and cellular proliferation) (Wang et al. 2015); still, some large-scale studies are already excluding Y- (and X-) chromosomal sequences (Tsherniak et al. 2017). There is little justifiable reason to exclude X- and Y-chromosome sequences from these screens. The Y-chromosome genes of greatest interest are all found within single-copy regions. Moreover, all promising hits from these screens (no matter their genomic location) will require follow-up investigation, providing the opportunity to exclude artifacts.

Human genetics is another promising, albeit more challenging, future area for the Y chromosome. The Y chromosome has been excluded from genome-wide association studies because its male-specific region does not normally recombine with a homologous partner during meiosis. Thus, recombination cannot be used to fine-map a signal of genetic association to a particular region of the Y chromosome. In addition, the Y chromosome has a distinct population structure compared to autosomal sequences due to its father-to-son inheritance (Jobling and Tyler-Smith 2017), which must be accounted for in any genetic association study. However, recent work suggests that associating Y-chromosome variation with quantitative traits and identifying the likely causal gene might be possible. In a series of two studies, European individuals carrying Y chromosomes of a particular lineage, haplogroup I1, were found to have increased risk of cardiovascular disease (Charchar et al. 2012; Eales et al. 2019). *UTY* was identified as the gene likely underlying this association, first by observing that *UTY* expression was altered in haplogroup I1 individuals in the relevant cell type, and then with experimental studies. A similar approach could be extended to other traits. Because a relatively small number of Y-chromosome genes is expressed in non-reproductive tissues, the list of candidate genes underlying any association would be small. Because a diverse set of Y chromosomes has been sequenced (Poznik et al. 2016), the set of mutations uniquely shared by Y chromosomes of any haplogroup can be enumerated. Even at autosomal loci, genetic association signals can rarely be narrowed to individual genes, meaning numerous clever approaches have been developed to assess the potential causality of candidate mutations (Spain and Barrett 2015; Farh et al. 2014; Zhou et al. 2018). Many of these could be applied to the Y chromosome. Despite the additional effort required to perform Y-haplogroup association studies, analyzing large-scale genotype-phenotype datasets (Bycroft et al. 2018) would likely reveal many new and surprising insights about Y-chromosome genes.

REFERENCES

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, Koutseva N, Zaghlul S, Graves T, Rock S, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* 508: 494–499.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562: 203–209.

Charchar FJ, Bloomer LD, Barnes TA, Cowley MJ, Nelson CP, Wang Y, Denniff M, Debiec R, Christofidou P, Nankervis S, et al. 2012. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. *Lancet* 379: 915–922.

Di Stazio MD, Collesi C, Vozzi D, Liu W, Myers M, Morgan A, Adamo PAD, Girotto G, Rubinato E, Giacca M, et al. 2018. TBL1Y: a new gene involved in syndromic hearing loss. *Eur J Hum Genet* 354: 466–474.

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167: 1853-1866.e17.

Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Nostrand ELV, Pratt GA, et al. 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* 70: 854-867.e9.

Eales JM, Maan AA, Xu X, Michoel T, Hallast P, Batini C, Zadik D, Prestes PR, Molina E, Denniff M, et al. 2019. Human Y Chromosome Exerts Pleiotropic Effects on Susceptibility to Atherosclerosis. *Arterioscler Thromb Vasc Biol* 39: 2386–2401.

Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, et al. 2014. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518: 337–43.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483: 82–86.

Jobling MA, Tyler-Smith C. 2017. Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet* 18: 485–497.

Malioutov D, Chen T, Airoldi E, Jaffe JD, Budnik B, Slavov N. 2019. Quantifying homologous proteins and proteoforms. *Mol Cell Proteomics* 18: 162–168.

McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, Bartel DP. 2019. The biochemical basis of microRNA targeting efficacy. *Science Y* 366: eaav1741. doi:10.1126/science.aav1741

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Sayres MAW, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* 48: 593–599.

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, Carninci P, Clatworthy M, et al. 2017. The Human Cell Atlas. *eLife* 6: e27041. doi:10.7554/eLife.27041

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

Spain SL, Barrett JC. 2015. Strategies for fine-mapping complex traits. *Hum Mol Genet* 24: R111-9. doi:10.1093/hmg/ddv260

Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, et al. 2017. Defining a Cancer Dependency Map. *Cell* 170: 564-576.e16.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* 350: 1096–1101.

Wilson MA, Makova KD. 2009. Evolution and survival on eutherian sex chromosomes. *PLoS Genet* 5: e1000568. doi:10.1371/journal.pgen.1000568

Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* 50: 1171–1179.

APPENDIX A. *Meioc* maintains an extended meiotic prophase I in mice

Authors:

Y. Q. Shirleen Soh[1,2], Maria M. Mikedis[1], Mina Kojima[1,2], Alexander K. Godfrey[1,2], Dirk de Rooij[1], David C. Page[1,2,5]

Affiliations:

[1] Whitehead Institute, Cambridge, MA, USA
[2] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, USA

Author Contributions:

Conceptualization: YQSS MMM MK DGdR DCP. Formal analysis: YQSS MMM MK AKG DGdR. Funding acquisition: DCP. Investigation: YQSS MMM MK DGdR. Methodology: YQSS MMM MK AKG. Supervision: DCP. Visualization: YQSS MMM MK AKG. Writing – original draft: YQSS MMM. Writing – review & editing: YQSS MMM MK AKG DGdR DCP.

(AKG designed and oversaw the computational analysis of RNA immunoprecipitation-and-sequencing (RIP-seq) data and assisted with data visualization.)

Published as:

Soh, Y. Q. S., Mikedis, M. M., Kojima, M., Godfrey, A. K., de Rooij, D. G., & Page, D. C. (2017). *Meioc* maintains an extended meiotic prophase I in mice. *PLoS Genet*, *13*(4), e1006704. http://doi.org/10.1371/journal.pgen.1006704

PLOS | GENETICS

RESEARCH ARTICLE

Meioc maintains an extended meiotic prophase I in mice

Y. Q. Shirleen Soh[1,2], Maria M. Mikedis[1], Mina Kojima[1,2], Alexander K. Godfrey[1,2], Dirk G. de Rooij[1], David C. Page[1,2,3]*

1 Whitehead Institute, Cambridge, MA, United States of America, **2** Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, United States of America, **3** Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, United States of America

* dcpage@wi.mit.edu

OPEN ACCESS

Citation: Soh YQS, Mikedis MM, Kojima M, Godfrey AK, de Rooij DG, Page DC (2017) *Meioc* maintains an extended meiotic prophase I in mice. PLoS Genet 13(4): e1006704. https://doi.org/10.1371/journal.pgen.1006704

Editor: Paula E. Cohen, Cornell University, UNITED STATES

Received: December 20, 2016

Accepted: March 20, 2017

Published: April 5, 2017

Data Availability Statement: All sequencing data are available from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE90702 and NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) under accession number SRP094112. Mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD005473.

Funding: YQSS was supported by a Howard Hughes Medical Institute International Student

Abstract

The meiosis-specific chromosomal events of homolog pairing, synapsis, and recombination occur over an extended meiotic prophase I that is many times longer than prophase of mitosis. Here we show that, in mice, maintenance of an extended meiotic prophase I requires the gene *Meioc*, a germ-cell specific factor conserved in most metazoans. In mice, *Meioc* is expressed in male and female germ cells upon initiation of and throughout meiotic prophase I. Mouse germ cells lacking *Meioc* initiate meiosis: they undergo pre-meiotic DNA replication, they express proteins involved in synapsis and recombination, and a subset of cells progress as far as the zygotene stage of prophase I. However, cells in early meiotic prophase—as early as the preleptotene stage—proceed to condense their chromosomes and assemble a spindle, as if having progressed to metaphase. *Meioc*-deficient spermatocytes that have initiated synapsis mis-express CYCLIN A2, which is normally expressed in mitotic spermatogonia, suggesting a failure to properly transition to a meiotic cell cycle program. MEIOC interacts with YTHDC2, and the two proteins pull-down an overlapping set of mitosis-associated transcripts. We conclude that when the meiotic chromosomal program is initiated, *Meioc* is simultaneously induced so as to extend meiotic prophase. Specifically, MEIOC, together with YTHDC2, promotes a meiotic (as opposed to mitotic) cell cycle program via post-transcriptional control of their target transcripts.

Author summary

Meiosis is the specialized cell division that halves the genetic content of germ cells to produce haploid gametes. This reductive division is preceded by a preparative phase of the cell cycle, meiotic prophase I, during which several meiosis-specific chromosomal events occur. Across sexually reproducing organisms, prophase of meiosis I is dramatically longer than mitotic prophase. However, it was not known in mammals how and why meiotic prophase I is extended. We have identified a mouse mutant in which this extended prophase I is disrupted: germ cells lacking *Meioc* initiate meiosis, but prematurely proceed to metaphase. Mutant male meiotic germ cells mis-express a cell cycle regulator that is normally expressed in mitotic male germ cells, suggesting that *Meioc* is required for germ

Research Fellowship (https://www.hhmi.org/) and the Abraham Siegel Fellowship of the Whitehead Institute (http://wi.mit.edu). MMM was supported by a Lalor Foundation Postdoctoral Fellowship (http://lalorfound.org). MK was supported by a National Science Foundation (NSF) Graduate Research Fellowship (https://www.nsf.gov). AKG was supported under a research grant by Biogen. This work was funded by the Howard Hughes Medical Institute (https://www.hhmi.org/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

cells to properly transition to a meiotic cell cycle program. Biochemical analyses of proteins and transcripts that associate with MEIOC protein suggest that MEIOC may promote the transition from a mitotic to meiotic cell cycle program by post-transcriptionally regulating target transcripts. Our studies indicate that in mammals, as in other sexually reproducing organisms, meiotic prophase I must be extended to allow time for meiotic chromosomal events to reach completion.

Introduction

Meiosis is a specialized cell division program that results in the halving of parental genetic material and the production of haploid gametes. This reductive division depends on a series of chromosomal events that occur specifically during meiotic but not mitotic prophase, including the loading of meiosis-specific cohesins on sister chromatids, alignment and synapsis of homologous chromosomes, and generation of covalent linkages between homologs via recombination. These meiotic chromosomal events occur during meiotic prophase I, which takes much longer than mitotic prophase. In yeast, it has been shown that completion of these chromosomal events requires the extended prophase I: yeast meiotic prophase I lasts 3.5 hours, compared to 15 minutes for mitotic prophase [1], and premature exit from prophase I results in recombination defects and chromosome missegregation [2].

Mammals similarly have an extended prophase I. In female mice, ovarian germ cells initiate meiosis around embryonic day 13.5 (E13.5), and arrest in the penultimate stage of prophase, diplotene, around the time of birth, one week after meiotic initiation [3,4]. In male mice, cohorts of testicular germ cells initiate meiosis continuously throughout post-pubertal life, each cohort taking two weeks from initiation to completion of meiotic prophase I [5]. In contrast, the typical mitotic prophase in mammalian cells lasts only minutes [6,7].

No mechanism for extension of meiotic prophase has yet been recognized in mammals. In other organisms, the extension of meiotic prophase is accomplished by meiosis-specific modifications of the cell cycle. In yeast, exit from meiotic prophase I is postponed via the suppression of mitotic cell cycle regulators by a meiosis-specific form of the anaphase-promoting complex [2]. In worm and fly, exit from meiotic prophase I is also actively suppressed by meiosis-specific factors via translational repression of, respectively, cyclins E and A [8,9].

Since meiotic initiation in both male and female mice is governed by the retinoic acid-induced gene *Stra8* [10,11], STRA8 activity might be at least indirectly related to prolonging prophase. Ovarian and testicular germ cells express *Stra8* shortly before entering meiotic prophase I [12,13], and *Stra8* is required for the chromosomal events of meiotic prophase I, including cohesion, synapsis, and recombination [14,15]. Consistent with a pivotal role in meiotic initiation, most genes involved in meiotic prophase I depend on *Stra8* for their expression. However, *Stra8* is only transiently expressed at the time of meiotic initiation, and therefore is unlikely to be the factor responsible for maintaining meiotic prophase I. We previously identified a subset of early meiotic genes that are expressed independently or partially independently of *Stra8*, and are induced concurrently or shortly after *Stra8* [16,17]. This subset of partially *Stra8*-independent early meiotic genes includes cohesins and synaptonemal complex proteins with known meiotic functions, and also *Meioc*, an uncharacterized gene formerly named *Gm1564*.

We examined MEIOC expression and find that it is expressed throughout meiotic prophase I in both testicular and ovarian germ cells; this expression profile suggests that its function begins early and persists throughout meiotic prophase I in both sexes. We examined mice

deficient for *Meioc*, and found that *Meioc*-deficient germ cells can initiate but do not complete meiotic prophase I. Instead, germ cells that have initiated meiosis proceed prematurely to an aberrant metaphase. *Meioc*-deficient germ cells that have initiated meiosis mis-express CCNA2, which is typically expressed in mitotic spermatogonia, suggesting a failure to properly transition to a meiotic cell cycle program. We propose that *Meioc* functions continuously throughout meiotic prophase I to prevent premature exit from prophase I, likely by promoting a meiotic (as opposed to mitotic) cell cycle program. Further, MEIOC interacts with an RNA helicase, YTHDC2, and binds a common set of germ cell transcripts, suggesting that MEIOC and YTHDC2 partner to regulate these transcripts.

Our observations that *Meioc*-deficient germ cells fail to complete meiotic prophase I and instead produce numerous abnormal metaphases are concordant with a recent study [18]. However, whereas Abby and colleagues propose that this phenotype results from arrested meiotic progression, our molecular analyses of cyclin expression and MEIOC-bound transcripts lead us to an alternate interpretation of the phenotype–that the precocious metaphases observed are a result of cell cycle mis-regulation. We propose a model for meiotic prophase I as comprised of multiple subprograms: these include the chromosomal program, whereby chromosomes synapse and undergo recombination, and a meiosis-specific cell cycle program, whereby cells are maintained in an extended prophase I to allow completion of the chromosomal program.

Results

Meioc is a conserved, germ cell-specific gene expressed during male and female meiotic prophase I

We had previously identified *Meioc* (*Gm1564*) as one of the earliest and most strongly induced transcripts upon meiotic initiation in the female germline [17]. In the study presented here, we identified full-length *Meioc* homologs in almost all vertebrate genomes examined. Furthermore, we found that *Meioc*'s conserved C-terminal domain, PF15189 (previously DUF4582), is present approximately once per genome in almost all metazoan genomes examined (S1 Fig). We were unable to identify orthologs of *Meioc* or matches to PF15189 in Diptera, including *Drosophila melanogaster*, which hints at *Meioc* being replaced functionally by alternate proteins or pathways in this lineage. Next, we examined *Meioc* expression in adult tissue panels from human, mouse, rat, and chicken, and found its expression to be highly testis-specific (S2 Fig). Thus, *Meioc* is a highly conserved gene whose expression pattern across diverse species is consistent with a role in meiosis.

To determine the precise cell types in which MEIOC is expressed, we generated a rabbit polyclonal antibody to a C-terminal fragment of MEIOC, which we verified to be specific using subsequently generated *Meioc*-deficient mice (S3 Fig). Immunohistochemistry for MEIOC on adult testis sections showed that MEIOC is expressed in spermatocytes, beginning in preleptotene and extending through most stages of meiotic prophase I, including leptotene, zygotene, and pachytene, but not during diplotene and diakinesis (the final stages of meiotic prophase I) or meiotic metaphase I (Fig 1A). MEIOC is absent in spermatogonia, in post-meiotic spermatids, and in somatic cells. Subcellular localization of MEIOC during early to mid-prophase I was predominantly cytoplasmic, but by late pachytene a fraction of MEIOC was nuclear. The prolonged expression of MEIOC contrasts starkly with that of STRA8, which is similarly induced in preleptotene cells, but then rapidly down regulated once cells enter leptotene (Fig 1B).

To determine if MEIOC is expressed at similar stages of meiotic prophase in the female, we immunostained for MEIOC on fetal ovary sections (Fig 1C). MEIOC was detected by E13.5,

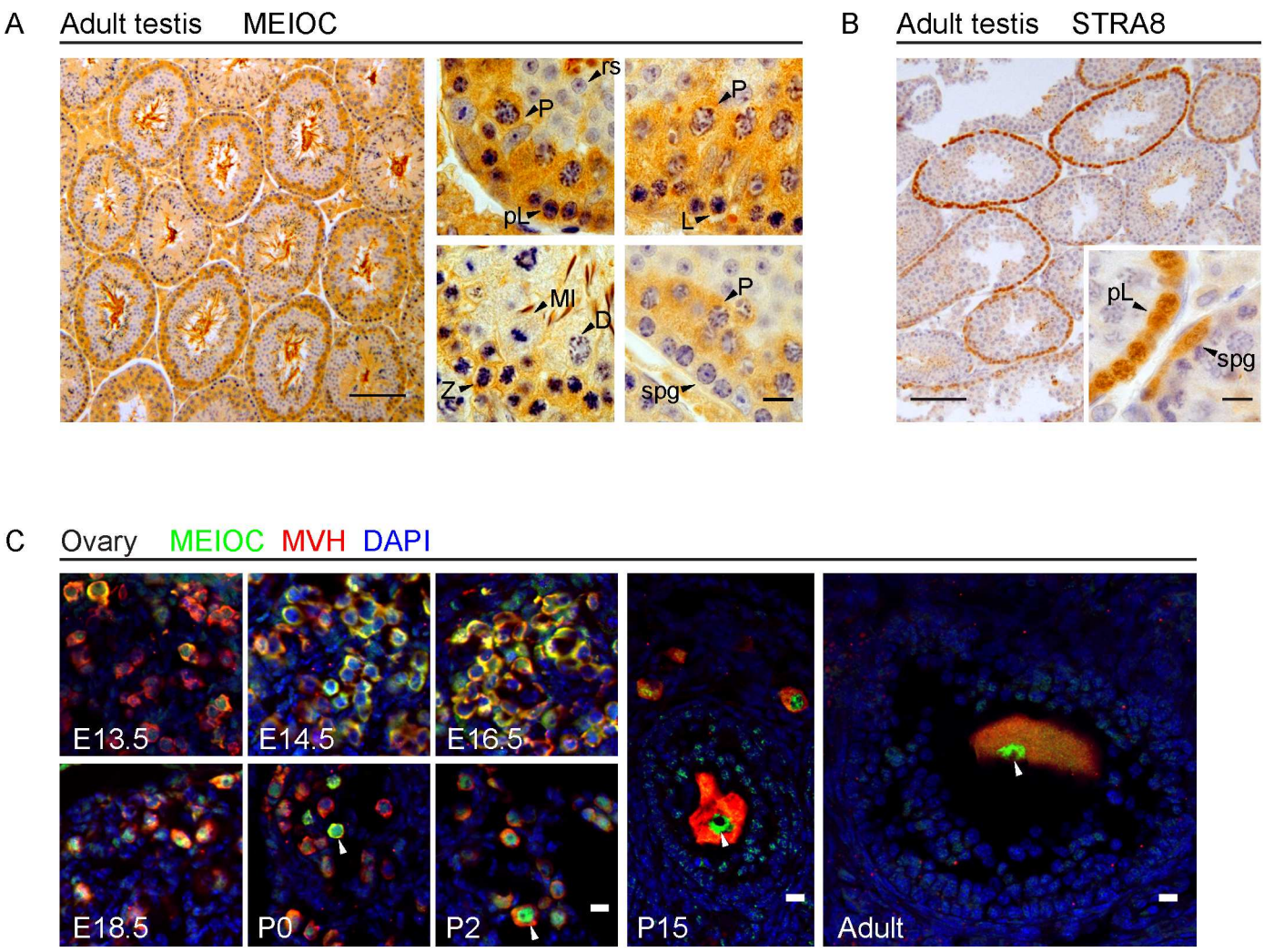


Fig 1. MEIOC is expressed throughout most of meiotic prophase I in the male and female germline. (A) Immunohistochemistry for MEIOC (brown) in adult testis, with hematoxylin counterstaining to enable identification of germ cell types by nuclear morphology [57]. Low magnification image shows MEIOC staining in the majority of meiotic cell populations. Background staining was also observed in mature sperm in center of the tubule. High magnification images show that MEIOC was detected in meiotic germ cells at preleptotene (pL), leptotene (L), zygotene (Z), and pachytene (P) stages; it was not detected in meiotic germ cells at diplotene (D) stage, in cells undergoing meiotic metaphase (MI), or in postmeiotic round spermatids (rs). Low magnification scale bar = 100 μm, high magnification scale bar = 10 μm. (B) Immunohistochemistry for STRA8 (brown) in adult testis, counterstained with hematoxylin. In contrast to MEIOC, STRA8 expression is limited to germ cells initiating meiosis (preleptotene and leptotene stages) as well as differentiating spermatogonia. Scale bar = 100 μm. (C) Immunofluorescence staining for MEIOC in ovary at E13.5, E14.5, E16.5, E18.5, P0, P2, P15, and adult (>8 weeks). Mouse Vasa Homolog (MVH) costaining identifies germ cells [58]. Nuclei stained by DAPI. MEIOC is detected in germ cells at all stages. From E13.5 to E18.5, corresponding to premeiotic to late pachytene stages, MEIOC is detected predominantly in cytoplasm. Towards the end of this period, MEIOC is detected in nucleus (arrowheads), and continues to be expressed in nuclei of germ cells at postnatal timepoints. Scale bar = 10 μm.

https://doi.org/10.1371/journal.pgen.1006704.g001

when germ cells are preparing to enter meiosis, and persists through leptotene, zygotene, pachytene stages of meiotic prophase, and dictyate arrest in the adult. In females, MEIOC is initially predominantly cytoplasmic, but becomes predominantly nuclear postnatally. Thus, MEIOC is similarly expressed in both sexes: beginning at meiotic initiation, and persisting through most of meiotic prophase in the male, and through to dictyate arrest in the female. To our knowledge, this combination of germ-cell-specific expression throughout most of meiotic prophase and predominantly cytoplasmic localization is unique to MEIOC. Our

characterization of MEIOC expression broadly agrees with results obtained using antibodies generated against full-length MEIOC [18], with a few exceptions: Abby and colleagues reported an exclusively cytoplasmic localization, whereas we observed MEIOC attaining nuclear localization towards the end of meiotic prophase.

Meioc-deficient mice are infertile

To explore the role of *Meioc* in germ cell differentiation and meiotic prophase, we generated *Meioc*-deficient mice using a targeting vector generated by the Knockout Mouse Project (KOMP) (S4 Fig). Results reported here were performed in mice 5 to 7 generations backcrossed to the C57BL/6 background unless otherwise stated. *Meioc*-deficient mice (*Meioc* -/-) had markedly smaller ovaries and testes than did wild-type control (*Meioc* +/-, and *Meioc* +/+) mice (S5A Fig), and they were infertile.

Meioc-deficient adult testes completely lacked post-meiotic germ cells (S5B and S5C Fig) and were dramatically depleted of cells in meiotic prophase I compared to littermate controls. To study progression through meiotic prophase I in a synchronous setting, we examined testes at 10 and 15 days after birth (P10 and P15) to follow the meiotic development of the first cohort of spermatogenic cells (Fig 2A). By P10 in wild-type testes, spermatogenic cells have initiated meiosis and progressed from preleptotene to the leptotene and zygotene stages of meiotic prophase. By P15, the most advanced spermatogenic cells have transitioned through zygotene and progressed to the pachytene stage. No later stages of meiosis, namely diplotene and meiotic metaphases, are observed at P10 and P15. In *Meioc*-deficient mutants, P10 and P15 testes contained cells with chromosomes condensed like those observed during metaphase. Meiotic metaphases are not expected until P20, and were not observed in our control P10 and P15 wild-type testes. *Meioc*-deficient testes also contained cells with abnormal condensed nuclei, and apoptotic cells. Mutant testes also contained leptotene and zygotene-stage spermatocytes, but were devoid of pachytene spermatocytes. TUNEL-positive cells were rare in wild-type adult testes but were abundant in *Meioc*-deficient adult testes, specifically in cells with condensed or apoptotic nuclei (S6 Fig). TUNEL staining was not observed in preleptotene, leptotene, zygotene-like, or metaphase-like cells of *Meioc*-deficient testes.

To determine if similar defects occur in females, we examined *Meioc*-deficient and wild-type ovaries. In contrast to wild-type adult ovaries, which contain follicles at various stages of maturation, adult ovaries of *Meioc*-deficient females contain no oocytes or follicles (S5D Fig). In wild-type fetal ovaries, germ cells progress from a premeiotic stage at E14.5 to zygotene or pachytene stages by E16.5 (Fig 2B). In *Meioc*-deficient ovaries, metaphase-like cells were observed as early as E14.5, and persist to E16.5. Most remaining germ cells exhibited premeiotic morphology, and few reached the leptotene or zygotene stages of prophase, even by E16.5. An independent *Meioc* knockout mouse line generated using the same KOMP vector on a mixed genetic background (C57BL/6 crossed to NMRI) exhibited similar histological phenotypes [18].

To pinpoint the timing of the primary defect in *Meioc*-deficient germ cells, we examined the time and stage at which aberrant metaphase-like cells first arise. In the testis, they are found adjacent to preleptotene, leptotene, and zygotene spermatocytes (Figs 2A and S7). The occurrence of metaphase-like cells adjacent to preleptotene cells in stage VIII tubules suggests that metaphase-like cells first arise shortly after the preleptotene stage, before recombination and synapsis would normally occur. Some germ cells proceed somewhat further, to the leptotene or zygotene stage, possibly because the primary defect that causes premature metaphase is not completely penetrant at the preleptotene stage. In the ovary, metaphase-like cells arise as early as E14.5, when most wild-type germ cells are still in the pre-meiotic stage. Thus, in both

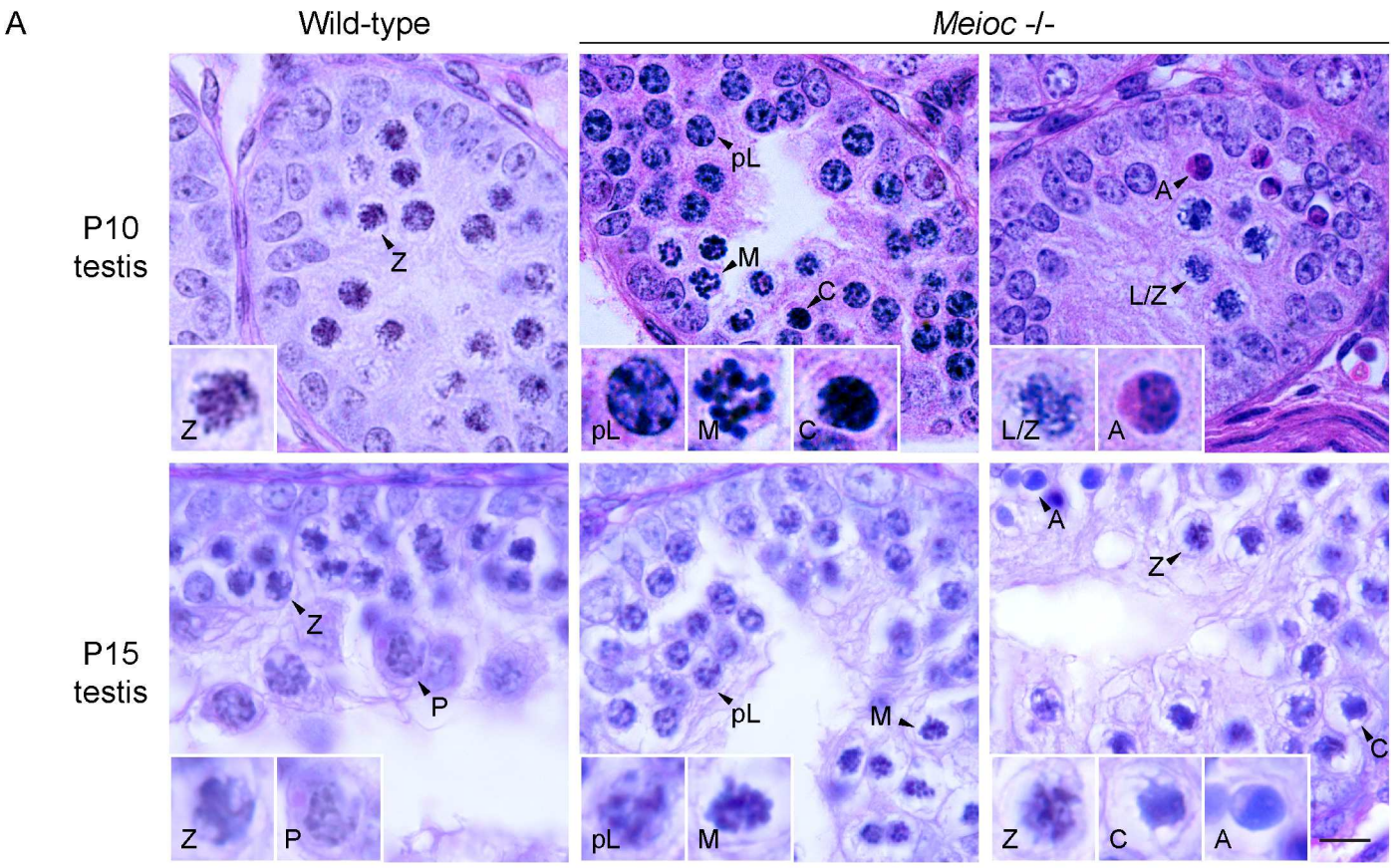


Fig 2. *Meioc*-deficient testicular and ovarian germ cells fail to progress through meiotic prophase, and instead exhibit metaphase-like chromosome condensation. (A) Hematoxylin and eosin stain of wild-type and *Meioc* -/- testes at P10 and P15. Germ cell types are identified by nuclear

morphology and position within tubules [57]. In wild-type P10 testis, germ cells have advanced to zygotene (Z) stage of meiotic prophase. In wild-type P15 testis, germ cells have advanced to an epithelial stage showing already two generations of spermatocytes: zygotene (Z) and pachytene (P) stages of meiotic prophase. In *Meioc* -/- P10 and P15 testes, germ cells in center of lumen and adjacent to preleptotene-stage cells exhibited metaphase-like chromosome condensation (M). Some germ cells also progress through leptotene (L) to late leptotene/early zygotene (L/Z). In addition, cells with condensed (C) and apoptotic nuclei (A) are observed. In *Meioc* -/- P15 testes, we observed no pachytene-stage cells. Scale bar = 10 μm. (B) Immunohistochemistry for MVH in fetal ovaries, counterstained with hematoxylin. In wild-type E14.5 ovary, germ cells exhibited a premeiotic morphology (pm). In *Meioc* -/- E14.5 ovary, some germ cells also exhibited a premeiotic morphology (pm), but other germ cells exhibited metaphase-like chromosome condensation (M). In wild-type E16.5 ovary, germ cells had progressed to late zygotene/early pachytene (P) stages of meiotic prophase. This was not observed In *Meioc* -/- E16.5 ovary; instead, germ cells either retained a premeiotic morphology (pm), or exhibited metaphase-like chromosome condensation (M). Scale bar = 10 μm.

https://doi.org/10.1371/journal.pgen.1006704.g002

sexes, the primary defect that causes premature metaphase occurs shortly after the decision to initiate meiosis, and prior to meiotic chromosomal events such as recombination and synapsis.

Meioc-deficient testicular and ovarian germ cells exhibit molecular markers of meiotic initiation and early meiotic prophase

To confirm that *Meioc*-deficient germ cells have initiated meiotic prophase I, we examined *Meioc*-deficient testes and ovaries for molecular markers of meiotic initiation and early meiotic prophase I.

Meiotic initiation requires *Stra8*, a retinoic acid-induced, germ cell-specific factor [14,15]. Germ cells in both wild-type and *Meioc*-deficient P10 testes and E14.5 ovaries express STRA8 (Fig 3A). One of the first events following the decision to initiate meiosis is premeiotic DNA replication. To detect DNA replication, we injected the thymidine analog EdU into wild-type and *Meioc*-deficient postnatal male mice, or into pregnant mothers carrying wild-type and *Meioc*-deficient fetal female mice, and harvested gonads two hours later. Both wild-type and *Meioc*-deficient P10 testes and E14.5 ovaries had numerous EdU and STRA8 double-positive cells (Fig 3A), indicating that they are able to undergo premeiotic DNA replication following the decision to enter meiosis.

Next, we examined *Meioc*-deficient germ cells for markers of the chromosomal program of meiotic prophase I, including homologous chromosome synapsis, recombination, and loading of meiotic cohesins.

We first stained for components of the synaptonemal complex: the axial element protein SYCP3, and transverse filament protein SYCP1 (Figs 3B–3D and S8A) [19,20]. In wild-type P15 testis sections and spreads, we observed SYCP3 and SYCP1 staining indicative of the leptotene, zygotene, and pachytene stages of meiotic prophase: SYCP3 staining was thin and thread-like in the leptotene stage, and progressively thickened as chromosomes synapsed through the pachytene stage. SYCP1 localized to synapsed regions of the chromosomes in zygotene and pachytene stage spermatocytes. In *Meioc*-deficient P15 testes, some germ cells exhibited SYCP3 and SYCP1 localization on chromosomes similar to leptotene and zygotene stages, but which were often accompanied by dense aggregates of SYCP3. Many germ cells displayed only SYCP3 aggregates. In the metaphase-like cells, SYCP3 localized to foci at the ends of chromosomes, likely the centromeres. This pattern of SYCP3 localization is similar to that observed in the first meiotic metaphases that normally appear beginning at P20 (Fig 3E) [21]. In wild-type E16.5 ovary sections, most germ cells were in zygotene and pachytene. In *Meioc*-deficient E16.5 ovary sections, no germ cells exhibited zygotene or pachytene-like SYCP3 staining. Instead, *Meioc*-deficient germ cells had either leptotene-like SYCP3 staining with some SYCP3 aggregates, or only SYCP3 aggregates (Figs 3B and S8A).

We next assayed *Meioc*-deficient cells for markers of meiotic recombination. Recombination is initiated by the formation of DNA double-strand breaks (DSBs) that are repaired by

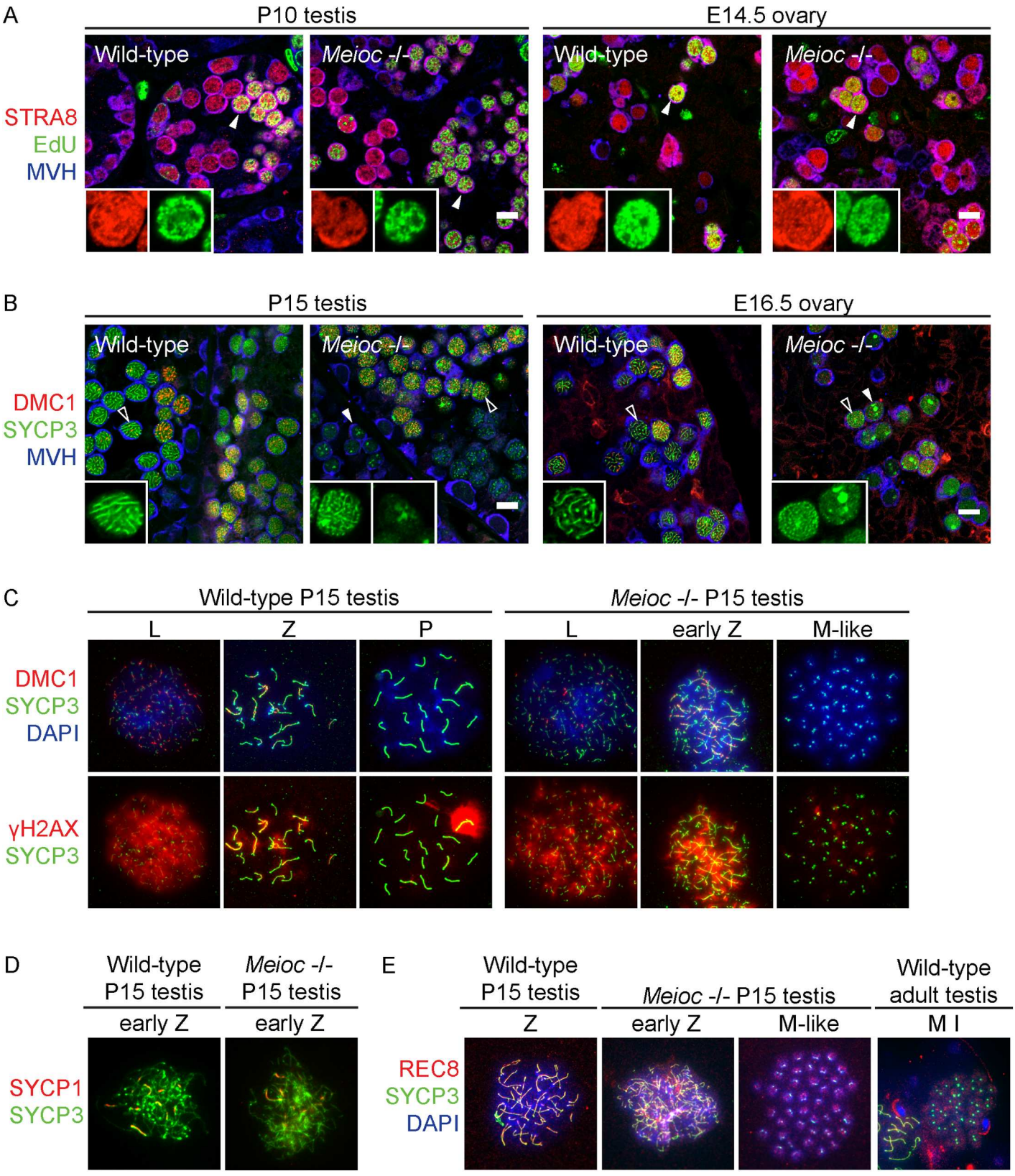


Fig 3. *Meioc*-deficient testicular and ovarian germ cells express molecular markers of meiotic prophase, but do not progress past early zygotene. (A) Immunofluorescence staining for STRA8, MVH, and EdU incorporation, in wild-type and *Meioc* -/- P10 testis and E14.5 ovary sections.

Insets: Higher magnification, STRA8 and EdU staining. In wild-type P10 testis, STRA8 expression is seen in a subset of MVH+ germ cells (arrowhead and inset), indicative of these cells initiating meiosis. STRA8+ germ cells are also observed in the *Meioc* -/- P10 testis. In wild-type and *Meioc* -/- E14.5 ovaries, STRA8 expression is visible in most MVH+ germ cells (arrowhead and inset), indicative of germ cells synchronously initiating meiosis. In wild-type and *Meioc* -/- ovaries of both sexes, some STRA8+ cells are also EdU+ (arrowhead and inset), reflecting premeiotic DNA synthesis. Scale bar = 10 μm. (B) Immunofluorescence staining for DMC1, SYCP3, and MVH, in wild-type and *Meioc* -/- P15 testis and E16.5 ovary sections. Insets: Higher magnification, SYCP3 staining. In wild-type P15 testis, we expected to observe germ cells in leptotene, zygotene, and pachytene (empty arrowhead and inset) stages of meiotic prophase. DMC1 expression and SYCP3 localization along the chromosomes is consistent with these stages. In *Meioc* -/- P15 testis, expression of both DMC1 and SYCP3 is seen, but the pattern of SYCP3 localization does not progress beyond what is typical of early zygotene, and is often accompanied by SYCP3 aggregates (empty arrowhead and inset). Additionally, some germ cells contain only SYCP3 aggregates (filled arrowhead and inset). In wild-type E16.5 ovary, we expected most germ cells to be in pachytene of meiotic prophase. DMC1 expression and SYCP3 localization along the chromosomes are consistent with pachytene stage (empty arrowhead and inset). In *Meioc* -/- E16.5 ovary, DMC1 expression and SYCP3 expression are also observed, but the pattern of SYCP3 localization does not progress beyond what is typical of early zygotene, and is often accompanied by SYCP3 aggregates (empty arrowhead and inset). Some germ cells contain only SYCP3 aggregates (filled arrowhead and inset). Scale bar = 10 μm. (C) Immunofluorescence staining for DMC1, γH2AX, and SYCP3 in chromosome spreads of wild-type and *Meioc* -/- germ cells from P15 testis. DNA stained by DAPI. In wild-type germ cells, we observed DMC1, γH2AX, and SYCP3 localization consistent with leptotene, zygotene, and pachytene stages of meiotic prophase. In some *Meioc* -/- germ cells, we observed DMC1, γH2AX, and SYCP3 staining indicative of leptotene and early zygotene stages. In metaphase-like cells, we observed SYCP3 localization at the ends of chromosomes, likely at centromeres. (D) Immunofluorescence staining for SYCP3 and SYCP1 in chromosome spreads of wild-type and *Meioc* -/- zygotene stage germ cells from P15 testis. In both wild-type and *Meioc* -/- germ cells, SYCP3 localizes along the entire length of chromosomes, and SYCP1 localizes to regions of synapsis. (E) Immunofluorescence staining for SYCP3 and REC8 in chromosome spreads of wild-type and *Meioc* -/- germ cells from P15 testis. DNA stained by DAPI. In both wild-type and *Meioc* -/- zygotene stage germ cells, SYCP3 and REC8 localize along the entire length of chromosomes. In metaphase-like cells, SYCP3 and REC8 localize, respectively, to the ends of chromosomes and to the condensed chromosomes. This localization of SYCP3 and REC8 is similar to that observed in wild-type metaphase I germ cells adult testes.

https://doi.org/10.1371/journal.pgen.1006704.g003

meiotic recombinase DMC1 [22,23]. Cells respond to DSBs by phosphorylating the histone variant H2AX, to yield γH2AX [24]. We assessed DSB formation by co-staining for DMC1 and γH2AX alongside SYCP3 in sections and spreads from wild-type and *Meioc*-deficient P15 testes and E16.5 ovaries (Figs 3B, 3C and S8A). In wild-type P15 testes and E16.5 ovaries, we observed DMC1 foci and γH2AX staining indicative of leptotene, zygotene, and pachytene stages of meiosis. In both *Meioc*-deficient P15 testes and E16.5 ovaries, DMC1 foci and γH2AX were also present in leptotene/zygotene-like cells. In metaphase-like cells, DMC1 foci are absent, but γH2AX staining suggests these cells suffer DNA damage.

Finally, we asked if cohesins are loaded onto chromosomes of *Meioc*-deficient germ cells. We immunostained for REC8, a meiotic cohesin [21,25], on spreads of meiotic cells from P15 testes (Fig 3E). In the leptotene/zygotene-like *Meioc*-deficient cells, REC8 localized along the lengths of chromosomes, much as in wild type. In metaphase-like cells, REC8 localizes to the condensed chromosomes, similar to what is observed in meiotic metaphases found in wild-type adult testes (Fig 3E) [21].

Quantification of cell spreads reveals that in P15 *Meioc*-deficient testes, metaphase-like and other abnormal germ cells (such as those with only clumpy SYCP3 staining) comprised about half of all germ cells (S8B Fig). In contrast, no meiotic metaphases were observed in wild-type P15 testes. *Meioc*-deficient testes also contained more leptotene stage germ cells but fewer zygotene stage germ cells than wild-type testes.

In summary, *Meioc*-deficient metaphase-like cells express and correctly localize proteins associated with synapsis and sister chromatid cohesion, demonstrating that the primary defect driving these cells to premature metaphase occurs after they have initiated meiosis. A subpopulation of cells is able to proceed with synapsis, cohesion, and recombination up to the leptotene/zygotene stages. Abby and colleagues focused their attention on the defects in this leptotene/zygotene cell population [18]. However, given our earlier histological analyses showing that metaphase-like cells first arise prior to leptotene and zygotene, it is unlikely that problems in synapsis and recombination cause the premature metaphases. The failure to proceed past the zygotene stage of synapsis and recombination is more likely a secondary consequence of the primary defect driving premature metaphase.

Meioc-deficient testicular and ovarian germ cells form univalent metaphases

Analysis of germ cells spreads showed that in *Meioc*-deficient metaphase-like cells, SYCP3 localized to the centromeres, and REC8 to the condensed chromosomes, similar to wild-type meiotic metaphases (Fig 3C and 3E). However, *Meioc*-deficient metaphase-like cells formed univalents instead of the bivalents formed in wild-type meiotic metaphases. Wild-type metaphase I cells form 20 bivalents, with 40 SYCP3 foci organized into 20 doublets, corresponding to 40 chromosomes organized into 20 homologous pairs. In contrast, *Meioc*-deficient metaphase-like cells retain 40 univalents, with 80 foci organized into 40 doublets, corresponding to 40 paired sister chromatids, with homologous chromosomes unpaired. The doublets of SYPC3 foci in the *Meioc*-deficient metaphase-like cells likely correspond to sister chromatid centromeres that have slightly separated, indicating a failure to maintain cohesion at sister centromeres. We did not observe any bivalents in the *Meioc* mutant amongst testis spreads from three P15 animals.

We asked if the *Meioc*-deficient cells with univalent chromosomes undergo molecular events associated with metaphase. In germ cells undergoing meiotic metaphase I in adult wild-type testes, chromosomes, visualized via DAPI, align at the equator of the cell to form a metaphase plate. The chromosomes are aligned by a bipolar spindle, formed by α-tubulin-positive microtubules emanating from opposite poles of the cell and attaching to the centromeres, marked by centromeric histone H3 variant CENPA (Fig 4). These features of metaphase are patently absent in wild-type P15 testes and wild-type E16.5 ovaries, where the chromosomes are not yet condensed, and centromeres localize along the nuclear envelope, as previously described [26]. Metaphase-like cells from *Meioc*-deficient P15 testes and E16.5 ovaries assemble a spindle, albeit a disorganized one that appears to emanate from a single pole. Their chromosomes do not assemble on a metaphase plate, and are instead scattered throughout the nucleus. In addition, *Meioc*-deficient metaphase-like germ cells undergo histone H3 phosphorylation and nuclear envelope breakdown, two events associated with wild-type metaphase (S9 Fig). In summary, metaphase-like cells from *Meioc*-deficient mice form spindles, phosphorylate histone H3 and undergo nuclear envelope breakdown much like wild-type meiotic metaphase cells. However, the chromosomes are in univalent rather than bivalent configuration, indicating a failure of the chromosomes to pair, likely as a result of prematurely proceeding to metaphase.

Aberrant expression of cyclins in meiotic *Meioc*-deficient testicular and ovarian germ cells

To gain insight into the molecular pathways that *Meioc* may regulate so as to extend meiotic prophase I, we performed RNA-seq on whole ovaries from E14.5 wild-type and *Meioc*-deficient fetuses (S1 Table). At this stage, *Meioc*-deficient ovaries did not exhibit TUNEL-positive apoptotic cells, which indicates that programmed cell death had not yet affected the size of the germ cell population (S6 Fig). We observed, in *Meioc*-deficient ovaries, 465 genes expressed at higher levels than wild type ($q < 0.01$) and 496 genes expressed at lower levels than wild type ($q < 0.01$); the two sets of genes were enriched for distinct functions (Table 1; S2 Table). Genes expressed at lower levels were enriched for involvement in the meiotic chromosomal program, which we interpreted as reflecting fewer cells entering meiotic prophase I in the mutant. Genes expressed at higher levels were enriched for factors typically associated with the mitotic cell cycle. Previously reported microarray analyses of *Meioc*-deficient gonads identified only 42 differentially expressed genes, of which 38 were expressed at lower levels [18]. Of these 38 genes, half were noted to be associated with meiosis. Those analyses failed to detect genes

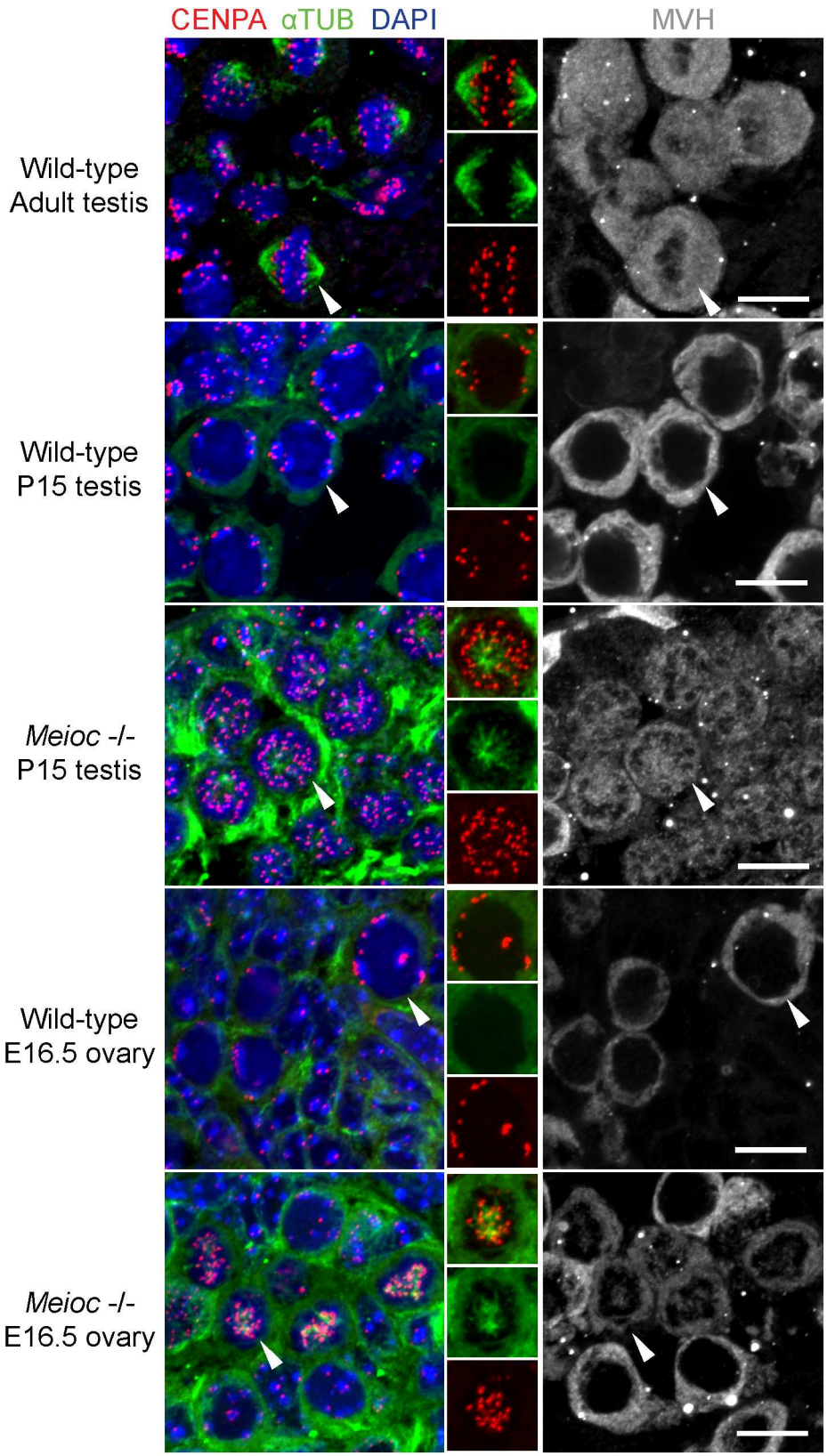


Fig 4. *Meioc*-deficient testicular and ovarian germ cells express molecular markers of metaphase. Immunofluorescence staining for CENPA and α-TUB in wild-type adult testicular germ cells in metaphase I, as

well as wild-type and *Meioc* -/- P15 testis and E16.5 ovary sections. Nuclei stained by DAPI; MVH immunostains germ cells. Inset: CENPA and α-TUB staining together, and separately. In wild-type adult testicular germ cells in metaphase I, CENPA localizes to the metaphase plate and a bipolar spindle is formed. In wild-type P15 testis and E16.5 ovary, CENPA localizes to periphery of nuclei in meiotic cells, and no spindle is observed. In *Meioc* -/- P15 testis and E16.5 ovary, CENPA does not localize to periphery of nuclei, and instead localizes to ends of a disorganized, radiating spindle. Scale bar = 10 μm.

https://doi.org/10.1371/journal.pgen.1006704.g004

expressed at higher levels, and thus did not identify the misregulation of mitotic cell cycle factors. Given that RNA-seq provides more sensitivity than microarray analysis [27], our RNA-seq analysis likely reveals a more complete snapshot of transcriptional changes in the absence of *Meioc*.

We explored the possibility that the premature metaphase entry observed in *Meioc*-deficient germ cells was associated with misregulation of cell cycle factors. Progression through the cell cycle is tightly controlled by cyclical fluctuations in expression of cyclins, which induce oscillatory activation of cyclin-dependent kinases. We therefore examined cyclin expression, focusing on determining whether *Meioc*-deficient germ cells express cyclins typical of mitosis or meiosis.

Cyclin A2 (CCNA2), which drives progression through mitotic S and G2-M [28], is expressed in the male germline in mitotic spermatogonia and preleptotene spermatocytes, and is normally down-regulated upon entry into leptotene [29,30]. We immunostained wild-type and *Meioc*-deficient P15 testes for CCNA2, as well as SYCP3, to identify cells in meiotic

Table 1. Top ten enriched GO categories for genes expressed at lower or higher levels in *Meioc*-/- ovaries.

Genes expressed at lower levels in Meioc-/- ovaries		
GO term	Fold Enrichment	Benjamini-corrected p-val
cell cycle process	2.75	9.80E-03
cell cycle	2.22	2.03E-02
meiosis	5.00	3.48E-02
M phase of meiotic cell cycle	5.00	3.48E-02
cell cycle phase	2.68	3.09E-02
meiotic cell cycle	4.88	2.55E-02
cofactor metabolic process	3.29	5.01E-02
meiosis I	7.99	4.59E-02
chromosome organization involved in meiosis	11.75	1.32E-01
synapsis	11.75	1.32E-01
Genes expressed at higher levels in Meioc-/- ovaries		
GO term	Fold Enrichment	Benjamini-corrected p-val
cell division	4.90	8.12E-10
mitotic cell cycle	4.76	7.39E-08
cell cycle phase	4.06	7.63E-08
cell cycle	2.96	3.39E-07
cell cycle process	3.50	8.64E-07
nuclear division	4.98	8.71E-07
mitosis	4.98	8.71E-07
M phase of mitotic cell cycle	4.88	1.09E-06
organelle fission	4.80	1.26E-06
M phase	3.95	1.90E-06

https://doi.org/10.1371/journal.pgen.1006704.t001

prophase. We confirmed that in the wild-type, CCNA2 is expressed in mitotic spermatogonia, but not in germ cells that had entered meiotic prophase, as evident by thread-like SYCP3 staining (Fig 5A). In contrast, in *Meioc*-deficient testes, CCNA2 is present in germ cells that exhibit SYCP3 staining typical of leptotene and zygotene. Thus, in *Meioc*-deficient testes, testicular germ cells in meiotic prophase aberrantly express CCNA2.

Cyclin A1 (CCNA1) is thought to replace CCNA2 during the meiotic cell cycle: it is expressed in meiotic spermatocytes from late pachytene through metaphase, and is required to initiate meiotic metaphase [31,32]. Using single molecule fluorescent in situ hybridization (smFISH), we observed *Ccna1* mRNA expression in wild-type P15 testes in late pachytene cells, but not spermatogonia, which instead expressed *Ccna2* (Fig 5B). In *Meioc*-deficient P15 testes, we failed to observe *Ccna1* expression in either meiotic or metaphase-like germ cells.

Cyclin Bs are essential for the G2/M transition [28]. Cyclin B1 and B2 (CCNB1, CCNB2) are expressed in both mitotically and meiotically dividing cells. In contrast, cyclin B3 (CCNB3) is expressed only during leptotene and zygotene in both males and females, and forms kinase-deficient complexes with CDK2, raising the possibility that CCNB3 could be inhibiting precocious cell cycle progression during early meiotic prophase I [33,34]. Using smFISH, we observed *Ccnb3* expression in meiotic cells of wild-type P15 testes and E16.5 ovaries as expected (Fig 5C), and also in meiotic germ cells from *Meioc*-deficient testes and ovaries.

In summary, we found that *Meioc*-deficient meiotic germ cells do not exclusively express either mitosis or meiosis-specific cyclins. They express meiosis-specific CCNB3, suggesting that they have initiated the meiotic cell cycle program, but they also aberrantly express CCNA2, which should be down-regulated during meiosis. Misexpression of CCNA2, accompanied by the broad up-regulation of genes associated with the mitotic cell cycle, leads us to conclude that although *Meioc*-deficient germ cells can initiate the meiotic chromosomal program, they fail to properly transition from a mitotic to meiotic cell cycle program. Based on these novel findings, not reported by Abby et al. [18], we propose that mis-regulation of the cell cycle is the primary cause of premature metaphases in the absence of *Meioc*.

MEIOC interacts with the mouse homolog of BGCN, a translational regulator in fly

To gain insight into how MEIOC functions at the molecular level to prevent premature exit from meiotic prophase I, we determined MEIOC's binding partners by performing an immunoprecipitation for MEIOC from testis lysates. Using quantitative mass spectrometry analysis, we identified one protein as interacting with MEIOC: YTHDC2 (enrichment over MEIOC immunoprecipitation in *Meioc*-deficient testes > 1.5, unique peptides >1; Table 2, S3 Table). We confirmed the interaction between MEIOC and YTHDC2 by immunoprecipitating each protein from adult testes and immunoblotting for the other (Fig 6A). MEIOC interaction with YTHDC2 was also previously observed [18].

YTHDC2 contains multiple domains that interact with nucleic acid—specifically, an R3H domain, an RNA helicase domain, and a YTH domain [35–37]—but its molecular function in mammalian cells remains poorly characterized. To gain insight into the function of YTHDC2, we looked for YTHDC2 orthologs in other species. We identified YTHDC2 orthologs in almost all metazoans examined (S10 Fig). In *Drosophila melanogaster*, the ortholog of mouse YTHDC2 is BGCN, which physically interacts with a partner, BAM, to regulate translation in germ cells [38]. Considering that we find no ortholog of MEIOC in the *Drosophila* genome (S1 Fig), mouse MEIOC may be interacting with YTHDC2 to perform a role analogous to that of BAM with BGCN in *Drosophila*. Based on this hypothesis, we might expect similar phenotypes

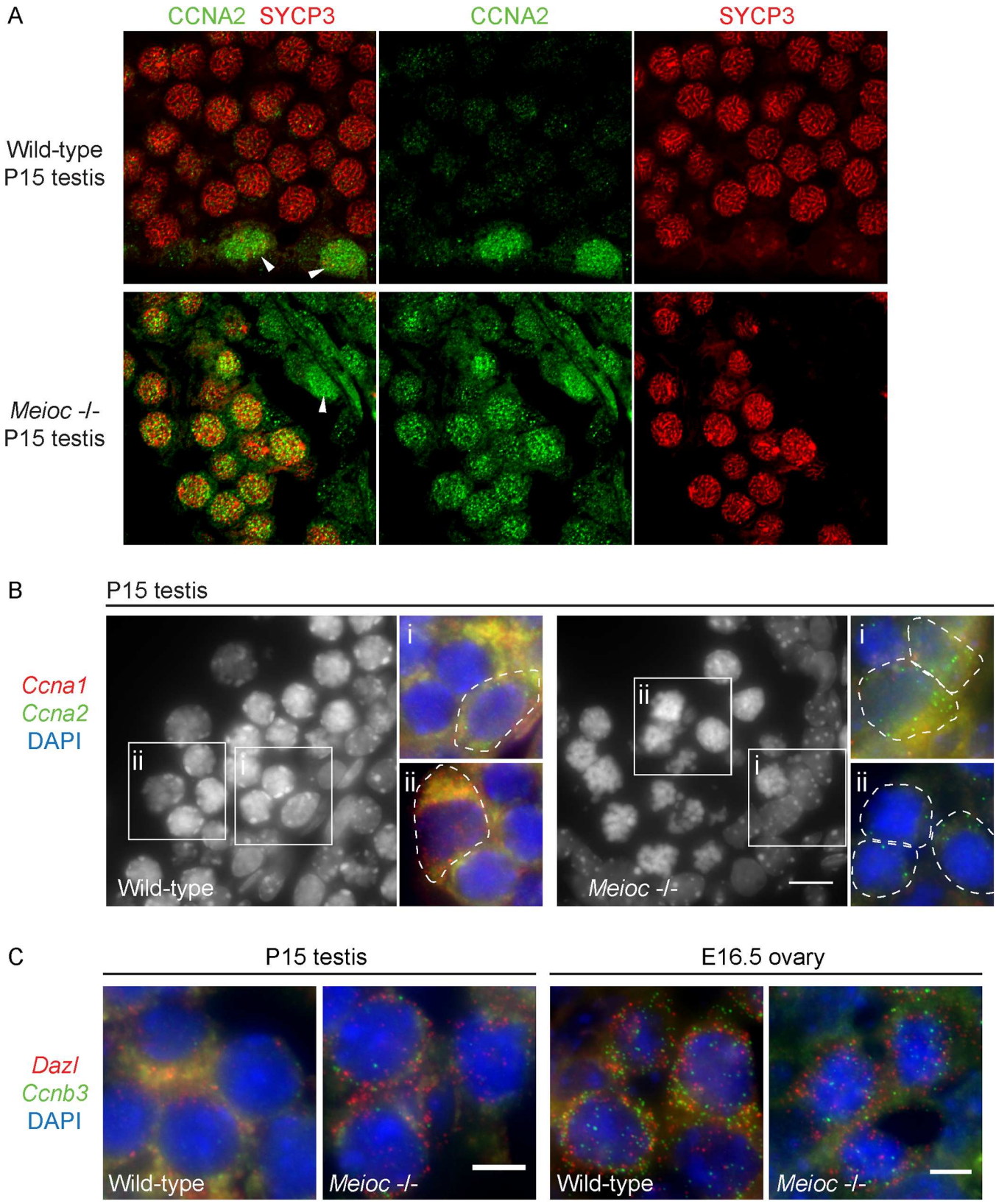


Fig 5. Expression of *cyclin A2* and *cyclin B3* in *Meioc*-deficient adult testis and ovary. (A) Immunofluorescence staining for CCNA2 and SYCP3 in wild-type and *Meioc* -/- P15 testis. In wild-type, CCNA2 is expressed in mitotic spermatogonia (arrowhead), and is not expressed in cells past the

leptotene stage of meiosis (SYCP3+ cells at zygotene stage of meiosis). In *Meioc* -/- testes, CCNA2 is misexpressed in SYCP3+ cells at zygotene-like stage. (B) Single molecule FISH staining for *Ccna1* and *Ccna2* in wild-type and *Meioc* -/- P15 testis. Low magnification image: DAPI staining of germ cells in testis tubule. High magnification images: DAPI is in blue, and single cells are outlined. Single transcripts are detected as individual red or green dots; diffuse staining is background. *Ccna2* (green dots) is detected in both wild-type and *Meioc* -/- testes in mitotic spermatogonia, identified by their position at base of tubule and by their nuclear morphology (wild-type and *Meioc* -/-, i). *Ccna2* is additionally detected in *Meioc* -/- germ cells that are in middle of lumen (likely meiotic, or else metaphasic cells, ii). *Ccna1* (red dots) is detected in late pachytene germ cells in wild-type testis, but not in *Meioc* -/- germ cells (wild-type and *Meioc* -/-, ii). (C) Single molecule FISH staining for *Ccnb3* and *Dazl* in late zygotene/early pachytene germ cells from wild-type and *Meioc* -/- P15 testis and E16.5 ovary. *Ccnb3* and *Dazl* are detected in both wild-type and *Meioc* -/- germ cells. Scale bar = 10 μm.

https://doi.org/10.1371/journal.pgen.1006704.g005

in *Meioc*-deficient and *Ythdc2*-deficient mice. *Ythdc2*-deficient male mice exhibit striking similarities to the *Meioc*-deficient mice: in both mutants, germ cells initiate but do not complete meiosis; instead, numerous abnormal metaphase-like cells are observed (A. Bailey, D. de Rooij, and M. Fuller, personal communication).

To determine if YTHDC2 protein expression is regulated by MEIOC, we immunostained wild-type and *Meioc*-deficient P15 testes for YTHDC2 (Fig 6B). In both wild-type and *Meioc*-deficient testes, YTHDC2 was present in the cytoplasm of meiotic germ cells, including leptotene, zygotene, and pachytene cells in the wild-type, and leptotene/zygotene-like cells in the mutant. Thus, in contrast to previous reports [18], we found that YTHDC2 expression is not dependent on *Meioc*.

MEIOC interacts with cell cycle-associated transcripts but not meiosis-specific transcripts

Given that YTHDC2 and MEIOC proteins localize to the cytoplasm, and that YTHDC2 contains multiple domains that interact with nucleic acid (specifically, an R3H domain, an RNA helicase domain, and a YTH domain) [35–37], we hypothesized that a YTHDC2/MEIOC complex binds to and post-transcriptionally regulates mRNA, like the *Drosophila* BGCN/BAM complex. Based on the observations that *Meioc*-deficient germ cells exhibit precocious progression into a metaphase-like state and misexpress cell cycle transcripts and mitotic cyclin CCNA2, we further hypothesized that this YTHDC2/MEIOC complex regulates transcripts involved in mitotic cell cycle progression.

We therefore investigated the transcripts to which both MEIOC and YTHDC2 bind via RNA immunoprecipitation and sequencing (RIP-seq). We performed MEIOC RIP-seq in wild-type P15 testes, along with the following controls: MEIOC RIP-seq in *Meioc*-deficient P15 testes, IgG RIP-seq controls in wild-type P15 testes, and RNA-seq from both wild-type

Table 2. Identification of MEIOC-interacting proteins by quantitative mass spectrometry.

Protein	Average TMT values relative to one replicate of MEIOC IP from *Meioc* -/- lysate (SD)			Number of unique peptides selected for fragmentation*
	IgG IP from wild-type lysate	MEIOC IP from wild-type lysate	MEIOC IP from *Meioc* -/- lysate	
MEIOC / GM1564	0.866 (0.033)	4.055 (0.673)	1.136 (0.193)	4
YTHDC2	0.631 (0.100)	2.028 (0.449)	1.069 (0.098)	9

*Sequences of unique peptides listed in S3 Table.

TMT quantification was obtained for two biological replicates in each of three immunoprecipitation conditions: (A) wild-type (C57BL/6) lysate with IgG antibody, (B) wild-type lysate with MEIOC antibody, or (C) *Meioc* -/- lysate with MEIOC antibody. Values were normalized to the signal from one replicate in condition C and then averaged across each condition. Shown here are all proteins represented by two or more peptides with a relative TMT value greater than 2 in condition B relative to condition C.

https://doi.org/10.1371/journal.pgen.1006704.t002

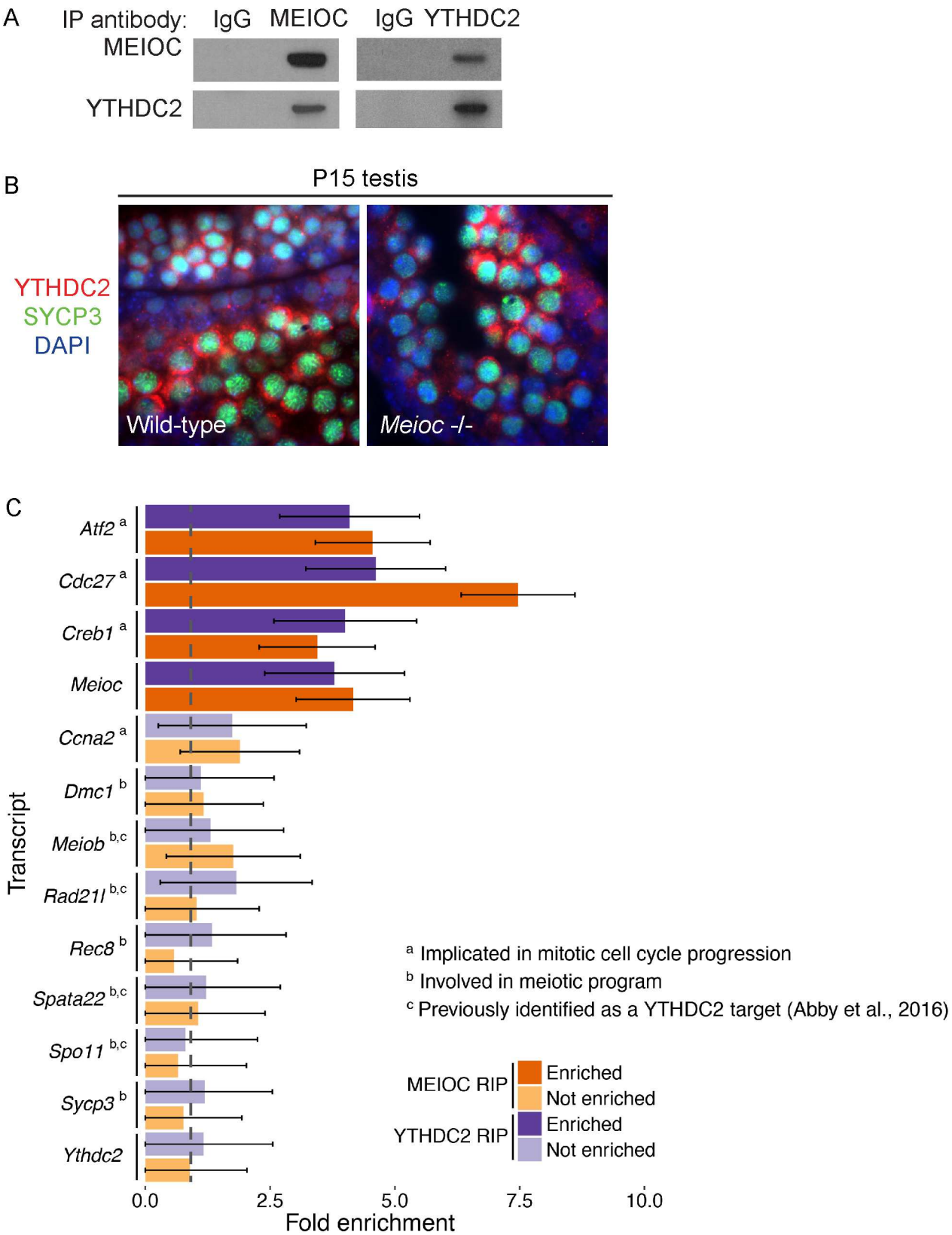


Fig 6. MEIOC co-immunoprecipitates with YTHDC2 and cell cycle-associated transcripts but not meiosis-specific transcripts. (A) Immunoprecipitation (IP) performed with anti-MEIOC or anti-YTHDC2 antibody and IgG control from adult testis lysates. IP was

followed by immunoblotting with either the anti-MEIOC antibody, or anti-YTHDC2 antibody. MEIOC and YTHDC2 were detected specifically in immunoprecipitation with either anti-MEIOC or anti-YTHDC2 antibody. (B) Immunofluorescence staining for YTHDC2 and SYCP3 in wild-type and *Meioc* -/- P15 testis. Nuclei stained by DAPI. YTHDC2 is expressed at comparable levels in zygotene and zygotene-like cells from both wild-type and *Meioc* -/- testes, respectively. (C) Fold enrichment for MEIOC-specific binding in P15 testis and YTHDC2-specific binding in P20 testis. For MEIOC, targets were identified via MEIOC RIP-seq and total RNA-seq analyses of wild-type and *Meioc* -/- testes as well as IgG RIP-seq from wild-type testes. For YTHDC2, targets were identified via YTHDC2 RIP-seq, IgG RIP seq, and total RNA-seq analyses of wild-type testes. Statistically significant enrichment was identified based on FDR < 0.05, FPKM > 1, and fold change > 3 for MEIOC, or fold change > 2 for YTHDC2. Of the transcripts that were enriched, some have been implicated in mitotic cell cycle progression. Transcripts that were not enriched were selected for analysis based on functions in the cell cycle (*Ccna2*), the canonical meiotic chromosomal program (*Dmc1*, *Rec8*, *Sycp3*), or previous reports of interaction with YTHDC2 (*Meiob*, *Rad21l*, *Spata22*, *Spo11*). Error bars represent standard error. Dashed grey line marks fold change of 1.

https://doi.org/10.1371/journal.pgen.1006704.g006

and *Meioc*-deficient testes to control for changes in mRNA abundances in wild-type and *Meioc*-deficient testes. We performed YTHDC2 RIP-seq in P20 testes using two independent YTHDC2 antibodies, along with the following controls: IgG RIP-seq in wild-type P20 testes, and RNA-seq in wild-type testes. We identified 626 transcripts that were enriched in immunoprecipitation with MEIOC (fold change $>$ 3, FDR $<$ 0.05, expressed at FPKM $>$ 1, S4 Table), and 80 transcripts enriched in immunoprecipitation with YTHDC2 (fold change $>$ 2, FDR $<$ 0.05, expressed at FPKM $>$ 1, S4 Table). Of these, 67 transcripts were identified as both MEIOC and YTHDC2 targets (a subset of results shown in Fig 6C). We validated a sampling of the MEIOC and YTHDC2 targets by RIP followed by quantitative PCR (qPCR; S11 Fig). While *Ccna2* was not a direct target of MEIOC or YTHDC2, bound transcripts included other cell-cycle related transcripts such as *Cdc27*, a component of the anaphase promoting complex [39], as well as *Creb1* and *Atf2*, transcription factors that can upregulate the expression of *Ccna2* [40–42]. In addition, both MEIOC and YTHDC2 interact with the *Meioc* transcript itself, but not with *Ythdc2* transcript.

In contrast to our model of the MEIOC/YTHDC2 complex as a regulator of meiotic prophase I exit, Abby and colleagues suggested that MEIOC and YTHDC2 function to stabilize transcripts involved in the chromosomal program of meiosis [18]. This conclusion was based, in part, on RIP data indicating that YTHDC2 bound four transcripts essential to the chromosomal program (*Spata22*, *Spo11*, *Meiob*, and *Rad21L*) [18]. This hypothesis predicts that MEIOC should also interact with these transcripts. We found no evidence, by either RIP-seq or RIP-qPCR, that MEIOC or YTHDC2 interacts with these transcripts (Fig 6C, S11 Fig). Furthermore, we could not demonstrate enrichment for additional canonical transcripts in the meiotic chromosomal program, such as *Dmc1*, *Rec8*, and *Sycp3* (Fig 6C; S4 Table), which were not identified as YTHDC2 targets by Abby and colleagues [18].

To determine whether the MEIOC/YTHDC2 complex promotes or inhibits expression of its targets, we returned to our RNA-seq dataset from E14.5 wild-type and *Meioc*-deficient ovaries. Given the remarkable similarity of *Meioc*-deficient phenotypes in males and females, we hypothesized that MEIOC/YTHDC2's targets from the testis would also be differentially expressed in the fetal ovary. We therefore compared MEIOC/YTHDC2's shared targets to our RNA-seq dataset from E14.5 wild-type and *Meioc*-deficient ovaries (S1 Table). Of the 67 MEIOC- and YTHDC2-bound mRNAs identified in the testis, 65 were expressed (FPKM $>$ 1) in the fetal ovary. Of these 65 MEIOC- and YTHDC2-bound transcripts, 28 (43%) were expressed differentially between E14.5 wild-type and *Meioc*-deficient ovaries. With the exception of the *Meioc* transcript itself, all 27 of these differentially expressed mRNAs were present at higher levels in the absence of MEIOC (S5 Table), suggesting that MEIOC and YTHDC2 destabilize their target mRNAs. These differentially expressed targets included the mitotic cell cycle regulators *Atf2*, *Cdc27*, and *Creb1*. Not all MEIOC/YTHDC2-bound mRNAs were observed to be differentially expressed. This may be because most MEIOC/YTHDC2-bound mRNAs were expressed in gonadal somatic cells as well as in germ cells, which may obscure

differential expression signals in RNA-seq data from whole gonads. Additionally, our MEIOC and YTHDC2 RIP experiments were performed using testis tissue, while our RNA-seq data was derived from fetal ovary; though they overlap, the sets of genes targeted by MEIOC and YTHDC2 in testis and ovary may not be identical.

In summary, we found that MEIOC and YTHDC2 bind transcripts that regulate the mitotic cell cycle, likely resulting in their destabilization. These observations are consistent with our hypothesis that MEIOC facilitates the switch from a mitotic to a meiotic cell cycle program. We find no evidence that MEIOC interacts with transcripts of the meiotic chromosomal program, and thus no reason to believe that it directly stabilizes such transcripts, as recently proposed by Abby and colleagues [18].

Discussion

An extended prophase I is a conserved feature of meiosis, and is critical for enabling completion of meiotic chromosomal events in yeast [2]. Our analyses of *Meioc*-deficient mice identify *Meioc* as a critical factor required for this extended prophase I in mice: in the absence of *Meioc*, both testicular and ovarian germ cells can initiate meiosis and embark on meiotic prophase I, but fail to progress past the zygotene stage. Instead, *Meioc*-deficient cells proceed precociously to metaphase. Our studies demonstrate that *Meioc* is required for an extended meiotic prophase I in mice, and reveal the extended prophase I as a critical and actively regulated feature of meiosis in a vertebrate system.

We posit that meiotic prophase I is comprised of various meiosis-specific subprograms, including a chromosomal program wherein chromosomes synapse and recombine, and a coordinately regulated cell cycle program that extends prophase I for the duration of the chromosomal program. We propose that when the chromosomal program of meiosis is initiated, the corresponding cell cycle program must be simultaneously implemented (Fig 7). Our previous findings demonstrated that *Stra8* is required for the meiotic chromosomal program [14,15]. Our present findings lead us to propose that *Meioc* is simultaneously required to promote the meiotic cell cycle program.

How can a germ cell ensure that it exits prophase into metaphase only when the meiotic chromosomal program is complete? We reasoned that the germ cell must transition from a mitotic cell cycle program (that of necessity is independent of the meiotic chromosomal program) to a meiotic cell cycle program in which prophase exit is dependent on meiotic chromosomal checkpoints. We hypothesize that *Meioc* is required for this transition. This model predicts that in the absence of *Meioc*, a germ cell that has already expressed key meiotic regulators (such as STRA8) and meiotic chromosomal proteins (such as SYCP3) will continue to run a mitotic cell cycle program. Due to an active mitotic cell cycle program, the meiotic cell will proceed to metaphase on a mitotic schedule, independent of the meiotic chromosomal checkpoints. It was previously observed that leptotene/zygotene stage spermatocytes are not competent to enter metaphase upon stimulation with okadaic acid [43]; this is likely because in wild-type cells, exit from prophase into metaphase is strictly dependent on the meiotic chromosomal checkpoints. In contrast, in *Meioc*-deficient germ cells, a persistent mitotic cell cycle program renders the cell cycle independent of the meiotic chromosomal events, and drives cells into metaphase as early as preleptotene, or shortly thereafter. Consistent with this idea of cell cycle mis-regulation, *Meioc*-deficient spermatocytes that have initiated meiotic prophase I misexpress Cyclin A2, which is normally expressed in mitotic spermatogonia and down-regulated by leptotene of meiotic prophase I.

A second possibility is that *Meioc* functions to establish a checkpoint for exit from meiotic prophase I. However, if lack of a checkpoint led to premature resumption of the meiotic cell

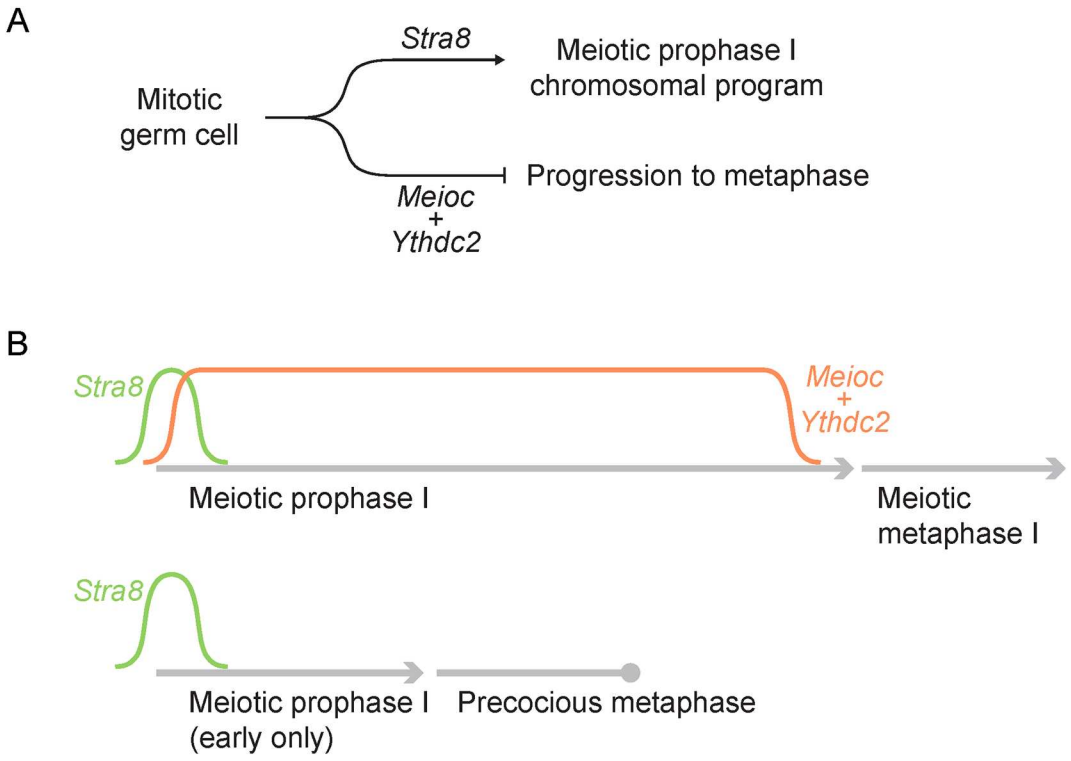


Fig 7. A proposed model of *Meioc*'s role in meiosis. (A) Mitotic germ cells transition into meiosis via expression of *Stra8*, which upregulates the meiotic chromosomal program. At the same time, *Meioc* and *Ythdc2* are required during meiotic prophase to inhibit progression into metaphase, thereby allowing meiotic prophase to proceed normally. (B) *Stra8* is expressed in early meiotic prophase, while *Meioc* is expressed throughout meiotic prophase. In wild-type germ cells, *Meioc* is downregulated before progression into metaphase. In the absence of *Meioc*, meiotic prophase is abbreviated, resulting in a precocious attempt at metaphase.

https://doi.org/10.1371/journal.pgen.1006704.g007

cycle, we might expect that the cell cycle resumed would be meiotic in nature, and thus primarily driven by cyclins typically expressed during the meiotic cell cycle, such as Cyclin A1. Instead, we find that Cyclin A2, not Cyclin A1, is expressed in *Meioc*-deficient germ cells. Therefore, MEIOC appears to govern the transition from a mitotic to a meiotic cell cycle program, in part or in whole by suppressing the mitotic program.

An alternate model has been proposed by Abby et al., wherein *Meioc* is required for stabilization of meiotic transcripts, such as those required for the chromosomal program [18]. In their model, the failure to stabilize these transcripts leads to lack of sufficient proteins required for the chromosomal events of meiosis, thus forcing cells to switch prematurely to metaphase. We find this model unsatisfying for the following reasons. First, we find no evidence that MEIOC and YTHDC2 bind transcripts that function in the chromosomal program. The conditions used for RIP experiments may explain the difference between our results and those of Abby et al. For immunoprecipitation of RNA, we used lysis conditions without reducing agents in order to maintain proteins' disulfide bonds. By contrast, Abby et al. used mild reducing conditions that could have relaxed disulfide bonds and potentially altered the proteins and transcripts with which YTHDC2 interacted. We propose that the non-reducing conditions used in this study are more likely to have captured the in vivo interactions of MEIOC and YTHDC2. In addition, the model proposed by Abby et al. does not explain how a failure to stabilize transcripts of the meiotic chromosomal program results in premature metaphase. In the vast majority of knock-outs of genes required for the meiotic chromosomal program (e.g.

Dmc1), germ cells arrest in meiosis and proceed to apoptosis, rather than attempting precocious metaphase [22,23].

Understanding the molecular function of MEIOC would aid in distinguishing these two alternative models. Our genetic and biochemical analyses suggest a role for MEIOC in post-transcriptional regulation of transcripts implicated in the cell cycle. First, the phenotype of *Meioc*-deficient mice is highly similar to that of male mice deficient for the mouse ortholog of *Bgcn* (A. Bailey, D. de Rooij, and M. Fuller, personal communication). *Drosophila* BGCN is an RNA helicase that acts in concert with an interacting partner, BAM, to repress translation in *Drosophila* germ cells [38,44,45]. The shared phenotype between mouse *Meioc* and *Ythdc2* suggests that they may act as interacting partners to regulate translation in the mouse germline, similar to BAM/BGCN in the fly. A putative mouse ortholog of fly *bam* had been previously identified, but mice lacking this gene exhibited no viability or fertility defects [46]. *Meioc*, while not orthologous to *Drosophila bam*, may be its functional analog in the mouse. Notably, we failed to identify an ortholog of *Meioc* in *Drosophila*, further supporting the notion that mouse MEIOC and *Drosophila* BAM substitute for each other in the two species. Consistent with the hypothesis that MEIOC and YTHDC2 function together, we find evidence that MEIOC physically interacts with YTHDC2. Further, MEIOC and YTHDC2 interact with overlapping sets of transcripts. These transcripts include genes associated with the mitotic cell cycle, but not with meiosis, bolstering our model that a MEIOC/YTHDC2 complex post-transcriptionally regulates transcripts associated with cell cycle progression. Transcripts that interact with MEIOC and YTHDC2 are up-regulated in the absence of MEIOC, suggesting that MEIOC/YTHDC2 functions to destabilize their target mRNAs. While MEIOC's PF15189 domain remains uncharacterized, YTHDC2 contains multiple domains that interact with nucleic acid. These domains include the R3H domain that binds single-stranded nucleic acid [36]; the DEAH box helicase domain that unwinds nucleic acids [35]; and the YTH domain that recognizes post-transcriptionally modified N6-methyladenosine (m6A) on RNA [37]. In particular, RNA helicases can regulate the stability of target transcripts by interacting with proteins that directly influence RNA stability/degradation, such as decapping enzymes, deadenylation complexes, and ribonucleases [35]. Helicases can further affect RNA stability by unfolding the RNA to make it accessible to these enzymes [35]. However, we do not yet know the extent to which these domains are active in the YTHDC2 protein, and how MEIOC may contribute to their activity. The precise molecular mechanism of MEIOC/YTHDC2 activity, and consequences for target transcripts, remain to be determined.

Mouse MEIOC and YTHDC2, and their *Drosophila* counterparts *bam* and *bgcn*, appear to have similar roles in gametogenesis based on post-transcriptional regulation of transcripts. However, the details of regulation differ between species, and even between sexes. Whereas MEIOC and YTHDC2 appear to regulate the meiotic cell cycle program, *bam* and *bgcn* function at earlier stages of *Drosophila* gametogenesis, prior to the decision to initiate meiosis. In *Drosophila* males, *bam* and *bgcn* are required for spermatogonia to cease proliferation and initiate spermatocyte differentiation and meiosis [47,48]. In the female, *bam* and *bgcn* function earlier to initiate the transit amplifying divisions [49,50]. Correspondingly, their target transcripts differ between the *Drosophila* sexes: BAM and BGCN repress *mei-P26* translation in the male, but not in the female [38]. Conversely, BAM represses translation of *nanos* in the female but not the male [44]. Furthermore, *mei-P26* and *nanos* are not components of the cell cycle program. Thus, while the involvement of the bam-bgcn and MEIOC-YTHDC2 complexes in gametogenesis via post-transcriptional regulation is conserved, their time of action, as well as the targets of translational repression, may vary according to sex and species.

A common pathway induces both initiation of the chromosomal program of meiotic prophase I, as well as *Meioc* expression, thus genetically linking the meiotic chromosomal

program with the meiotic cell cycle program. We previously demonstrated through an in vivo genetic knock-out mouse model that *Stra8* is required for initiation of the meiotic chromosomal program in both ovarian and testicular germ cells [14,15]. More recently, further in vivo studies of *Stra8*-deficient ovaries showed that *Stra8* is required for full induction of *Meioc* expression: *Meioc* expression is 4-fold higher in wild-type fetal ovaries than in *Stra8*-deficient ovaries [17]. Since *Stra8* is induced by RA [10,11], *Meioc* expression in fetal ovarian germ cells is thus also at least partially dependent on RA signaling. Contrary to these results, Abby et al. concluded that *Meioc* expression is completely independent of RA signaling in both ovarian and testicular germ cells, based on data from fetal gonads cultured with RA or an RAR inverse agonist as well as postnatal testes from pups exposed to the RAR inverse agonist [18]. This discrepancy in results in ovarian germ cells suggests that the in vivo genetic model may more accurately reflect the endogenous biology than a culture system, especially when dealing with a relatively modest (4-fold) change in gene expression. Therefore, similar in vivo examination of whether RA and *Stra8* contribute to *Meioc* expression in testicular germ cells is still needed.

Our study leads us to propose that successful meiosis in mice requires coordination of a meiosis-specific cell cycle program with the elaborate chromosomal program of prophase I. Further studies will elucidate how *Meioc*, in partnership with *Ythdc2*, promotes the transition to a meiosis-specific cell cycle program at the time germ cells initiate the meiotic chromosomal program.

Materials and methods

Ethics statement

All experiments involving mice were performed in accordance with the guidelines of the Massachusetts Institute of Technology (MIT) Division of Comparative Medicine, which is overseen by MIT's Institutional Animal Care and Use Committee (IACUC). The animal care program at MIT/Whitehead Institute is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International (AAALAC), and meets or exceeds the standards of AAALAC as detailed in the Guide for the Care and Use of Laboratory Animals. The MIT IACUC approved this research (no. 0714-074-17).

Generation of anti-MEIOC antibody

A polyclonal antibody against MEIOC was raised in rabbits against C-terminal peptide CHESINSSNPMNQRGETSKH (YenZym Antibodies, LLC), and affinity purified using the antigenic peptide (SulfoLink Immobilization Kit for Peptides, ThermoScientific).

Generation of *Meioc* mutant alleles

The *Meioc* gene was targeted for homologous recombination in v6.5 embryonic stem (ES) cells with a targeting vector for a knockout-first allele of *Meioc* (obtained from the Knockout Mouse Project Repository, vector PG00048_X_6_E03) (S4 Fig). Resultant colonies were tested for correct integration by Southern blot analysis of a KpnI/XhoI restriction digest. Three independent, verified ES cell clones were injected into C57BL/6 recipient blastocysts, and germline transmission was obtained with all three clones. The 'knockout-first' allele is denoted 3lox or 3L as it retains 3 loxP sites. In the 3lox allele, the open reading frame is disrupted by the active lacZ reporter. The 3lox allele was subject to Flp recombination by breeding mice bearing the 3lox allele to ACTB:FLPe B6J mice (Jackson laboratory no. 005703). The resultant allele is a conditional allele, denoted 2lox or 2L. The *lacZ* and *Neo* genes are excised, leaving exon 3 flanked by loxP sites. The 2lox allele was subject to Cre recombination by breeding mice

bearing the 2lox allele to $Mvh^{Cre\text{-}mOrange}$ mice [51]. The resultant allele is a knockout allele, denoted 1lox, 1L or *Meioc* -. Cre recombination excises exon 3, and is predicted to result in a frame shift and generate a premature stop codon subsequent to exon 2. All three alleles were genotyped by PCR (detailed in S4 Fig).

Mice and sample collection

We analyzed both *Meioc* 3L/3L and *Meioc*-deficient (*Meioc* -/-; *Meioc* 1L/1L) mice. *Meioc* 3L/3L and *Meioc*-deficient mice or embryos were generated by heterozygote matings. For wild-type controls, we used littermates that were either heterozygote for the mutant and wild-type allele or homozygous for the wild-type allele. *Meioc* 3L/3L mice were of mixed 129S4 and C57BL/6 background. *Meioc*-deficient mice were backcrossed to the C57BL/6 strain for at least 5 generations; all data shown in figures are from mice 5 to 7 generations backcrossed.

EdU incorporation

Mice, or pregnant mothers, were injected with 4μg/μl of EdU dissolved in PBS, for a final dose of 20μg/g. Samples were collected 2 h after EdU injection.

Histology

Testes were fixed overnight in Bouin's solution, embedded in paraffin, sectioned, and stained with hematoxylin and eosin. Sections were examined using a light microscope, and germ cell types were identified by their location, nuclear size, and chromatin pattern (Russell et al., 1990).

Immunostaining of sections

Postnatal or adult testes, or embryonic ovaries, were fixed one of three ways: in 4% paraformaldehyde (PFA) overnight followed by embedding in paraffin, in Bouins solution for 2 h followed by embedding in paraffin, or in 4% PFA for 1 h following by freezing in OCT (Sakura Finetek, Torrance, CA). Paraffin or frozen blocks were sectioned. Paraffin sections were dewaxed, rehydrated, and subject to antigen retrieval by heating in citrate buffer (10mM sodium citrate, 0.05% Tween 20, pH6.0). Frozen sections were thawed and washed in PBS. Sections were then blocked in 5% normal donkey serum, incubated with primary antibodies at 4˚C overnight, washed with PBS, incubated with the secondary antibody at room temperature for 1 h, and washed with PBS. Details for primary antibodies and their corresponding fixation and incubation conditions are detailed in S6 Table. For fluorescent detection, fluorophore-conjugated secondary antibodies were used at 1:250 (Jackson Immunoresearch Laboratories or Invitrogen), and sections were mounted in ProLong Gold Antifade reagent with DAPI (Thermo Fisher Scientific). For colorimetric detection, ImmPRESS peroxidase-conjugated secondary antibodies were used (Vector Laboratories), followed by detection using DAB substrate (Vector Laboratories). TUNEL staining was performed on PFA-fixed sections embedded in paraffin using the DeadEnd Colorimetric TUNEL System (Promega) according to the manufacturer's intstructions. Slides were then counterstained with hematoxylin, dehydrated, and mounted in Permount (Thermo Fisher Scientific). EdU was detected as per manufacturer's protocol (Click-iT EdU Alexa Fluor 488 Imaging Kit) after secondary antibody incubation and wash.

Immunostaining of chromosome spreads

Spreads were prepared from male and female meiotic germ cells as previously described [52] with some modifications. Male germ cells in suspension were obtained by mechanically

disrupting seminiferous tubules. Germ cells were spun down and resuspended in hypobuffer (30mM TrisHCl pH8.2, 50mM sucrose, 17mM sodium citrate) for 7 min at room temperature, then spun down again and resuspended in 100mM sucrose. Cell suspensions were placed on slides wetted with 1% PFA/0.15% TritonX-100. Female germ cells were obtained by first incubating embryonic ovaries in hypobuffer for 15 min, then mechanically disrupting the ovaries in 100mM sucrose. Dispersed cells were then placed on slides wetted with 1% PFA/0.2% TritonX-100. In both cases, slides were air dried, washed in 0.4% Photo-Flo, and stored at -80C until use. For immunofluorescence staining, frozen sections were thawed and washed in PBS. Sections were then blocked in 3% BSA/1% normal donkey serum/0.05% Triton-X, incubated with primary antibodies at 4˚C overnight, washed with PBS, incubated with the secondary antibody at room temperature for 1 h, and washed with PBS. Detailed information on primary antibodies and incubation conditions is provided in S6 Table. Fluorophore-conjugated secondary antibodies were used at 1:250 (Jackson Immunoresearch Laboratories or Invitrogen), and sections were mounted in ProLong Gold Antifade reagent with DAPI (Life Technologies).

Single molecule fluorescent in situ hybridization

Probe design, synthesis, and coupling were as previously described [53]. Probe sequences are provided in S7 Table. Samples were prepared and hybridization performed as previously described [17,53]. Germ cells were identified by smFISH for *Dazl* and/or nuclear morphology by DAPI staining.

RNA-seq

We performed RNA-seq on whole ovaries dissected away from mesonephros from E14.5 wild-type and *Meioc* 3L/3L fetuses. Each genotype was represented by three biological replicates of one pair of ovaries each. Total RNA (~1 μg) was extracted from ovaries using Trizol (Invitrogen) according to the manufacturer's protocol. Libraries were prepared using the Illumina TruSeq RNA Sample Preparation Kit. Libraries were multiplexed and sequenced on the Illumina HiSeq 2000 platform to obtain 40-base-pair single reads. RNA-seq data have been deposited in NCBI GEO under accession number GSE90702 and NCBI SRA under accession number SRP094112. Reads were aligned to the mouse genome (mm10) using TopHat v2.0.11 using default settings, and differential expression analysis was performed using Cufflinks v.2.2.1 [54] with the RefSeq transcript annotation. Enriched GO categories were identified using DAVID [55].

Immunoprecipitation for immunoblotting and mass spectrometry

To prepare lysates for immunoprecipitation followed by immunoblotting, one testis from a 3-month-old C57BL/6 male was homogenized in lysis buffer (25mM Tris-HCl pH7.5, 150mM NaCl, 1.5mM $MgCl_2$, 1mM dithiothreitol (DTT), 0.4% Triton X-100) supplemented with EDTA-free protease inhibitor (Roche Diagnostics) and 250U Benzonase nuclease (EMD Millipore), incubated at 4˚C with rotation for 30 min, and then centrifuged at 20,000 *g* for 15 min at 4˚C. For immunoprecipitation, the soluble lysate from each testis was pre-cleared for 2 h at 4˚C with Dynabeads Protein G (Thermo Fisher Scientific) prior to a 4˚C overnight incubation with antibody-bound Dynabeads. Beads were prepared by three brief washes in PBS with 0.1% Tween 20 (PBST) followed by resuspension in PBST and incubation with 5 μg of anti-MEIOC antibody (antibody generation described above) or normal rabbit IgG (Santa Cruz Biotechnology) for 2 h at room temperature. Following the overnight incubation, beads were washed three times with lysis buffer containing 150mM NaCl and transferred to a new tube.

To prepare lysates for mass spectrometry, immunoprecipitations were performed as described above with slight modifications: lysates were prepared from testes of P15 mice, and antibodies were crosslinked to the beads by a 30 min incubation with 5mM bis(sulfosuccinimidyl)suberate. Immunoprecipitations were performed in one of three conditions: wild-type (C57BL/6) lysate with IgG antibody, wild-type lysate with MEIOC antibody, or *Meioc*-deficient lysate with MEIOC antibody. Each condition was represented by two biological replicates, with one testis pair per replicate. The immunoprecipitates were washed three times in wash buffer (25mM Tris-HCl pH7.5, 150mM NaCl, 1.5mM $MgCl_2$, 1mM DTT), then washed twice with PBS.

Immunoblotting

Immunoprecipitated proteins were denatured in sample buffer for 10 min at 70˚C, resolved on a NuPAGE 4–12% Bis-Tris gel (Thermo Fisher Scientific), and transferred to a nitrocellulose membrane. The membrane was blocked in 5% BSA/Tris-buffered saline containing 0.1% Tween-20 (TBST) for 1 h at room temperature, incubated overnight at 4˚C with a primary antibody solution prepared in 5% BSA/TBST, and incubated for 1 h at room temperature with a 1:5,000 dilution of peroxidase-conjugated anti-rabbit IgG (Jackson Immunoresearch) prepared in 5% BSA/TBST. Proteins on the membrane were detected by the addition of Lumi-Light Western Blotting Substrate (Roche). Antibodies used for immunoblotting were MEIOC (1:2,000) and YTHDC2 (1:1,000; Bethyl Laboratories A303-026A).

Mass spectrometry

Immunoprecipitates were washed with 100mM NH_4HCO_3 and reduced (10 mM DTT, 56˚C for 45 min) and alkylated (50 mM iodoacetamide, in the dark at room temperature for 1 h). Proteins were subsequently digested with trypsin (sequencing grade, Promega, Madison, WI) at an enzyme/substrate ratio of 1:50 at room temperature overnight in 100 mM NH_4HCO_3 pH8. Trypsin activity was quenched by adding formic acid to a final concentration of 5%. Peptides were desalted using C18 SpinTips (Protea, Morgantown, WV) then vacuum centrifuged to near dryness and stored at −80˚C. Peptide labeling with TMT 6plex (Thermo Fisher Scientific) was performed per manufacturer's instructions. Samples were dissolved in 70 μL ethanol and 30 μL of 500 mM triethylammonium bicarbonate, pH8.5, and the TMT reagent was dissolved in 30 μL of anhydrous acetonitrile. The solution containing peptides and TMT reagent was vortexed and incubated at room temperature for 1 h. Samples labeled with the six different isobaric TMT reagents were combined and concentrated to completion in a vacuum centrifuge. The peptides were separated by reverse phase HPLC using an EASY- nLC1000 system (Thermo Fisher Scientific) over a 140-min gradient followed by nanoelectrospray using a QExactive mass spectrometer (Thermo Fisher Scientific). The mass spectrometer was operated in a data-dependent mode. The parameters for the full scan MS were: resolution of 70,000 across 350–2000 *m/z*, AGC $3e^6$, and maximum IT 50 ms. The full MS scan was followed by MS/MS for the top 10 precursor ions in each cycle with a NCE of 32 and dynamic exclusion of 30 s. Raw mass spectral data files (.raw) were searched using Proteome Discoverer (Thermo Fisher Scientific) and Mascot version 2.4.1 (Matrix Science). Mascot search parameters were: 10 ppm mass tolerance for precursor ions; 10mmu for fragment ion mass tolerance; 2 missed cleavages of trypsin. Fixed modifications were carbamidomethylation of cysteine and TMT 6plex modification of lysines and peptide N-termini; variable modification was oxidized methionine. Only peptides with a Mascot score greater than or equal to 25 and an isolation interference less than or equal to 30 were included in the quantitative data analysis. TMT quantification was obtained using Proteome Discoverer and isotopically corrected per manufacturer's

instructions. Mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [56] with the dataset identifier PXD005473.

RNA immunoprecipitation and sequencing (RIP-seq)/qPCR

MEIOC RIP-seq and IgG RIP-seq were carried out on P15 testes from wild-type C57BL/6 male mice (N = 2 per RIP-seq type). MEIOC RIP-seq was also carried out on P15 testes from wild-type and *Meioc*-deficient littermates (N = 2 per genotype). YTHDC2 RIP-seq and IgG RIP-seq were carried out on P20 testes from wild-type C57BL/6 male mice (N = 2 per RIP-seq type). To prepare lysates, testis pairs were isolated and lysed under non-reducing conditions (50mM Tris-HCl, pH7.4, 100mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate) supplemented with 40U/mL RNAseOUT (Thermo Fisher Scientific) and EDTA-free protease inhibitor (Roche Diagnostics). Lysates were incubated at 4˚C with rotation for 15–25 min and cleared using Ultrafiltration Spin Columns, 0.45 μm cutoff (EMD Millipore). Dynabeads Protein G were washed twice with lysis buffer and resuspended in lysis buffer at the original volume. The soluble lysate from each testis pair was pre-cleared for 1 h at 4˚C with 100 μl of Dynabeads Protein G (Thermo Fisher Scientific), and 40–80 μL was set aside as the input control. Beads were prepared by incubating 5 μg of anti-MEIOC antibody (antibody generation described above), one of two anti-YTHDC2 antibodies (Santa Cruz Biotechnology sc-249370 or Bethyl Laboratories A303-026A), or normal rabbit or goat IgG (Santa Cruz Biotechnology) per 100 μL Dynabeads with rotation for 45–60 min at room temperature. For immunoprecipitation, 570 μL lysate was incubated with 100 μL antibody-bound Dynabeads with rotation for 2 h at 4˚C. The beads were then washed six times for 5 min with rotation in wash buffer (50mM Tris-HCl, pH7.4, 300mM NaCl, 1mM EDTA, 1% NP-40, 0.1% SDS, and 0.5% sodium deoxycholate). A subset of the immunoprecipitate was then set aside for immunoblotting to verify successful immunoprecipitation of MEIOC and YTHDC2 (immunoblotting described above). The RNA from immunoprecipitates and input control was released by adding an additional 0.125% SDS and 250 mg/mL Proteinase K (Thermo Fisher Scientific) and incubating for 30 min with shaking at 37˚C. RNA was isolated via extraction with acid phenol:chloroform: IAA, pH4.5 (Thermo Fisher Scientific) using phase lock gel tubes (5 PRIME) according to the manufacturer's protocol. Extracted RNA was supplemented with GlycoBlue (Thermo Fisher Scientific) to 37.5 μg/mL and sodium acetate, pH5.5, to 0.1M. RNA was precipitated overnight at -20˚C in two volumes of 100% ethanol, pelleted by spinning for 20 min at 16,000 *g* at room temperature, washed once with 80% ethanol, dried, and resuspended in 25 μl water. For each sample, 5 μL of RNA was kept for qPCR analysis and the remaining RNA was used for sequencing library preparation via the SMARTer Stranded RNA-Seq Kit (ClonTech). Libraries from each RNA immunoprecipitation experiment (MEIOC or YTHDC2 RIP, IgG control RIP, and input control) were multiplexed and sequenced on the Illumina MiSeq platform. MEIOC RIP libraries were sequenced with 52-base-pair single-end reads. YTHDC2 RIP libraries were sequenced with 26-base-pair paired-end reads. Sequencing data have been deposited in NCBI GEO under accession number GSE90702 and NCBI SRA under accession number SRP094112. For qPCR analysis, RNA was reverse transcribed using Superscript VILO Master Mix (Thermo Fisher Scientific) and analyzed in triplicate using Power SYBR Green PCR Master Mix (Thermo Fisher Scientific) according to the manufacturer's protocol on a 7500 Fast Real-Time PCR System (Applied Biosystems). Primers for qPCR analyses are listed in S8 Table. Results were analyzed using *Actb* expression as a non-target normalization control and calculating the fold change over the IgG control RIP.

DESeq analysis of RIP-seq data

Prior to mapping, reads were trimmed for a minimum quality score of 20 and the first three bases of the first sequencing read, which were added during SMARTer Stranded library preparation, were removed using Cutadapt v1.8. Reads were aligned to the mouse genome (mm10) via TopHat v2.0.13 using default parameters and supplying the RefSeq transcript annotation. Alignments were converted to counts using HTSeq v0.6.1p1, using the "–a" option to skip reads whose alignment quality indicated non-unique alignments (i.e., alignment quality <50). DESeq2 v1.10.1 was then used to estimate RIP-seq enrichments resulting from MEIOC or YTHDC2 binding. DESeq2's default procedure was applied to normalize read counts across all samples. Data were analyzed with multi-factor designs to estimate protein-specific binding over controls. For YTHDC2 RIP-seq data, log2(read counts) for each gene was modeled as a linear combination of the gene-specific effects of three variables: binding to YTHDC2 ("YTHDC2"), binding to IgG ("IgG"), and batch ("batch") (S9A Table). The last variable captured differences due to the YTHDC2 antibody used and sequencing batch. This model identified transcripts that were enriched in YTHDC2 RIP-seq datasets generated using both antibodies. MEIOC analyses included RIP-seq experiments performed on wild-type and knockout samples. Read-count differences between wild-type and *Meioc*-deficient RIP-seq samples thus reflect both the effects of MEIOC protein binding and gene expression differences due to the *Meioc* genotype. To estimate the former independently of the latter, wild-type and *Meioc*-deficient RIP-seq and RNA-seq data were analyzed jointly, modeling log2 (read counts) as a linear combination of five variables: genotype, binding to MEIOC protein ("Meioc.specific"), binding to MEIOC antibody ("Meioc.nonspecific"), binding to IgG antibody ("IgG"), and sequencing batch (S9B Table). (For this analysis, RNA-seq data were summarized as gene-level read counts obtained from HTSeq, processed identically to the RIP-seq samples.) Enrichments (FDR < 0.05; fold change > 3 for MEIOC; fold change > 2 for YTHDC2) are reported as the fold changes between samples with and without protein-specific binding, independent of the effects of non-specific binding and sequencing batch. These were obtained from the *results* function in DESeq2 supplying the argument: *contrast = c("YTHDC2", 1, 0) or contrast = c ("Meioc.specific", 1, 0)*. For RIP-associated RNA-seq data, FPKMs were obtained using Cuffnorm v2.2.1 [54].

Supporting information

S1 Fig. MEIOC is conserved in vertebrates. Alignment of electronic predictions of MEIOC orthologs. We searched for homologs of mouse MEIOC (NP_001121048.1) by querying the RefSeq protein database by blastp, and the translated NCBI nucleotide collection database by tblastn. Both methods yielded similar results. We restricted the search to the following representative species: *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*, *Monodelphis domestica*, *Homo sapiens*, *Pan troglodytes*, *Anolis carolinensis*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Branchiostoma floridae*, *Ciona intestinalis*, *Strongylocentrotus purpuratus*, *Bombyx mori*, *Caenorhabditis elegans*, *Nematostella vectensis*, *Petromyzon marinus*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*. Homologs of mouse MEIOC (>80% query coverage and >30% identity) were aligned by Clustal Omega and visualized by Jalview (shown in figure). The box denotes a conserved domain annotated by PFAM as PF15189. Additional matches to MEIOC, restricted to the region corresponding to PF15189, were found in *Danio rerio*, *Branchiostoma floridae*, *Ciona intestinalis*, *Strongylocentrotus purpuratus*, *Bombyx mori*, *Caenorhabditis elegans*, and *Nematostella vectensis*. We were unable to identify matches to either full-length mouse MEIOC or the region corresponding to PF15189 in *Petromyzon marinus*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*.
(TIF)

S2 Fig. Testicular expression of *Meioc* is conserved. Expression of *Meioc* homologs in tissues from human, mouse, rat, and chicken as measured by RNAseq. RNAseq data of tissue panel from various species from Merkin et al., 2012. Expression of *Meioc* is measured in fragments per kilobase per million reads (FPKM). Amongst the species and tissues sampled, *Meioc* expression is predominantly in the testis. ND = no data. Note that the chromosomal events of meiotic prophase occur in the female during fetal stages, so we do not necessarily expect *Meioc* expression in the adult ovary.
(TIF)

S3 Fig. Immunofluorescence for MEIOC in wild-type and *Meioc*-deficient E16.5 ovary and P15 testis. Rabbit anti-MEIOC antibodies were generated to a peptide corresponding to the terminal 20 amino acids of mouse MEIOC (CHESINSSNPMNQRGETSKH). Germ cell-specific staining was observed in wild-type ovary and testis, but was absent in *Meioc*-/- ovary and testis. Sections were co-stained for MVH, to identify germ cells, and with DAPI to mark nuclei.
(TIF)

S4 Fig. Generation of *Meioc* mutant alleles.
(A) The *Meioc* gene was targeted for homologous recombination with a targeting vector for a knockout-first allele of *Meioc* (obtained from the KOMP Repository, vector PG00048_X_6_E03). Briefly, a 0.8 kb region containing exon 3 of the *Meioc* gene was replaced with a lacZ reporter, Neo selection marker, and exon 3, flanked by FRT (green triangles) and loxP (red triangles) sites. K: KpnI restriction site; X: XhoI restriction site. a, b, c, d, e: genotyping primers described in (F, G).
(B) The homologously targeted allele, denoted 3lox as it retains 3 loxP sites. The homologously targeted allele yields a 10.8 kb K/X fragment, whereas the wild-type allele yields a 18.9 kb K/X fragment. In the 3lox allele, *Meioc* is expected to be disrupted by the active lacZ reporter.
(C) Conversion of the 3lox allele to a conditional allele, denoted 2lox, by Flp recombination. The lacZ and Neo genes are excised, leaving exon 3 flanked by loxP sites.
(D) Conversion of the 2lox allele to a knockout allele, denoted 1lox, or *Meioc*-, by Cre recombination. Exon 3 of Meioc is excised. Both *Meioc* 3lox/3lox and *Meioc* 1lox/1lox (*Meioc* -/-) mice are considered *Meioc*-deficient.
(E) Southern blot confirmation of correctly targeted ES cell clones using a KpnI/XhoI restriction digest, and a probe 3' of the 3' homology arm.
(F) PCR assays for genotyping of wild-type (+/+), 3lox (3L), 2lox (2L), and 1lox (1L or -) alleles.
(G) Germline transmission of various *Meioc* alleles verified using indicated PCR assays.
(TIF)

S5 Fig. Histological analyses of *Meioc*-deficient adult testis and ovary.
(A) Wild-type and *Meioc* 3L/3L P30 testis and ovary.
(B, C) Hematoxylin and eosin-stained sections of adult (>8 weeks) testes from (B) wild-type and *Meioc* 3L/3L mice and (C) wild-type and *Meioc* -/- male mice. *Meioc*-deficient testes completely lacked postmeiotic germ cells, and were depleted for meiotic germ cells. The extent of this depletion varied among mice of mixed background: in some individuals, germ cells did not progress past preleptotene (prior to meiotic prophase), while in others, germ cells advanced to the zygotene stage of meiotic prophase. To obtain a reproducible phenotype, we backcrossed the *Meioc* mutant alleles onto the C57BL/6 background. In backcrossed mice, we consistently found that germ cells advanced to the zygotene stage. All experiments reported in

the main text were performed in mice backcrossed to the C57BL/6 background between five to seven generations (96.9–99.2% of genome expected to be of C57BL/6 origin), unless otherwise noted. All results were obtained using both *Meioc* 3L/3L and *Meioc* -/- mice, and phenotypes were consistent between the two alleles. pL–preleptotene spermatocyte, L–leptotene spermatocyte, Z–zygotene spermatocyte, P–pachytene spermatocyte, D–diplotene spermatocyte, ML–metaphase-like, rSt–round spermatid, St–spermatid, spz–spermatozoa.
(D) Hematoxylin and eosin-stained sections of adult ovaries from wild-type and *Meioc* -/- female mice. Wild-type adult ovaries contain oocytes contained within follicles at various stages of maturation (arrowheads). *Meioc* -/- adult ovaries are devoid of oocytes.
(TIFF)

S6 Fig. TUNEL analyses of *Meioc*-deficient adult testis and ovary.
(A) Wild-type and 1L/1L adult testis. In *Meioc*-/- adult testis, TUNEL staining was readily detected in cells with condensed (C) or apoptotic (A) nuclei. TUNEL staining was not detected in preleptotenes (pL) or in cells with metaphase-like chromosome condensation (M). TUNEL-positive cells were rarely detected in wild-type adult testes. Scale bar = 10 μm.
(B) Wild-type and 1L/1L E14.5 ovary. Low magnification images: most cells in both wild-type and *Meioc*-/- ovaries (o) were TUNEL-negative. High magnification image: a few TUNEL-positive cells were detected in the wild-type ovary. m, mesonephros. Scale bar = 10 μm (low magnification images) or 3.3 μm (high magnification image).
(TIF)

S7 Fig. Quantification of tubules containing metaphase-like cells. Percentage of tubule cross-sections containing preleptotene (pL), leptotene (L), zygotene (Z) and pachytene (P) cells in P15 *Meioc* -/- and control testes. We also determined the percentage of tubules containing metaphase-like cells, cells with condensed nuclei, or apoptotic cells. When a tubule contained, for example, a metaphase-like cell, we noted the stage of meiotic prophase found in that tubule. Each vertical column represents counts from one animal.
(TIFF)

S8 Fig. *Meioc*-deficient testicular and ovarian germ cells express molecular markers of meiotic prophase.
(A) Immunofluorescence staining for DMC1, γH2AX, and SYCP3 in chromosome spreads from wild-type and *Meioc* -/- germ cells from E16.5 ovaries. DNA stained by DAPI. In wild-type germ cells, we observed DMC1, γH2AX, and SYCP3 localization consistent with leptotene, and zygotene stages of meiotic prophase. In *Meioc* -/- germ cells, the most advanced stage of meiotic prophase we observed was leptotene stage. Although metaphase-like cells were observed in histological sections, we were unable to identify any metaphase-like cells in spreads.
(B) Frequencies of leptotene, zygotene, pachytene, or metaphase-like germ cells, or germ cells with other abnormal morphology, in cell spreads from P15 *Meioc* -/- and wild-type testes.
(TIFF)

S9 Fig. *Meioc*-deficient testicular and ovarian germ cells express molecular markers of metaphase. Immunofluorescence staining for LAMIN and pH3 in wild-type and *Meioc*-/- P15 testis and E16.5 ovary sections. Nuclei are stained by DAPI. In wild-type P15 testis and E16.5 ovary, meiotic germ cell nuclei are still intact, as detected by LAMIN staining, and no pH3 is observed. In *Meioc*-/- P15 testis and E16.5 ovary, LAMIN is not detected in germ cells which have condensed their nuclei and are pH3+. LAMIN and pH3 staining of wild-type adult testicular germ cells in metaphase I are shown for comparison.
(TIF)

S10 Fig. YTHDC2 is conserved in verterbrates. Alignment of electronic predictions of YTHDC2 orthologs. We searched for homologs of mouse YTHDC2 (NP_001156485) by querying the RefSeq protein database by blastp. We restricted the search to the following representative species: *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*, *Monodelphis domestica*, *Homo sapiens*, *Pan troglodytes*, *Anolis carolinensis*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Branchiostoma floridae*, *Ciona intestinalis*, *Strongylocentrotus purpuratus*, *Bombyx mori*, *Caenorhabditis elegans*, *Nematostella vectensis*, *Petromyzon marinus*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*. Homologs of mouse YTHDC2 ($\geq$75% query coverage and $\geq$25% identity) were aligned by Clustal Omega and visualized by Jalview (shown in figure). In *Drosophila melanogaster*, the homolog was annotated as benign gonial cell neoplasm (BGCN; NP_523832.2). Additional matches to YTHDC2 were found in *Danio rerio*, *Branchiostoma floridae*, *Ciona intestinalis*, *Strongylocentrotus purpuratus*, *Bombyx mori*, *Caenorhabditis elegans*, and *Nematostella vectensis*. We were unable to identify matches to full-length mouse YTHDC2 in *Petromyzon marinus* and *Saccharomyces cerevisiae*.
(TIF)

S11 Fig. MEIOC and IgG RIP-qPCR. A subset of targets and non-targets, identified via RIP-Seq and enrichment analysis, were verified via qPCR of the same P15 RIP samples analyzed via sequencing (N = 2). All ΔΔCt values were normalized to *Actb* qPCR results and displayed as fold change over IgG RIP-qPCR. Error bars represent s.e.m. Overall trends of target abundance in MEIOC RIP compared to IgG RIP are consistent with RIP-seq results. However, statistical analysis (one-tailed, paired Student t-test) did not show the statistical enrichment of any target in the MEIOC RIP ($p>0.05$ for all targets).
(TIF)

S1 Table. Gene expression levels and fold changes of wild-type and *Meioc* -/- E14.5 ovaries.
(XLSX)

S2 Table. GO categories enriched in genes expressed at higher or lower levels in E14.5 *Meioc* -/- ovaries.
(XLSX)

S3 Table. Unique peptides enriched in MEIOC immunoprecipitation (IP), identified via quantitative mass spectrometry. Samples A are IgG IP from wild-type lysates; samples B are MEIOC IP from wild-type lysates; and samples C are MEIOC IP from *Meioc*-deficient lysates.
(XLSX)

S4 Table. Enrichment in MEIOC RIP-seq and YTHDC2 RIP-seq from postnatal testis. MEIOC targets were defined as exhibiting a fold change >3, FDR<0.05, and FPKM>1; YTHDC2 targets were defined as exhibiting a fold change >2, FDR<0.05, and FPKM>1.
(XLSX)

S5 Table. MEIOC and YTHDC2 targets from postnatal testis that are differentially expressed in *Meioc*-deficient and wild-type E14.5 ovaries. MEIOC and YTHDC2 targets were identified from the RIP-seq analysis.
(XLSX)

S6 Table. Antibodies and experimental conditions for immunofluorescence stainings performed in this study.
(DOCX)

S7 Table. Single molecule FISH probes used in this study.
(XLSX)

S8 Table. List of qPCR primers used in this study.
(DOCX)

S9 Table. Coding of sample types for DESeq analysis of YTHDC2 and MEIOC RIP.
(DOCX)

Acknowledgments

We thank Gregoriy Dokshin, Peter Nicholls, Katherine Romer, and Marsha Wibowo for help with experiments; Tsutomu Endo, and Kyomi Igarashi for advice on experiments and analyses; Mary Goodheart for mouse husbandry; Kyomi Igarashi for genotyping; the Koch Institute ES Cell and Transgenics Facility for gene targeting and blastocyst injection to make the *Meioc* mouse line; the Whitehead Institute Genome Technology Core and Ting-Jan Cho for sequencing; Amanda del Rosario at the Koch Institute Biopolymers and Proteomics Facility for mass spectrometry and for contributing the mass spectrometry-related methods; and Iain Cheeseman, Terry Orr-Weaver, Peter Nicholls and Bluma Lesch for critical reading of the manuscript.

Author Contributions

Conceptualization: YQSS MMM MK DGdR DCP.

Formal analysis: YQSS MMM MK AKG DGdR.

Funding acquisition: DCP.

Investigation: YQSS MMM MK DGdR.

Methodology: YQSS MMM MK AKG.

Supervision: DCP.

Visualization: YQSS MMM MK AKG.

Writing – original draft: YQSS MMM.

Writing – review & editing: YQSS MMM MK AKG DGdR DCP.

References

1. Padmore R, Cao L, Kleckner N. Temporal comparison of recombination and synaptonemal complex formation during meiosis in S. cerevisiae. Cell. 1991; 66: 1239–1256. PMID: 1913808
2. Okaz E, Argüello-Miranda O, Bogdanova A, Vinod PK, Lipp JJ, Markova Z, et al. Meiotic prophase requires proteolysis of M phase regulators mediated by the meiosis-specific APC/C^{Ama1}. Cell. 2012; 151: 603–18. https://doi.org/10.1016/j.cell.2012.08.044 PMID: 23101628
3. Borum K. Oogenesis in the mouse. Exp Cell Res. 1961; 24: 495–507. PMID: 13871511
4. Speed RM. Meiosis in the foetal mouse ovary. Chromosoma. 1982; 85: 427–437. PMID: 6180868
5. Oakberg EF. Duration of spermatogenesis in the mouse and timing of stages of the cycle of the seminiferous epithelium. Am J Anat. 1956; 99: 507–16. https://doi.org/10.1002/aja.1000990307 PMID: 13402729
6. Liskay RM. Absence of a measurable G2 phase in two Chinese hamster cell lines. Proc Natl Acad Sci USA. 1977; 74: 1622–1625. PMID: 266201
7. Leblond CP, El-Alfy M. The eleven stages of the cell cycle, with emphasis on the changes in chromosomes and nucleoli during interphase and mitosis. Anat Rec. 1998; 252: 426–43. PMID: 9811221
8. Biedermann B, Wright J, Senften M, Kalchhauser I, Sarathy G, Lee M-H, et al. Translational repression of cyclin E prevents precocious mitosis and embryonic gene activation during *C. elegans* meiosis. Dev Cell. 2009; 17: 355–64. https://doi.org/10.1016/j.devcel.2009.08.003 PMID: 19758560

9. Sugimura I, Lilly MA. Bruno inhibits the expression of mitotic cyclins during the prophase I meiotic arrest of *Drosophila* oocytes. Dev Cell. 2006; 10: 127–35. https://doi.org/10.1016/j.devcel.2005.10.018 PMID: 16399084
10. Koubova J, Menke DB, Zhou Q, Capel B, Griswold MD, Page DC. Retinoic acid regulates sex-specific timing of meiotic initiation in mice. Proc Natl Acad Sci USA. 2006; 103: 2474–9. https://doi.org/10.1073/pnas.0510813103 PMID: 16461896
11. Bowles J, Knight D, Smith C, Wilhelm D, Richman J, Mamiya S, et al. Retinoid signaling determines germ cell fate in mice. Science. 2006; 312: 596–600. https://doi.org/10.1126/science.1125691 PMID: 16574820
12. Menke DB, Koubova J, Page DC. Sexual differentiation of germ cells in XX mouse gonads occurs in an anterior-to-posterior wave. Dev Biol. 2003; 262: 303–312. PMID: 14550793
13. Zhou Q, Nie R, Li Y, Friel P, Mitchell D, Hess RA, et al. Expression of stimulated by retinoic acid gene 8 (*Stra8*) in spermatogenic cells induced by retinoic acid: an in vivo study in vitamin A-sufficient postnatal murine testes. Biol Reprod. 2008; 79: 35–42. https://doi.org/10.1095/biolreprod.107.066795 PMID: 18322276
14. Anderson EL, Baltus AE, Roepers-Gajadien HL, Hassold TJ, de Rooij DG, van Pelt AMM, et al. *Stra8* and its inducer, retinoic acid, regulate meiotic initiation in both spermatogenesis and oogenesis in mice. Proc Natl Acad Sci USA. 2008; 105: 14976–80. https://doi.org/10.1073/pnas.0807297105 PMID: 18799751
15. Baltus AE, Menke DB, Hu Y-C, Goodheart ML, Carpenter AE, de Rooij DG, et al. In germ cells of mouse embryonic ovaries, the decision to enter meiosis precedes premeiotic DNA replication. Nat Genet. 2006; 38: 1430–4. https://doi.org/10.1038/ng1919 PMID: 17115059
16. Koubova J, Hu Y-C, Bhattacharyya T, Soh YQS, Gill ME, Goodheart ML, et al. Retinoic acid activates two pathways required for meiosis in mice. PLoS Genet. 2014; 10: e1004541. https://doi.org/10.1371/journal.pgen.1004541 PMID: 25102060
17. Soh YQS, Junker JP, Gill ME, Mueller JL, van Oudenaarden A, Page DC. A gene regulatory program for meiotic prophase in the fetal ovary. PLoS Genet. 2015; 11: e1005531. https://doi.org/10.1371/journal.pgen.1005531 PMID: 26378784
18. Abby E, Tourpin S, Ribeiro J, Daniel K, Messiaen S, Moison D, et al. Implementation of meiosis prophase I programme requires a conserved retinoid-independent stabilizer of meiotic transcripts. Nat Commun. 2016; 7: 10324. https://doi.org/10.1038/ncomms10324 PMID: 26742488
19. Moens PB, Spyropoulos B. Immunocytology of chiasmata and chromosomal disjunction at mouse meiosis. Chromosoma. 1995; 104: 175–182. PMID: 8529457
20. Meuwissen RL, Offenberg HH, Dietrich AJ, Riesewijk A, van Iersel M, Heyting C. A coiled-coil related protein specific for synapsed regions of meiotic prophase chromosomes. EMBO J. 1992; 11: 5091–100. PMID: 1464329
21. Eijpe M, Offenberg H, Jessberger R, Revenkova E, Heyting C. Meiotic cohesin REC8 marks the axial elements of rat synaptonemal complexes before cohesins SMC1β and SMC3. J Cell Biol. 2003; 160: 657–70. https://doi.org/10.1083/jcb.200212080 PMID: 12615909
22. Yoshida K, Kondoh G, Matsuda Y, Habu T, Nishimune Y, Morita T. The mouse *RecA*-like gene *Dmc1* Is required for homologous chromosome synapsis during meiosis. Mol Cell. 1998; 1: 707–718. PMID: 9660954
23. Pittman DL, Cobb J, Schimenti KJ, Wilson LA, Cooper DM, Brignull E, et al. Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for *Dmc1*, a germline-specific RecA homolog. Mol Cell. 1998; 1: 697–705. PMID: 9660953
24. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. DNA Double-stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139. J Biol Chem. 1998; 273: 5858–5868. PMID: 9488723
25. Lee J, Iwai T, Yokota T, Yamashita M. Temporally and spatially selective loss of Rec8 protein from meiotic chromosomes during mammalian meiosis. J Cell Sci. 2003; 116: 2781–90. https://doi.org/10.1242/jcs.00495 PMID: 12759374
26. Scherthan H. Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. J Cell Biol. 1996; 134: 1109–1125. PMID: 8794855
27. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010; 11: 220. https://doi.org/10.1186/gb-2010-11-12-220 PMID: 21176179
28. Satyanarayana A, Kaldis P. Mammalian cell-cycle regulation: several Cdks, numerous cyclins and diverse compensatory mechanisms. Oncogene. 2009; 28: 2925–39. https://doi.org/10.1038/onc.2009.170 PMID: 19561645

29. Ravnik SE, Wolgemuth DJ. Regulation of meiosis during mammalian spermatogenesis: the A-type cyclins and their associated cyclin-dependent kinases are differentially expressed in the germ-cell lineage. Dev Biol. 1999; 207: 408–18. https://doi.org/10.1006/dbio.1998.9156 PMID: 10068472

30. Ravnik SE, Wolgemuth DJ. The developmentally restricted pattern of expression in the male germ line of a murine *cyclin A*, *cyclin A2*, suggests roles in both mitotic and meiotic cell cycles. Dev Biol. 1996; 173: 69–78. https://doi.org/10.1006/dbio.1996.0007 PMID: 8575639

31. Liu D, Matzuk MM, Sung WK, Guo Q, Wang P, Wolgemuth DJ. Cyclin A1 is required for meiosis in the male mouse. Nat Genet. 1998; 20: 377–80. https://doi.org/10.1038/3855 PMID: 9843212

32. Sweeney C, Murphy M, Kubelka M, Ravnik S, Hawkins C, Wolgemuth D, et al. A distinct cyclin A is expressed in germ cells in the mouse. Development. 1996; 122: 53–64. PMID: 8565853

33. Nguyen TB, Manova K, Capodieci P, Lindon C, Bottega S, Wang X-Y, et al. Characterization and expression of mammalian cyclin b3, a prepachytene meiotic cyclin. J Biol Chem. 2002; 277: 41960–9. https://doi.org/10.1074/jbc.M203951200 PMID: 12185076

34. Refik-Rogers J, Manova K, Koff A. Misexpression of cyclin B3 leads to aberrant spermatogenesis. Cell Cycle. 2014; 5: 1966–1973. PMID: 16929180.

35. Bourgeois CF, Mortreux F, Auboeuf D. The multiple functions of RNA helicases as drivers and regulators of gene expression. Nat Rev Mol Cell Biol. 2016; 17: 426–38. https://doi.org/10.1038/nrm.2016.50 PMID: 27251421

36. Grishin N, Pabo CO, Sauer RT, Harrison SC, Musco G, al. et, et al. The R3H motif: a domain that binds single-stranded nucleic acids. Trends Biochem Sci. 1998; 23: 329–330. PMID: 9787637

37. Wang X, He C. Reading RNA methylation codes through methyl-specific binding proteins. RNA Biol. 2014; 11: 669–72. https://doi.org/10.4161/rna.28829 PMID: 24823649

38. Insco ML, Bailey AS, Kim J, Olivares GH, Wapinski OL, Tam CH, et al. A self-limiting switch based on translational control regulates the transition from proliferation to differentiation in an adult stem cell lineage. Cell Stem Cell. 2012; 11: 689–700. https://doi.org/10.1016/j.stem.2012.08.012 PMID: 23122292

39. Barford D. Understanding the structural basis for controlling chromosome division. Philos Trans A Math Phys Eng Sci. 2015; 373.

40. Shimizu M, Nomura Y, Suzuki H, Ichikawa E, Takeuchi A, Suzuki M, et al. Activation of the rat cyclin A promoter by ATF2 and Jun family members and its suppression by ATF4. Exp Cell Res. 1998; 239: 93–103. https://doi.org/10.1006/excr.1997.3884 PMID: 9511728

41. Desdouets C, Matesic G, Molina CA, Foulkes NS, Sassone-Corsi P, Brechot C, et al. Cell cycle regulation of cyclin A gene expression by the cyclic AMP-responsive transcription factors CREB and CREM. Mol Cell Biol. 1995; 15: 3301–9. PMID: 7760825

42. Hayakawa J, Mittal S, Wang Y, Korkmaz KS, Adamson E, English C, et al. Identification of promoters bound by c-Jun/ATF2 during rapid large-scale gene activation following genotoxic stress. Mol Cell. 2004; 16: 521–535. https://doi.org/10.1016/j.molcel.2004.10.024 PMID: 15546613

43. Cobb J, Cargile B, Handel MA. Acquisition of Competence to Condense Metaphase I Chromosomes during Spermatogenesis. Dev Biol. 1999; 205: 49–64. https://doi.org/10.1006/dbio.1998.9101 PMID: 9882497

44. Li Y, Minor NT, Park JK, McKearin DM, Maines JZ. Bam and Bgcn antagonize *Nanos*-dependent germline stem cell maintenance. Proc Natl Acad Sci USA. 2009; 106: 9304–9. https://doi.org/10.1073/pnas.0901452106 PMID: 19470484

45. Shen R, Weng C, Yu J, Xie T. eIF4A controls germline stem cell self-renewal by directly inhibiting BAM function in the *Drosophila* ovary. Proc Natl Acad Sci USA. 2009; 106: 11623–8. https://doi.org/10.1073/pnas.0903325106 PMID: 19556547

46. Tang H, Ross A, Capel B. Expression and functional analysis of *Gm114*, a putative mammalian ortholog of *Drosophila bam*. Dev Biol. 2008; 318: 73–81. https://doi.org/10.1016/j.ydbio.2008.03.001 PMID: 18423593

47. McKearin DM, Spradling AC. *bag-of-marbles*: a *Drosophila* gene required to initiate both male and female gametogenesis. Genes Dev. 1990; 4: 2242–2251. PMID: 2279698

48. Gonczy P, Matunis E, DiNardo S. *bag-of-marbles* and *benign gonial cell neoplasm* act in the germline to restrict proliferation during *Drosophila* spermatogenesis. Development. 1997; 124: 4361–4371. PMID: 9334284

49. Ohlstein B, McKearin D. Ectopic expression of the *Drosophila* Bam protein eliminates oogenic germline stem cells. Development. 1997; 124: 3651–3662. PMID: 9342057

50. McKearin D, Ohlstein B. A role for the *Drosophila* Bag-of-marbles protein in the differentiation of cystoblasts from germline stem cells. Development. 1995; 121: 2937–47. PMID: 7555720

51. Hu Y-C, de Rooij DG, Page DC. Tumor suppressor gene *Rb* is required for self-renewal of spermatogonial stem cells in mice. Proc Natl Acad Sci USA. 2013; 110: 12685–12690. https://doi.org/10.1073/pnas.1311548110 PMID: 23858447

52. Peters AH, Plug AW, van Vugt MJ, de Boer P. A drying-down technique for the spreading of mammalian meiocytes from the male and female germline. Chromosome Res. 1997; 5: 66–8. PMID: 9088645

53. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008; 5: 877–9. https://doi.org/10.1038/nmeth.1253 PMID: 18806792

54. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7: 562–78. https://doi.org/10.1038/nprot.2012.016 PMID: 22383036

55. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008; 4: 44–57. PMID: 19131956

56. Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016; 44: D447–56. https://doi.org/10.1093/nar/gkv1145 PMID: 26527722

57. Russell LD. Histological and histopathological evaluation of the testis. Cache River Press; 1990.

58. Fujiwara Y, Komiya T, Kawabata H, Sato M, Fujimoto H, Furusawa M, et al. Isolation of a DEAD-family protein gene that encodes a murine homolog of *Drosophila* vasa and its specific expression in germ cell lineage. Proc Natl Acad Sci USA. 1994; 91: 12258–12262. PMID: 7991615

APPENDIX B. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals

Authors:

Sahin Naqvi[1,2], Alexander K. Godfrey[1,2], Jennifer F. Hughes[1], Mary L. Goodheart[1,3], Richard N. Mitchell[4], David C. Page[1,2,5]

Affiliations:

[1] Whitehead Institute, Cambridge, MA, USA
[2] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA, USA
[4] Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Author Contributions:

SN, AKG, JFH, and DCP designed the study. JFH procured cyno tissue samples. MLG procured mouse and rat tissue samples, with assistance from SN. SN processed tissue samples and performed computational analyses, with assistance from AKG. RNM performed histological evaluations on human tissue sections. DCP supervised work. SN and DCP wrote the paper.

Published as:

Naqvi, S., Godfrey, A. K., Hughes, J. F., Goodheart, M. L., Mitchell, R. N., & Page, D. C. (2019). Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science*, *365*(6450), eaaw7317. http://doi.org/10.1126/science.aaw7317

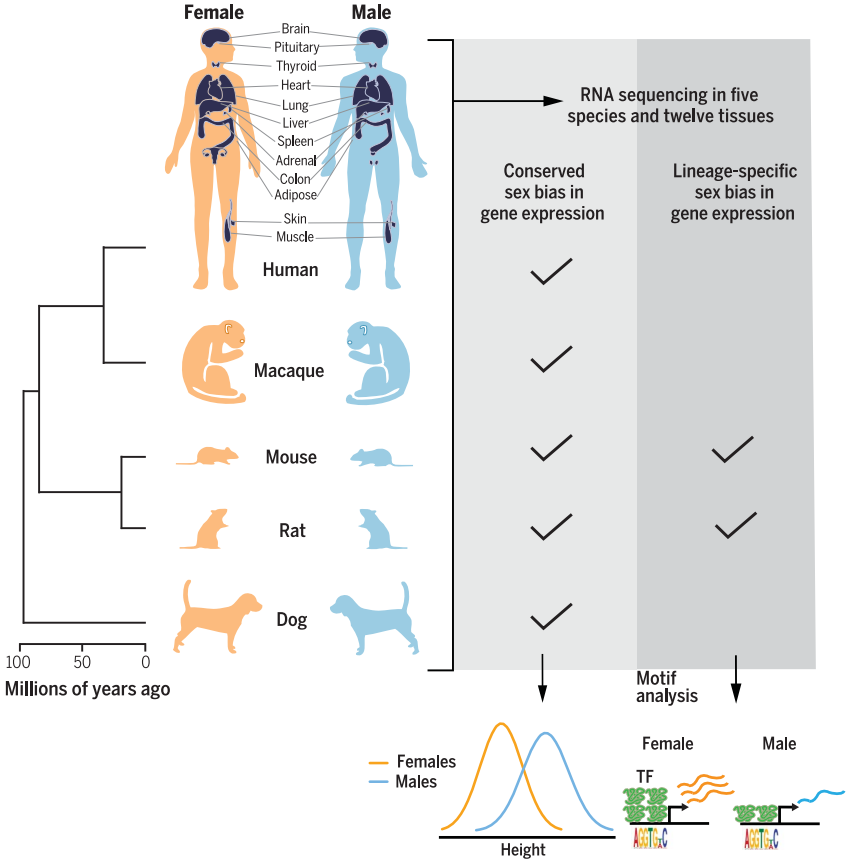
RESEARCH ARTICLE SUMMARY

COMPARATIVE GENETICS

Conservation, acquisition, and functional impact of sex-biased gene expression in mammals

Sahin Naqvi, Alexander K. Godfrey, Jennifer F. Hughes, Mary L. Goodheart, Richard N. Mitchell, David C. Page*

INTRODUCTION: Sex differences are widespread in humans and other mammals. For example, the distribution of height or body size is shifted upwards in males relative to females, and sex differences are found in the immune and cardiovascular systems as well as in metabolism. However, little is known about how gene expression differs between the sexes in a broad range of mammalian tissues and species. A catalog of such sex-biased gene expression could help us understand phenotypic sex differences. Assessing the extent to which sex-biased gene expression is conserved across the body could also have important implications for the use of nonhuman mammals as models of sex-biased human biology.



RNA sequencing of male and female samples in 12 tissues and five species reveals the functional impact and mechanistic underpinnings of sex-biased gene expression. A survey of sex differences in gene expression using RNA sequencing data (left) leads to the discovery of both conserved (species-shared) and lineage- or species-specific sex biases in expression across the genome. Genes with conserved sex bias contribute to the sex difference in mean height in humans and other mammals, whereas lineage-specific changes can be partially explained by gains and losses of motifs for sex-biased TFs.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaw7317

RATIONALE: To identify both conserved and lineage- or species-specific sex differences in gene expression, we sequenced RNA from male and female samples in 12 tissues in each of four nonhuman mammals (cynomolgus macaque, mouse, rat, and dog) and analyzed these data jointly with publicly available data from postmortem male and female human tissues. To assess the impact of sex-biased gene expression on the sex difference in mean human height, we applied methods that integrate the effects of genetic variation on both gene expression and phenotype (height in this case). We sought to understand which transcription factors (TFs) contribute to evolutionary changes in sex bias by analyzing motifs gained or lost concurrently with lineage- or species-specific changes in sex bias.

RESULTS: Linear modeling revealed ~3000 genes with conserved (species-shared) sex bias in gene expression, most of which was tissue specific. The cumulative effects of conserved sex bias explain ~12% of the sex difference in mean human height, and cases such as that of *LCORL*, a TF with conservation of both female-biased expression and genetic association with height, suggest a contribution to sex differences in body size beyond humans. However, most sex-biased gene expression (~77%) was specific to single species or subsets of species, implying that it arose more recently during evolution. We identified 83 instances where TFs showed sex-biased expression in the same tissue, in which their motifs were associated with gain or loss of sex bias at other genes, accounting for a significant portion (~27%) of lineage-specific changes in sex bias.

CONCLUSION: By conducting a 12-tissue, five-species survey of sex differences in gene expression, we found that although conserved sex bias in gene expression exists throughout the body, most sex bias has been acquired more recently during mammalian evolution. Height is likely subject to opposing selective pressures in males and females; our study thus documents how such selective forces can result in sex-biased expression which, when layered upon genetic pathways acting identically in males and females, can lead to trait distributions shifted between the sexes. Our findings also suggest that, in many cases, molecular sex differences observed in humans may not be mirrored in nonhuman mammals. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: dcpage@wi.mit.edu
Cite this article as S. Naqvi *et al.*, *Science* **365**, eaaw7317 (2019). DOI: 10.1126/science.aaw7317

RESEARCH ARTICLE

COMPARATIVE GENETICS

Conservation, acquisition, and functional impact of sex-biased gene expression in mammals

Sahin Naqvi[1,2], Alexander K. Godfrey[1,2], Jennifer F. Hughes[1], Mary L. Goodheart[1,3], Richard N. Mitchell[4], David C. Page[1,2,3]*

Sex differences abound in human health and disease, as they do in other mammals used as models. The extent to which sex differences are conserved at the molecular level across species and tissues is unknown. We surveyed sex differences in gene expression in human, macaque, mouse, rat, and dog, across 12 tissues. In each tissue, we identified hundreds of genes with conserved sex-biased expression—findings that, combined with genomic analyses of human height, explain ~12% of the difference in height between females and males. We surmise that conserved sex biases in expression of genes otherwise operating equivalently in females and males contribute to sex differences in traits. However, most sex-biased expression arose during the mammalian radiation, which suggests that careful attention to interspecies divergence is needed when modeling human sex differences.

Males and females exhibit differences across a wide range of biological processes. Studies in humans have documented sex differences in anthropometric traits (*1*), energy metabolism (*2*), brain morphology (*3*), and immune (*4*) and cardiac (*5*) function. Sex differences are also evident in the incidence, prevalence, and mortality across diseases, including autoimmune disorders (*6*), cardiovascular diseases (*7*), and autism (*8*). Sex differences are common in other mammals besides humans, many of which are models of sex-biased human traits and diseases (*9*). For example, males are larger than females in most mammalian species (*10*), whereas sex differences in brain structures (*11*) and immune (*12*) and cardiac (*13*) function have been observed in rodents. These phenotypic sex differences are likely associated with, and may be caused by, sex differences in gene activity or function.

The sex chromosomes are one source of sex differences in gene activity. The Y chromosome harbors male-specific genes (*14*), some broadly expressed (*15*). Incomplete inactivation of the second X chromosome in females results in female-biased expression of some X-linked genes (*16*). However, given the scale and complexity of gene networks, and the greater number of autosomal genes, it is unlikely that sexually dimorphic expression of sex-linked genes accounts for all phenotypic sex differences in mammals. Understanding the molecular origins of these sex differences therefore requires a genome-wide, multitissue, and comparative approach to sex biases in gene expression.

Our understanding of sex bias in mammalian gene expression is lacking in three regards. First, the degree to which sex-biased expression is conserved across the mammalian lineage and the extent of conservation in different tissues and organ systems are unknown. Multitissue studies of sex bias in gene expression focused on humans (*17*, *18*) or mice (*19*). Multispecies studies in *Drosophila* (*20*–*24*) examined RNA from whole carcasses or gonads, whereas studies in mammals that examined nonreproductive tissues focused on single tissues (*25*, *26*). Second, little is known about how sex differences in gene expression across the body cumulatively result in phenotypic sex differences. Sex-biased expression of the autosomal genes *VGLL3* (*27*) and *IL-33* (*28*), as well as the X-linked gene *TLR7* (*29*), appears to contribute to sexually dimorphic immune phenotypes. However, most complex traits are polygenic and underpinned by variation in hundreds or even thousands of genes (*30*). Third, apart from single-gene studies in *Drosophila* (*31*), lineage-specific regulatory changes that drive the evolution of sex-biased expression remain unexplored. Progress has been made in understanding mechanisms of X-linked dosage compensation (*32*, *33*), the lack of which can lead to sex-biased expression on the X chromosome, but additional mechanisms likely contribute to genome-wide sex-biased gene expression. Thus, previous studies sought to understand the extent of sex-biased expression across either tissues (*17*–*19*) or species (*25*, *26*), or they explored its phenotypic impact (*27*–*29*) or underlying evolutionary mechanisms (*31*) for individual genes. Assessing sex-biased expression across tissues and species, together with its cumulative contribution to phenotypic sex differences, would advance our understanding of molecular differences between males and females.

Results

A five-species, 12-tissue survey of sex differences in gene expression

To assess sex differences in nonhuman mammals, we collected RNA sequencing data from three males and three females from cynomolgus macaque (*Macaca fascicularis*, cyno), mouse (*Mus musculus*), rat (*Rattus norvegicus*), and dog (*Canis familiaris*). Together with humans, these five species, whose last common ancestor lived 80 to 100 million years ago, span the evolution of the Boreoeutheria, including all placental mammals except Afrotheria and Xenarthra (which include the elephant and anteater, respectively). We sampled 12 tissues from each individual: adipose, adrenal gland, brain, colon, heart, liver, lung, muscle, pituitary, skin, spleen, and thyroid. These tissues represent many organ systems and all three germ layers (Fig. 1A). We designed tissue collection and processing procedures to minimize biological and technical variation (*34*) (table S1). We used our RNA sequencing (RNA-seq) data to systematically improve the transcriptome annotations of each nonhuman mammalian species, which we then assessed using the percentage of reads from independent studies that mapped to our annotations versus existing annotations (e.g., a 16% increase in read mapping rate in dog) (fig. S1).

To assess sex differences in humans, we analyzed RNA-seq data from the Genotype-Tissue Expression Consortium (GTEx, v6p release) (*35*). To reduce the possibility of sex biases in cell-type composition, pathology, or other factors driving our results, we performed stringent quality control for samples from each of the 12 target tissues using individual- and sample-level metadata from GTEx and our own evaluation of histological images (*34*) (table S2). We adjusted gene expression values using top principal components to remove variation due to hidden technical or biological confounders. In three tissues (adipose, brain, and skin) for which expression data from purified cell populations is available, there is a correlation between sample-level cell-type proportions estimated by CIBERSORT (*36*) and top principal component loadings (fig. S2). Although this approach controls for variation in cell-type composition in the human samples, we acknowledge that some sex biases, especially those specific to nonhuman mammals, could reflect sex differences in cell-type composition.

We removed outlier samples (*34*) to obtain 740 human and 277 nonhuman RNA-seq samples (see table S3 for human sample sizes by sex and tissue). We clustered all nonhuman samples and a randomly chosen subset of human samples, using the expression levels of 12,939 one-to-one orthologous protein-coding genes.

[1]Whitehead Institute, Cambridge, MA 02142, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, MA 02142, USA. [4]Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.
*Corresponding author. Email: dcpage@wi.mit.edu

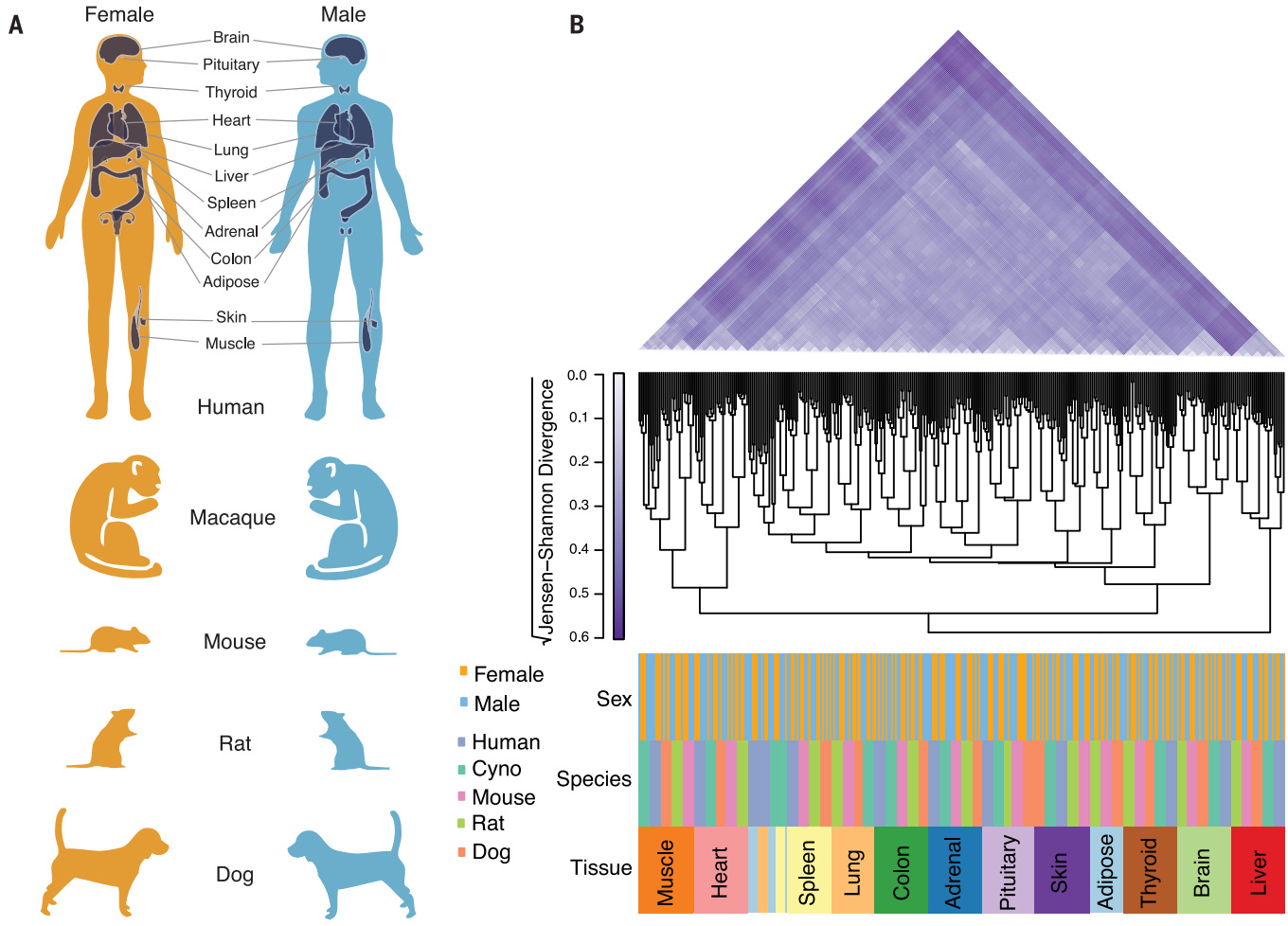


Fig. 1. Five-species, 12-tissue survey of sex differences in gene expression. (**A**) Schematic of study design, with tissues chosen for analysis in all five species highlighted in humans. (**B**) Hierarchical clustering of 349 RNA-seq samples. (Top) Pairwise estimates of Jensen-Shannon divergence (JSD) between pairs of samples. Six random human samples per tissue, in addition to all nonhuman samples, were included for display purposes. (Middle) Tree dendrogram obtained by hierarchical clustering (average linkage) based on pairwise JSD values. (Bottom) Sample labels by tissue, species, and sex.

With the exception of human adipose tissue and lung, which cluster closely together, samples cluster first by tissue and then by species (Fig. 1B). This tissue-dominated clustering agrees with prior studies (*37*, *38*) and indicates consistent sampling of tissues across species and also that the nonhuman data generated in this study are comparable to the human data from GTEx (*35*). There are no cases where samples cluster by sex before tissue or species, indicating that species effects dominate over sex effects. Nevertheless, sex contributes significantly to gene expression variation as pairwise within-sex distances in each tissue-species combination are significantly lower than pairwise between-sex distances (fig. S3).

Both our reanalysis of GTEx data and our analysis of our own data replicated published estimates of sex bias in six human and mouse tissues (*27*, *39*–*45*) (Pearson's correlation coefficient r = 0.29 to 0.92) (figs. S4 and S5). These results indicate that expression values are comparable across species and yield reproducible estimates of sex bias.

Conserved sex-biased gene expression exists across the body

Within each tissue, we used a linear mixed model to identify genes that showed a consistent sex bias [false discovery rate (FDR) 4.5%, as estimated by permutation of male and female sample labels] across species while controlling for differences in expression variability and sample size between species. Further, we required that genes show a fold change ≥ 1.05 in the same direction in at least four of the five species studied. We assume that such genes likely had a conserved sex bias in the common ancestor of Boreoeutheria (example in Fig. 2A). Of 113,853 expressed gene-tissue pairs, 3885 pairs (corresponding to 3161 genes) show a conserved sex bias. We used a rank-based statistic to confirm that gene-tissue pairs with conserved sex bias also have low P values for sex bias in each of the individual species (fig. S6). Conserved sex bias is generally of modest magnitude (~90% of sex-biased gene-tissue pairs had a less than twofold change between the sexes) (fig. S7) but reproducible in independent datasets (Pearson's r = 0.18 to 0.78) (fig. S8). The number of genes with conserved sex bias per tissue varies from 128 in colon to 805 in pituitary (Fig. 2B and table S4) and is not correlated with tissue sample size or rates of between-species gene expression divergence (Pearson's r = 0.093 and 0.0083 and P = 0.77 and 0.97, respectively) (fig. S9). A naïve approach, requiring $P < 0.05$ in at least four of five species for each tissue, found a smaller number of gene-tissue pairs with conserved sex bias but revealed between-tissue patterns that were correlated with results from the linear mixed model (fig. S10). Of genes with conserved sex bias in any of the 12 tissues examined, 562 genes (18%) are sex-biased in more than one tissue (Fig. 2C). In cases of multitissue sex bias, the bias is significantly more likely to be in the same direction in multiple tissues (P = 0.00035, two-sided Fisher's exact test) (Fig. 2D). Thus, conserved sex bias in gene expression is mostly tissue-specific, but a significant minority of genes shows concordant sex bias across multiple tissues, implying that some regulatory factors result in similar profiles of sex-biased expression in multiple tissues or cell types.

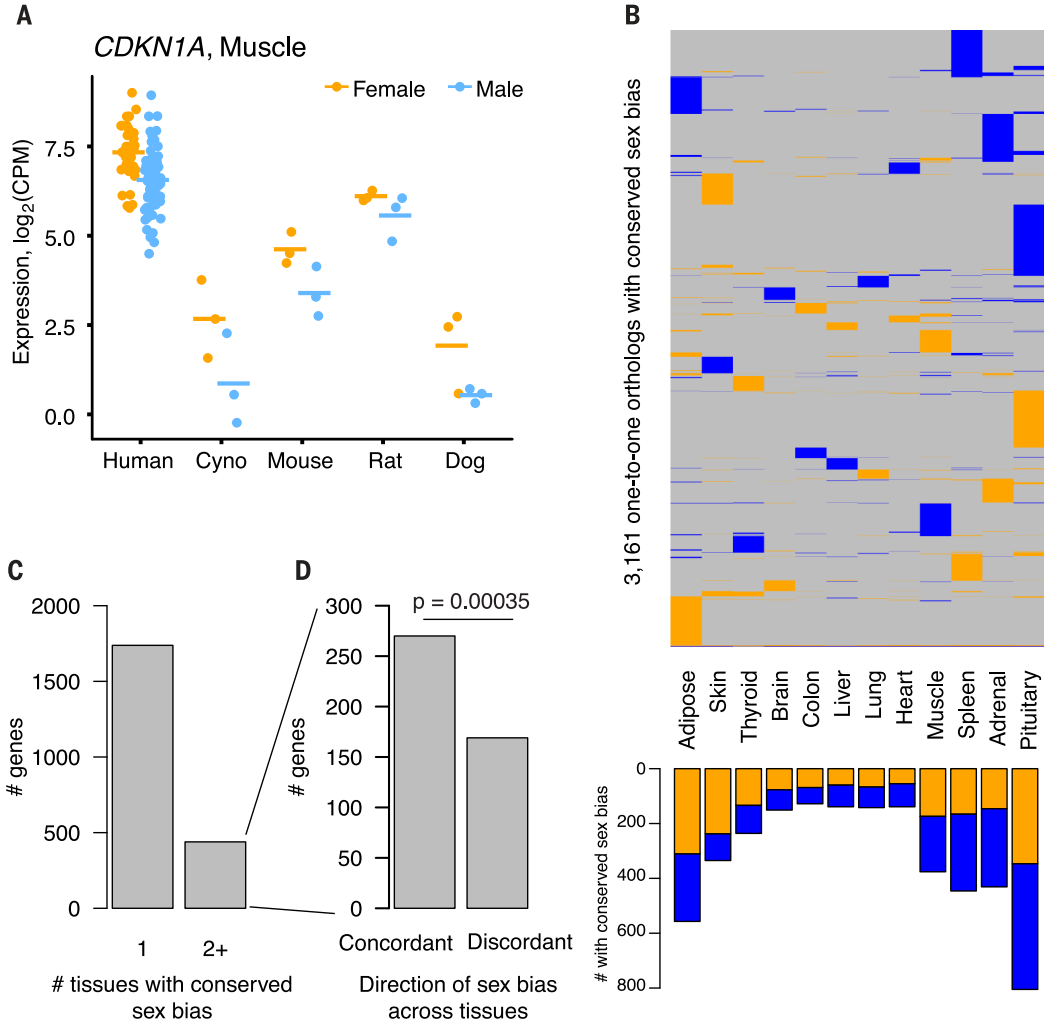


Fig. 2. Conserved sex bias in gene expression across the body. (**A**) Example of gene with conserved female-biased expression. CPM, counts per million. (**B**) Heatmap of conserved male (blue) and female (orange) sex bias across genes (rows) and tissues (columns). (**C**) The *y* axis represents number of genes with conserved sex bias in one (left) or multiple (right) tissues. (**D**) Of genes with conserved sex bias in multiple tissues, the number concordant (same direction) or discordant (opposite direction) in multiple tissues is plotted. Significance as assessed by two-sided Fisher's exact test comparing to equal proportions.

We considered the extent to which genes with conserved sex-biased expression were enriched for sex linkage. All assayed Y-linked genes are male-biased (fig. S11A), as expected, whereas X-linked genes are significantly enriched for conserved female bias (2.1- to 10.2-fold increase relative to autosomes two-sided Fisher's exact test) (fig. S11B). The enrichment for X-linked genes is driven by genes that escape X inactivation in females. In turn, the enrichment for X-escape genes is largely driven by the subset of X-escape genes that have a nonrecombining Y-linked homolog in mammals (two-sided Fisher's exact test) (fig. S11B). Despite these enrichments, most (85 to 95%, depending on the tissue) genes with conserved sex bias are autosomal (fig. S12). We compared the magnitude of sex bias between autosomal and X-linked genes using independent, publicly available datasets (*27*, *39*, *40*, *43*, *44*) (seven mouse, three human) to avoid ascertainment bias. X-linked genes show significantly higher magnitudes of sex bias in four of the ten datasets (adjusted $P < 0.05$, two-sided Wilcoxon rank-sum test) (fig. S13). Thus, the sex chromosomes, primarily as a result of harboring genes with both X- and Y-linked homologs, contribute a small but significant fraction of conserved sex bias in gene expression.

Most sex bias in gene expression has arisen since the last common ancestor of boreoeutherian mammals

We investigated sex-biased gene expression specific to subsets of the five species, mindful that differences in statistical power between species could result in false positive calls of lineage-specific sex bias. For example, a gene with true primate-specific male bias might falsely appear to have a human-specific male bias if its expression is significantly biased in humans but does not reach statistical significance in cyno. At the same time, false positive calls of sex bias in single species will by necessity appear to be species specific. We used mashr (*46*) to model the covariation in sex bias across tissues and species and to more confidently determine the lineage of sex bias. We repeated the mashr procedure using permuted male/female sample labels to empirically estimate the FDR for any given set of sex-biased genes (*34*). This increased the number of rodent-specific gains of sex bias in most tissues (fig. S14). After using mashr to estimate sex bias in each tissue-species combination, we assigned each sex-biased gene-tissue pair (other than those with conserved sex bias) to one of 12 lineage-specific categories by parsimony: primate-specific gains or losses, rodent-specific gains or losses, gains specific to one of the five species, multiple gains or losses, and more complex patterns of sex bias inconsistent with single gains or losses (examples in Fig. 3A and table S4). In each category, we used the permutation-estimated FDR to estimate the

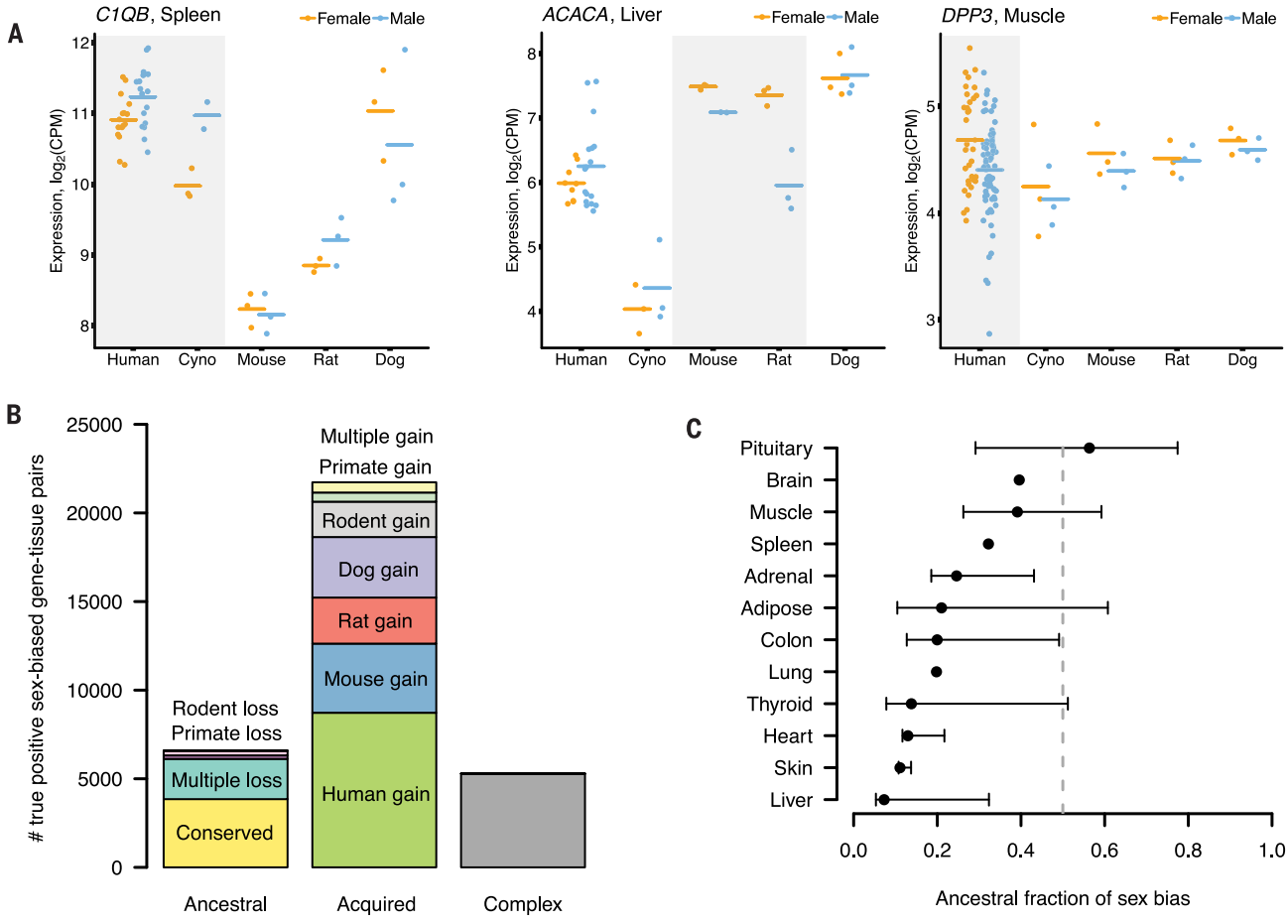


Fig. 3. Most sex bias in gene expression has arisen since the last common ancestor of Boreoeutheria. (**A**) Examples of genes with lineage-specific sex bias. (**B**) Number of true-positive sex-biased gene-tissue pairs (*y* axis) in each evolutionary class was calculated as the difference between the total number discovered across all tissues using true or permuted sex labels. Evolutionary classes defined in main text are designated as ancestral, acquired, or complex relative to last common ancestor of Boreoeutheria (the five species considered here). (**C**) Comparisons of ancestral to acquired sex biases as in (B), but performed in each tissue separately. Upper and lower confidence intervals represent fraction of sex bias estimated to be ancestral when counting all complex events as ancestral or acquired, respectively.

number of true positive sex-biased gene-tissue pairs.

We assessed how much of the sex-biased gene expression observed in the five species was present in the common ancestor of Boreoeutheria (i.e., ancestral). Instances of ancestrally sex-biased expression included gene-tissue pairs that we previously identified as having a "conserved" sex bias as well as gene-tissue pairs that lost sex bias in the primate or rodent lineages or in multiple lineages. Instances of acquired sex bias included gene-tissue pairs with primate-, rodent-, or species-specific sex bias, as well as multiple gains of sex bias. By this logic, 6539 (23%) of sex-biased gene-tissue pairs were likely sex-biased in the common ancestor, and 22,194 (77%) likely acquired sex bias after divergence from a common ancestor. An additional 8495 gene-tissue pairs exhibited more complex patterns and could not be confidently assigned as ancestrally sex-biased or acquired (Fig. 3B). If all such "complex" events were ancestral, the ancestral fraction of sex bias would be 40%, whereas if they were acquired, the fraction would be 18%. Performing these calculations in each tissue separately, we found that ancestral sex bias constituted the minority of total sex bias in all tissues except the pituitary (Fig. 3C). We also quantified the fraction of ancestral bias using a range of fold-change cutoffs up to 1.5 and found that, for all cases, ancestral sex bias was in the minority (fig. S15). Repeating this analysis with conserved sex bias called by mashr rather than the linear mixed model yielded similar results, with both methods detecting similar numbers of gene-tissue pairs with conserved sex bias (fig. S15). We conclude that most sex bias in gene expression in nonreproductive tissues arose during, rather than before, the boreoeutherian radiation.

Sex-biased gene expression is associated with reduced selective constraint

We assessed the degree of selective constraint operating on sex-biased gene expression. Reasoning that genes functioning across many tissues and cell types face increased selective constraint on gene expression levels, we compared the breadth of expression of genes with and without sex bias, in each tissue. Sex-biased genes showed significantly lower expression breadth than genes with no bias, with the exception of lung, where sex-biased genes were more broadly expressed (adjusted $P < 0.05$, two-sided Wilcoxon rank-sum test) (Fig. 4A). These differences in expression breadth could either be downstream consequences of, or have predated, the observed sex bias. We thus analyzed expression breadth in chicken, an evolutionary outgroup to mammals, reasoning that patterns found in both human and chicken were likely present in the common mammalian ancestor before the acquisition of sex bias. Again, sex-biased genes in mammals showed almost uniformly lower expression breadth in chicken than unbiased genes (adjusted $P < 0.05$, two-sided Wilcoxon rank-sum test) (fig. S16).

To assess conservation of expression levels in a tissue-specific manner, we used estimates of

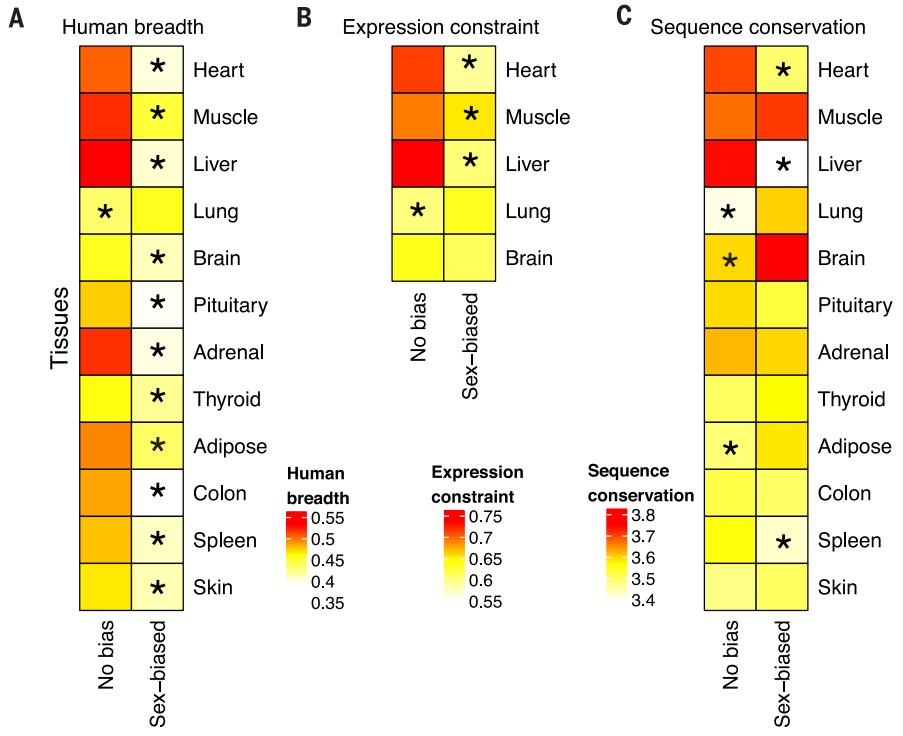


Fig. 4. Sex-biased gene expression is associated with reduced selective constraint. In each tissue, genes were binned as showing no sex bias or showing sex bias of any evolutionary type. Human breadth (**A**) was calculated on the basis of median expression values in the 12 selected GTEx tissues (*34*), expression constraint (**B**) represents the genome-wide percentile, and sequence conservation (**C**) is calculated as the mean coding phyloP score (*34*). In each heatmap, the group median of the indicated gene-level trait is plotted; asterisks indicate a Benjamini-Hochberg–adjusted $P < 0.05$ from a two-sided Wilcoxon rank-sum test, placed on the group ("No bias" or "Sex-biased") with the lower value of the gene-level trait.

mammalian gene expression-level constraint learned from 16 species (*47*) and seven tissues, five of which were also assessed in our study. As with expression breadth, sex-biased genes showed lower constraint than unbiased genes in heart, muscle, and liver but higher constraint in lung (adjusted $P < 0.05$, two-sided Wilcoxon rank-sum test) (Fig. 4B).

We observed that genes sex-biased in heart, spleen, and liver showed lower sequence conservation than unbiased genes, whereas genes sex-biased in adipose, brain, and lung showed higher sequence conservation than unbiased genes (adjusted $P < 0.05$, two-sided Wilcoxon rank-sum test) (Fig. 4C). Thus, some sex-biased genes are relatively strongly constrained at the sequence level, perhaps because they perform important or pleiotropic functions. However, considering both expression levels and sequence conservation, our findings indicate that sex-biased gene expression is primarily associated with reduced selective constraint, from before the divergence of the boreoeutherian lineages.

Conserved sex bias in autosomal gene expression contributes to sex differences in mammalian height and body size

Males are larger than females in most mammalian taxa (*10*). Human males are, on average, 10 to 15 cm (7 to 13%) taller than females, but the distributions of height in males and females overlap substantially (Fig. 5A). The genetic architecture of human height is polygenic and largely shared between the sexes. A recent meta-analysis reported 712 genome-wide significant loci (*48*), and only a handful of sex-specific associations with height have been discovered (*49*, *50*); recent human studies reported a between-sex genetic correlation of 0.96 (*50*, *51*). Studies in humans and other mammals have concluded that height is likely subject to opposing selective pressures between the sexes, where increased height enhances reproductive success in males and decreased height favors reproductive success in females (*52*, *53*).

Could sex-biased gene expression contribute to the sex difference in height and body size observed in humans and other mammals? To link variation in gene expression to variation in height, we turned to transcriptome-wide association studies (TWAS) that integrate an expression quantitative trait loci (eQTL) study from a given tissue or cell type and genome-wide association studies (GWAS) of a given trait (*54*).

If sex-biased gene expression contributes to sex differences in height, genes with male-biased expression levels should mostly identify height-increasing effects, as measured by TWAS, whereas female-biased genes should identify height-decreasing effects. We considered genes with genome-wide significant associations for height, as annotated in the NHGRI-EBI GWAS catalog (*55*). We used TWAS to combine height GWAS statistics from a meta-analysis (*48*) of data from the UK Biobank (*56*) and GIANT consortium (*57*) (which we verified were correlated; Pearson's $r = 0.83$, $P < 2.2 \times 10^{-16}$) (fig. S17) with reference eQTL panels from 43 different tissues generated using data from GTEx (*35*). TWAS z-scores largely agree in sign across the 43 tissues (figs. S18 and S19; see table S5 for all TWAS z-scores). We therefore combined z-scores for each gene across tissues by meta-analysis. Sixty-two genome-wide significant height genes have both computed TWAS z-scores and conserved sex bias in at least one tissue. Genes with conserved male-biased expression have more-positive TWAS z-scores than genes with conserved female-biased expression (mean z-score difference = 18, $P = 0.023$, group permutation test), but this difference was not seen when analyzing genes with human-specific or primate-specific sex bias (Fig. 5B). Expanding our analyses to include TWAS results for all genes allowed for greater stringency by only considering TWAS z-scores calculated for the same tissue in which sex-biased expression was observed. Five hundred sixty gene-tissue pairs have both computed TWAS z-scores and conserved sex bias; these are distributed across all 12 tissues, with the largest numbers in muscle, adipose, and pituitary (fig. S20), and they are enriched for metabolic functions (adjusted P value < 0.05, two-sided Fisher's exact test) (table S6). Gene-tissue pairs with conserved male bias have more-positive TWAS z-scores than those with conserved female bias (mean z-score difference = 0.7, $P = 0.039$, group permutation test), but this difference was not seen when considering gene-tissue pairs with human- or primate-specific sex bias (Fig. 5C). Together, these results indicate that genes with conserved male-biased expression show height-increasing effects, whereas genes with conserved female-biased expression show height-decreasing effects.

We sought to quantify the fraction of sex difference in height explained by conserved sex bias in gene expression, focusing on cases where the sex bias was in the same tissue as the TWAS z-score (i.e., Fig. 5C). We estimated the contribution of conserved sex bias to the height sex difference with two approaches. One approach used a physical scale with the effect sizes of eQTLs in GWAS, and the other examined a relative fold-change on the basis of TWAS z-scores (*34*) (fig. S21). The two approaches yielded similar estimates of the contribution of conserved sex-biased gene expression: ~1.6 cm, or 12%, of the observed sex difference in mean height.

Genes with conserved male and female bias show the largest difference in height TWAS z-scores, suggesting a contribution to sex differences in size in other mammals. Indeed, all five species assessed in this study exhibit sex differences in size (fig. S22). Consider the transcription factor *LCORL*, which shows conserved female

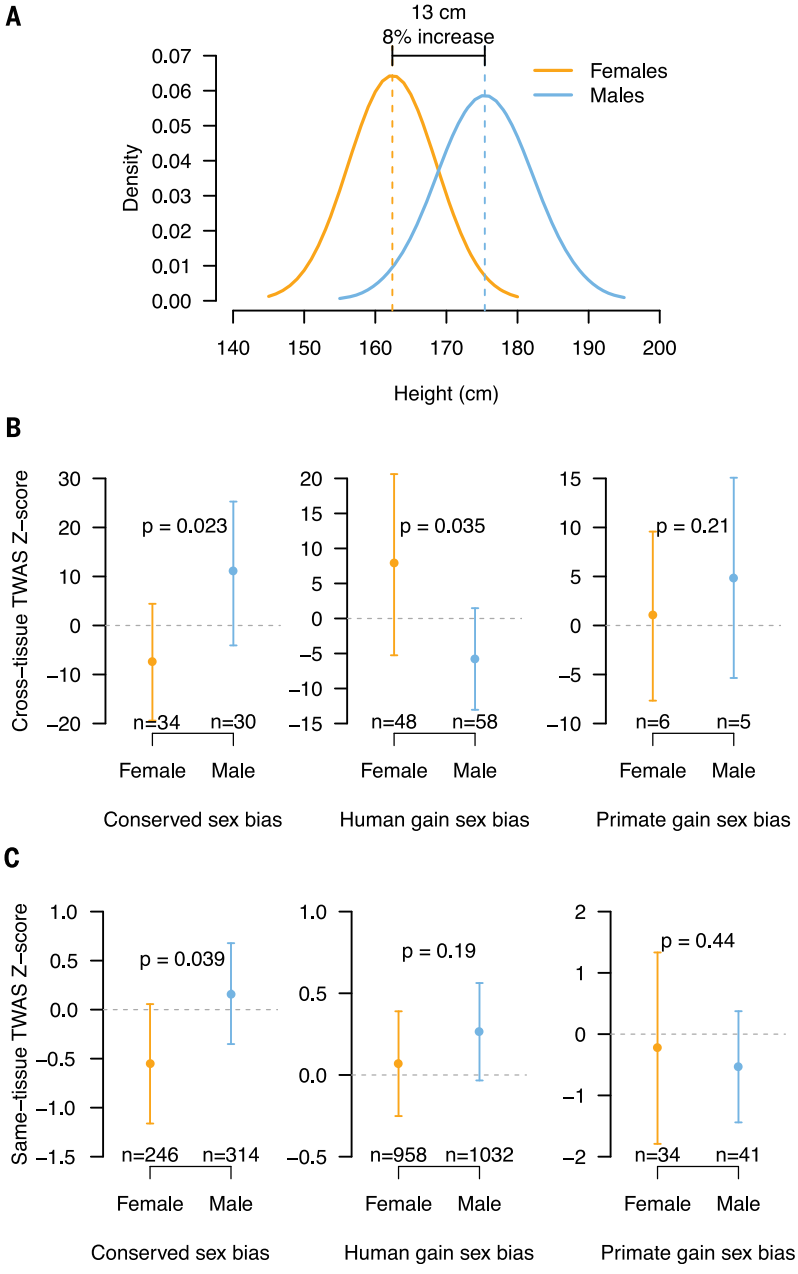


Fig. 5. Conserved sex bias in autosomal gene expression contributes to sex differences in human height. (**A**) Overlapping but shifted distributions of male and female heights. Theoretical normal distributions using published means and standard deviations of male and female heights in individuals of European ancestry from the United Kingdom (*53*). (**B**) TWAS *z*-scores for genome-wide significant height genes with either female (orange) or male (blue) bias in one of 12 tissues, either conserved across mammals (left), specific to humans (middle), or specific to primates (right). For each gene, TWAS *z*-scores were meta-analyzed across 48 GTEx tissues. (**C**) TWAS *z*-scores for gene-tissue pairs with either female (orange) or male (blue) bias, either conserved across mammals (left), specific to humans (middle), or specific to primates (right), in all cases in same tissue as computed TWAS *z*-score. Points represent group means; whiskers represent 95% confidence intervals. *P* value for mean difference calculated by 1000 permutations of male and female point labels.

bias in the pituitary (fig. S23A) and is height-decreasing in humans (cross-tissue TWAS *z*-score = −28.7). Although *LCORL* lacks a predictive TWAS model in the pituitary, an allele at the *LCORL* locus associated with increased expression in the pituitary is associated with decreased height (fig. S23B). Notably, genetic variation at the *LCORL* locus has been associated with height or body size in dogs (*58*), cattle (*59*), and horses (*60*). Reanalysis of publicly available RNA-seq data (*61*) shows that *LCORL* is one of the most strongly female-biased autosomal genes in the cattle pituitary (1.6-fold higher in females) (fig. S23C and table S7). An allele associated with increased body size in horses is associated with decreased *LCORL* expression in hair root (*62*), indicating that the negative association between *LCORL* expression and height likely extends beyond humans. These observations suggest that female-biased expression of *LCORL* contributes to sex differences in size in multiple species. Beyond *LCORL*, studies have observed significant overlap in genome-wide-significant height loci between humans (*57*), dogs (*58*), and cattle (*59*) (table S8), suggesting a broader contribution of conserved sex-biased gene expression to sex differences in body size in a range of mammals.

Conserved and acquired sex biases in gene expression are similarly enriched for specific biological pathways and show similar magnitudes of bias

Our results indicate that although sex-biased gene expression overall shows signs of lowered selective constraint, conserved sex bias contributes to sex differences in height or body size. This raises the possibility that more-recently acquired sex bias in expression has little or no functional impact. To assess this possibility, we compared genes with either conserved or acquired sex bias with respect to (i) overrepresentation in specific biological pathways via Gene Ontology (GO) category enrichment and (ii) the magnitude of their sex bias.

We observed similar degrees of enrichments for biological pathways among conserved and acquired sex biases. Genes with conserved male bias in pituitary are enriched for cyclic adenosine monophosphate signaling, which functions in response to stress (*63*). Genes with conserved female bias in colon and thyroid are enriched for adaptive immune pathways, whereas genes with conserved female bias in adipose tissue are enriched for mitochondrial translation and ribosomal RNA processing. At the same time, genes with acquired male bias in the liver, adipose tissue, and heart are enriched for functions related to fatty acid metabolism, regulation of hormone secretion, and nucleotide metabolism, respectively, and genes with acquired female bias in the liver are enriched for extracellular matrix organization (adjusted $P < 0.05$, two-sided Fisher's exact test) (table S6).

We compared the magnitude of conserved and acquired sex bias using 10 independent human and mouse datasets to minimize differences due to ascertainment; we found no significant differences (adjusted $P > 0.05$, two-sided Wilcoxon rank-sum test) (fig. S24). Considering that conserved sex bias, although generally small in magnitude, can nevertheless contribute to sex differences in height, these results suggest that acquired sex bias could also be functionally consequential. Additional studies are needed to demonstrate the functional impact of acquired sex-biased gene expression.

Evolutionary turnover of motifs for sex-biased transcription factors reflects lineage-specific changes in sex bias

One mechanism by which sex-biased expression could evolve is via sex-biased transcription factors (TFs). For example, male-biased TF expression in muscle would result in higher TF activity in male muscle. Genes that acquired motifs for this TF in, for example, the primate lineage would then show a primate-specific sex bias in muscle. To test this idea, we searched for motifs enriched in the promoters of sex-biased genes with lineage-specific changes in a given tissue, relative to their unbiased orthologs. We repeated this analysis with random, equally sized sets of genes showing no lineage-specific sex bias in order to calculate an empirical *P* value for motif enrichment (*34*). Because of the nonparametric nature of this *P* value calculation and our desire to analyze enriched motifs inclusively as a set, we considered motifs at a 10% FDR. We found 83 instances in which such motifs matched predicted binding sites of TFs with sex bias in the same tissue (Fig. 6A and table S9). This was significantly more (P = 0.014, tissue permutation test) than the ~67 instances of matches expected when randomly assigning the tissue of sex bias for each TF (Fig. 6B), which, combined with the 10% FDR for motif discovery, yields ~6.7 matches expected by chance. By quantifying the enrichment of each motif in its corresponding set of sex-biased orthologs (*34*), we estimated that these 83 instances account for the lineage-specific sex bias of 6073 gene-tissue pairs, or 27% of all lineage-specific sex bias. Furthermore, 13 TFs showed matches to enriched motifs in more than one tissue, significantly more than the approximately one such TF expected by chance (P = 0.032, tissue permutation test), as determined by the motif discovery FDR and permuting the tissue of TF sex bias as described above (fig. S25). This suggests that gains and losses of motifs for sex-biased TFs could, in some cases, coordinate the evolution of sex-biased gene expression across multiple tissues or cell types.

To confirm that genes with lineage-specific gains or losses of motifs for sex-biased TFs are TF-bound in living cells, we leveraged publicly available data from chromatin immunoprecipitation sequencing (ChIP-seq) in human and mouse (table S9). Although these assays were almost invariably performed in a different cell type than the tissue of TF sex bias and motif gain, we reasoned that sex-biased genes with gained motifs in a given tissue should nevertheless show enrichment for TF ChIP-seq signal. Eleven of 15 cases with available data showed significant enrichment of ChIP-seq peaks in the promoters of genes with a gain or loss of sex bias and the relevant motif, relative to a background set of

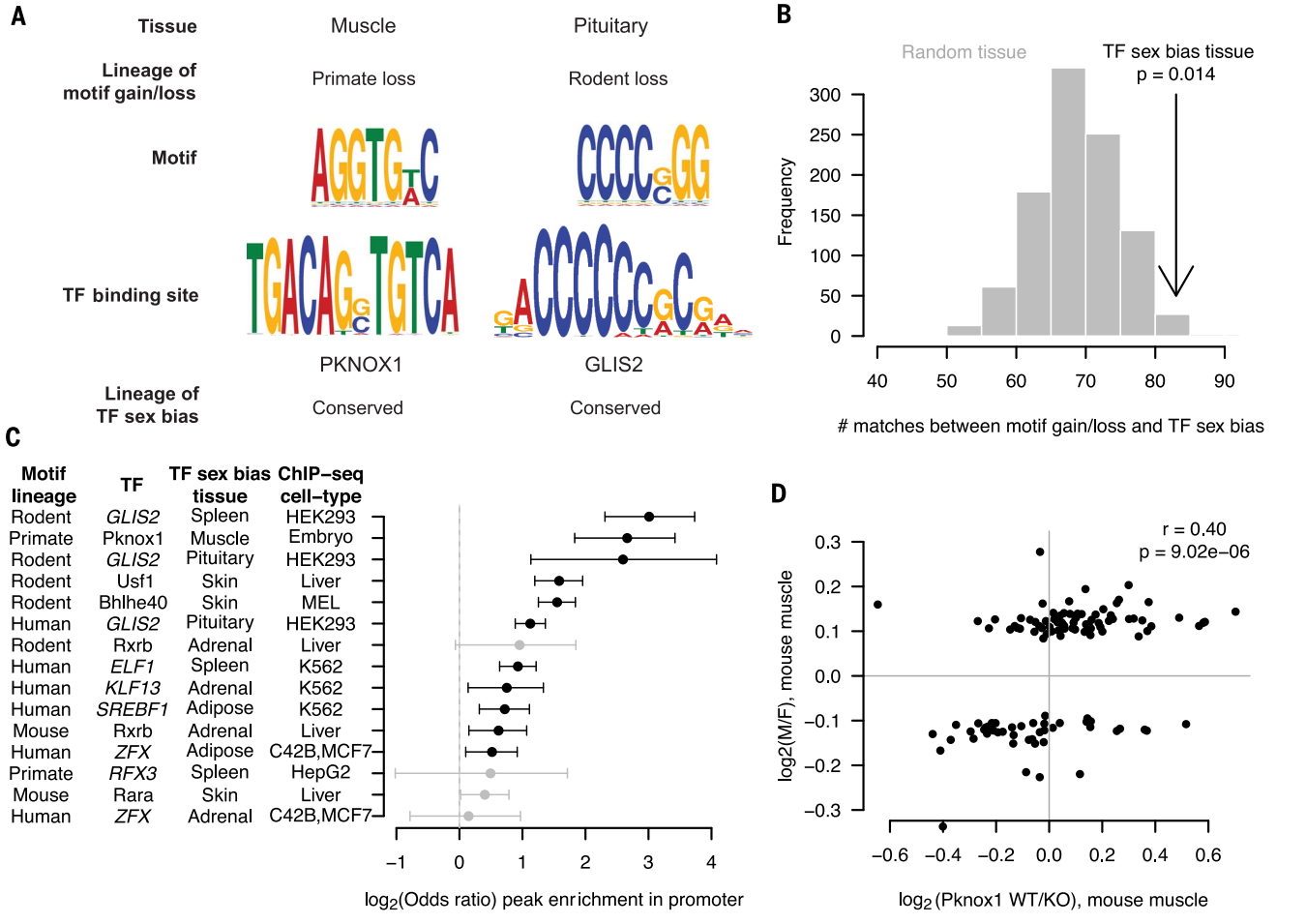


Fig. 6. Evolutionary turnover of motifs for sex-biased transcription factors is associated with gains and losses of sex bias. (**A**) Representative gained or lost motifs in promoters of genes with lineage-specific gains or losses of sex bias (top) aligned with motifs for sex-biased TFs in same tissue (bottom). The lineage of sex bias gain or loss is indicated above each motif; the sex-biased TF and lineage of its sex bias are indicated below. (**B**) Total number of matches between gained or lost motifs and sex-biased TFs when considering tissue of TF sex bias (black) or randomly chosen tissues (gray). (**C**) Enrichment of ChIP-seq peaks in promoters of genes with lineage-specific sex biases containing gained or lost motifs for the TF. The sex-biased TF, along with tissue of sex bias and motif gain or loss and cell type in which ChIP-seq was performed, are indicated to left. The $\log_2$ odds ratio for genes with lineage-specific sex bias and containing the motif as compared with a background set of genes with no motif is shown on the *x* axis, with 95% confidence intervals by Fisher's exact test. (**D**) Effect of *Pknox1* knockout (*x* axis) (*64*) versus sex bias (*y* axis), both in mouse muscle, for genes that show loss of sex bias in primate lineage and contain a motif for PKNOX1 in mouse.

genes lacking the motif (two-sided Fisher's exact test), which is consistent with this prediction (Fig. 6C). Thus, the evolutionary gains and losses of motifs we observed likely correspond to gains and losses of binding by cognate TFs.

If gains and losses of motifs for sex-biased TFs contribute to lineage-specific changes in sex bias in their target genes, there should be directional agreement between the activating or repressive effect of the TF, the sex bias of the TF, and the sex bias of the target gene. For example, target genes activated (or repressed) by a male-biased TF should be male (or female) biased, and the opposite should be true for female-biased TFs. Rigorously testing this prediction requires experimental manipulation of the TF in the tissue where lineage-specific changes in sex bias are observed. Such data are available for PKNOX1, a homeobox TF with conserved male-biased expression in muscle (*64*) (fig. S26). Genes with loss of muscle-specific sex bias in the primate lineage show depletion (at a stringent 5% FDR) of PKNOX1-matching motifs relative to mouse, rat, and dog (Fig. 5A, examples of PKNOX1 targets in fig. S26). Genes with a PKNOX1-matching motif show significant positive correlation between the effect of *Pknox1* knockout (*64*) and the effect of sex on muscle gene expression (Pearson's $r = 0.40$, $P = 9.02 \times 10^{-6}$) (Fig. 6D). Thus, both ChIP-seq and TF knockout data confirm that gains and losses of regulation by sex-biased TFs have contributed to the evolution of sex bias.

Discussion

Comparative studies of sex-biased gene expression have implications for the use of nonhuman mammals as models of sex-biased human traits or diseases. Conserved sex bias in gene expression across the body indicates that certain molecular sex differences in humans are amenable to study in a wide range of mammalian model organisms. However, in many cases, nonhuman models may not adequately recreate the human sex differences in question. This is supported by two lines of evidence: (i) in each tissue, samples cluster by species rather than sex, and (ii) most sex bias in gene expression has arisen recently and is thus not shared between most mammals. For example, genetic variants that decrease expression of the TF *KLF14* in human adipose tissue tend to increase insulin resistance and risk for type 2 diabetes only in females, but elimination of *Klf14* expression in mouse adipose tissue leads to analogous phenotypes in both sexes (*65*). Nonhuman mammals may still be useful as models of physiological or systems-level sex differences, but caution should be exercised when extrapolating specific molecular findings to humans.

We find that conserved sex bias in autosomal gene expression explains ~12% of the sex difference in mean human height, whereas all common single-nucleotide polymorphisms are thought to explain 60% of the heritability of height (*66*). Although these two numbers are not directly comparable (the former relates to between-group differences, the latter to between-individual variation), these height heritability estimates suggest that additional genes and instances of sex bias relevant to sex differences in height remain to be discovered. Deletions of the *SHOX* gene, located in the pseudoautosomal region of the human X and Y chromosomes, contribute to short stature in Turner syndrome (*67*), whereas increases in sex chromosome number (and thus *SHOX* dosage) increase height (*68*). Although the height GWAS used here excluded the pseudoautosomal regions, precluding analysis of *SHOX*, targeted studies indicate that *SHOX* dosage is positively correlated with height (*67*, *68*). In light of reports that expression of *SHOX* is male-biased in multiple tissues (*16*), it may be that *SHOX* contributes a fraction of the sex difference in height [discussed further in (*69*)].

Studies of selection on height have illustrated how males and females can have different optimal values for a quantitative trait, with increased height favored in males and decreased height favored in females (*53*). Our finding that conserved sex bias in gene expression contributes to sex differences in height suggests one way in which such optimal values can be reached—through the acquisition and maintenance of sex-biased gene expression (*70*). Thus, although some conserved sex bias in gene expression may have arisen through selectively neutral processes, opposing selective forces between the sexes appear to have been at work here. Height is also subject to balancing selection, in which extreme variation in either direction negatively impacts reproductive fitness (*53*). A recent study in *Drosophila* found a strong signature of balancing selection at loci with opposite fitness effects in females and males, establishing that sexual antagonism and balancing selection can coincide (*71*). Future studies may identify mechanisms that reduce fitness at the extremes of the height distributions in both sexes. Whereas our study focused exclusively on height, genetic pleiotropy may broaden the reach of our findings. Sex-biased gene expression resulting from opposing selective pressures on male and female height could result in sex differences in phenotypes yet to be identified.

Our results also illustrate one way in which sex-biased gene expression can lead to phenotypic sex differences: autosomal genes, operating identically in males and females to influence a trait, can be expressed more abundantly in one sex. Although most genetic variation influencing height acts identically between the sexes, pronounced sex-specific genetic effects have been demonstrated for waist-to-hip ratio, body mass index (*72*–*74*), thyroid hormone levels (*75*), and obsessive-compulsive disorder (*76*). Fully accounting for such sex differences in genetic architecture in association mapping (*77*), and integrating this information with sex biases in gene expression, may reveal additional mechanisms underlying phenotypic sex differences.

Our finding of sex-biased TFs underlying lineage-specific changes in sex bias provides molecular insight into mechanisms underlying the evolution of sex-biased gene expression in nonreproductive mammalian tissues. We focused on regulatory changes in promoter regions because of the lack of tissue-specific enhancer annotations in cyno, rat, or dog. However, single-gene studies in *Drosophila* indicate that sex-biased gene expression can evolve through more complex changes in cis-regulatory elements at larger genomic distances from their target gene (*31*). Studying gains and losses of TF binding motifs in promoters, although an important first step, is a simplifying approach. It is thus necessary to catalog both tissue and species specificity of mammalian enhancers to enable detailed analyses of the cis-regulatory changes driving gains or losses of sex-biased gene expression during mammalian evolution. Most of the sex-biased TFs we identified as contributing to lineage-specific evolution of sex bias are autosomal, indicating that their sex biases could arise as a result of trans-regulatory effects of sex chromosomes or sex hormones. Distinguishing between these two possibilities is an important future direction for research.

Materials and methods summary

Human (GTEx) samples were filtered on the basis of cause of death, medical history, and notes from GTEx pathologists (*35*), and additional detailed evaluations were conducted on samples with available histology. Samples from cynomolgus macaque, mouse, rat, and dog were collected within 1 hour of euthanizing healthy animals, and only tissues from nonestrous females were used. RNA extraction, library preparation, and RNA-seq of nonhuman mammals were performed in batches randomized with respect to tissue, species, and sex. Analyzing the combined human and nonhuman dataset, we used a linear mixed model (*78*) to identify genes showing consistent sex bias across species in each tissue, and we used mashr to identify lineage-specific changes in sex bias. These analyses were repeated with permuted male/female sample labels to empirically estimate FDRs. Magnitudes of sex bias were assessed in independent datasets by reanalyzing raw data, where available. For lineage-specific sex biases in each tissue, motif analysis (*79*, *80*) was used to identify TF binding sites enriched in the set of sex-biased orthologs relative to the unbiased orthologs. Height-increasing or -decreasing effects of gene-tissue pairs were determined by combining publicly available TWAS predictive models (*54*) based on eQTL information from GTEx with height GWAS summary statistics from the UK Biobank (*56*), GIANT consortium (*57*), and a meta-analysis of the two studies (*48*).

REFERENCES AND NOTES

1. J. C. K. Wells, Sexual dimorphism of body composition. *Best Pract. Res. Clin. Endocrinol. Metab.* **21**, 415–430 (2007). doi: 10.1016/j.beem.2007.04.007; pmid: 17875489
2. H. J. Green, I. G. Fraser, D. A. Ranney, Male and female differences in enzyme activities of energy metabolism in vastus lateralis muscle. *J. Neurol. Sci.* **65**, 323–331 (1984). doi: 10.1016/0022-510X(84)90095-9; pmid: 6238135
3. A. N. V. Ruigrok *et al.*, A meta-analysis of sex differences in human brain structure. *Neurosci. Biobehav. Rev.* **39**, 34–50 (2014). doi: 10.1016/j.neubiorev.2013.12.004; pmid: 24374381

4. S. L. Klein, K. L. Flanagan, Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016). doi: 10.1038/nri.2016.90; pmid: 27546235
5. C. S. Hayward, W. V. Kalnins, R. P. Kelly, Gender-related differences in left ventricular chamber function. *Cardiovasc. Res.* **49**, 340–350 (2001). doi: 10.1016/S0008-6363(00)00280-7; pmid: 11164844
6. S. T. Ngo, F. J. Steyn, P. A. McCombe, Gender differences in autoimmune disease. *Front. Neuroendocrinol.* **35**, 347–369 (2014). doi: 10.1016/j.yfrne.2014.04.004; pmid: 24793874
7. V. Regitz-Zagrosek *et al*., Gender in cardiovascular diseases: Impact on clinical manifestations, management, and outcomes. *Eur. Heart J.* **37**, 24–34 (2016). doi: 10.1093/eurheartj/ehv598; pmid: 26530104
8. D. M. Werling, D. H. Geschwind, Sex differences in autism spectrum disorders. *Curr. Opin. Neurol.* **26**, 146–153 (2013). doi: 10.1097/WCO.0b013e32835ee548; pmid: 23406909
9. H. Olson *et al*., Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul. Toxicol. Pharmacol.* **32**, 56–67 (2000). doi: 10.1006/rtph.2000.1399; pmid: 11029269
10. P. Lindenfors, J. L. Gittleman, K. Jones, in *Evolutionary Studies of Sexual Size Dimorphism* (Oxford Univ. Press, 2007), pp. 16–26.
11. R. A. Gorski, J. H. Gordon, J. E. Shryne, A. M. Southam, Evidence for a morphological sex difference within the medial preoptic area of the rat brain. *Brain Res.* **148**, 333–346 (1978). doi: 10.1016/0006-8993(78)90723-0; pmid: 656937
12. R. S. Scotland, M. J. Stables, S. Madalli, P. Watson, D. W. Gilroy, Sex differences in resident immune cell phenotype underlie more efficient acute inflammatory responses in female mice. *Blood* **118**, 5918–5927 (2011). doi: 10.1182/blood-2011-03-340281; pmid: 21911834
13. K. M. Shioura, D. L. Geenen, P. H. Goldspink, Sex-related changes in cardiac function following myocardial infarction in mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **295**, R528–R534 (2008). doi: 10.1152/ajpregu.90342.2008; pmid: 18550865
14. H. Skaletsky *et al*., The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003). doi: 10.1038/nature01722; pmid: 12815422
15. D. W. Bellott *et al*., Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014). doi: 10.1038/nature13206; pmid: 24759411
16. T. Tukiainen *et al*., Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017). doi: 10.1038/nature24265; pmid: 29022598
17. M. Melé *et al*., The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015). doi: 10.1126/science.aaa0355; pmid: 25954002
18. M. Gershoni, S. Pietrokovski, The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* **15**, 7 (2017). doi: 10.1186/s12915-017-0352-z; pmid: 28173793
19. X. Yang *et al*., Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16**, 995–1004 (2006). doi: 10.1101/gr.5217506; pmid: 16825664
20. J. M. Ranz, C. I. Castillo-Davis, C. D. Meiklejohn, D. L. Hartl, Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300**, 1742–1745 (2003). doi: 10.1126/science.1085881; pmid: 12805547
21. Y. Zhang, D. Sturgill, M. Parisi, S. Kumar, B. Oliver, Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* **450**, 233–237 (2007). doi: 10.1038/nature06323; pmid: 17994089
22. S. Grath, J. Parsch, Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of *Drosophila* evolution. *Genome Biol. Evol.* **4**, 346–359 (2012). doi: 10.1093/gbe/evs012; pmid: 22321769
23. R. Assis, Q. Zhou, D. Bachtrog, Sex-biased transcriptome evolution in *Drosophila*. *Genome Biol. Evol.* **4**, 1189–1200 (2012). doi: 10.1093/gbe/evs093; pmid: 23097318
24. J. C. Perry, P. W. Harrison, J. E. Mank, The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. *Mol. Biol. Evol.* **31**, 1206–1219 (2014). doi: 10.1093/molbev/msu072; pmid: 24526011
25. B. Reinius *et al*., An evolutionarily conserved sexual signature in the primate brain. *PLOS Genet.* **4**, e1000100 (2008). doi: 10.1371/journal.pgen.1000100; pmid: 18566661
26. R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, Y. Gilad, Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010). doi: 10.1101/gr.099226.109; pmid: 20009012
27. Y. Liang *et al*., A gene network regulated by the transcription factor VGLL3 as a promoter of sex-biased autoimmune diseases. *Nat. Immunol.* **18**, 152–160 (2017). doi: 10.1038/ni.3643; pmid: 27992404
28. A. E. Russi, M. E. Ebel, Y. Yang, M. A. Brown, Male-specific IL-33 expression regulates sex-dimorphic EAE susceptibility. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1520–E1529 (2018). doi: 10.1073/pnas.1710401115; pmid: 29378942
29. M. Souyris *et al*., *TLR7* escapes X chromosome inactivation in immune cells. *Sci. Immunol.* **3**, eaap8855 (2018). doi: 10.1126/sciimmunol.aap8855; pmid: 29374079
30. E. A. Boyle, Y. I. Li, J. K. Pritchard, An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017). doi: 10.1016/j.cell.2017.05.038; pmid: 28622505
31. T. M. Williams *et al*., The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *Drosophila*. *Cell* **134**, 610–623 (2008). doi: 10.1016/j.cell.2008.06.052; pmid: 18724934
32. C. M. Disteche, Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* **46**, 537–560 (2012). doi: 10.1146/annurev-genet-110711-155454; pmid: 22974302
33. C. M. Disteche, Dosage compensation of the sex chromosomes and autosomes. *Semin. Cell Dev. Biol.* **56**, 9–18 (2016). doi: 10.1016/j.semcdb.2016.04.013; pmid: 27112542
34. Materials and methods are available as supplementary materials.
35. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). doi: 10.1038/nature24277; pmid: 29022597
36. A. M. Newman *et al*., Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015). doi: 10.1038/nmeth.3337; pmid: 25822800
37. J. Merkin, C. Russell, P. Chen, C. B. Burge, Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012). doi: 10.1126/science.1228186; pmid: 23258891
38. D. Brawand *et al*., The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011). doi: 10.1038/nature10532; pmid: 22012392
39. D. M. Werling, N. N. Parikshak, D. H. Geschwind, Gene expression in human brain implicates sexually dimorphic pathways in autism spectrum disorders. *Nat. Commun.* **7**, 10717 (2016). doi: 10.1038/ncomms10717; pmid: 26892004
40. M. E. Lindholm *et al*., The human skeletal muscle transcriptome: Sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *FASEB J.* **28**, 4571–4581 (2014). doi: 10.1096/fj.14-255000; pmid: 25016029
41. M. S. Newman, T. Nguyen, M. J. Watson, R. W. Hull, H.-G. Yu, Transcriptome profiling reveals novel BMI- and sex-specific gene expression signatures for human cardiac hypertrophy. *Physiol. Genomics* **49**, 355–367 (2017). doi: 10.1152/physiolgenomics.00122.2016; pmid: 28500252
42. N. Viguerie *et al*., Determinants of human adipose tissue gene expression: Impact of diet, sex, metabolic status, and cis genetic regulation. *PLOS Genet.* **8**, e1002959 (2012). doi: 10.1371/journal.pgen.1002959; pmid: 23028366
43. R. Marin *et al*., Convergent origination of a *Drosophila*-like dosage compensation mechanism in a reptile lineage. *Genome Res.* **27**, 1974–1987 (2017). doi: 10.1101/gr.223727.117; pmid: 29133310
44. B. Li *et al*., A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq. *Sci. Rep.* **7**, 4200 (2017). doi: 10.1038/s41598-017-04520-z; pmid: 28646208
45. M. D. Franco *et al*., Transcriptome of normal lung distinguishes mouse lines with different susceptibility to inflammation and to lung tumorigenesis. *Cancer Lett.* **294**, 187–194 (2010). doi: 10.1016/j.canlet.2010.01.038; pmid: 20189714
46. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019). doi: 10.1038/s41588-018-0268-8; pmid: 30478440
47. J. Chen *et al*., A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.* **29**, 53–63 (2019). doi: 10.1101/gr.237636.118; pmid: 30552105
48. L. Yengo *et al*., Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018). doi: 10.1093/hmg/ddy271; pmid: 30124842
49. T. Tukiainen *et al*., Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLOS Genet.* **10**, e1004127 (2014). doi: 10.1371/journal.pgen.1004127; pmid: 24516404
50. J. Sidorenko, I. Kassam, K. Kemper, J. Zeng, L. Lloyd-Jones, G. W. Montgomery, G. Gibson, A. Metspalu, T. Esko, J. Yang, A. F. McRae, P. M. Visscher, The effect of X-linked dosage compensation on complex trait variation. bioRxiv 433870 [Preprint]. 3 October 2018. https://doi.org/10.1101/433870
51. K. Rawlik, O. Canela-Xandri, A. Tenesa, Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biol.* **17**, 166 (2016). doi: 10.1186/s13059-016-1025-x; pmid: 27473438
52. G. Stulp, L. Barrett, Evolutionary perspectives on human height variation. *Biol. Rev. Camb. Philos. Soc.* **91**, 206–234 (2016). doi: 10.1111/brv.12165; pmid: 25530478
53. J. S. Sanjak, J. Sidorenko, M. R. Robinson, K. R. Thornton, P. M. Visscher, Evidence of directional and stabilizing selection in contemporary humans. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 151–156 (2018). doi: 10.1073/pnas.1707227114; pmid: 29255044
54. A. Gusev *et al*., Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016). doi: 10.1038/ng.3506; pmid: 26854917
55. J. MacArthur *et al*., The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45** (D1), D896–D901 (2017). doi: 10.1093/nar/gkw1133; pmid: 27899670
56. P. R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, A. L. Price, Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018). doi: 10.1038/s41588-018-0144-6; pmid: 29892013
57. A. R. Wood *et al*., Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014). doi: 10.1038/ng.3097; pmid: 25282103
58. J. J. Hayward *et al*., Complex disease and phenotype mapping in the domestic dog. *Nat. Commun.* **7**, 10460 (2016). doi: 10.1038/ncomms10460; pmid: 26795439
59. A. C. Bouwman *et al*., Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* **50**, 362–367 (2018). doi: 10.1038/s41588-018-0056-5; pmid: 29459679
60. H. Signer-Hasler *et al*., A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLOS ONE* **7**, e37282 (2012). doi: 10.1371/journal.pone.0037282; pmid: 22615965
61. M. Seo *et al*., Comprehensive identification of sexually dimorphic genes in diverse cattle tissues using RNA-seq. *BMC Genomics* **17**, 81 (2016). doi: 10.1186/s12864-016-2400-4; pmid: 26818975
62. J. Metzger, R. Schrimpf, U. Philipp, O. Distl, Expression levels of LCORL are associated with body size in horses. *PLOS ONE* **8**, e56497 (2013). doi: 10.1371/journal.pone.0056497; pmid: 23418579
63. N. Stroth, Y. Holighaus, D. Ait-Ali, L. E. Eiden, PACAP: A master regulator of neuroendocrine stress circuits and the cellular stress response. *Ann. N. Y. Acad. Sci.* **1220**, 49–59 (2011). doi: 10.1111/j.1749-6632.2011.05904.x; pmid: 21388403
64. T. Kanzleiter *et al*., Pknox1/Prep1 regulates mitochondrial oxidative phosphorylation components in skeletal muscle. *Mol. Cell. Biol.* **34**, 290–298 (2014). doi: 10.1128/MCB.01232-13; pmid: 24216763
65. K. S. Small *et al*., Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat. Genet.* **50**, 572–580 (2018). doi: 10.1038/s41588-018-0088-x; pmid: 29632379
66. J. Yang *et al*., Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010). doi: 10.1038/ng.608; pmid: 20562875
67. E. Rao *et al*., Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nat. Genet.* **16**, 54–63 (1997). doi: 10.1038/ng0597-54; pmid: 9140395
68. T. Ogata, N. Matsuo, Sex chromosome aberrations and stature: Deduction of the principal factors involved in the determination of adult height. *Hum. Genet.* **91**, 551–562 (1993). doi: 10.1007/BF00205079; pmid: 8340109
69. A. K. San Roman, D. C. Page, A strategic research alliance: Turner syndrome and sex differences. *Am. J. Med. Genet. C. Semin. Med. Genet.* **181**, 59–67 (2019). doi: 10.1002/ajmg.c.31677; pmid: 30790449

70. J. Parsch, H. Ellegren, The evolutionary causes and consequences of sex-biased gene expression. *Nat. Rev. Genet.* **14**, 83–87 (2013). doi: 10.1038/nrg3376; pmid: 23329110
71. F. Ruzicka *et al*., Genome-wide sexually antagonistic variants reveal long-standing constraints on sexual dimorphism in fruit flies. *PLOS Biol.* **17**, e3000244 (2019). doi: 10.1371/journal.pbio.3000244; pmid: 31022179
72. J. C. Randall *et al*., Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLOS Genet.* **9**, e1003500 (2013). doi: 10.1371/journal.pgen.1003500; pmid: 23754948
73. T. W. Winkler *et al*., The influence of age and sex on genetic associations with adult body size and shape: A large-scale genome-wide interaction study. *PLOS Genet.* **11**, e1005378 (2015). doi: 10.1371/journal.pgen.1005378; pmid: 26426971
74. D. Shungin *et al*., New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015). doi: 10.1038/nature14132; pmid: 25673412
75. E. Porcu *et al*., A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function. *PLOS Genet.* **9**, e1003266 (2013). doi: 10.1371/journal.pgen.1003266; pmid: 23408906
76. E. A. Khramtsova *et al*., Sex differences in the genetic architecture of obsessive-compulsive disorder. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* 10.1002/ajmg.b.32687 (2018). doi: 10.1002/ajmg.b.32687; pmid: 30456828
77. E. Y. Kang *et al*., An Association Mapping Framework To Account for Potential Sex Difference in Genetic Architectures. *Genetics* **209**, 685–698 (2018). pmid: 29752291
78. M. E. Ritchie *et al*., *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015). doi: 10.1093/nar/gkv007; pmid: 25605792
79. T. L. Bailey, DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011). doi: 10.1093/bioinformatics/btr261; pmid: 21543442
80. R. C. McLeay, T. L. Bailey, Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010). doi: 10.1186/1471-2105-11-165; pmid: 20356413
81. S. Naqvi, A. Godfrey, J. Hughes, M. Goodheart, R. Mitchell, D. Page, Conservation, acquisition, and functional impact of sex-biased gene expression in mammalian tissues [Data set], Version 1, Zenodo (2019); http://doi.org/10.5281/zenodo.2658829.

ACKNOWLEDGMENTS

We thank D. W. Bellott, L. Chmatal, and R. Ransohoff for critically reading the manuscript. **Funding:** S.N. and A.K.G. were supported by a research grant from Biogen, Inc. This work was supported by Biogen, Whitehead Institute, National Institutes of Health (grants R01HG007852 and U01HG007857), Howard Hughes Medical Institute, generous gifts from Brit and Alexander d'Arbeloff and Arthur W. and Carol Tobin Brill. **Author contributions:** S.N., A.K.G., J.F.H., and D.C.P. designed the study. J.F.H. procured cyno tissue samples. M.L.G. procured mouse and rat tissue samples, with assistance from S.N. S.N. processed tissue samples and performed computational analyses, with assistance from A.K.G. R.N.M. performed histological evaluations on human tissue sections. D.C.P. supervised work. S.N. and D.C.P. wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Raw RNA-seq data (.fastq files) have been deposited in the Gene Expression Omnibus (GSE125483). Processed data (expression count and TPM matrices and sample metadata) and code required to reproduce the analyses are available at http://pagelab.wi.mit.edu/page/papers/Naqvi_et_al_2019 and at Zenodo (*81*).

SUPPLEMENTARY MATERIALS

19 January 2019; accepted 12 June 2019
10.1126/science.aaw7317

Science

Conservation, acquisition, and functional impact of sex-biased gene expression in mammals

Sahin Naqvi, Alexander K. Godfrey, Jennifer F. Hughes, Mary L. Goodheart, Richard N. Mitchell and David C. Page

Science **365** (6450), eaaw7317.
DOI: 10.1126/science.aaw7317

The genetics of sexual dimorphism

In mammals, many species exhibit sex-specific phenotypes that differ between males and females. Although attention has been directed to the effects of the X and Y sex chromosomes, we do not understand how sex affects the rest of the genome. Naqvi *et al.* examined gene expression in 12 tissues in male and female humans, mice, rats, dogs, and cynomolgus macaques and identified diversity in gene expression between the sexes. Examining sex-biased gene expression in human height identified opposing male or female bias. Although conservation of differential sex-specific gene expression among species was observed, specific genes differed in the sexes among species and lineages suggesting the evolution of species- or lineage-specific sex-biased expression.

Science, this issue p. eaaw7317