bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Full Title

"Aligning RNA-Seq reads to a sex chromosome complement informed reference genome increases ability to detect sex differences in gene expression"

Short Title

"*Sex chromosome complement informed alignment*"

Authors

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

Affiliations

1. School of Life Sciences,

2. Center for Evolution and Medicine,

3. Center for Mechanisms of Evolution, The Biodesign Institute

Arizona State University, Tempe AZ 85282 USA

4. Department of Genetics, University of North Carolina, Chapel Hill NC 27599 USA

*Corresponding author

Melissa A. Wilson

School of Life Sciences | Arizona State University | PO Box 874501 | Tempe, AZ 85287-4501

mwilsons@asu.edu

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Author contributions

¶ These authors contributed equally to this work

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

VVV: Investigation, Writing - Review and Editing

JA: Investigation, Writing - Review and Editing

MAW: Conceptualization, Supervision, Visualization, Resources, Project Administration, Writing - Original Draft Preparation, Writing - Review and Editing, Funding Acquisition

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Background: Human X and Y chromosomes share an evolutionary origin and, as a consequence, sequence similarity. We investigated whether sequence homology between the X and Y chromosomes affects alignment of RNA-Seq reads, and estimates of differential expression. We tested the effects of using reference genomes informed by the sex chromosome complement of the sample's genome on 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 transcript abundance in the brain cortex, and whole blood from three genetic female (46, XX) and three genetic male (46, XY) samples, using both HISAT and STAR read aligners. 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. 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.

Conclusions: We show that regardless of the choice of read aligner, using an alignment protocol informed by the sex chromosome complement of the sample results in higher expression estimates on the X chromosome in both genetic male and genetic female samples and an increased number of unique genes being called as differentially expressed between the sexes.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Key words: RNA-Seq, sex chromosomes, differential expression, transcriptome, mapping, alignment

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 brain cortex and whole blood of three genetic female (46, XX) and three 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 recommend future studies requiring aligning reads to a reference genome should consider the sex chromosome complement of their samples prior to running default pipelines.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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) 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 [10]. The sex chromosomes, X and Y, previously recombined along their entire lengths, but due to recombination suppression from Y chromosome-specific inversions [9,12], 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 [10,13,14] (Figure 1A). The PAR1 is a remnant of the XAR [10] 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] (Figure 1A). The XTR formed from an X chromosome to the Y chromosome duplication event following the human-chimpanzee divergence [10,16]. Thus, the evolution of the X and Y

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

105 chromosomes has resulted in a pair of chromosomes that are diverged, but still share some regions
106 of high sequence similarity.

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

123 Here, we tested the effect of sex chromosome complement informed read alignment to the
124 quantified levels of gene expression and the ability to detect sex-biased gene expression. We
125 utilized data from the GTEx project, focusing on two tissues, whole blood and brain cortex, which
126 are known to exhibit sex differences in gene expression [24–26]. Many genes have been reported
127 to be differentially expressed between male and female brain samples [5–7]. Differential

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

128 expression in blood samples between males and females has also been documented [5,6]. We used
129 brain cortex and whole blood tissues from three genetic male (46, XY) and three genetic female
130 (46, XX) individuals for a total of twelve samples evenly distributed among tissues and genetic
131 sex. We aligned all samples to a default reference genome that includes both the X and Y
132 chromosomes and to a reference genome that is informed on the sex chromosome complement of
133 the genome: Male XY samples were aligned to a reference genome that includes both the X and
134 Y chromosome where the Y chromosome PAR1 and PAR2 are hard-masked with Ns (Figure 1C)
135 so that reads will align uniquely to the X PAR sequences. Conversely, female XX samples were
136 aligned to a reference genome where the entirety of the Y chromosome is hard-masked (Figure
137 1C). We tested two different read aligners, HISAT [27] and STAR [28], to account for variation
138 between alignment methods and measured differential expression using Limma/Voom [29]. We
139 found that using a sex chromosome complement informed reference genome for aligning RNA-
140 Seq reads increased X chromosome expression estimates in both male XY and female XX samples
141 and uniquely identified differentially expressed genes.
142

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

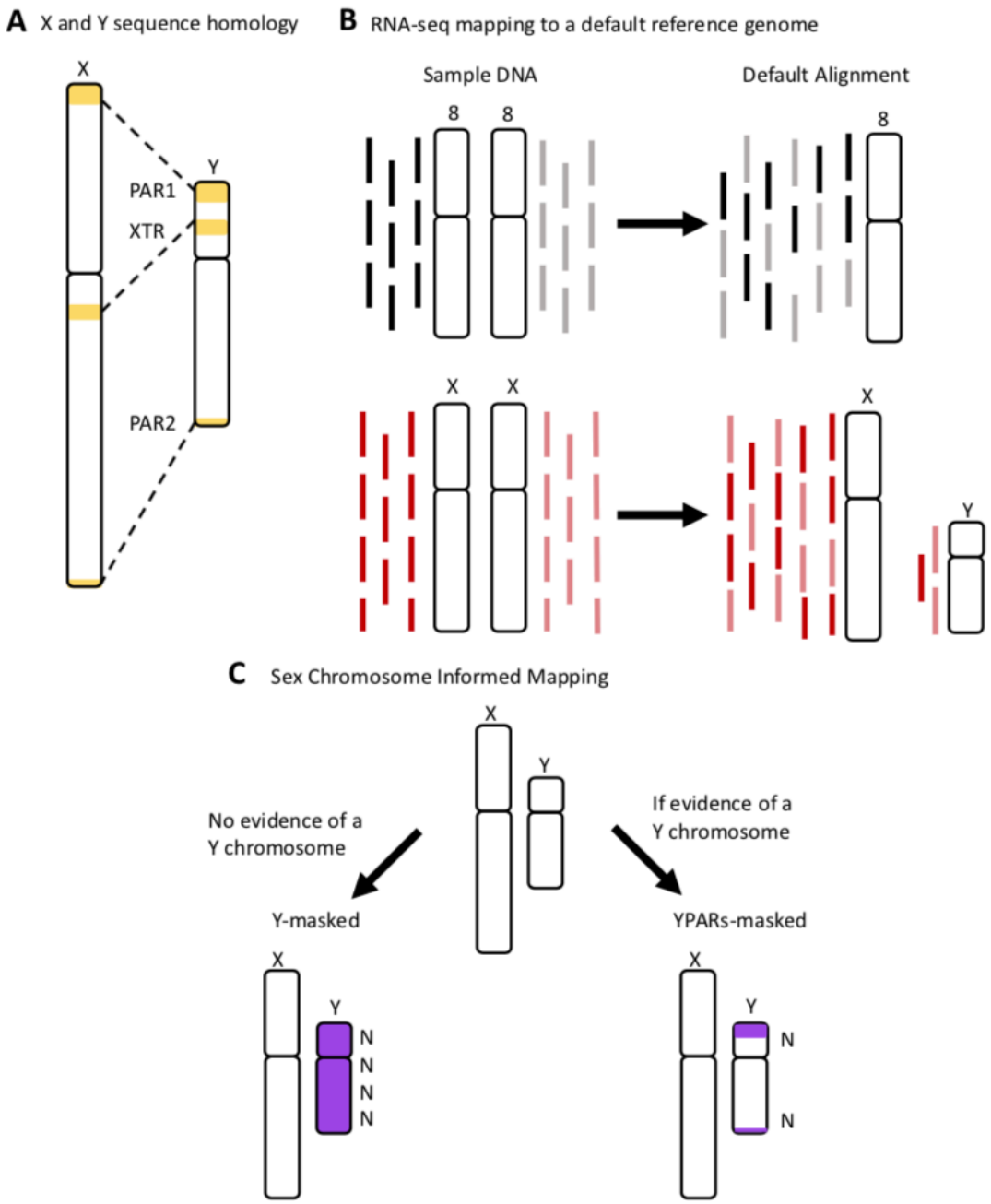


Figure 1. Homology between the human X and Y chromosomes where misaligning could occur. A) High sequence homology exists between the human X and Y chromosomes in three regions: 100% sequence identity for the pseudoautosomal regions (PARs), PAR1 and PAR2 and ~99% sequence homology in the X-transposed 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

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

149 Mb extending from X:155,701,383 to 156,030,895. The X chromosome PAR1 and PAR2 are
150 identical in sequence to the Y chromosome PAR1 Y:10,001 - 2,781,479 and PAR2 Y:56,887,903
151 - 57,217,415. **B)** Using a standard alignment approach will result in reads misaligning between
152 regions of high sequence homology on the sex chromosomes. **C)** Using a reference genome that
153 is informed by the genetic sex of the sample may help to reduce misaligning between the X and Y
154 chromosomes. In humans, samples without evidence of a Y chromosome should be aligned to a
155 Y-masked reference genome and samples with evidence of a Y should be aligned to a YPARs-
156 masked reference genome.
157
158 **Methods**
159 *Building sex chromosome complement informed reference genomes*
160 All GRCh38.p10 unmasked genomic DNA sequences, including autosomes 1-22, X, Y,
161 mitochondrial DNA (mtDNA), and contigs were downloaded from ensembl.org release 92 [13].
162 The default reference genome here includes all 22 autosomes, mtDNA, the X chromosome, the Y
163 chromosome, and contigs. For the two sex chromosome complement informed reference
164 assemblies, we included all 22 autosomes, mtDNA, and contigs from the default reference and a)
165 one with the Y chromosome either hard-masked for the "Y-masked reference genome" or b) one
166 with the pseudoautosomal regions, PAR1 and PAR2, hard-masked on the Y chromosome for
167 "YPARs-masked reference genome" (Figure 1C). Hard-masking with Ns will force reads to not
168 align to those masked regions in the genome. Masking the entire Y chromosome for the sex
169 chromosome complement informed reference genome, Y-masked, was accomplished by changing
170 all the Y chromosome nucleotides [ATGC] to Ns using sed command in linux. YPARs-masked
171 was created by hard-masking the Y PAR1: 6001-2699520 and the Y PAR2: 154931044-

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

155260560 regions. The GRCh38.p10 Y PAR1 and Y PAR2 chromosome start and end location was 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 YPARs-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 [27] hisat2-build -f option and STAR version 2.5.2 [28] using option --genomeDir and --sjdbGTFfile. Reference genome indexing was followed by picard tools version 1.119 CreateSequenceDictionary [30], which created a dictionary for each reference genome (Pipeline 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 brain cortex and whole blood tissues from 3 genetic female (46, XX) and 3 genetic male (46, XY) individuals [19,31] (Additional file 1). 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. 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, USP9Y, ZFX, ZFY, UTX, UTY, XIST, and SRY (Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

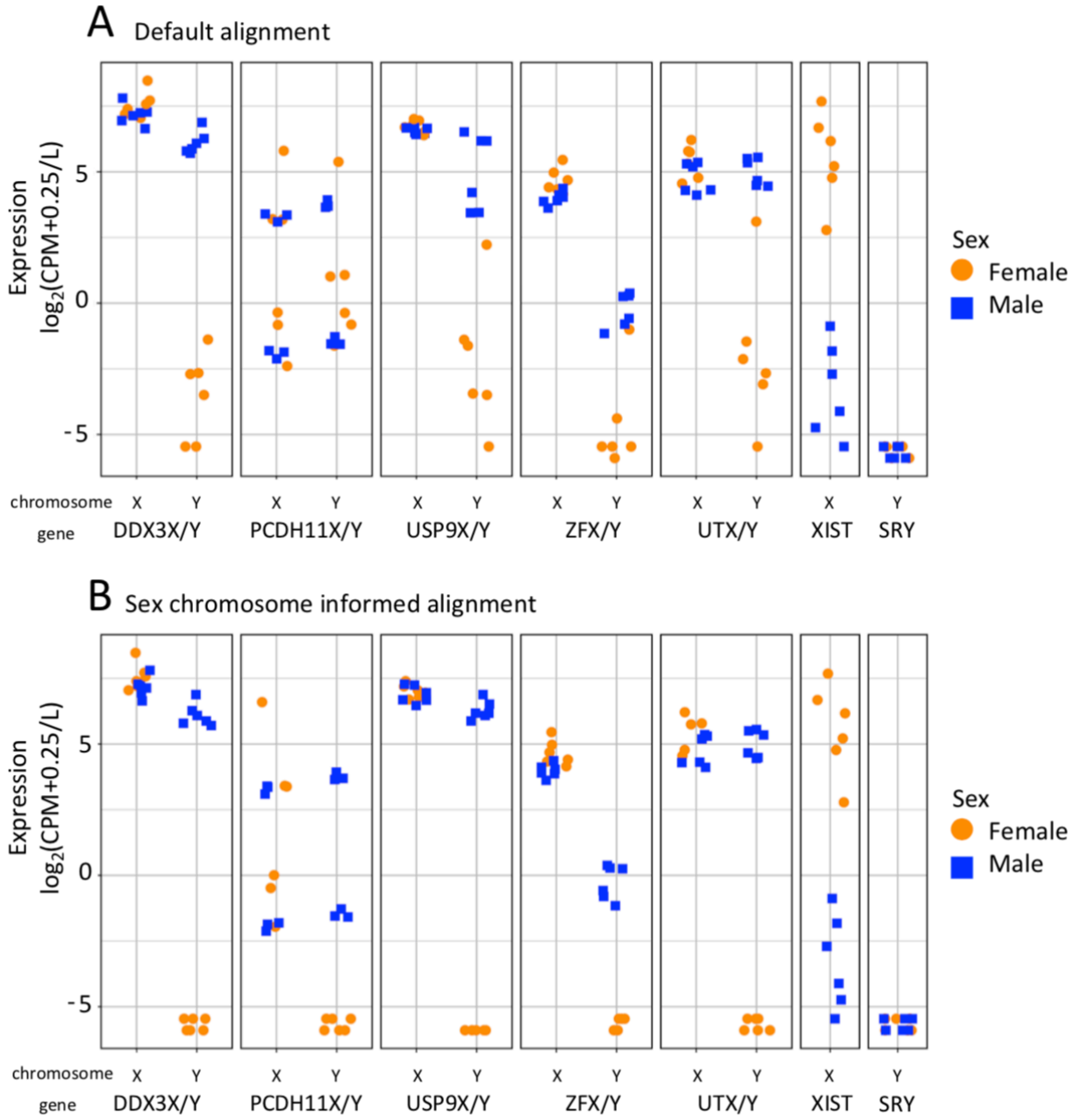


Figure 2. Genetic sex of RNA-Seq samples. We investigated gene expression, $\log_2$(CPM+0.25/L), of XY homologous genes, XIST, and SRY in all samples analyzed here from genetic males (blue squares) and genetic females (orange circles) **A)** when aligned to a default reference genome, and **B)** when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

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 [32]. Quality of the samples' raw sequencing reads was examined using FastQC [33] and MultiQC [34]. Subsequently, adapter sequences were removed using Trimmomatic version 0.36 [35]. 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 genome and reads from male (46, XY) individuals were aligned to the YPARs-masked reference genome. Read alignment was performed using HISAT version 2.1.0 [27], 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 [28], 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).

Processing of RNA-Seq alignment files

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Aligned RNA-Seq samples from HISAT and STAR were output in Sequence Alignment Map (SAM) format and converted to Binary Alignment Map (BAM) using bamtools version 2.4.0 [36]. Summaries on the BAM files, including the number of reads mapped, duplicate reads and unaligned reads were computed using bamtools version 2.4.0 package [36] (Additional file 2). RNA-Seq BAM files were indexed, sorted, duplicates were marked, and read groups added using bamtools, samtools, and Picard [36–38]. All RNA-Seq BAM files were indexed using the default reference genome using Picard ReorderSam [38], 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 will require that all samples have the same chromosomes, this is why we hard-mask rather than delete. 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 alignment.

Gene expression level quantification

Read counts for each gene across all autosomes, sex chromosomes, mtDNA, and contigs were generated using featureCounts [39] 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 2,283 genes annotated on the X chromosome and a total of 56,571 genes

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

245 across the entire genome for GRCh38 version of the human reference genome [13]. Only primary
246 alignments were counted and specified using the --primary option in featureCounts.
247
248 *Differential expression*
249 Differential expression analysis was performed using the limma/voom pipeline [29,40] which has
250 been shown to be a robust differential expression software package [41,42]. Quantified read counts
251 from each sample generated from featureCounts were combined into a count matrix, each row
252 representing a unique gene id and each column representing the gene counts for each unique
253 sample. This was repeated for each tissue type, brain cortex and whole blood, and was read into R
254 using the DGEList function in the R limma package [29,40]. A sample-level information file
255 related to the genetic sex of the sample, male or female, and the reference genome used for
256 alignment, default or sex chromosome complement informed, was created and corresponds to the
257 columns of the count matrix described above.
258 Using edgeR [43], raw counts were normalized to adjust for compositional differences
259 between the RNA-Seq libraries using the voom normalize quantile function which normalizes the
260 reads by the method of trimmed mean of values (TMM) [29]. Counts were then transformed to
261 $\log_2$(CPM+0.25/L), where CPM is counts per million, L is library size, and 0.25 is a prior count to
262 avoid taking the log of zero [43]. For this dataset, the average library size is about 15 million,
263 therefore L is ~15. Thus, the minimum $\log_2$(CPM+0.25/L) value for each sample, representing
264 zero transcripts, is $\log_2$(0+0.25/15) = -5.90.
265 A minimum of 1 CPM, or the equivalent of 0 in $\log_2$(CPM+2/L), in at least 3 samples per
266 comparison was required for the gene to be kept for downstream analysis. A CPM value of 1 was
267 used in our analysis to separate expressed genes from unexpressed genes, meaning that in a library

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

size of ~15 million reads, there are at least 15 counts in that sample. After filtering for a minimum CPM, 22,695 out of the 56,571 quantified genes were retained for the brain samples and 14,944 for whole blood, as differential expression between the sexes was run separately for each tissue. A linear model was fitted to the DGEList-object, which contains the normalized gene counts for each sample, using the limma lmfit function which will fit a separate model to the expression values for each gene [29].

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 brain cortex and whole blood for contrasts of pairwise comparisons between the sexes and between reference genomes used for alignment. Pairwise contrasts were generated using limma makecontrasts function [29]. We identified genes that exhibited significant expression differences defined using an adjusted p-value cutoff that is less than 0.05 (5%) to account for multiple testing in pairwise comparisons between conditions using limma/voom decideTests vebayesfit [29,40] (Pipeline available on GitHub, https://github.com/SexChrLab/XY_RNAseq).

GO analysis

We examined 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 (Additional file 5). We investigated gene ontology enrichment for lists of genes that were identified as showing overexpression in one sex versus the other sex for brain cortex and

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

whole blood samples (adjusted p-value < 0.05). We used the GOrilla webtool, which utilizes a hypergeometric distribution to identify enriched GO terms [44]. A modified Fisher exact p-value cutoff < 0.001 was used to select significantly enriched terms [44].

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 for when reads were aligned to a reference genome informed on the sex chromosome complement (Additional file 2). Reads mapped across the whole genome, excluding the sex chromosomes, either stayed the same or increased slightly when samples were aligned to a reference genome informed on the sex chromosome complement. 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 it is of similar size to chromosome X. Reads aligning to chromosome 8 didn't vary tremendously or at all between aligning to the default versus sex chromosome complement informed reference for both brain cortex and whole blood tissues (1,479,544 versus 1,479,584 average number of reads mapped in default versus sex chromosome complement informed, respectively when aligned using HISAT; Additional file 2). A female XX brain cortex sample showed the largest chromosome 8 mapped read increase (334 more reads with HISAT; 608 more reads with STAR) when aligned to a sex chromosome complement informed reference genome using either aligner (Additional file 2). The sample that showed the highest decrease in mapped reads was a male XY brain cortex sample, which showed 5 fewer reads on chromosome 8 when aligned to a reference genome informed on the sex chromosome complement using HISAT read aligner. There was no decrease in chromosome 8 mapped reads when aligned

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

to a sex chromosome complement informed reference genome using STAR read aligner for male or female whole blood samples. Overall, we observed little difference in chromosome 8 mapped reads but were interested in how reads changed on the sex chromosomes, X and Y, when aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome.

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

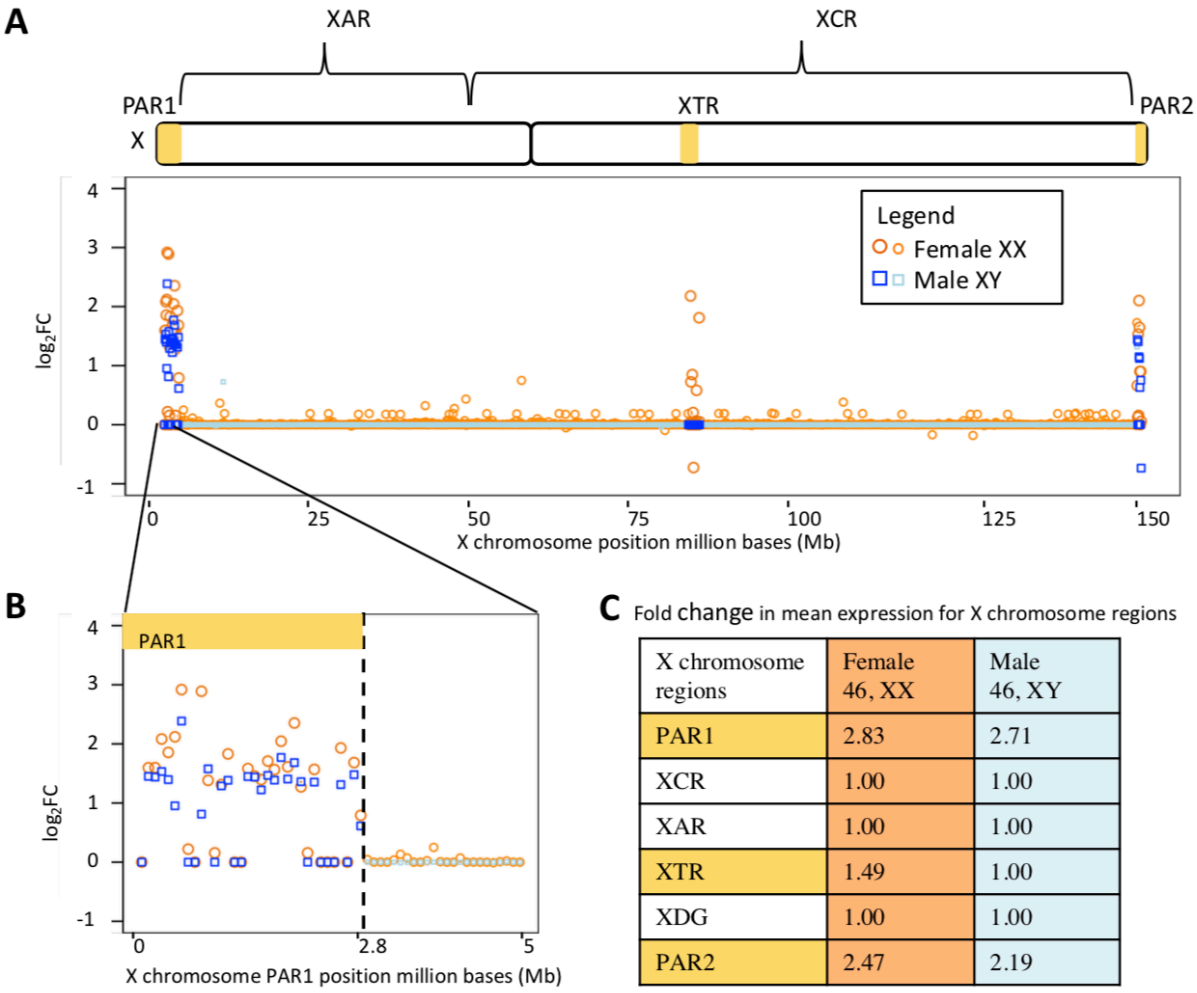
We found that when reads were aligned to a reference genome informed by the sex chromosome complement for both male XY and female XX brain cortex and whole blood samples, reads on the X chromosome increased by ~0.13% (1,627 increase in reads out of an average 1,238,463 reads for chromosome X) when aligned using HISAT. Reads on the Y chromosome decreased 100% (46,852 reads on average) in female XX samples and by ~63.43% (52,327 reads on average) in male XY samples for brain cortex and whole blood when aligned using HISAT (Additional file 2). Similar increases in X chromosome and decreases in Y chromosome reads when aligned to a sex chromosome complement informed reference was observed for when STAR was used as the read aligner for both male XY and female XX brain cortex and whole blood samples (Additional file 2). While we observed tens of thousands of reads aligning to new locations when sex chromosome complement was taken into account, we also investigated the effect on expression estimates.

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

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We explored the effect of changes in read alignment on gene expression. There is an increase in X chromosome PAR1 and PAR2 expression when reads were aligned to a reference genome informed on the sex chromosome complement and this is true for both male XY and female XX samples from the brain cortex (Figure 3) and whole blood samples using either HISAT or STAR as the read aligner (Additional file 6). We found an average of 2.83 fold increase in expression (1.21 $\log_2$ fold increase) in PAR1 expression for female XX brain cortex samples and 2.71 fold increase in expression (0.95 $\log_2$ fold change increase) in PAR1 for male XY brain cortex samples using HISAT read aligner (Figure 3, Additional file 6, Additional file 7). XTR in female XX brain cortex samples showed a 1.49 fold increase in expression (0.22 $\log_2$ fold increase) and no change in male XY brain cortex samples. PAR2 showed an average of 2.47 fold increase (0.81 $\log_2$ fold change increase) for female XX brain cortex samples and 2.19 fold increase (0.58 $\log_2$ fold change increase) in PAR2 for male XY brain cortex samples using HISAT read aligner with similar results for STAR read aligner (Figure 3, Additional file 6, Additional file 7). Complete lists of the $\log_2$(CPM+0.25/L) values for each X chromosomal gene and each gene within the whole genome for male XY and female XX brain cortex and whole blood samples for when reads were aligned using the different read aligners and reference genomes are in Additional file 3 and Additional file 4.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



C Fold change in mean expression for X chromosome regions

X chromosome regions	Female 46, XX	Male 46, XY
PAR1	2.83	2.71
XCR	1.00	1.00
XAR	1.00	1.00
XTR	1.49	1.00
XDG	1.00	1.00
PAR2	2.47	2.19

Figure 3. X chromosome RNA-Seq alignment differences. We plot $\log_2$ fold change (FC) across **A)** the entire X chromosome and **B)** the first 5 million bases (Mb) and show **C)** average fold change in large genomic regions on the X chromosome between 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

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

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

Intriguingly, regions outside the PARs on the X chromosome, and across the genome, showed little to no increase in expression when aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome (Additional file 7). X and Y homologous genes showed little to no increase in expression when aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome with the exception of PCDH11X gene (Additional file 8). PCDH11X showed a 1.50 fold increase (0.58 $\log_2$ fold increase) in female XX brain cortex samples with a similar increase, 1.47 fold, (0.55 $\log_2$ fold increase) in female XX whole blood samples using HISAT read aligner. A similar increase in expression for PCDH11X was observed for female XX brain cortex and whole blood samples when STAR was used as the read aligner (Additional file 8). PCDH11X showed no differences in expression between reference genomes for male XY brain cortex and whole blood samples when using either HISAT or STAR (Additional file 8). With noticeable increases in X chromosome gene expression and decreases in Y chromosome expression when aligned to a sex chromosome complement informed reference genome, we next investigated how this would affect gene differential expression between the sexes.

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

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

386 Generally, when comparing gene expression differences between the sexes, we find more genes
387 are differentially expressed on the sex chromosomes, and fewer, or the same, are differentially
388 expressed on the autosomes when taking sex chromosome complement into account. At an
389 adjusted p-value of 0.05 and aligning with HISAT, we find 12 new genes (all on the Y
390 chromosome) that are only called as differentially expressed between the sexes in the brain cortex
391 when aligned to reference genomes informed on the sex chromosome complement (Figure 4).
392 Additionally, we find 49 genes (48 autosomal and 1 X-linked) that are called as differentially
393 expressed when using a default reference genome for aligning reads that are no longer called as
394 differentially expressed between the sexes in the brain cortex when aligning to a reference genome
395 informed by sex chromosome complement (Figure 4; Additional file 9). We observed similar
396 trends in changes for differential expression between male XY and female XX whole blood
397 samples using either HISAT or STAR as the aligner (Additional file 9, Additional file 10).
398

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

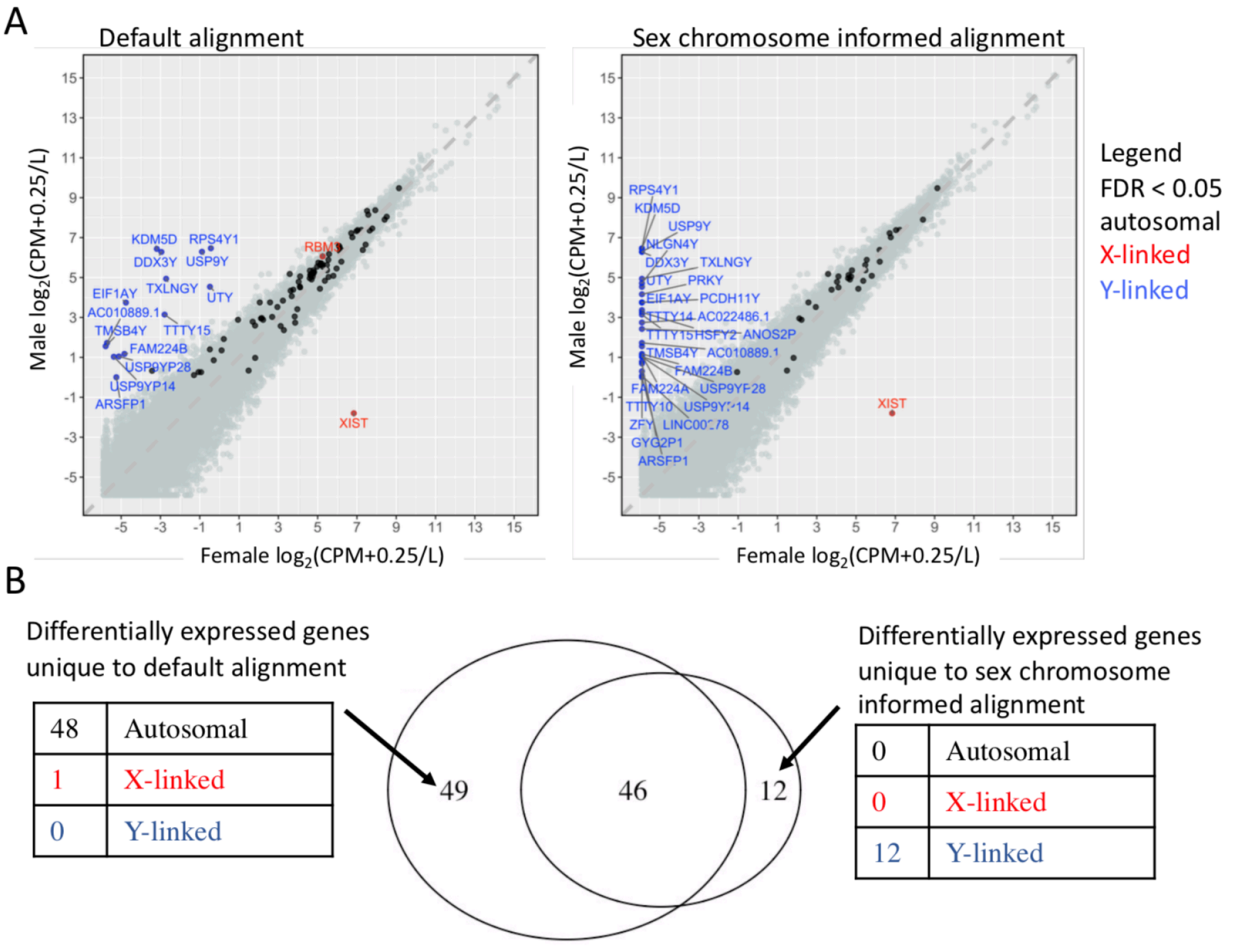


Figure 4. Sex chromosome complement informed alignment calls more sex-linked genes as being differentially expressed. A) Sex differences in gene expression, $\log_2$(CPM+0.25/L), between the three samples from genetic males and females are shown when aligning all samples to the default reference genome (left) and a reference genome informed on the sex chromosome complement (right). Each point represents a gene. Genes that are differentially expressed, adjusted p-value < 0.05 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 aligned to the default genome, and genes that are called as differentially expressed when aligned to a sex chromosome complement informed genome. When aligned to a default reference genome there were 95 genes that were differentially expressed between male XY and

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

410 female XX brain cortex samples. Of these 95 genes, 49 genes were uniquely called as differentially
411 expressed when aligned to the default reference genome but were not called as differentially
412 expressed when aligned to a sex chromosome complement informed reference genome. Of the 49
413 genes, 48 are autosomal and 1 is X-linked, RBM3. When samples were aligned to a reference
414 genome informed on the sex chromosome complement 58 genes were called as differentially
415 expressed between the sexes of which 12 were uniquely called in the sex chromosome complement
416 informed alignment. All 12 uniquely called differentially expressed genes when samples were
417 aligned to a sex chromosome complement informed reference genome are Y-linked.
418
419 *Increased concordance between aligners when informed by sex chromosome complement*
420 When using a default reference genome, and calling genes as differentially expressed between the
421 sexes, of all genes when aligned using either HISAT or STAR as the aligner, only 46% of genes
422 called as differentially expressed overlapped between the aligners, HISAT and STAR for brain
423 cortex. There is 67% concordance in genes being called as differentially expressed for whole blood
424 between aligners, HISAT and STAR, when samples were aligned to a default reference genome.
425 In contrast, when aligning samples to a reference genome informed by the sex chromosome
426 complement, we find increased concordance in all tissues for the set of genes called as
427 differentially expressed between the sexes when aligned using either HISAT or STAR (Additional
428 file 9).
429
430 *Using sex-linked genes alone is inefficient for determining the sex chromosome complement of a*
431 *sample*

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The sex of each sample 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), and XIST and SRY gene expression in male XY and female XX brain cortex and whole blood (Figure 2). Both males XY and females XX 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 male XY and female XX brain cortex and whole blood 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 and XIST gene expression (Figure 2). Male XY samples showed expression for DDX3X and DDX3Y (greater than 5 $\log_2$(CPM+2/L). Female XX samples showed expression for XIST (greater than 2.5 $\log_2$(CPM+2/L) and male XY samples showed little to no expression for XIST (less than 0 $\log_2$(CPM+2/L). In contrasts to the default reference genome, when aligned to a sex chromosome complement informed reference genome, samples cluster distinctly by sex for DDX3Y, PCDH11Y, USP9Y, ZFY, UTY, and XIST all showing at least a 2.5 $\log_2$(CPM+2/L) difference between the sexes (Figure 2).

Discussion

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 nothing masked [45]. Neither Ensembl or Gencode human reference genomes are correct for aligning both XX and XY samples. The sex

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

chromosome complement of the sample should be taken into account when aligning RNA-Seq reads to reduce misaligning sequences.

Measurements of X chromosome expression increase for both male XY and female XX brain cortex and whole blood samples when aligned to a sex chromosome complement informed reference genome versus aligning to a default reference genome (Figure 3). There was a minimum 1.6 fold increase in expression for the genes in PAR1 and PAR2 for all samples (male and female) when aligned to a sex chromosome complement informed reference genome (Figure 3; Additional file 7). We see that the XTR has a minimum 1.35 fold increase in expression for female brain cortex and whole blood samples, when aligned to a sex chromosome complement informed reference genome compared to a default reference genome, but no change in males. This is because XTR is not hard-masked in the YPARs-masked reference genome, which is used to align male XY samples. The XTR shares 98.78% homology between X and Y but no longer recombines between X and Y [15] (Figure 1A) and is therefore, because of this divergence, not hard-masked when aligning male XY samples. A comprehensive table for mean expression values for regions of the X chromosome and X chromosome gene expression in default and sex chromosome complement informed alignment is in additional file 7 and additional file 3, respectively. Given striking increases in measurements of X chromosome expression, we further investigated the effect of sex chromosome complement informed alignment on differential gene expression between the sexes.

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 (Figure 2; Figure 4). However, when aligned to a reference genome informed by the sex chromosome complement, female XX samples no longer showed Y-linked gene expression,

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and more Y-linked genes were called as being differentially expressed between the sexes. 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 (Figure 4). Furthermore, these Y-linked genes serve in various important biological processes, thus altering the functional interpretation of the sex differences. GO enrichment analysis of genes that are more highly expressed in brain cortex male samples than females, when samples were aligned to a default reference genome, were genes involved in germline cell cycle switching and mitotic to meiotic cell cycle (Additional file 5). However, when these samples were aligned to a sex chromosome complement informed reference genome, genes upregulated in males were enriched for positive regulation of transcription from RNA polymerase II promoter in response to heat stress and GO component of specific granule lumen, which are secretory vesicles of the immune system [44].

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 [42,46]. When using a default reference genome for alignment we find discordance between HISAT and STAR in which genes called as differentially expressed between the sexes. However, we find increased concordance between HISAT and STAR, when using a sex chromosome complement informed reference genome (Additional file 9). Differences between HISAT and STAR could be contributed to differences in default parameters for handling multi-aligning reads [27]. We show here that the choice of read aligner, HISAT and STAR, will have less variance on differential expression results if a sex chromosome complement informed reference genome is used for aligning RNA-Seq reads.

Ideally, one would use DNA to confirm presence or absence of the Y chromosome, but if DNA sequence was not generated, one would need to determine the sex of the sample by assessing expression estimates for Y-linked genes. To more carefully investigate the ability to use gene

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (Figure 2, Additional file 11). SRY is predominantly expressed in the testis [47,48] 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 (Figure 2, Additional file 11). 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 brain cortex and whole blood samples. XIST expression is important in the X chromosome inactivation process [49] 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 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, so only presence/absence of Y-linked reads alone is insufficient to determine sex chromosome complement of the sample (Figure 2, Additional file 11). That said, 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

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

523 increasingly important as karyotypically XY individuals are known to have lost the Y chromosome
524 in particular tissues sampled, as shown in Alzheimer Disease [50], age-related macular
525 degeneration [51], and in the blood of aging individuals [52]. Self-reported sex may not match the
526 sex chromosome complement of the samples, even in karyotypic individuals.
527
528 **Conclusion**
529 Here we show that aligning RNA-Seq reads to a sex chromosome complement informed reference
530 genome will change the results of the analysis compared to aligning reads to a default reference
531 genome. We have previously observed that a sex chromosome complement informed alignment is
532 important for DNA as well [53]. A sex chromosome complement informed approach is needed for
533 a sensitive and specific analysis of gene expression on the sex chromosomes [1]. A sex
534 chromosome complement informed reference alignment resulted in increased X chromosome
535 expression for both male XY and female XX samples. We further found different genes called as
536 differentially expressed between the sexes, and identified sex differences in gene pathways that
537 were missed when samples were aligned to a default reference genome. We additionally identified
538 that there is greater concordance between read aligners, HISAT and STAR, in the genes that are
539 called as differentially expressed between the sexes when samples were aligned to a sex
540 chromosome complement informed reference genome. The accurate alignment of the short RNA-
541 Seq reads to the reference genome is essential to drawing reliable conclusions from differential
542 expression data analysis on the sex chromosomes. We strongly urge future human RNA-Seq
543 analysis to carefully consider the genetic sex of the sample when aligning reads, and have provided
544 a framework for doing so (https://github.com/SexChrLab/XY_RNAseq).
545

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Acknowledgments

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 VVV, ARCS 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. We thank Heini Natri and Angela Taravella for comments on the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

554 **Additional files**

555 **Additional file 1. Sample IDs.** RNA-Seq brain cortex and whole blood tissue samples from 3
556 genetic female (46, XX) and 3 genetic male (46, XY) individuals were downloaded from the
557 Genotype-Tissue Expression (GTEx) project [19,31] for a total of 12 RNAseq tissue samples.
558
559 **Additional file 2. Table of mapped read statistics for default and sex chromosome**
560 **complement informed alignment.** Total reads mapped in brain cortex and whole blood female
561 XX and male XY samples for when reads were aligned to a default reference genome and for when
562 reads were aligned to a reference genome informed on the sex chromosome complement. Total
563 mapped reads for HISAT and STAR as the read aligner. The difference in reads mapped between
564 sex chromosome complement informed and default alignment is shown to the right of the mapped
565 reads statistics for each tissue and aligner used. Chromosome Y and chromosome X show the
566 highest degree of difference between default and sex chromosome complement informed
567 alignment were chromosome Y always decreased in total mapped reads and chromosome X always
568 increased in total mapped reads when using a sex chromosome complement informed alignment.
569
570 **Additional file 3. X chromosome gene expression values per sample, aligner and reference**
571 **genome used for alignment.** CPM values for male XY and female XX brain cortex and whole
572 blood samples when aligned to a default and sex chromosome complement informed reference
573 genome for chromosome X. A text format of CPM values are available on GitHub,
574 https://github.com/SexChrLab/XY_RNAseq.
575
576 **Additional file 4. Whole genome gene expression values per sample, aligner and reference**

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

577 **genome used for alignment.** CPM values for male XY and female XX brain cortex and whole
578 blood samples when aligned to a default and sex chromosome complement informed reference
579 genome for the whole genome (1-22, MT, X, Y and non-chromosomal). A text format of CPM
580 values are available on GitHub, https://github.com/SexChrLab/XY_RNAseq.
581
582 **Additional file 5.** Gene enrichment analysis of genes that are more highly expressed in one sex
583 verses the other sex for when samples were aligned to a default or sex chromosome complement
584 informed reference genome using either HISAT or STAR.
585
586 **Additional file 6. X chromosome expression differences between default and sex chromosome**
587 **complement informed alignment.** X chromosome gene expression differences between default
588 and sex chromosome complement informed alignment. Increase in expression when aligned to a
589 sex chromosome complement informed reference genome is a $\log_2$ fold change (FC) > 0. A
590 decrease in expression when aligned to a sex chromosome complement informed reference
591 genome is $\log_2$ FC < 0. Female XX samples are indicated by red and pink circles for PAR1, XTR,
592 and PAR2 genes and for all other X chromosome genes respectively. Blue and light blue squares
593 represent male XY samples. Blue squares indicate which gene points are in PAR1, XTR, and
594 PAR2 and light blue squares are for genes outside of those regions. Differences in X chromosome
595 expression between reference genomes for male XY and female XX samples aligned using HISAT
596 for the whole X chromosome and the first 5Mb are shown for the brain cortex (**A** and **B**,
597 respectively), and the whole blood (**C** and **D**, respectively). Differences in X chromosome
598 expression between reference genomes for male XY and female XX samples aligned using STAR
599 for the whole X chromosome and the first 5Mb are shown for the brain cortex (**E** and **F,**

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

600 respectively), and the whole blood (**G** and **H**, respectively).

601

602 **Additional file 7. X chromosome regions mean and median expression values.** X chromosome
603 regions PAR1, PAR2, XTR, XDG, XAR, XCR mean and median CPM expression for male XY
604 and female XX brain cortex and whole blood samples when aligned to a default or sex chromosome
605 complement informed reference genome for HISAT and STAR.

606

607 **Additional file 8. Gene expression for XY homologous genes.** X chromosome expression for 26
608 X and Y homologous genes. Difference in gene expression for when male XY and female XX
609 brain cortex and whole blood samples were aligned to a default and sex chromosome complement
610 informed reference genome. Little to no difference in gene expression between default and sex
611 chromosome complement informed reference genome alignment was observed for 25 of the 26 X
612 and Y homologous genes for both male XY and female XX brain cortex and whole blood samples
613 using either HISAT or STAR. PCDH11X showed a 1.50 and 1.47 fold increase in expression for
614 brain cortex and whole blood, respectively in female XX samples using HISAT read aligner with
615 similar results for STAR. XY male brain cortex and whole blood samples showed little to no
616 differences in gene expression between reference genomes for the 26 X and Y homologous genes
617 using either HISAT or STAR.

618

619 **Additional file 9. Differentially expressed genes between the sexes that were uniquely and**
620 **jointly called between reference genomes.** Genes that are differentially expressed between the
621 sexes, male XY and female XX, for brain cortex and whole blood samples. Differentially
622 expressed genes that are uniquely called when using either the default or sex chromosome

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

623 complement informed reference genome and differentially expressed genes that jointly called
624 between the reference genomes.
625
626 **Additional file 10. Gene expression differences between male XY and female XX samples.**
627 Sex differences in gene expression for brain cortex and whole blood samples for when samples
628 were aligned to a default reference genome and a to a reference genome informed on the sex
629 chromosome complement. Showing sex differences in gene expression between reference
630 genomes used for alignment and for when samples were aligned using HISAT and STAR.
631
632 **Additional file 11. Genetic sex of RNA-Seq samples.** Gene expression $\log_2$(CPM+0.25/L) for
633 select XY homologous genes and XIST and SRY for when reads were aligned to a default
634 reference genome **A)** and **C)** using HISAT and STAR, respectively and for when reads were
635 aligned to a sex chromosome complement informed reference genome **B)** and **D)** using HISAT
636 and STAR, respectively. Male XY brain cortex and whole blood samples are shown in blue squares
637 and female XX brain and blood samples shown in red circles.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

638 **References**

639 1. Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human
640 complex traits. Nat Rev Genet. 2019;20: 173–190.

641 2. Arnold AP, Chen X, Itoh Y. What a Difference an X or Y Makes: Sex Chromosomes, Gene
642 Dose, and Epigenetics in Sexual Differentiation. Handbook of Experimental Pharmacology.
643 2012. pp. 67–88.

644 3. Traglia M, Bseiso D, Gusev A, Adviento B, Park DS, Mefford JA, et al. Genetic
645 Mechanisms Leading to Sex Differences Across Common Diseases and Anthropometric
646 Traits. Genetics. 2017;205: 979–992.

647 4. Raznahan A, Parikshak NN, Chandran V, Blumenthal JD, Clasen LS, Alexander-Bloch AF,
648 et al. Sex-chromosome dosage effects on gene expression in humans. Proc Natl Acad Sci U
649 S A. 2018;115: 7398–7403.

650 5. Goldstein JM, Holsen L, Handa R, Tobet S. Fetal hormonal programming of sex differences
651 in depression: linking women's mental health with sex differences in the brain across the
652 lifespan. Front Neurosci. 2014;8: 247.

653 6. Gershoni M, Pietrokovski S. The landscape of sex-differential transcriptome and its
654 consequent selection in human adults. BMC Biol. 2017;15: 7.

655 7. Shi L, Zhang Z, Su B. Sex Biased Gene Expression Profiling of Human Brains at Major
656 Developmental Stages. Sci Rep. 2016;6. doi:10.1038/srep21181

657 8. Arnold AP, Chen X. What does the "four core genotypes" mouse model tell us about sex

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

658 differences in the brain and other tissues? Front Neuroendocrinol. 2009;30: 1–9.

659 9. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science.
660 1999;286: 964–967.

661 10. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, et al. The DNA
662 sequence of the human X chromosome. Nature. 2005;434: 325–337.

663 11. Charlesworth B. The evolution of sex chromosomes. Science. 1991;251: 1030–1033.

664 12. Pandey RS, Wilson Sayres MA, Azad RK. Detecting evolutionary strata on the human x
665 chromosome in the absence of gametologous y-linked sequences. Genome Biol Evol.
666 2013;5: 1863–1871.

667 13. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017.
668 Nucleic Acids Res. 2017;45: D635–D642.

669 14. Charchar FJ, Svartman M, El-Mogharbel N, Ventura M, Kirby P, Matarazzo MR, et al.
670 Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). Genome
671 Res. 2003;13: 281–286.

672 15. Veerappa AM, Padakannaya P, Ramachandra NB. Copy number variation-based
673 polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-
674 transposed region (XTR) in the Y chromosome. Funct Integr Genomics. 2013;13: 285–293.

675 16. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The
676 male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.
677 Nature. 2003;423: 825–837.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

678 17. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq
679 data. Am J Hum Genet. 2013;93: 641–651.

680 18. Webster TH, Couse M, Grande BM, Karlins E, Phung TN, Richmond PA, et al. Identifying,
681 understanding, and correcting technical biases on the sex chromosomes in next-generation
682 sequencing data [Internet]. doi:10.1101/346940

683 19. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot
684 analysis: multitissue gene regulation in humans. Science. 2015;348: 648–660.

685 20. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw
686 KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat
687 Genet. 2013;45: 1113–1120.

688 21. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, et al.
689 Transcriptome and genome sequencing uncovers functional variation in humans. Nature.
690 2013;501: 506–511.

691 22. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome
692 Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538: 201–206.

693 23. Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, et al. Landscape of X
694 chromosome inactivation across human tissues. Nature. 2017;550: 244–248.

695 24. Li R, Singh M. Sex differences in cognitive impairment and Alzheimer’s disease. Front
696 Neuroendocrinol. 2014;35: 385–403.

697 25. de Perrot M, Licker M, Bouchardy C, Usel M, Robert J, Spiliopoulos A. Sex differences in

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

698 presentation, management, and prognosis of patients with non-small cell lung carcinoma. J
699 Thorac Cardiovasc Surg. 2000;119: 21–26.

700 26. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human
701 genomics. The human transcriptome across tissues and individuals. Science. 2015;348:
702 660–665.

703 27. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory
704 requirements. Nat Methods. 2015;12: 357–360.

705 28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
706 universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

707 29. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model
708 analysis tools for RNA-seq read counts. Genome Biol. 2014;15: R29.

709 30. Picard Toolkit [Internet]. Broad Institute, GitHub Repository; 2018. Available:
710 http://broadinstitute.github.io/picard/

711 31. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:
712 580–585.

713 32. Sequence Read Archive. Downloading SRA data using command line utilities. National
714 Center for Biotechnology Information (US); 2011; Available:
715 https://www.ncbi.nlm.nih.gov/books/NBK158899/

716 33. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence
717 Data [Internet]. [cited 11 Sep 2017]. Available:

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

718 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

719 34. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for
720 multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048.

721 35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
722 data. Bioinformatics. 2014;30: 2114–2120.

723 36. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API
724 and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27: 1691–1692.

725 37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
726 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

727 38. Picard Tools - By Broad Institute [Internet]. [cited 7 May 2019]. Available:
728 http://broadinstitute.github.io/picard/

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

731 40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
732 RNA-seq data with DESeq2 [Internet]. 2014. doi:10.1101/002832

733 41. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting
734 differential expression in RNA-seq studies. Brief Bioinform. 2015;16: 59–70.

735 42. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An
736 extended review and a software tool. PLoS One. 2017;12: e0190152.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

737 43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
738 expression analysis of digital gene expression data [Internet]. Bioinformatics. 2010. pp.
739 139–140. doi:10.1093/bioinformatics/btp616

740 44. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and
741 visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10: 48.

742 45. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al.
743 GENCODE: the reference human genome annotation for The ENCODE Project. Genome
744 Res. 2012;22: 1760–1774.

745 46. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A
746 survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17: 13.

747 47. Albrecht KH, Young M, Washburn LL, Eicher EM. Sry expression level and protein
748 isoform differences play a role in abnormal testis development in C57BL/6J mice carrying
749 certain Sry alleles. Genetics. 2003;164: 277–288.

750 48. Turner ME, Ely D, Prokop J, Milsted A. Sry, more than testis determination? [Internet].
751 American Journal of Physiology-Regulatory, Integrative and Comparative Physiology.
752 2011. pp. R561–R571. doi:10.1152/ajpregu.00645.2010

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

755 50. Dumanski JP, Lambert J-C, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, et al. Mosaic
756 Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. Am J Hum Genet.

bioRxiv preprint doi: https://doi.org/10.1101/668376; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2016;98: 1208–1219.

51. Grassmann F, Kiel C, den Hollander AI, Weeks DE, Lotery A, Cipriani V, et al. Y chromosome mosaicism is associated with age-related macular degeneration. Eur J Hum Genet. 2019;27: 36–41.

52. 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: 657–663.

53. Webster TH, Couse M, Grande BM, Karlins E, Phung TN, Richmond PA, et al. Identifying, understanding, and correcting technical biases on the sex chromosomes in next-generation sequencing data [Internet]. 2018. doi:10.1101/346940