Olney et al. Biology of Sex Differences (2020) 11:42
https://doi.org/10.1186/s13293-020-00312-9

Biology of Sex Differences

RESEARCH Open Access

Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data

Kimberly C. Olney[1,2], Sarah M. Brotman[1,3], Jocelyn P. Andrews[1,4], Valeria A. Valverde-Vesling[1] and Melissa A. Wilson[1,2,5*]

Abstract

Background: Human X and Y chromosomes share an evolutionary origin and, as a consequence, sequence similarity. We investigated whether the sequence homology between the X and Y chromosomes affects the alignment of RNA-Seq reads and estimates of differential expression. We tested the effects of using reference genomes and reference transcriptomes informed by the sex chromosome complement of the sample's genome on the measurements of RNA-Seq abundance and sex differences in expression.

Results: The default genome includes the entire human reference genome (GRCh38), including the entire sequence of the X and Y chromosomes. We created two sex chromosome complement informed reference genomes. One sex chromosome complement informed reference genome was used for samples that lacked a Y chromosome; for this reference genome version, we hard-masked the entire Y chromosome. For the other sex chromosome complement informed reference genome, to be used for samples with a Y chromosome, we hard-masked only the pseudoautosomal regions of the Y chromosome, because these regions are duplicated identically in the reference genome on the X chromosome. We analyzed the transcript abundance in the whole blood, brain cortex, breast, liver, and thyroid tissues from 20 genetic female (46, XX) and 20 genetic male (46, XY) samples. Each sample was aligned twice: once to the default reference genome and then independently aligned to a reference genome informed by the sex chromosome complement of the sample, repeated using two different read aligners, HISAT and STAR. We then quantified sex differences in gene expression using featureCounts to get the raw count estimates followed by Limma/Voom for normalization and differential expression. We additionally created sex chromosome complement informed transcriptome references for use in pseudo-alignment using Salmon. Transcript

(Continued on next page)

* Correspondence: mwilsons@asu.edu
[1]School of Life Sciences, Arizona State University, PO Box 874501, Tempe, AZ 85287-4501, USA
[2]Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85282, USA
Full list of author information is available at the end of the article

(Continued from previous page)
abundance was quantified twice for each sample: once to the default target transcripts and then independently to target transcripts informed by the sex chromosome complement of the sample.

Conclusions: We show that regardless of the choice of the read aligner, using an alignment protocol informed by the sex chromosome complement of the sample results in higher expression estimates on the pseudoautosomal regions of the X chromosome in both genetic male and genetic female samples, as well as an increased number of unique genes being called as differentially expressed between the sexes. We additionally show that using a pseudo-alignment approach informed on the sex chromosome complement of the sample eliminates Y-linked expression in female XX samples.

Keywords: RNA-Seq, Sex chromosomes, Differential expression, Transcriptome, Mapping, Alignment, Pseudo-alignment, Quantification

Author summary

The human X and Y chromosomes share an evolutionary origin and sequence homology, including regions of 100% identity; this sequence homology can result in reads misaligning between the sex chromosomes, X and Y. We hypothesized that misalignment of reads on the sex chromosomes would confound estimates of transcript abundance if the sex chromosome complement of the sample is not accounted for during the alignment step. For example, because of shared sequence similarity, X-linked reads could misalign to the Y chromosome. This is expected to result in reduced expression for regions between X and Y that share high levels of homology. For this reason, we tested the effect of using a default reference genome versus a reference genome informed by the sex chromosome complement of the sample on estimates of transcript abundance in human RNA-Seq samples from the whole blood, brain cortex, breast, liver, and thyroid tissues of 20 genetic female (46, XX) and 20 genetic male (46, XY) samples. We found that using a reference genome with the sex chromosome complement of the sample resulted in higher measurements of X-linked gene transcription for both male and female samples and more differentially expressed genes on the X and Y chromosomes. We additionally investigated the use of a sex chromosome complement informed transcriptome reference index for alignment-free quantification protocols. We observed no Y-linked expression in female XX samples only when the transcript quantification was performed using a transcriptome reference index informed on the sex chromosome complement of the sample. We recommend that future studies requiring aligning RNA-Seq reads to a reference genome or pseudo-alignment with a transcriptome reference should consider the sex chromosome complement of their samples prior to running default pipelines.

Background

Sex differences in aspects of human biology, such as development, physiology, metabolism, and disease susceptibility, are partially driven by sex-specific gene regulation [1–4]. There are reported sex differences in gene expression across human tissues [5–7] and while some may be attributed to hormones and environment, there are documented genome-wide sex differences in expression based solely on the sex chromosome complement [8]. However, accounting for the sex chromosome complement of the sample in quantifying gene expression has been limited due to shared sequence homology between the sex chromosomes, X and Y, that can confound gene expression estimates.

The X and Y chromosomes share an evolutionary origin: mammalian X and Y chromosomes originated from a pair of indistinguishable autosomes ~ 180–210 million years ago that acquired the sex-determining genes [9–11]. The human X and Y chromosomes formed in two different segments: (a) the one that is shared across all mammals called the X-conserved region (XCR) and (b) the X-added region (XAR) that is shared across all eutherian animals [11]. The sex chromosomes, X and Y, previously recombined along their entire lengths, but due to recombination suppression from Y chromosome-specific inversions [10, 12], now only recombine at the tips in the pseudoautosomal regions (PAR) PAR1 and PAR2 [9–11]. PAR1 is ~ 2.78 million bases (Mb), and PAR2 is ~ 0.33 Mb; these sequences are 100% identical between X and Y [11, 13, 14] (Fig. 1a). The PAR1 is a remnant of the XAR Ross et al. [11] and shared among eutherians, while the PAR2 is recently added and human-specific [14]. Other regions of high sequence similarity between X and Y include the X-transposed region (XTR) with 98.78% homology [15] (Fig. 1a). The XTR formed from an X chromosome to Y chromosome duplication event following the human-chimpanzee divergence [11, 16]. Thus, the evolution of the X and Y chromosomes has resulted in a pair of chromosomes that are diverged, but still share some regions of high sequence similarity.

To infer which genes or transcripts are expressed, RNA-Seq reads can be aligned to a reference genome.

The abundance of reads mapped to a transcript is reflective of the amount of expression of that transcript. RNA-Seq methods rely on aligning reads to an available high-quality reference genome sequence, but this remains a challenge due to the intrinsic complexity in the transcriptome of regions with a high level of homology [17]. By default, the GRCh38 version of the human reference genome includes both the X and Y chromosomes, which is used to align RNA-Seq reads from both male XY and female XX samples. It is known that sequence reads from DNA will misalign along the sex chromosomes affecting downstream analyses [18]. However, this has not been tested using RNA-Seq data and the effects on differential expression analysis are not known. Considering the increasing number of human RNA-Seq consortium datasets (e.g., the Genotype-Tissue Expression project (GTEx) [19], The Cancer Genome Atlas (TCGA) [20], Geuvadis project [21], and Simons Genome Diversity Project [22]), there is an urgent need to understand how aligning to a default reference genome that includes both X and Y may affect estimates of gene expression on the sex chromosomes [2, 23]. We hypothesize that regions of high sequence similarity will result in misaligning of RNA-Seq reads and reduced expression estimates (Fig. 1a, b).

Here, we tested the effect of sex chromosome complement informed read alignment on the quantified levels of gene expression and the ability to detect sex-biased gene expression. We utilized data from the GTEx project, focusing on five tissues, whole blood, brain cortex, breast, liver, and thyroid, which are known to exhibit sex differences in gene expression [5, 24–27]. Many genes have been reported to be differentially expressed between male and female brain samples [5–7], and differential expression in blood samples between males and females has also been documented [5, 6]. An analysis of all GTEx tissue samples reported that breast mammary gland tissues are the most sex differentially expressed tissue [5]. It has also been reported that there are sex disparities in thyroid cancer [28] and liver cancer [29, 30] suggesting possible sex differences in gene expression. We used whole blood, brain cortex, breast, liver, and thyroid tissues from 20 genetic male (46, XY) and 20 genetic female (46, XX) individuals for a total of 200 samples evenly distributed among tissues. Male and female samples, for each tissue, were age-matched between the sexes and only included samples of age 55 to 70. We aligned all samples to a default reference genome that includes both the X and Y chromosomes and to a reference genome that is informed on the sex chromosome complement of the genome: Male XY samples were aligned to a reference genome that includes both the X and Y chromosome, where the Y chromosome PAR1 and PAR2 are hard-masked with Ns (Fig. 1c) so that reads will align uniquely to the X PAR sequences. Conversely, female XX samples were aligned to a reference genome where the entirety of the Y chromosome is hard-masked (Fig. 1c). We tested two different read aligners, HISAT [31] and STAR [32], to account for variation between alignment methods and measured differential expression using Limma/Voom [33]. We found that using a sex chromosome complement informed reference genome for aligning RNA-Seq reads increased expression estimates on the pseudoautosomal regions of the X chromosome in both male XY and female XX samples and uniquely identified differentially expressed genes.

We additionally investigated the effect of transcriptome references on pseudo-alignment methods. We quantified abundance using Salmon [34] in male and female brain cortex samples twice, once to a default reference transcriptome index that includes both the X and Y chromosome-linked transcripts and to a reference transcriptome index that is informed on the sex chromosome complement of the sample. We found that using a sex chromosome complement informed reference transcriptome index for RNA-Seq pseudo-alignment quantification eliminated Y-linked expression estimates in female XX samples that were observed in the default approach.

Regardless of the alignment or pseudo-alignment approach, we recommended carefully considering the annotations of the sex chromosomes in the references used, as these will affect quantifications and differential expression estimates, especially of sex chromosome-linked genes.

Methods

Building sex chromosome complement informed reference genomes

All GRCh38.p10 unmasked genomic DNA sequences, including autosomes 1-22, X, Y, mitochondrial DNA (mtDNA), and contigs were downloaded from ensembl.org release 92 [13]. The default reference genome here includes all 22 autosomes, mtDNA, the X chromosome, the Y chromosome, and contigs. For the two sex chromosome complement informed reference assemblies, we included all 22 autosomes, mtDNA, and contigs from the default reference and (a) one with the Y chromosome either hard-masked for the "Y-masked reference genome" or (b) one with the pseudoautosomal regions, PAR1 and PAR2, hard-masked on the Y chromosome for "YPAR-masked reference genome" (Fig. 1c). Hard-masking with Ns will force reads to not align to those masked regions in the genome. Masking the entire Y chromosome for the sex chromosome complement informed reference genome, Y-masked, was accomplished by changing all the Y chromosome nucleotides [ATGC] to N using a sed command in linux. YPAR-masked was created by hard-masking the Y PAR1: 6001-2699520 and the Y PAR2:

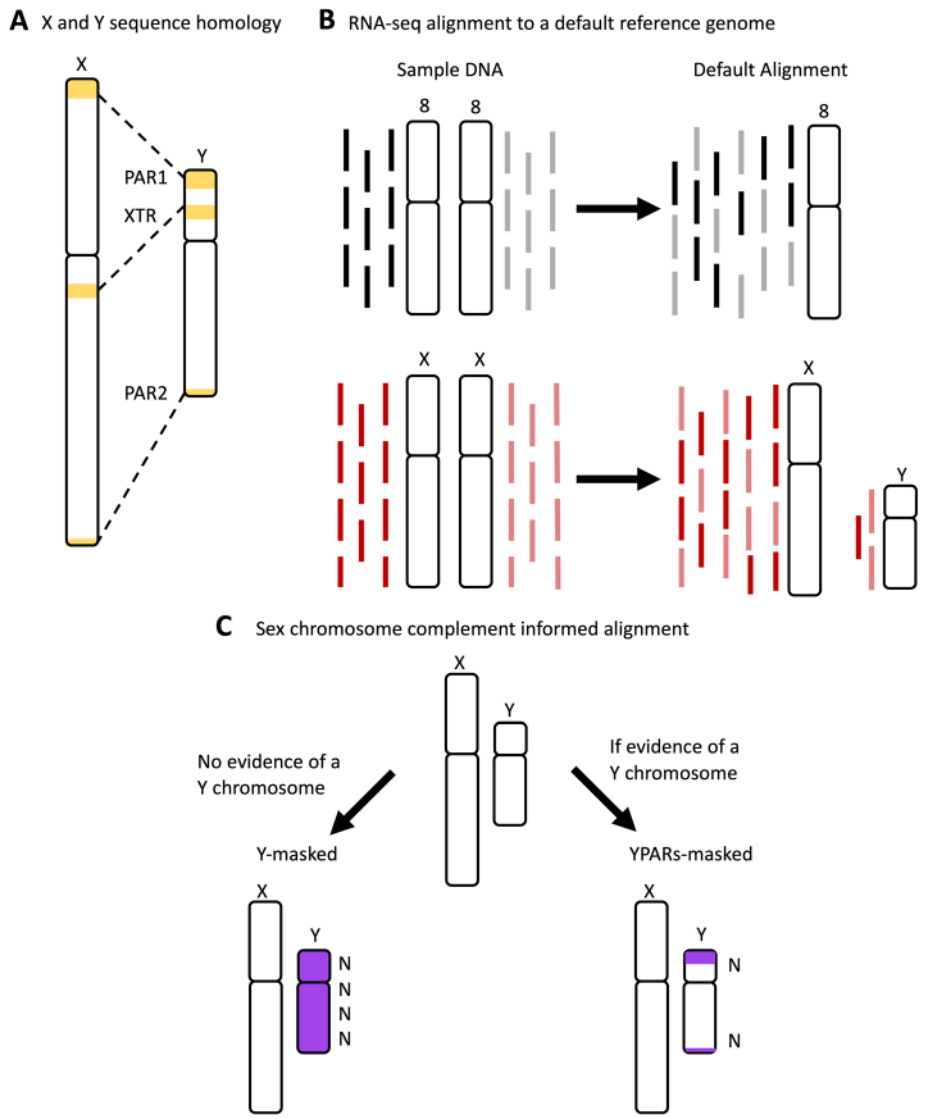


Fig. 1 Homology between the human X and Y chromosomes where misaligning could occur. **a** High sequence homology exists between the human X and Y chromosomes in three regions: 100% sequence identity for the pseudoautosomal regions (PARs), PAR1, and PAR2, and ~ 99% sequence homology in the X-transposed region (XTR). The X chromosome PAR1 is ~ 2.78 million bases (Mb) extending from X:10,001 to 2,781,479 and the X chromosome PAR2 is ~ 0.33 Mb extending from X:155,701,383 to 156,030,895. The X chromosome PAR1 and PAR2 are identical in sequence to the Y chromosome PAR1 Y:10,001 to 2,781,479 and PAR2 Y:56,887,903 to 57,217,415. **b** Using a standard alignment approach will result in reads misaligning between regions of high sequence homology on the sex chromosomes. **c** Using a reference genome that is informed by the genetic sex of the sample may help to reduce misaligning between the X and Y chromosomes. In humans, samples without evidence of a Y chromosome should be aligned to a Y-masked reference genome, and samples with evidence of a Y should be aligned to a YPAR-masked reference genome

154931044-155260560 regions. The GRCh38.p10 Y PAR1 and Y PAR2 chromosome start and end locations were defined using Ensembl GRCh38 Y PAR definitions [13]. After creating the Y chromosome PAR1- and PAR2-masked fasta files, we concatenated all the Y chromosome regions together to create a YPAR-masked reference genome. After creating the GRCh38.p10 default reference genome and the two sex chromosome complement informed reference genomes, we indexed the reference genomes and created a dictionary for each using HISAT version 2.1.0 [31] hisat2-build -f option and STAR version 2.5.2 [32], using option --genomeDir and --sjdbGTFfile. Reference genome indexing was followed by Picard tools version 1.119 CreateSequenceDictionary [35], which created a dictionary for each reference genome (Pipeline available on GitHub, https://github.com/SexChrLab/XY_RNAseq).

Building sex chromosome complement informed transcriptome index

Ensembl's GRCh38.p10 cDNA reference transcriptome fasta consists of transcript sequences resulting from Ensemble gene predictions. Ensembl's cDNA was downloaded from ensembl.org release 92 [13]. The default transcriptome reference includes 199,234 transcripts which include autosomal, mtDNA, X chromosome, Y chromosome, and contig transcripts. The default Ensembl cDNA does not contain Y chromosome PAR-linked transcript sequences; it only contains the X chromosome PAR sequence transcripts. For the sex chromosome complement informed reference transcriptome index, we included all 22 autosomes, mtDNA, X, and contigs from the default cDNA transcriptome and we hard-masked all available Y chromosome-linked transcript sequences. Hard-masking the Y chromosome-linked transcripts was accomplished by changing all the Y chromosome nucleotides [ATGC] to N using a sed command in linux. After downloading the GRCh38.p10 default reference transcriptome and creating the Y-masked sex chromosome complement informed reference transcriptome fasta files, we then generated a decoy-aware transcriptome for each transcriptome reference. For generating the default decoy-aware reference transcriptome, we used the default genome as the decoy sequence. This was accomplished by concatenating the default genome fasta to the end of the default transcriptome fasta to populate the decoy file with the chromosome names, as suggested by Salmon [34]. The default transcriptome fasta and the default decoy file were then used to create the mapping-based index using the Salmon version 1.2.0 index function [34]. The Y-masked decoy-aware transcriptome fasta was generated by concatenating the Y-masked genome fasta to the end of the Y-masked transcriptome fasta to populate the decoy file with the chromosome names. The Y-masked transcriptome fasta and the decoy file were then used as inputs for generating the Y-masked mapping-based index using the salmon index function. For both the default and the Y-masked mapping-based index, a *k*-mer of 31 was used as this was suggested to work well for reads of 75 bp.

In addition to the Ensembl reference, we investigated the effects of a sex chromosome complement reference transcriptome index using the gencode transcript reference fasta GRCh38.p12 that contains 206,694 transcripts which include autosomal, mtDNA, X, Y, and contigs. The gencode transcriptome reference includes both the X and Y PAR transcripts [36]. Following the same parameters for the Ensembl decoy-aware transcriptome, we created two gencode sex chromosome complement decoy-aware transcriptome references, in addition to a default gencode decoy-aware transcriptome reference. The pipeline is available on GitHub, https://github.com/SexChrLab/XY_RNAseq.

RNA-Seq samples

From the Genotyping-Tissue Expression (GTEx) Project data, we downloaded SRA files for whole blood, brain cortex, breast, liver, and thyroid tissues from 20 separate genetic female (46, XX) and 20 separate genetic male (46, XY) individuals [19, 37] that were age-matched between the sexes and ranged from age 55 to 70 (Additional file 1 & 2). Age matching exactly was accomplished using the matchit function in the R package MatchIt [38]. The GTEx data is described and available through dbGaP under accession phs000424.v6.p1; we received approval to access this data under dbGaP accession #8834. GTEx RNA-Seq samples were sequenced to 76-bp reads, and the median coverage was ~ 82 million (M) reads [37]. Although information about the genetic sex of the samples was provided in the GTEx summary downloads, it was additionally investigated by examining the gene expression of select genes that are known to be differentially expressed between the sexes or are known X-Y homologous genes: *DDX3X*, *DDX3Y*, *PCDH11X*, *PCDH11Y*, *USP9X*, *USP9Y*, *ZFX*, *ZFY*, *UTX*, *UTY*, *XIST*, and *SRY* (Fig. 2; Additional file 3 & 4).

RNA-Seq trimming and quality filtering

RNA-Seq sample data was converted from sequence read archive (sra) format to the paired-end FASTQ format using the SRA toolkit [39]. Quality of the samples' raw sequencing reads was examined using FastQC [40] and MultiQC. Subsequently, adapter sequences were removed using Trimmomatic version 0.36 [41]. More specifically, reads were trimmed to remove bases with a quality score less than 10 for the leading strand and less than 25 for the trailing strand, applying a sliding window of 4 with a mean PHRED quality of 30 required in the window and a minimum read length of 40 bases.

RNA-Seq read alignment

Following trimming, paired RNA-Seq reads from all samples were aligned to the default reference genome. Unpaired RNA-Seq reads were not used for alignment. Reads from the female (46, XX) samples were aligned to the Y-masked reference genome, and reads from male (46, XY) individuals were aligned to the YPAR-masked reference genome. Read alignment was performed using HISAT version 2.1.0 [31], keeping all parameters the same, only changing the reference genome used, as described above. Read alignment was additionally performed using STAR version 2.5.2 [32], where all samples were aligned to a default reference genome and to a reference genome informed on the sex chromosome complement, keeping all parameters the same (Pipeline available on GitHub, https://github.com/SexChrLab/XY_RNAseq).

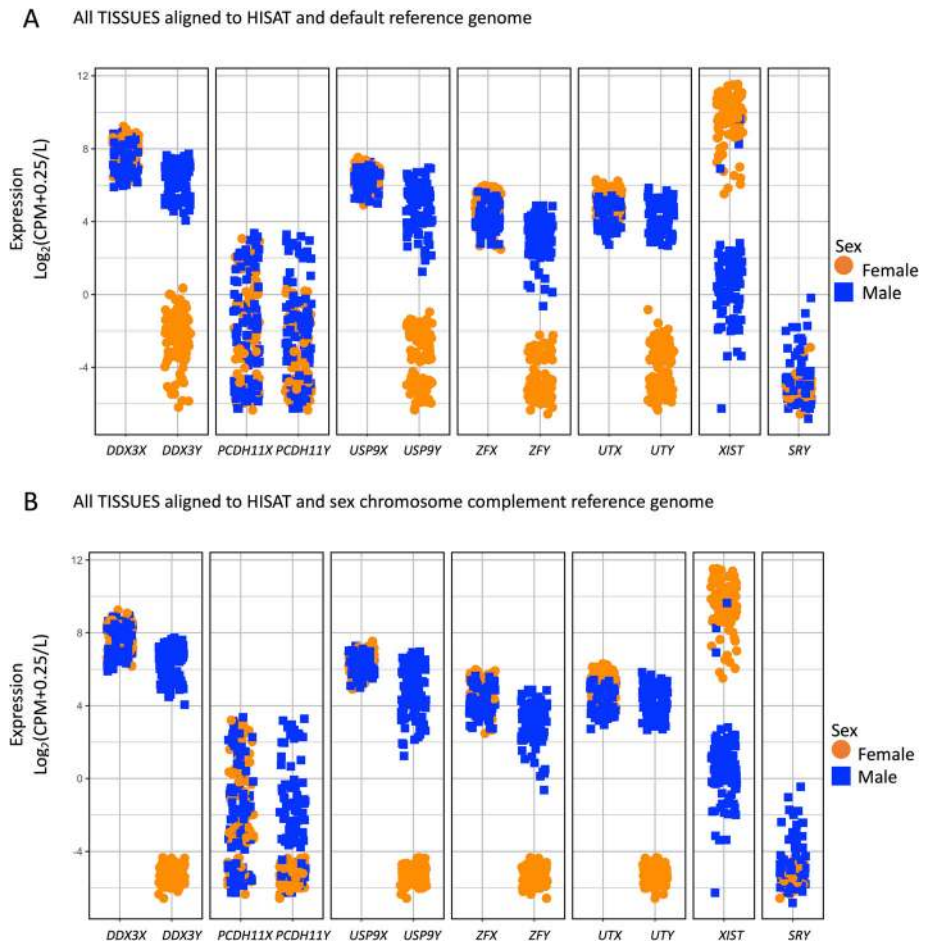


Fig. 2 Genetic sex of RNA-Seq samples. We investigated the gene expression, $\log_2$(CPM + 0.25/*L*), of XY homologous genes (*DDX3X/Y*, *PCDH11X/Y*, *USP9X/Y*, *ZFX/Y*, *UTX/Y*); *XIST*; and *SRY* in all samples from all tissues analyzed here from genetic males (blue squares) and genetic females (orange circles) **a** when aligned to a default reference genome and **b** when aligned to a sex chromosome complement informed reference genome, using HISAT as the read aligner

Processing of RNA-Seq alignment files

Aligned RNA-Seq samples from HISAT and STAR were output in Sequence Alignment Map (SAM) format and converted to Binary Alignment Map (BAM) format using bamtools version 2.4.0 [42]. Summaries on the BAM files including the number of reads mapped were computed using bamtools version 2.4.0 package [43]. RNA-Seq BAM files were indexed and sorted, duplicates were marked, and read groups added using bamtools, samtools, and Picard [35, 42, 43]. All RNA-Seq BAM files were indexed using the default reference genome using Picard ReorderSam [35]; this was done so that all samples would include all chromosomes in the index files. Aligning XX samples to a Y-masked reference genome using HISAT indexes would result in no Y chromosome information in the aligned BAM and BAM index bai files. For downstream analysis, some tools require that all samples have the same chromosomes, which is why we hard-masked rather than removed. Reindexing the BAM files to the default reference genome does not alter the read alignment and thus does not alter our comparison between default and sex chromosome complement informed alignments.

Gene expression level quantification

Read counts for each gene across all autosomes, sex chromosomes, mtDNA, and contigs were generated using featureCounts version 1.5.2 [44] for all aligned and processed RNA-Seq BAM files. Female XX samples when aligned to a sex chromosome complement informed reference genome will show zero counts for Y-linked genes, but will still include those genes in the raw counts file. This is an essential step for downstream differential expression analysis between males and females to keep the total genes the same between the sexes for comparison. Only rows that matched gene feature type in Ensembl Homo_sapiens.GRCh38.89.gtf gene annotation [13] were included for read counting. There are

2283 genes annotated on the X chromosome and a total of 56,571 genes across the entire genome for GRCh38 version of the human reference genome [13]. Only the primary alignments were counted and specified using the --primary option in featureCounts.

RNA-Seq quantification for transcriptome index

Transcript quantification for trimmed paired RNA-Seq brain cortex samples was estimated twice, once to a default decoy-aware reference transcriptome index and once to a sex chromosome complement informed decoy-aware reference transcriptome index using Salmon with the –validateMappings flag. Salmon's –validateMappings adopts a scheme for finding protentional mapping loci of a read using a chain algorithm introduced in minimap2 [45]. Transcript quantification for female (46, XX) samples was estimated using a Y-masked reference transcriptome index, and male (46, XY) transcript quantification was estimated using a Y PAR masked reference transcriptome index when the Y PAR sequence information was available for the transcriptome build. This was repeated for both the Ensembl and the gencode cDNA transcriptome builds, keeping all parameters the same, only changing the reference transcriptome index used, as described above.

DGEList object

Differential expression analysis was performed using the limma/voom pipeline [33] which has been shown to be a robust differential expression software package [46, 47] for both reference-based and pseudo-alignment quantification. Quantified read counts from each sample for the reference-based quantification which were generated from featureCounts were combined into a count matrix, each row representing a unique gene ID and each column representing the gene counts for each unique sample. This was repeated for each tissue type and read into R using the DGEList function in the R limma package [48]. A sample-level information file related to the genetic sex of the sample, male or female, and the reference genome used for alignment, default or sex chromosome complement informed, was created and corresponds to the columns of the count matrix described above.

The pseudo-aligned transcript read counts from each brain cortex sample quantified using Salmon were combined into a count matrix using tximport [49] with each row representing a unique transcript ID and each column representing the transcript counts for each unique sample. To create length scaled transcripts per million (TPM) values to pass into limma, tximport function lengthScaledTPM was employed [49]. The reference assembly annotation file was read into R using tximport function makeTxDbFromGFF. Following this, a key of the transcript ID corresponding to the gene ID was created using the keys function [49]. Gene-level TPM values were then generated using the tx2gene function. This was repeated for the Ensembl and the gencode default and sex chromosome complement informed transcriptome quantification estimates.

Multidimensional scaling

Multidimensional Scaling (MDS) was performed using the DGEList-object containing gene expression count information for each sample. MDS plots were generated using the plotMDS function in the R limma package [33]. The distance between each pair of samples is shown as the $\log_2$ fold change between the samples. The analysis was done for each tissue separately using all shared common variable genes for dimensions (dim) 1 and 2 and dim 2 and 3. Samples that did not cluster with reported sex or clustered in unexpected ways in either dim1, 2, or 3 were removed from all downstream analysis (Additional file 5). MDS plots for each tissue containing the samples that were used for quality control are located in Additional file 6. Briefly, one male XY whole blood did not cluster with any of the other samples and was removed. One female XX breast sample clustered with the opposite sex and was thus removed. In the brain cortex, three male XY brain cortex samples that did not cluster neatly with the other male XY samples in dim 1 and 2 were thus removed. Another male brain cortex sample, although clustered with other male samples, had the lowest number of sequencing remaining after trimming for quality, 23.9 M, and thus was also removed. To keep the number of samples in each sex roughly equal, four female XX brain cortex samples were randomly selected for removal. For liver and thyroid tissue, no samples appeared to cluster in any unexpected ways and thus no liver or thyroid tissue samples were removed. For all aligners, the first component of variation in the MDS plot is explained by the sex of the sample (Fig. 3).

Differential expression

Using edgeR [50], raw counts were normalized to adjust for compositional differences between the RNA-Seq libraries using the voom normalize quantile function, which normalizes the reads by the method of trimmed mean of values (TMM) [33]. Counts were then transformed to $\log_2(\mathrm{CPM} + 0.25/L)$, where CPM is the counts per million, L is the library size, and 0.25 is a prior count to avoid taking the log of zero [50]. For this dataset, the average library size is about 79.76 million; therefore, L is 79.76. Thus, the minimum $\log_2(\mathrm{CPM} + 0.25/L)$ value for each sample, representing zero transcripts, is $\log_2(0 + 0.25/15) = -8.32$.

A mean minimum of 1 CPM, or the equivalent of 0 in $\log_2(\mathrm{CPM} + 2/L)$, in at least one sex per tissue

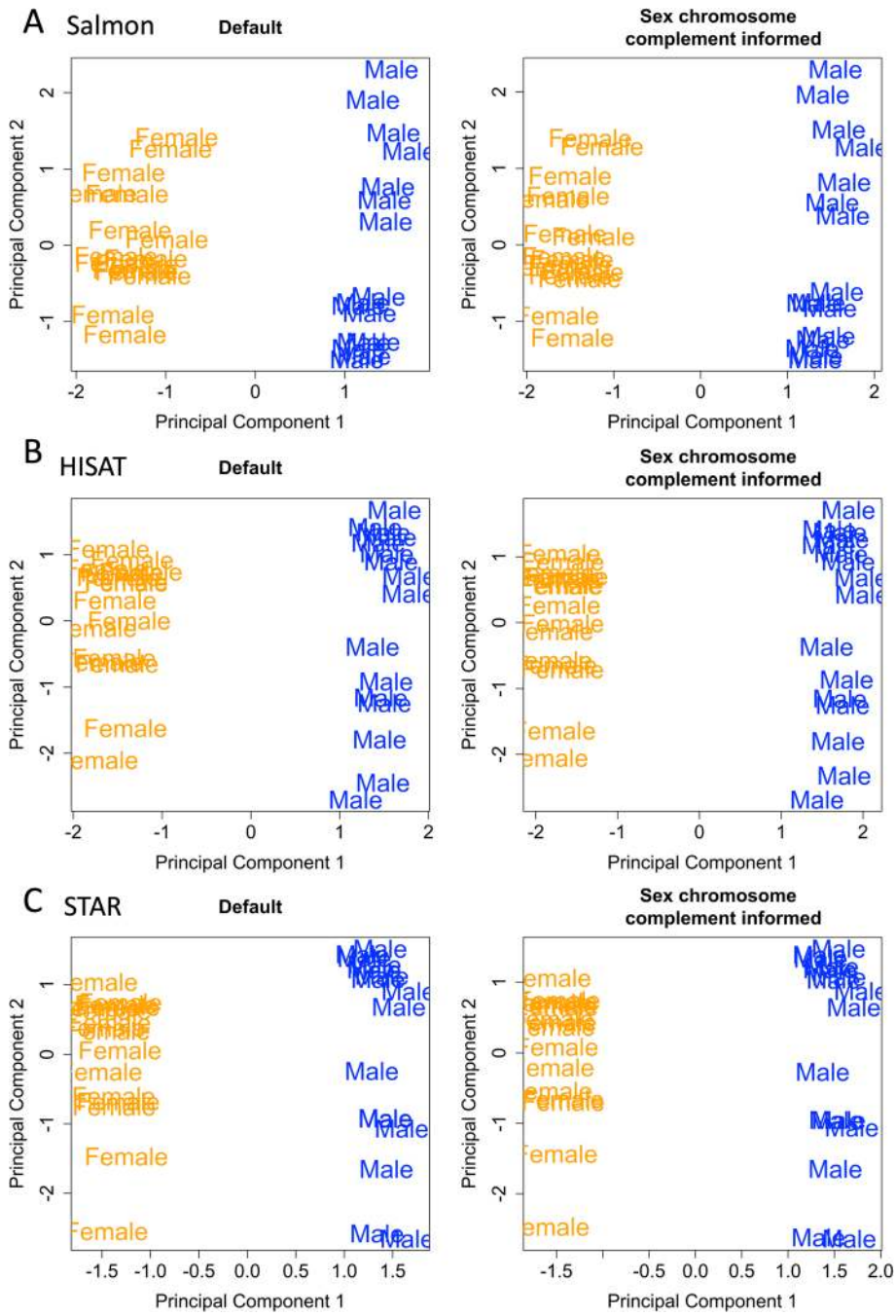


Fig. 3 Multidimensional scaling for the top 100 most variable genes. We investigated multidimensional scaling for the top 100 common variable genes in the brain cortex samples. **a** Salmon pseudo-alignment with Ensembl transcriptome reference, **b** HISAT read aligner, and **c** STAR read aligner when quantifying using both the default and the sex chromosome complement informed references. Most variation in the data is explained by the sex of the sample

comparison was required for the gene to be kept for downstream analysis. A CPM value of 1 was used in our analysis to separate expressed genes from unexpressed genes, meaning that in a library size of ~ 79.76 million reads, there are at least an average of 79 counts in at least one sex. After filtering for a minimum CPM, 53,804 out of the 56,571 quantified genes were retained for the whole blood samples, 53,822 for brain cortex, 54,184 for breast, 53,830 for liver, and 53,848 for thyroid. A linear model was fitted to the DGEList-object, which contains the filtered and normalized gene counts for each sample, using the limma lmfit function which will fit a

separate model to the expression values for each gene [33].

For differential expression analysis, a design matrix containing the genetic sex of the sample (male or female) and which reference genome the sample was aligned to (default or sex chromosome complement informed) was created for each tissue type for contrasts of pairwise comparisons between the sexes. Pairwise contrasts were generated using limma makecontrasts function [33]. We identified genes that exhibited significant expression differences defined using an Benjamini-Hochberg adjusted p value cutoff that is less than 0.01 (1%) to account for multiple testing in pairwise comparisons between conditions using limma/voom decideTests vebayesfit [33]. A conservative adjusted p value cutoff of less than 0.01 was chosen to be highly confident in the genes that were called as differentially expressed when comparing between reference genomes used for alignment. Pipeline is available on GitHub, https://github.com/SexChrLab/XY_RNAseq.

GO analysis

We examined the differences and similarities in gene enrichment terms between the differentially expressed genes obtained from the differential expression analyses of the samples aligned to the default and sex chromosome complement informed reference genomes, to investigate if the biological interpretation would change depending on the reference genome the samples were aligned to. We investigated the Gene Ontology enrichment for lists of genes that were identified as showing overexpression in one sex versus the other sex for whole blood, brain cortex, breast, liver, and thyroid samples (adjusted p value < 0.01). We used the GOrilla webtool, which utilizes a hypergeometric distribution to identify enriched GO terms [51]. A modified Fisher exact p value cutoff < 0.001 was used to select significantly enriched terms [51].

Results

RNA-Seq reads aligned to autosomes do not vary much between reference genomes

We compared total mapped reads when reads were aligned to a default reference genome and to a reference genome informed on the sex chromosome complement. Reads mapped across the whole genome, including the sex chromosomes, decreased when samples were aligned to a reference genome informed on the sex chromosome complement, paired t test p value < 0.05 (Additional files 7, 8 and 9). This was true regardless of the read aligner used, HISAT or STAR, or of the sex of the sample, XY or XX. To test the effects of realignment on an autosome, we selected chromosome 8, because of its similar size to chromosome X. Overall, there is a slight mean increase in reads mapping to chromosome 8 when samples are aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome (Additional file 9). For female XX samples, the mean increase in reads mapping for chromosome 8 was 42.2 reads for the whole blood, 50.25 for the brain cortex, 109.9 for the breast, 68.5 for the liver, and 98.2 for the thyroid (Additional file 9), which was significant using a paired t test, p value < 0.05 in all tissues (Additional file 9). Male XY samples also showed a mean increase in reads mapping for chromosome 8. The mean increase in reads mapping to chromosome 8 for male whole blood samples was 0.84, 2.38 for brain cortex, 5.58 for breast, 3.2 for liver, and 5 for thyroid (Additional file 9). There was a significant increase, p value < 0.05 paired t test, for reads mapping to chromosome 8 for male brain cortex, breast, liver, and thyroid samples. There was no significant increase in reads mapping for male whole blood for chromosome 8 (Additional file 9).

Reads aligned to the X chromosome increase in both XX and XY samples when using a sex chromosome complement informed reference genome

We found that when reads were aligned to a reference genome informed by the sex chromosome complement for both male XY and female XX tissue samples, reads on the X chromosome increased by ~ 0.12% when aligned using HISAT. For all tissues and both sexes, we observe an average increase of 1991 reads on chromosome X. We observe an increase in reads mapping to the X chromosome for all tissues and for each sex, which was significant using a paired t test, p value < 0.05 (Additional file 9). Reads on the Y chromosome decreased 100% (67,033 reads on average) across all female XX samples and by ~ 57.32% (69,947 reads on average) across all male XY samples when aligned using HISAT (Additional file 7 & 9). Similar increases in X chromosome and decreases in Y chromosome reads when aligned to a sex chromosome complement informed reference were observed when STAR was used as the read aligner for both male XY and female XX samples (Additional file 8 & 9).

Aligning to a sex chromosome complement informed reference genome increases the X chromosome PAR1 and PAR2 expression

We next explored the effect of changes in read alignment on gene expression. There was an increase in pseudoautosomal region, PAR1 and PAR2, expression when reads were aligned to a reference genome informed on the sex chromosome complement for both male XY and female XX samples (Additional file 10 & 11). We found an average of 2.73 $\log_2$ fold increase in the expression in PAR1 for female XX brain cortex samples and

2.75 $\log_2$ fold increase in the expression in PAR1 for male XY brain cortex samples using HISAT (Fig. 4). The X-transposed region (XTR) in female XX brain cortex samples showed a 1.22 $\log_2$ fold increase in the expression and no change in male XY brain cortex samples. PAR2 showed an average of 2.13 $\log_2$ fold increase for female XX brain cortex samples and 2.19 $\log_2$ fold increase in PAR2 for male XY brain cortex samples using HISAT, with similar results for STAR read aligner (Additional file 10 & 11). Complete lists of the $\log_2(\mathrm{CPM} + 0.25/L)$ values for each X chromosomal gene and each gene within the whole genome for male XY and female XX samples are in Additional file 12 available on Dryad for download under https://doi.org/10.5061/dryad.xksn02vbv.

Regions outside the PARs and XTR show little difference in expression between reference genomes

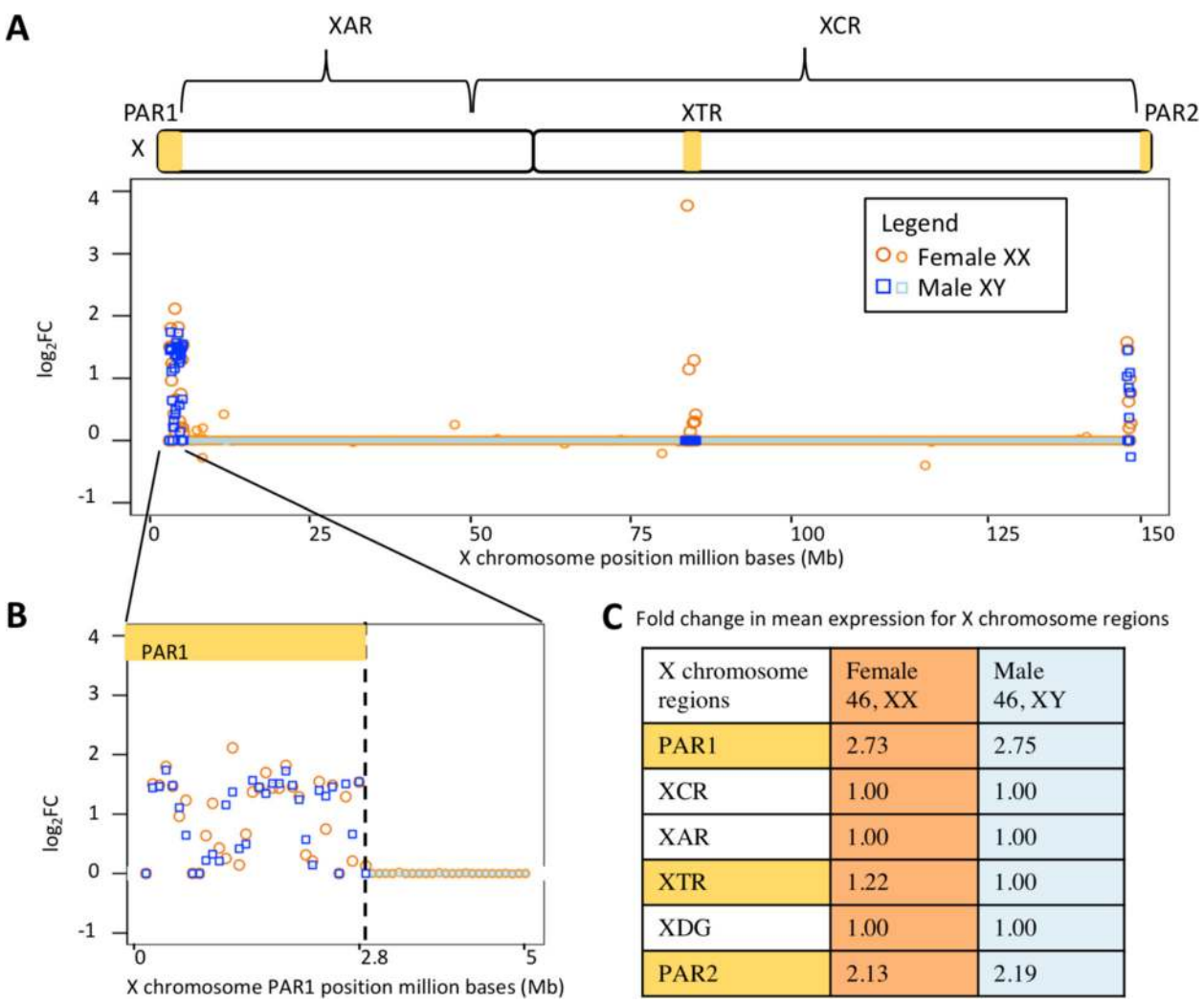
Intriguingly, regions outside the PARs on the X chromosome were less affected by the choice of the reference genome. Across the entire X-conserved region, we observed practically no change in estimates of gene expression between the default and sex chromosome complement informed references (e.g., a 0.99 $\log_2$ fold in male thyroid samples and 1.00 $\log_2$ fold change in female brain cortex samples, essentially showing no difference (Additional file 10 & 11)). Additionally, X and Y homologous genes (*AMELX*, *ARSD*, *ARSE*, *ARSF*, *CASK*, *GYG2*, *HSFX1*, *HSFX2*, *NLGN4X*, *OFD1*, *PCDH11X*, *PRKX*, *RBMX*, *RPS4X*, *SOX3*, *STS*, *TBL1X*, *TGIF2LX*, *TMSB4X*, *TSPYL2*, *USP9X*, *VCX*, *VCX2*, *VCX3A*, *VCX3B*, *ZFX*) showed little to no increase in the expression when aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome (Additional file 13). *PCDH11X* showed the highest increase in the expression for all tissues regardless of the read aligner. The $\log_2$ fold increase in thr expression for *PCDH11X* for female samples when aligned using HISAT was 0.4, 0.28, 0.33, 0.16, and 0.16 for whole blood, brain cortex, breast, liver, and thyroid, respectively. Other X and Y homologous genes sometimes increased in the expression depending on the tissue, and sometimes, there was no change in the expression (Additional file 13). Next to *PCDH11X*, the most increase in expression in an X and Y homologous genes was *VCX3B*, *NLGN4X*, and *VCX3A*. *NLGN4X* in whole blood showed a 0.14 $\log_2$ fold increase when aligned using HISAT. *VCX3B* showed a 0.2 $\log_2$ fold increase in the brain, *NLGN4X* showed a 0.04 $\log_2$ fold increase in the breast, *VCX3A* showed a 0.07 $\log_2$ fold increase in the liver, and *VCX3B* showed a 0.04 $\log_2$ fold increase in the thyroid, when aligned using HISAT (Additional file 13).

A sex chromosome complement informed reference genome increases the ability to detect sex differences in gene expression

We next investigated how this would affect the gene differential expression between the sexes. Generally, we find that more genes are differentially expressed on the sex chromosomes between the sexes when the sex chromosome complements are taken into account. The number of differentially expressed genes on the autosomes remained the same or increased. At a conservative Benjamini-Hochberg adjusted p value of < 0.01 and aligning with HISAT, we find 4 new genes (3 Y-linked and 1 X-linked) that are only called as differentially expressed between the sexes in the brain cortex when aligned to reference genomes informed on the sex chromosome complement (Fig. 5; Additional file 14). We observed similar trends in changes for differential expression between male XY and female XX for whole blood, breast, liver, and thyroid samples using either HISAT or STAR as the aligner (Additional file 14). For example, in the whole blood, 3 additional genes are called as being differentially expressed between the sexes using HISAT, while 1 additional gene is called differentially expressed when aligned using STAR. Additionally, when taking sex chromosome complement into account, the number of genes called as differentially expressed between the sexes for the breast samples increased by 13 genes (8 autosomal, 3 X-linked and 2 Y-linked) using HISAT and by 8 genes using STAR (6 autosomal and 2 X-linked) (Additional file 14 & 15). For all tissues, no genes were uniquely called as being differentially expressed between the sexes when aligned to a default reference genome compared to a reference genome informed on the sex chromosome complement (Additional file 14 & 15). Rather, only when samples were aligned to a sex chromosome complement did we observe an increase in the genes called as being differentially expressed (Fig. 5; Additional file 14 & 15).

Increase in gene enrichment pathways when samples are aligned to a sex chromosome complement informed reference genome

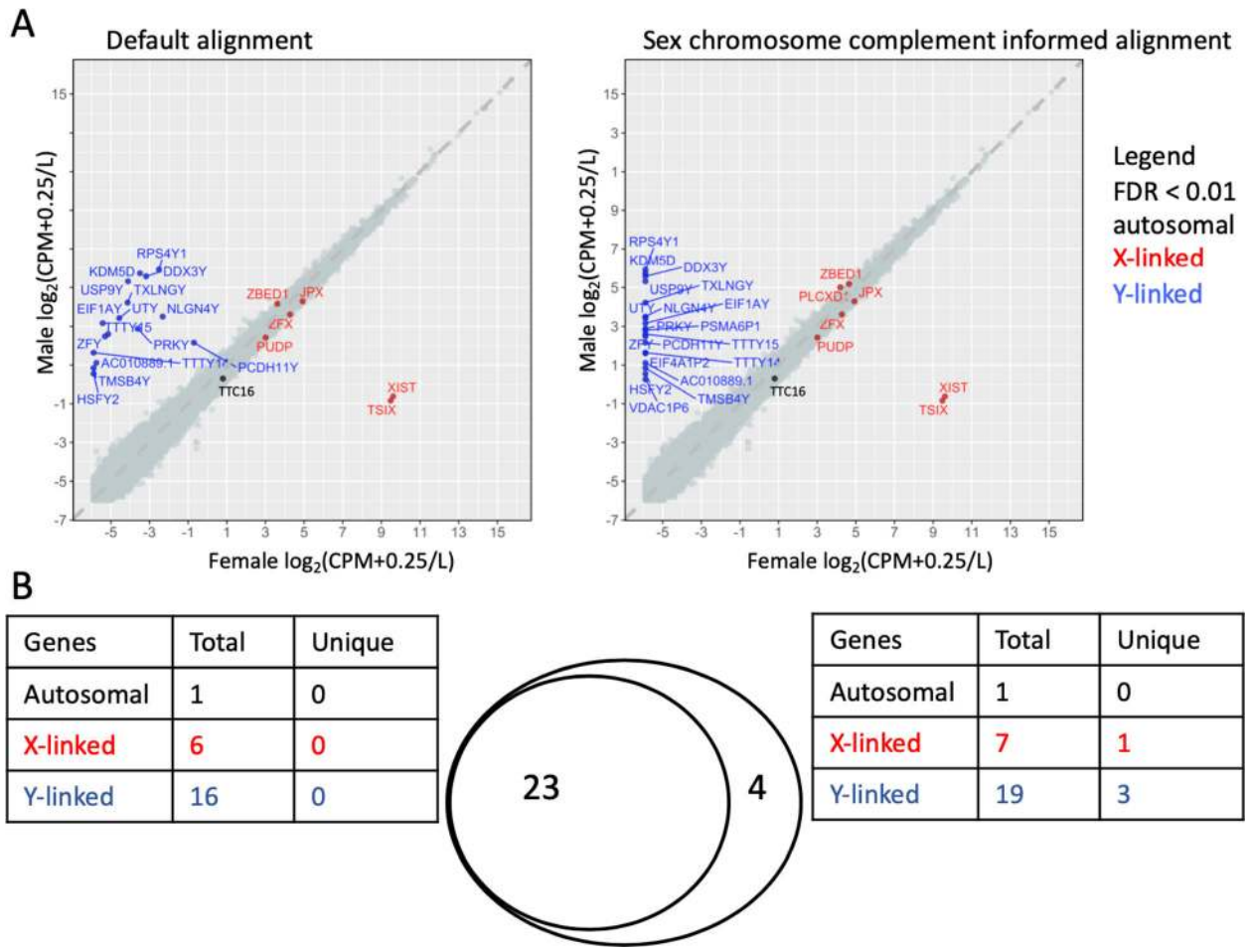
A sex chromosome complement informed reference genome increases the ability to detect genes as differentially expressed between the sexes and thus alters gene enrichment results. When the thyroid samples were aligned using a sex chromosome complement informed reference genome using HISAT, genes upregulated in male XY samples still show enrichment for positive regulation of transcription from RNA polymerase II (found when aligning to a default reference genome), but additionally find postsynaptic membrane assembly, postsynaptic membrane organization, and vocalization behavior (Additional file 16). These additional GO enrichments in the male XY thyroid samples involve



X chromosome regions	Female 46, XX	Male 46, XY
PAR1	2.73	2.75
XCR	1.00	1.00
XAR	1.00	1.00
XTR	1.22	1.00
XDG	1.00	1.00
PAR2	2.13	2.19

Fig. 4 X chromosome RNA-Seq alignment differences in the brain cortex. We plot $\log_2$ fold change (FC) across **a** the entire X chromosome and **b** the first 5 million bases (Mb) and show **c** the average fold change in large genomic regions on the X chromosome between the aligning brain cortex using HISAT to the default genome and aligning to a sex chromosome complement informed reference genome. For $\log_2$ FC, a value less than zero indicates that the gene showed higher expression when aligned to a default reference genome, while values above zero indicate that the gene shows higher expression when aligned to a reference genome informed by the sex chromosome complement of the sample. Samples from genetic females are plotted in orange circles, while samples from males are plotted in blue squares. Darker shades indicate which gene points are in PAR1, XTR, and PAR2 while lighter shades are used for genes outside of those regions

NRXN1 and *NLGN4Y* genes; both of these genes are located on the Y chromosome. GO enrichment analysis of genes that are more highly expressed in female liver compared to male liver samples, when samples were aligned to a default reference genome using HISAT, were genes involved in modification histone lysine demethylation (Additional file 16). However, when these samples were aligned to a sex chromosome complement informed reference genome, genes upregulated in females were enriched for histone lysine demethylation as well as negative regulation of endopeptidase activity, negative regulation of peptidase activity, cytoplasmic actin-based contraction involved in cell motility (Additional file 16). These additional GO enrichments in the female XX liver samples include the involvement of *KDM6A*, *DDX3X*, and *VIL1*. *KDM6A* and *DDX3X* are X-linked, and *VIL1* is on chromosome 2. Whole blood, brain cortex, male liver, and female thyroid samples showed no difference in GO enrichment pathways when using a default reference genome compared to a sex chromosome complement reference genome for alignment when using HISAT with similar results for STAR as the read aligner (Additional file 17). Thus, while there will not always be a difference, aligning to a sex chromosome complement informed reference genome can increase the ability to detect enriched pathways.



Genes	Total	Unique
Autosomal	1	0
X-linked	6	0
Y-linked	16	0

Genes	Total	Unique
Autosomal	1	0
X-linked	7	1
Y-linked	19	3

Fig. 5 Sex chromosome complement informed alignment calls more sex-linked genes as being differentially expressed. **a** Sex differences in the gene expression, $\log_2$(CPM + 0.25/*L*), between the twenty samples from genetic males and females are shown when aligning all samples to the default reference genome (left) and a reference genome informed on the sex chromosome complement (right) for the brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted *p* value < 0.01, are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. **b** We show the overlap between genes that are called as differentially expressed when all samples are aligned to the default genome, and genes that are called as differentially expressed when aligned to a sex chromosome complement informed genome. When samples were aligned to a reference genome informed on the sex chromosome complement, 27 genes were called as differentially expressed between the sexes, of which 4 were uniquely called in the sex chromosome complement informed alignment. There were no genes that were uniquely called as differentially expressed when aligned to a default reference genome

Using sex-linked genes alone is inefficient for determining the sex chromosome complement of a sample

The sex of each sample used in this analysis was provided in the GTEx manifest. We investigated the expression of genes that could be used to infer the sex of the sample. We studied X and Y homologous genes (*DDX3X/Y*, *PCDH11X/Y*, *USP9X/Y*, *ZFX/Y*, *UTX/Y*), *XIST*, and *SRY* gene expression in male and female whole blood, brain cortex, breast, liver, and thyroid (Fig. 2; Additional file 3 & 4). Both males and females are expected to show expression for the X-linked homologs, whereas only XY samples should show expression of the Y-linked homologs. Further, *XIST* expression should only be observed in XX samples and *SRY* should only be expressed in samples with a Y chromosome. Using the default reference genome for aligning samples, we observed a small number of reads aligning to the Y-linked genes in female XX samples, but also observed clustering by sex for *DDX3Y*, *USP9Y*, *ZFY*, and *UTY* gene expression (Fig. 2). Male XY samples showed expression for *DDX3X*, *DDX3Y*, *USP9X*, *ZFX*, and *UTX* (greater than 5 $\log_2$(CPM + 2/*L*). Female XX samples showed expression for *XIST* (greater than 4.0 $\log_2$(CPM + 2/*L*), and male XY samples showed little to no expression for *XIST* (less than 0 $\log_2$(CPM + 2/*L*) with the exception of 2 male whole blood samples and 1 male liver sample, which showed greater than 5 $\log_2$(CPM + 2/*L*) expression). In contrast to the default reference genome, when aligned to a sex chromosome

complement informed reference genome, samples cluster more distinctly by sex for *DDX3Y*, *USP9Y*, ZFY, and *UTY*, all showing at least a 4 $\log_2$(CPM + 2/*L*) difference between the sexes (Fig. 2; Additional file 3 & 4). *SRY* is predominantly expressed in the testis [52, 53] and typically one would expect *SRY* to show male-specific expression. In our set, we did not observe *SRY* expressed in any sample, and so it could not be used to differentiate between XX and XY samples (Fig. 2, Additional file 3 & 14). In contrast, the X-linked gene *XIST* was differentially expressed between genetic males and genetic females in both genome alignments (default and sex chromosome complement informed) for the whole blood, brain cortex, breast, liver, and thyroid samples with the exception of 3 male XY samples. *XIST* expression is important in the X chromosome inactivation process [54] and serves to distinguish samples with one X chromosome from those with more than one X chromosome [23]. However, this does not inform about whether the sample has a Y chromosome or not. For X-Y homologous genes, we do not find sex differences in the read alignment with either default or sex chromosome complement informed for the X-linked homolog. When aligned to a default reference genome, female XX samples showed some expression for homologous Y-linked genes, but only the presence/absence of Y-linked reads alone is insufficient to determine sex chromosome complement of the sample (Fig. 2, Additional file 3).

No Y-linked transcript expression in female XX samples when quantification was estimated using a transcriptome index informed on the sex chromosome complement

A pseudo-alignment shows similar effects of the reference to that of an alignment approach (Fig. 5, Additional files 18 & 19). We observed no Y-linked expression in female XX samples when transcript quantification was estimated using a Y-masked sex chromosome complement reference transcriptome index. This was true for both the Ensembl and gencode pseudo-alignment with a sex chromosome complement reference transcriptome index (Additional files 18 & 19). Interestingly, there was a large difference between the Ensembl and gencode reference files. The transcript IDs in the transcriptome cDNA fasta and the transcript IDs in the annotation file are not one-to-one for the Ensembl assembly [55]. There are 190,432 transcript sequences in the Ensembl cDNA fasta file, but there are 199,234 transcripts in the Ensembl annotation file. Notably, Ensembl's cDNA reference transcriptome fastas do not contain known transcripts such as the XIST transcripts [56]. The Ensembl reference transcriptome fasta also does not contain the Y PAR transcript sequences, it only contains the X PAR transcript sequences. In contrast, the gencode cDNA reference transcriptome fasta and annotation file both contain 206,694 sequences, including the Y PARs. Regardless of using an Ensembl or gencode transcriptome, female XX sample shows Y-linked expression when using a default reference transcriptome index for pseudo-alignment; however, the changes necessary for making a sex chromosome complement informed reference are different for the two builds.

Discussion

For accuracy, the sex chromosome complement of the sample should be taken into account when aligning RNA-Seq reads to reduce misaligning sequences. Neither Ensembl nor Gencode human reference genomes are correct for aligning both XX and XY samples. The Ensembl GRCh38 human reference genome includes all 22 autosomes, mtDNA, the X chromosome, the Y chromosome with the Y PARs masked, and contigs [13]. The Gencode hg19 human reference genome includes everything with no sequences masked [36].

Measurements of X chromosome expression increase for both male XY and female XX whole blood, brain cortex, breast, liver, and thyroid samples when aligned to a sex chromosome complement informed reference genome versus aligning to a default reference genome (Fig. 4). While we see increases in measured expression for PAR1 and PAR2 genes in both males and females, we only observe a difference in measured XTR expression in females. This is because while the PARs are 100% identical between the X and Y and so one copy (here we mask the Y-linked copy) should be masked, the XTR is not hard-masked in the YPAR-masked reference genome. The XTR is not identical between the X and Y; it shares 98.78% homology between X and Y but no longer recombines between X and Y [15] (Fig. 1a), and because of this divergence, it is therefore not hard-masked when aligning male XY samples. Tukiainen et al. [23] and others have shown that PAR1 genes have a male bias in expression [23]. Our findings here support this regardless if the samples were aligned to a default or a sex chromosome complement reference genome (Additional file 11 & 12). Differential expression results changed when using a sex chromosome complement informed alignment compared to using a default alignment. When aligned to a default reference genome, due to sequence similarity, some reads from female XX samples aligned to the Y chromosome (Figs. 2 and 5). However, when aligned to a reference genome informed by the sex chromosome complement, female XX samples no longer showed Y-linked gene expression, and more Y-linked genes were called as being differentially expressed between the sexes (Figs. 2 and 5; Additional file 12 & 15). This suggests that if using a default reference genome for aligning RNA-Seq reads, one would miss some Y-linked genes as differentially expressed between the sexes (Fig. 5). Furthermore, these Y-linked genes serve

in various important biological processes, thus altering the functional interpretation of the sex differences (Additional file 16 & 17). Only when samples were aligned to a sex chromosome complement reference genome did we observe more genes called as differentially expressed between the sexes (Additional file 14). An increase in genes called differentially expressed additionally alters the GO analysis results (Additional files 16 & 17). When samples were aligned to a default reference genome, we sometimes missed the GO pathways or misinterpreted which were the top pathways.

The choice of read aligner has long been known to give slightly differing results of differential expression due to the differences in the alignment algorithms [46, 57]. Differences between HISAT and STAR could be contributed to differences in default parameters for handling multi-aligning reads [31]. We show that regardless of choice of read aligner, HISAT or STAR, we observe similar results. Sample size has also long been known to alter differential expression analysis [58–60]. We therefore additionally replicated our findings in a smaller sample size of 3 male XY compared to 3 female XX samples for whole blood and brain cortex tissue and where the samples were randomly selected and confirmed the results from the larger sample size (Additional file 20).

In addition to reference-based quantification, we tested whether quantifying sex-linked reads with a pseudo-aligner would be affected by using a sex chromosome complement reference. Previous studies have shown that reference-based alignment is not necessary for high-quality estimation of transcript levels [61]. However, we observed expression estimates for Y-linked transcripts in female XX samples when using a default reference transcriptome index for pseudo-alignment quantification estimates. In contrast, when a sex chromosome complement informed reference transcriptome index was used, we observed no Y-linked expression in female XX samples. Salmon, and other alignment-free tools such as Kallisto [62] and Sailfish [63], builds an index of *k*-mers from a reference transcriptome. The *k*-mer transcriptome index is used to group pseudoalignments belonging to the same set of transcripts to directly estimate the expression of each transcript. A *k*-mer alignment-free approach is faster and less demanding than alignment protocols [61]; however, a sex chromosome complement informed transcriptome index should be carefully considered because even a *k*-mer approach is not sensitive to regions that are 100% identical in sequence. Additionally, alignment-free methods are not as robust in quantifying expression estimates for small RNAs and lowly expressed genes [64].

The choice of reference transcriptome or reference genome can also give slightly differing results of differential expression due to the difference in which transcripts are included in the transcriptome [55]. The Ensembl cDNA does not include the Y PAR-linked transcripts whereas the gencode transcriptome fasta includes both the X and Y PARs. The Ensembl transcriptome does not include non-coding RNAs, such as *XIST* transcripts. The *XIST* gene is called as being upregulated in the female XX samples for all tissues, and all comparisons except for when transcript expression were estimated using the Ensembl reference transcriptome (Additional file 15, 18, & 19). Given the current builds, for RNA-Seq projects interested in sex chromosome-linked transcript expression, we suggest that researchers use a gencode sex chromosome complement informed reference transcriptome index.

Ideally, one would use DNA to confirm the presence or absence of the Y chromosome, but if DNA sequence was not generated, one would need to confirm the genetic sex of the sample by assessing expression estimates for X-linked and Y-linked genes. To more carefully investigate the ability to use gene expression to infer sex chromosome complement of the sample, we examined the gene expression for a select set of X-Y homologous genes, as well as *XIST* and *SRY* that are known to be differentially expressed between the sexes (Fig. 2, Additional file 13). The samples broadly segregated by sex for Y-linked gene expression using default alignment. However, the pattern was messy for each individual Y-linked gene. Thus, if inferring sex from RNA-Seq data, we recommend using the estimated expression of multiple X-Y homologous genes and *XIST* to infer the genetic sex of the sample. Samples should be aligned to a default reference genome first to look at the expression for several Y-specific genes to determine if the sample is XY or XX. Then, samples should be realigned to the appropriate sex chromosome complement informed reference genome. Independently assessing sex chromosome complement of samples becomes increasingly important as karyotypically XY individuals are known to have lost the Y chromosome in particular tissues sampled, as shown in Alzheimer disease [65], age-related macular degeneration [66], and in the blood of aging individuals [67], but should not have *XIST* expression. However, *XIST* may not be a sufficient marker alone to infer sex chromosome complement, especially in cancer in samples from XX individuals, where the inactive X can become reactivated [68]. Self-reported sex may not match the sex chromosome complement of the samples, even in karyotypic individuals.

Conclusion

Here, we show that aligning RNA-Seq reads to a sex chromosome complement informed reference genome will change the results of the analysis compared to aligning reads to a default reference genome. We previously observed that a sex chromosome complement informed alignment is important for DNA as well [18]. A sex

chromosome complement informed approach is needed for a sensitive and specific analysis of gene expression on the sex chromosomes [2]. A sex chromosome complement informed reference alignment resulted in increased expression of the PARs of the X chromosome for both male XY and female XX samples. We further found different genes called as differentially expressed between the sexes and identified sex differences in gene pathways that were missed when samples were aligned to a default reference genome.

Perspectives and significance

The accurate alignment and pseudo-alignment of the short RNA-Seq reads to the reference genome or reference transcriptome is essential for drawing reliable conclusions from differential expression data analysis on the sex chromosomes. We strongly urge studies using RNA-Seq to carefully consider the genetic sex of the sample when quantifying reads and provide a framework for doing so in the future (https://github.com/SexChrLab/XY_RNAseq).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13293-020-00312-9.

Additional file 1: Sample IDs. RNA-Seq whole blood, brain cortex, breast, liver, and thyroid tissue samples from 20 genetic female (46, XX) and 20 genetic male (46, XY) individuals were downloaded from the Genotype-Tissue Expression (GTEx) project [19] for a total of 200 RNA-Seq tissue samples.

Additional file 2: Histogram of sample reported age. For each tissue, whole blood, brain cortex, breast, liver, and thyroid, male XY and female XX samples were age matched perfectly between age 55 to 70. Females are shown in blue and males are shown in lime green. Since the samples were aged perfectly the histogram bars show only the overlap of female and male samples is a mix color of the blue and lime green.

Additional file 3: Genetic sex of RNA-Seq samples when aligned using STAR. Gene expression $\log_2$(CPM + 0.25/L) for select XY homologous genes (*DDX3X/Y, PCDH11X/Y, USP9X/Y, ZFX/Y, UTX/Y*) and *XIST* and *SRY* when reads were aligned to a default reference genome **A)**, and for **B)** when reads were aligned to a sex chromosome complement informed reference using STAR. Male XY whole blood, brain cortex, breast, liver, and thyroid samples are shown in blue squares and female XX in orange circles.

Additional file 4: Genetic sex of RNA-Seq samples per tissue. Gene expression $\log_2$(CPM + 0.25/L) for select XY homologous genes (*DDX3X/Y, PCDH11X/Y, USP9X/Y, ZFX/Y, UTX/Y*) and *XIST* and *SRY* when reads were aligned to a default reference genome **A)**, and for **B)** when reads were aligned to a sex chromosome complement informed reference using HISAT and **C)** and **D)**, for when the reads were aligned using STAR. Male XY whole blood, brain cortex, breast, liver, and thyroid samples are shown in blue squares and female XX in orange circles.

Additional file 5: List of samples that were removed from downstream analysis. Samples that did not cluster with the reported sex or clustered in unexpected ways were removed from the differential expression analysis. One male XY whole blood, 4 female XX and 4 male XY brain cortex, and one female XX breast sample were removed.

Additional file 6: Multidimensional Scaling plots. We investigated multidimensional scaling for all shared common variable genes for dimensions 1 and 2, and for dimensions 2 and 3 in each tissue. The most variation in each tissue is explained by the aligner **C.aligner**. The second most variation in each tissue is explained by the sex of the sample **A.sex**.

Additional file 7: HISAT mapped reads bar plot. Mean difference in expression for average total reads mapped for each tissue and each sex when aligned to a sex chromosome informed versus a default reference genome. Paired t-test to test for significant difference in total reads mapped for the whole transcriptome, chromosome 8, and chromosome X. Nonparametric Wilcox single rank sum test was used to test for significant difference in total reads mapped on the Y chromosome for male samples in each tissue separately. Red * indicate a significant, p-value <0.05, difference in average mapped reads, NS is no significant differences.

Additional file 8: STAR mapped reads bar plot. Mean difference in expression for average total reads mapped for each tissue and each sex when aligned to a sex chromosome informed versus a default reference genome. Paired t-test to test for significant difference in total reads mapped for the whole transcriptome, chromosome 8, and chromosome X. Nonparametric Wilcox single rank sum test was used to test for significant difference in total reads mapped on the Y chromosome for male samples in each tissue separately. Red * indicate a significant, p-value <0.05, difference in average mapped reads, NS is no significant differences.

Additional file 9: Paired t-test for mapped reads in default compared to sex chromosome complement reference genome. Mean difference in expression for average total reads mapped for each tissue and each sex when aligned to a sex chromosome informed versus a default reference genome. Paired t-test to test for significant difference in total reads mapped for the whole transcriptome (WT), chromosome 8, and chromosome X. Nonparametric Wilcox single rank sum test was used to test for significant difference in total reads mapped on the Y chromosome for male samples in each tissue separately.

Additional file 10: X chromosome expression differences between default and sex chromosome complement informed alignment. X chromosome gene expression differences between default and sex chromosome complement informed alignment. Increase in expression when aligned to a sex chromosome complement informed reference genome is a $\log_2$ fold change (FC) > 0. A decrease in expression when aligned to a sex chromosome complement informed reference genome is $\log_2$ FC < 0. Female XX samples are indicated by red and pink circles for PAR1, XTR, PAR2 genes, and for all other X chromosome genes respectively. Blue and light blue squares represent male XY samples. Blue squares indicate which gene points are in PAR1, XTR, and PAR2, and light blue squares are for genes outside of those regions. Differences in X chromosome expression between reference genomes default and sex chromosome complement for male XY and female XX samples aligned using HISAT for the whole X chromosome and the first 5 Mb are shown for the whole blood (**A** and **B**, respectively), brain cortex (**E** and **F**, respectively), breast (**I** and **J**, respectively), liver (**M** and **N**, respectively), and thyroid (**Q** and **R**, respectively). Differences in X chromosome expression between reference genomes for male XY and female XX samples aligned using STAR for the whole X chromosome and the first 5 Mb are shown for the whole blood (**C** and **D**, respectively), brain cortex (**G** and **H**, respectively), breast (**K** and **L**, respectively), liver (**O** and **P**, respectively), and thyroid (**S** and **T**, respectively).

Additional file 11: X chromosome regions mean and median expression values. X chromosome regions PAR1, PAR2, XTR, XDG, XAR, XCR mean and median CPM expression for male XY and female XX samples for each tissue separately when aligned to a default or sex chromosome complement informed reference genome using either HISAT and STAR. Paired t-test was used to test for significant differences in expression. XTR and XAR show a significant increase, p-value <0.05, in female expression for each tissue type. XTR and XAR additionally show a significant increase, p-value <0.05, in male expression for liver and thyroid. PAR2 shows a significant increase, p-value <0.05, in female liver expression. Additionally reported fold change in mean expression when using a sex chromosome complement informed compared to a default reference genome. The mean fold change in expression either increased or stayed the same ranging from 2.8 to 0.999 fold increase in expression. Finally, mean male over mean female expression was reported for each X

chromosome region for each tissue. Mean male over mean female expression decreases for XTR when using a sex chromosome complement reference genome for each tissue.

Additional file 12: Whole genome gene expression values per sample, aligner and reference genome used for alignment. CPM values for male XY and female XX whole blood, brain cortex, breast, liver and thyroid samples when aligned to a default and sex chromosome complement informed reference genome for the whole genome (1-22, mtDNA, X, Y and non-chromosomal).

Additional file 13: Gene expression for XY homologous genes. X chromosome expression for 26 X and Y homologous genes (*AMELX*, *ARSD*, *ARSE*, *ARSF*, *CASK*, *GYG2*, *HSFX1*, *HSFX2*, *NLGN4X*, *OFD1*, *PCDH11X*, *PRKX*, *RBMX*, *RPS4X*, *SOX3*, *STS*, *TBL1X*, *TGIF2LX*, *TMSB4X*, *TSPYL2*, *USP9X*, *VCX*, *VCX2*, *VCX3A*, *VCX3B*, *ZFX)*. Difference in gene expression for when male XY and female XX samples were aligned to a default and sex chromosome complement informed reference genome for each tissue. Little to no difference in gene expression between default and sex chromosome complement informed reference genome alignment was observed for 25 of the 26 X and Y homologous genes for both male XY and female XX samples using either HISAT or STAR. The $\log_2$ fold increase in expression for *PCDH11X* when aligned using HISAT was 0.4, 0.28, 0.33, 0.16, and 0.16 for whole blood, brain cortex, breast, liver, and thyroid, respectively. The greatest increase in expression was observed for *PCDH11X* in female whole blood at a $\log_2$ fold increase of 0.4.

Additional file 14: Differentially expressed genes between the sexes that were uniquely and jointly called between reference genomes. Genes that are differentially expressed between the sexes, male XY and female XX, for whole blood, brain cortex, breast, liver, and thyroid samples. Differentially expressed genes that are uniquely called when using either the default or sex chromosome complement informed reference genome and differentially expressed genes that were jointly called between the reference genomes.

Additional file 15: Gene expression differences between male XY and female XX samples. Sex differences in gene expression for whole blood, brain cortex, breast, liver, and thyroid samples for when samples were aligned to a default reference genome and to a reference genome informed on the sex chromosome complement. Showing sex differences in gene expression between reference genomes used for alignment and for when samples were aligned using HISAT and STAR.

Additional file 16: GO analysis of differentially expressed genes in female and male samples with HISAT aligner. Gene enrichment analysis of genes that are more highly expressed in one sex versus the other sex for each tissue, whole blood, brain cortex, breast, liver and thyroid, when samples were aligned to a default or sex chromosome complement informed reference genome using HISAT.

Additional file 17: GO analysis of differentially expressed genes in female and male samples with STAR aligner. Gene enrichment analysis of genes that are more highly expressed in one sex versus the other sex for each tissue, whole blood, brain cortex, breast, liver and thyroid, when samples were aligned to a default or sex chromosome complement informed reference genome using STAR.

Additional file 18: Sex chromosome complement informed transcriptome reference eliminates Y-linked expression in female XX samples. A) Sex differences in gene expression, $\log_2$(CPM + 0.25/L), between the sixteen samples from genetic males and females are shown when aligning all samples to the default Ensembl reference transcriptome (left) and a reference transcriptome informed on the sex chromosome complement (right) for brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p-value <0.01 are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. **B)** We show overlap between genes that are called as differentially expressed when all samples are pseudo-aligned to the default transcriptome, and genes that are called as differentially expressed when pseudo-aligned to a sex chromosome complement informed transcriptome reference. When samples were aligned to a reference transcriptome informed on the sex chromosome complement, 14 genes were called as differentially expressed between the sexes. *PLCXD1* was uniquely called as differentially expressed when aligned to a default reference genome. **Ensembl sex chromosome complement informed transcriptome reference eliminates Y-linked expression in female XX samples. A)** Sex differences in gene expression, $\log_2$(CPM + 0.25/L), between the sixteen samples from genetic males and females are shown when aligning all samples to the default Ensembl reference transcriptome (left) and a reference transcriptome informed on the sex chromosome complement (right) for brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p-value <0.01 are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. **B)** We show overlap between genes that are called as differentially expressed when all samples are pseudo-aligned to the default transcriptome, and genes that are called as differentially expressed when pseudo-aligned to a sex chromosome complement informed transcriptome reference. When samples were aligned to a reference transcriptome informed on the sex chromosome complement, 14 genes were called as differentially expressed between the sexes. *PLCXD1* was uniquely called as differentially expressed when aligned to a default reference genome.

Additional file 19: Gencode sex chromosome complement informed transcriptome reference eliminates Y-linked expression in female XX samples. A) Sex differences in gene expression, $\log_2$(CPM + 0.25/L), between the sixteen samples from genetic males and females are shown when aligning all samples to the default gencode reference transcriptome (left) and a reference transcriptome informed on the sex chromosome complement (right) for brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p-value <0.01 are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. **B)** We show overlap between genes that are called as differentially expressed when all samples are pseudo-aligned to the default transcriptome, and genes that are called as differentially expressed when pseudo-aligned to a sex chromosome complement informed transcriptome reference. When samples were aligned to a reference transcriptome informed on the sex chromosome complement, 17 genes were called as differentially expressed between the sexes. *ZBED1* was uniquely called as differentially expressed when aligned to a default reference genome.

Additional file 20:. 3 male XY and 3 female XX brain cortex and whole blood differential expression analysis. Replicated analysis in a smaller sample size of 3 male XY compared to 3 female XX samples for whole blood and brain cortex tissue. Samples were randomly selected, and confirm the results from the larger sample size.

Acknowledgements
We thank Heini Natri for the comments on the manuscript.

Authors' contributions
KCO: supervision; formal analysis; investigation; visualization; writing—original draft preparation; writing—review and editing. SMB: formal analysis; investigation; writing—original draft preparation; writing—review and editing. JPA: formal analysis; investigation; writing—review and editing. VAVV: investigation; writing—review and editing. MAW: conceptualization; supervision; visualization; resources; project administration; writing—original draft preparation; writing—review and editing; funding acquisition. The author(s) read and approved the final manuscript.

Funding
This research was supported by startup funds from the School of Life Sciences and the Biodesign Institute at Arizona State University to MAW, School of Life Sciences Undergraduate Research (SOLUR) funding to SMB, IMSD funding to VAVV, and ARCS Spetzler Scholar funding to KCO. This study was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM124827 to MAW. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
The RNA-Seq datasets analyzed during the current study are available from the GTEx project through dbGaP under accession phs000424.v6.p1; we received approval to access this data under dbGaP accession #8834. All codes used are available on GitHub: https://github.com/SexChrLab/XY_RNAseq.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Life Sciences, Arizona State University, PO Box 874501, Tempe, AZ 85287-4501, USA. [2]Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85282, USA. [3]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA. [4]College of Osteopathic Medicine of the Pacific, Western University of Health Sciences, Pomona, CA 91766, USA. [5]Center for Mechanisms of Evolution, The Biodesign Institute, Arizona State University, Tempe, AZ 85282, USA.

Received: 5 February 2020 Accepted: 17 June 2020
Published online: 21 July 2020

References
1. Arnold AP, Chen X, Itoh Y. What a difference an X or Y makes: sex chromosomes, gene dose, and epigenetics in sexual differentiation. Handb. Exp. Pharmacol. 2012;67–88.
2. Khramtsova E, Davis L, Stranger B. The role of sex in the genomics of human complex traits. Nat Rev Genet. 2018;*20*.
3. Raznahan A, Parikshak NN, Chandran V, Blumenthal JD, Clasen LS, Alexander-Bloch AF, Zinn AR, Wangsa D, Wise J, Murphy DGM, et al. Sex-chromosome dosage effects on gene expression in humans. Proc Natl Acad Sci U S A. 2018;*115*:7398–403.
4. Traglia M, Bseiso D, Gusev A, Adviento B, Park DS, Mefford JA, Zaitlen N, Weiss LA. Genetic mechanisms leading to sex differences across common diseases and anthropometric traits. Genetics. 2017;*205*:979–92.
5. Gershoni M, Pietrokovski S. The landscape of sex-differential transcriptome and its consequent selection in human adults. BMC Biol. 2017;*15*:7.
6. Goldstein JM, Holsen L, Handa R, Tobet S. Fetal hormonal programming of sex differences in depression: linking women's mental health with sex differences in the brain across the lifespan. Front Neurosci. 2014;*8*.
7. Shi L, Zhang Z, Su B. Sex biased gene expression profiling of human brains at major developmental stages. Sci Rep. 2016;*6*:21181.
8. Arnold AP, Chen X. What does the "four core genotypes" mouse model tell us about sex differences in the brain and other tissues? Front Neuroendocrinol. 2009;*30*:1–9.
9. Charlesworth B. The evolution of sex chromosomes. Science. 1991;*251*:1030–3.
10. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science. 1999;*286*:964–7.
11. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al. The DNA sequence of the human X chromosome. Nature. 2005;*434*:325–37.
12. Pandey RS, Wilson Sayres MA, Azad RK. Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. Genome Biol Evol. 2013;*5*:1863–71.
13. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, et al. Ensembl 2017. Nucleic Acids Res. 2017;*45*:D635–42.
14. Charchar FJ, Svartman M, El-Mogharbel N, Ventura M, Kirby P, Matarazzo MR, Ciccodicola A, Rocchi M, D'Esposito M, Graves JAM. Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). Genome Res. 2003;*13*:281–6.
15. Veerappa AM, Padakannaya P, Ramachandra NB. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Funct Integr Genomics. 2013;*13*:285–93.
16. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;*423*:825–37.
17. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;*93*:641–51.
18. Webster TH, Couse M, Grande BM, Karlins E, Phung TN, Richmond PA, Whitford W, Wilson MA. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. GigaScience. 2019;*8*.
19. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;*348*:648–60.
20. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, KRM S, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet, 45. 2013:1113–20.
21. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;*501*:506–11.
22. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;*538*:201–6.
23. Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, et al. Landscape of X chromosome inactivation across human tissues. BioRxiv. 2016;073957.
24. Li R, Singh M. Sex Differences in Cognitive Impairment and Alzheimer's Disease. Front Neuroendocrinol. 2014;35(3):385–403 https://doi.org/10.1016/j.yfrne.2014.01.002.
25. de Perrot M, Licker M, Bouchardy C, Usel M, Robert J, Spiliopoulos A. Sex Differences in Presentation, Management, and Prognosis of Patients with Non-Small Cell Lung Carcinoma. J Thorac Cardiovasc Surg. 2000;119(1):21–6 https://doi.org/10.1016/s0022-5223(00)70213-3.
26. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, et al. Human Genomics. The Human Transcriptome across Tissues and Individuals. Science (New York, N.Y.). 2015;348(6235):660–5 https://doi.org/10.1126/science.aaa0355.
27. Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C, Roberts CT. Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. Front Genet. 2016;7:183 https://doi.org/10.3389/fgene.2016.00183.
28. Rahbari R, Zhang L, Kebebew E. Thyroid cancer gender disparity. Future Oncol Lond Engl. 2010;*6*:1771–9.
29. Natri HM, Wilson MA, Buetow KH. Distinct molecular etiologies of male and female hepatocellular carcinoma. BMC Cancer. 2019;*19*:951.
30. Naugler WE, Sakurai T, Kim S, Maeda S, Kim K, Elsharkawy AM, Karin M. Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. Science. 2007;*317*:121–4.
31. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;*12*:357–60.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;*29*:15–21.
33. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
34. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;*14*:417–9.
35. broadinstitute/picard (Broad Institute) 2020.
36. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res. 2012;*22*:1760–74 https://doi.org/10.1101/gr.135350.111.
37. Consortium T. Gte. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;*348*:648–60.
38. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software. June 2011;42(8):1–28. https://doi.org/10.18637/jss.v042.i08.
39. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011;*39*:D19–21.
40. Andrews S. Babraham Bioinformatics – FastQC: a quality control tool for high throughput sequence data; 2010.
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;*30*:2114–20.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. The

sequence alignment/map format and SAMtools. Bioinforma Oxf Engl. 2009; *25*:2078–9.
43. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;*27*:1691–2.
44. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma Oxf Engl. 2014;*30*:923–30.
45. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;*34*:3094–100.
46. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: an extended review and a software tool. PLoS One. 2017;*12*: e0190152.
47. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;*16*: 59–70.
48. Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. Genome Biol. 2014. https://doi.org/10.1186/s13059-014-0550-8.
49. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;*4*:1521.
50. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma Oxf Engl. 2010;*26*:139–40.
51. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;*10*:48.
52. Albrecht KH, Young M, Washburn LL, Eicher EM. Sry expression level and protein isoform differences play a role in abnormal testis development in C57BL/6J mice carrying certain Sry alleles. Genetics. 2003;*164*:277–88.
53. Turner ME, Ely D, Prokop J, Milsted A. Sry, more than testis determination? Am J Physiol-Regul Integr Comp Physiol. 2011;*301*:R561–71.
54. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005;*434*:400–4.
55. Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;*16*:97.
56. Eyras E, Caccamo M, Curwen V, Clamp M. ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res. 2004;*14*:976–87.
57. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;*17*:13.
58. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA N Y N. 2014;*20*:1684–96.
59. Lamarre S, Frasse P, Zouine M, Labourdette D, Sainderichin E, Hu G, Le Berre-Anton V, Bouzayen M, Maza E. Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. Front Plant Sci. 2018;*9*:108.
60. Zhao S, Li C-I, Guo Y, Sheng Q, Shyr Y. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. BMC Bioinformatics. 2018;*19*:191.
61. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;*18*:186.
62. Bray, N., Pimentel, H., Melsted, P., and Pachter, L. (2015). Near-optimal RNA-Seq quantification. ArXiv150502710 Cs Q-Bio.
63. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462–4.
64. Wu DC. Yao J, Ho KS, Lambowitz AM. Wilke CO. Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics. 2018;*19*:510.
65. Dumanski JP, Lambert J-C, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, Lindgren CM, Campion D, Dufouil C. European Alzheimer's Disease Initiative Investigators, et al. Mosaic loss of chromosome Y in blood is associated with Alzheimer disease. Am J Hum Genet. 2016;*98*:1208–19.
66. Grassmann F, Kiel C, den Hollander AI, Weeks DE, Lotery A, Cipriani V, Weber BHF. International Age-related Macular Degeneration Genomics Consortium (IAMDGC). Y chromosome mosaicism is associated with age-related macular degeneration. Eur J Hum Genet EJHG. 2019;*27*:36–41.
67. Forsberg LA. Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men. Hum Genet. 2017;*136*: 657–63.
68. Chaligné R, Popova T, Mendoza-Parra M-A, Saleem M-AM, Gentien D, Ban K, Piolot T, Leroy O, Mariani O, Gronemeyer H, et al. The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. Genome Res. 2015;*25*:488–503.

Publisher's Note