

Concerning the eXclusion in human genomics: the choice of sex chromosome representation in the human genome drastically affects the number of identified variants

Brendan J. Pinto ^{1,2,3,*} Brian O'Connor,⁴ Michael C. Schatz,⁵ Samantha Zarate,⁵ Melissa A. Wilson^{1,2,6,*}

¹School of Life Sciences, Arizona State University, Tempe, AZ 85282, USA

²Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85282, USA

³Department of Zoology, Milwaukee Public Museum, Milwaukee, WI 53233, USA

⁴Sage Bionetworks, Seattle, WA 98121, USA

⁵Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

⁶The Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, AZ 85282, USA

*Corresponding author: School of Life Sciences, Arizona State University, PO Box 874501, Tempe, AZ 85287-4501, USA. Email: brendan.pinto@asu.edu; *Corresponding author: School of Life Sciences, Arizona State University, PO Box 874501, Tempe, AZ 85287-4501, USA. Email: mwilsons@asu.edu

Over the past 30 years, a community of scientists has pieced together every base pair of the human reference genome from telomere to telomere. Interestingly, most human genomics studies omit more than 5% of the genome from their analyses. Under “normal” circumstances, omitting any chromosome(s) from an analysis of the human genome would be a cause for concern, with the exception being sex chromosomes. Sex chromosomes in eutherians share an evolutionary origin as an ancestral pair of autosomes. In humans, they share 3 regions of high-sequence identity (~98–100%), which, along with the unique transmission patterns of the sex chromosomes, introduce technical artifacts in genomic analyses. However, the human X chromosome bears numerous important genes, including more “immune response” genes than any other chromosome, which makes its exclusion irresponsible when sex differences across human diseases are widespread. To better characterize the possible effect of the inclusion/exclusion of the X chromosome on variants called, we conducted a pilot study on the Terra cloud platform to replicate a subset of standard genomic practices using both the CHM13 reference genome and the sex chromosome complement-aware reference genome. We compared the quality of variant calling, expression quantification, and allele-specific expression using these 2 reference genome versions across 50 human samples from the Genotype-Tissue Expression consortium annotated as females. We found that after correction, the whole X chromosome (100%) can generate reliable variant calls, allowing for the inclusion of the whole genome in human genomics analyses as a departure from the status quo of omitting the sex chromosomes from empirical and clinical genomics studies.

Key words: sex chromosomes; sex differences; gene expression; medical genomics; reference genome; variant calling; eXclusion

Introduction

The X and Y chromosomes in placental mammals share an evolutionary origin as an ancestral pair of autosomes (Graves 2008). Due to this shared ancestry and subsequent chromosomal rearrangements, the X and Y chromosomes in humans are highly divergent yet share regions of high-sequence identity (~98–100%; Fig. 1a), which introduces regions of varying ploidy across this chromosomal pair. Although this is well understood biologically, it introduces technical artifacts within modern genomic analyses that require correction to prevent potentially erroneous conclusions (Webster et al. 2019; Carey et al. 2022). Although these technical artifacts have remained ignored in many empirical and clinical studies, they have been used as a justification to ignore the sex chromosomes on a grand scale and, therefore, the importance of sex-linked variation to human health is likely to be greatly underestimated (Wise et al. 2013; Khrantsova et al. 2019; Natri

et al. 2019; Köferle et al. 2022; Inkster et al. 2023; Sun et al. 2023). In this study, we aim to better grasp the scope of data lost by excluding or misrepresenting the sex chromosomes in human genomics. We urge empiricists and clinicians to confront these issues moving forward to simultaneously increase the number of genome-wide association studies and reduce the number of autosome-wide association scan studies currently being published (Sun et al. 2023).

The typical human genome contains a diploid count of 46 chromosomes ($2n = 46$), but reference genome-based analyses require a haploid representation of each chromosome for correct inference (e.g. $n = 23$). In humans, the reference genome complement includes haploid representations for each autosome ($n = 22$) but not the sex chromosomes X and Y ($n = 2$); thus, the human reference genome contains an $n = 24$ chromosome representation (Fig. 1b). The X and Y deviate from autosomal expectations in that (1) not all individuals possess a Y chromosome, making the

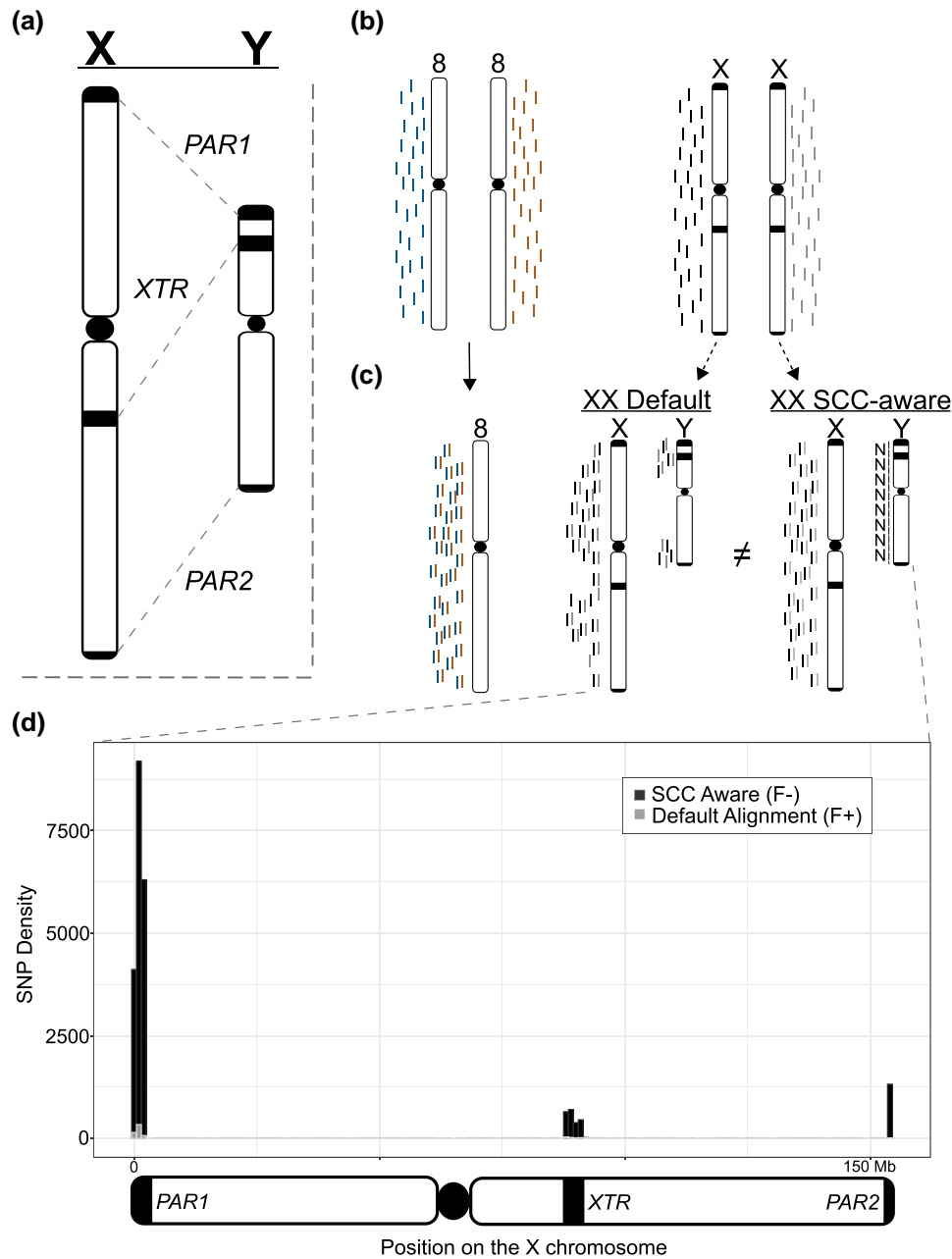


Fig. 1. Overview of technical artifacts on sex chromosomes for read mapping and variant calling. a) Overview of regions of high-sequence identity between the X and the Y chromosomes. b) NGS reads originating from a karyotypically diploid XX individual. c) How reads from an XX-karyotype individual align to the Default reference and how masking the Y chromosome in these cases improves read MQ in these regions. d) Connecting changes in read mapping to differences in called variants across the X chromosome in the analysis presented in this paper. Dark/black regions can be viewed as presumed false negatives (variants missed with the Default reference), while light/gray areas can be viewed as false-positive calls (variants unique to the Default reference). Overlapping variant calls between the 2 reference genomes have been removed. The 3 regions that contain the most incorrect calls using the Default reference are the 2 PARs (beginning and end of the plot) and the XTR (just right of the plot center). SNPs are binned into 1 Mb windows.

mapping of all reads to the Y chromosome erroneous in XX (or X0 for example) samples and (2) the X and Y retain regions of high-sequence similarity (maintaining between 98 and 100% sequence identity) due to their shared ancestry (Olney et al. 2020; Rhie et al. 2022). Thus, particular regions on the X and Y chromosomes violate the assumption that reference genome representation for linear alignments be uniformly haploid.

According to the most recent telomere-to-telomere (T2T) human reference genome (CHMv2.0), the X chromosome makes up 5.04% of the total genome size and contains approximately the same percentage of annotated genes (Nurk et al. 2022). Thus,

many published studies in humans blatantly ignore 5% or more of the human genome when conducting routine genomic analyses (Wise et al. 2013; Koboldt 2020; Zverinova and Guryev 2021). Indeed, despite recent advances in methodology to control for known technical artifacts inherent when analyzing the sex chromosomes (e.g. Webster et al. 2019), little progress has been made to further incorporate the X chromosome into broader biological analyses (Wise et al. 2013; Carey et al. 2022; Sun et al. 2023).

We set out to identify the extent of technical artifacts introduced by using the most complete human genome assembly currently available. Specifically, we aimed to better understand the

benefits of accurately representing the sex chromosome complement (SCC) when conducting standard genomic analyses. To parse the effects of the T2T-CHM13 reference genome on downstream analyses, we conducted parallel analyses using the GenBank default reference genome (Default) and a sex chromosome complement-aware reference (SCC-aware) using whole-genome resequencing and RNAseq data for 50 individuals from the Genotype-Tissue Expression (GTEx) project. We found that every analysis suffered in some capacity (in either accuracy, robustness, or both) because of not using the reference genome appropriate for the data. In line with observations from previous simulation studies, we found an overwhelming number of new variants called, using an SCC-aware reference, that are missed when using a Default reference (Oill 2022) and that are focused in regions with higher-sequence similarity to the Y chromosome than a large part of the X chromosome [i.e. both pseudoautosomal regions (PARs) and the X-transposed region (XTR)]. These differences are substantial and constitute ~5% of the total variants on the X chromosome.

Methods

Computational overview

All primary analyses for this project were conducted on the Terra platform (Schatz et al. 2022), which interfaces multiple biomedical genomic databases with Google Cloud (GCP) through the NIH Cloud Platform Interoperability Effort. As such, all analyses detailed below were written in Workflow Description Language; they are available for reuse here (https://github.com/DrPintoThe2nd/XYalign_AC3) and are also available for integration into others' custom Terra workspaces via Dockerstore (<https://dockstore.org>). Further, all analyses were conducted in a single, stable Docker container (Merkel 2014) including the following software and their dependencies (in alphabetical order): BamTools [v2.5.2] (Barnett et al. 2011), BMAP [v38.96] (Bushnell 2014), BCFtools [v1.15.1] (Li 2011), BEDTools [v2.30.0] (Quinlan and Hall 2010), bwa [v0.7.17] (Li and Durbin 2009), gatk4 [v4.2.6.1] (McKenna et al. 2010), HISAT2 [v2.2.1] (Kim et al. 2019), OpenSSL [v1.1.1q] (OpenSSL Project 2003), pandas [v1.4.3] (McKinney 2010), RTGTools [v3.12.1] (Cleary et al. 2015), salmon [v1.9.0] (Patro et al. 2017), SAMBLASTER [v0.1.26] (Faust and Hall 2014), SAMtools [v1.15.1] (Li and Durbin 2009), and Trim Galore! [v0.6.7] (Martin 2011; <https://doi.org/10.5281/zenodo.5127899>). This Docker is publicly available for reuse (https://hub.docker.com/r/drpintothe2nd/ac3_xysupp).

Data description

We selected a subset of 50 individuals annotated as female ($N = 50$, 46, XX) from the GTEx project (Aguet et al. 2020). All samples were consistent with 46, XX-karyotype, except for one, which we discarded because of issues related to anomalous read depth (adjusted $N = 49$). Each individual possessed a minimum of whole-genome resequencing data and RNAseq data for the same tissue; we chose the nucleus accumbens region of the basal ganglia because brain regions tend to have a high number of expressed genes (Li et al. 2017), and there is little difference in how distinct tissues are affected by reference genome mapping (Olney et al. 2020).

Variant calling

Because genomic data are stored on the cloud in a compressed alignment format (either CRAM or BAM, depending on data type), we first converted these files to unaligned read files, filtered

PCR duplicates, and trimmed them using SAMtools, BMAP, and Trim Galore!, respectively. We used bwa (DNA) and HISAT2 (RNA) to realign them to 2 different configurations of the recently published T2T human reference genome (CHM13v2.0; Nurk et al. 2022). The first configuration of the reference used was the default version downloaded from GenBank (Default), while the other was prepared as an XX-karyotype-specific reference genome by hard-masking the Y chromosome (SCC-aware) using XYalign (Webster et al. 2019). This type of approach also improves variant calling in XY samples (Oill 2022; Rhie et al. 2022). As the downloaded genome version does not include a mitogenome sequence, both the Default and the SCC-aware reference genomes were spiked with the mitogenome from the GRCh38 reference to help prevent mtDNA reads from mismatching to our regions of interest. We called variants on chromosome 8 and the X chromosome using GATK's HaplotypeCaller and GenotypeGVCFs functions. We filtered to select only biallelic variants with ≥ 4 alleles present in called genotypes ($AN \geq 4$), high mapping quality ($MQ > 40.0$), a minimum quality by the depth of 7 ($QD > 7.0$), and a total read depth of ≥ 10 but $\leq 2,500$ ($DP \geq 10.0$ && $DP \leq 2,500.0$). We parsed and interrogated the resultant VCF files using BCFtools, RTGTools, and BEDTools to better characterize the technical artifacts involved in mapping to the Default vs SCC-aware reference genomes.

RNAseq analyses

We analyzed the effects of reference genome on 2 common RNAseq data analysis, gene expression analysis, and allele-specific expression (ASE) analysis using salmon and GATK, respectively. We generated Default and SCC-aware reference transcriptomes for salmon analysis by extracting transcripts from the Default and SCC-aware genomes from the RefSeq annotation file using gffread [v0.12.1] (Pertea and Pertea 2020). We soft-masked an alternate version of the Default genome using RepeatModeler [v2.0.3] (Flynn et al. 2020) to facilitate the generation of index decoys via the generateDecoyTranscriptome.sh script accompanying salmon software distribution. We ran salmon using the trimmed RNAseq reads for each individual for each reference transcriptome using the `-gcBias` and `-validateMappings` flags. For ASE, we split the filtered, genotyped VCF for each individual using BCFtools and combined each individual VCF file with their realigned RNAseq data using GATK's ASEReadCounter function. We compared the results between reference genomes as a deviation from a 1:1 relationship. For ASE, we also compared the efficacy of variant calling and alignment on the total number of transcripts identified as allele-specific.

Results

Sex chromosome-aware reference augments variant calling

On a broad scale, we identified that the SCC-aware reference alignment increased the number of properly paired reads mapped for many individuals (mean: +6,551; +0.0008%) and decreased in mapped reads with an MQ of 0 ($MQ = 0$) in every individual (mean: -605,396; -1.05%) (Supplementary Tables 1 and 2). These changes in read mapping resulted in changes in the total number of biallelic, single nucleotide variants (SNPs) among all 49 individuals. In contrast, on chromosome 8, the total number of variants called were nearly identical between the 2 reference genome configurations, 719,826 variants and 719,824 variants for the Default and SCC-aware reference, respectively. At a per-individual scale, this course held, with the average number

Table 1. Numerical differences in variant calling outcomes on chromosome 8 and the X chromosome between SCC-aware and default reference alignment.

Category	Chrom	Default	SCC-aware	% change (SCC/D)
Total SNPs	8	719,826	719,824	-0.0002%
Per-indiv. avg SNPs	8	178,885	178,882	-0.002%
Per-indiv. ref allele	8	540,938	540,939	0.0002%
Total SNPs	X	475,763	498,297	4.74%
Per-indiv. avg SNPs	X	98,877	105,413	6.61%
Per-indiv. ref allele	X	376,884	392,882	4.07%

Numbers are quality-filtered biallelic SNPs for chromosome 8 (top) and the X chromosome (bottom).

Table 2. Dissection of differences in variant calling within regions of interest across the x chromosome.

Category	Default	SCC-aware	% change (SCC/D)	Added (F-)	Lost (F+)
Total SNPs	475,763	498,297	+4.74%	23,279	745
Non-PAR/ XTR	453,822	453,882	+0.01%	150	90
PARs	2,790	23,161	+730.14%	20,931	560
XTR	19,151	21,254	+10.98%	2,198	95
Genic SNPs	162,989	171,351	+5.13%	8,600	238
Non-PAR/ XTR	157,957	157,979	+0.01%	45	23
PARs	1,390	9,235	+564.39%	8,042	197
XTR	3,642	4,137	+13.59%	513	18

Numbers are quality-filtered biallelic SNPs across various regions on the X chromosome (top) and within genic regions only across various regions on the X chromosome (bottom).

of variants being 178,885 and 178,882 variants, respectively (Fig. 1; Table 1). However, this impartiality was not replicated on the X chromosome, where we found a sharp increase of 22,534 total SNPs (from 475,763 to 498,297) when using the SCC-aware reference configuration. This deviation also held for each individual in our study, with an average increase in the number of called SNPs from 98,877 to 105,413 (Table 1; Supplementary Table 3).

Across a large part of the X chromosome (~95%), we found little variation between the 2 reference genome configurations (Fig. 1d; Table 2). Indeed, as a large part of the X chromosome shares little sequence identity between the X and the Y chromosomes, very few areas generate read mapping conflict between them, even for XX samples (Fig. 1). In the 3 regions of high-sequence similarity (PAR1, XTR, and PAR2), changes in the total number of SNPs called between reference configurations ranged from an 11% increase to a 730% increase in the XTR and PARs, respectively (Table 2). Indeed, we saw an increase in called variants in both genic (PARs: +564.39%; XTR: +13.59%) and intergenic (PARs: +894.71%; XTR: +10.37) regions (Table 2). Thus, while differences across a large part of the X chromosome are negligible, the differences in the number of called SNPs in the XTR and 2 PARs are significant in relation to both autosomes or to the rest of the X chromosome.

Default reference distorts gene expression quantification on the X

Somewhat contrary to the exceptional differences between variant calling with different reference genome configurations, differences between gene expression quantification are more subtle, yet still apparent (Fig. 2). For gene expression quantification, we

calculated transcripts per kilobase million (TPM) and found that differences between expression levels are greatest in PAR1, followed by PAR2, and then the rest of the chromosome (Supplementary Fig. 3a). However, contrary to expectations we find little change in expression values within the XTR (Supplementary Fig. 3a). Also contrary to expectations, we find no relationship between observed expression differences and transcript length (Supplementary Fig. 3b) or expression level (Supplementary Fig. 3c and d).

When examining ASE levels, or the allele balance ratio, we see an opposite pattern—where the higher expressed a transcript is, the more skewed the Default alignment data become on the X chromosome. We observed that allele balance values are generally inflated using the Default reference (Fig. 2). Importantly, we see a premature summit, or abbreviated climb, from allele balance values from 0.5 to 1.0, when using the Default reference genome—where allele balance values >0.9 get rounded up to 1.0 (Fig. 2). Because there is an extra alignment step in ASE analysis relative to regular expression quantification (i.e. variant calling), we attempted to parse which aspects of ASE analysis are most affected by which segment of the analysis. We paired each potential variant calling output (VCF file) with each potential RNAseq alignment output (BAM file) by rerunning the analysis in a “round-robin”, or “all-vs-all”, format. We found that the VCF file (and thus the reference genome used for variant calling) chosen to run ASE had the greatest influence on the number of recovered biallelic transcripts (Table 3).

Discussion

As expected, there were negligible differences in all analyses between results on chromosome 8 between Default and SCC-aware reference genomes (Fig. 2; Tables 1 and 3). However, the differences on the X chromosome were substantial (Figs. 1 and 2; Tables 1–3). The most numerous differences between the Default and SCC-aware reference genomes were the sheer number of (presumed) false negatives when using the Default reference, i.e. variants called using the SCC-aware reference but missed with the Default reference (Table 2). There were also (presumed) false positives, variants called with the Default reference that were absent in the SCC-aware reference; however, these made up a small fraction of the observed differences (Table 2). To expand on this concept, we calculated the major allele frequencies for all sites in both the Default and SCC-aware VCFs (Supplementary Fig. 1) and then filtered out variants that overlap between the 2 (Supplementary Fig. 2). We expected that if 1 spectrum contained an increase in false positives the major allele frequency would skew more heavily towards 1.0 (an increase in singleton calls). Indeed, this is exactly what we observed in both PAR regions and the XTR (Supplementary Figs. 1 and 2).

Although the PARs make-up only ~1.77% of the X chromosome, they contain ~5% of both genic (5.39%) and indiscriminate (all) SNPs (4.65%) within our sampled individuals. However, using the Default reference genome, these numbers are unfathomably low for both genic (0.85%) and indiscriminate SNPs (0.59%). This pattern also holds, albeit mediated by genetic divergence between X and Y alleles relative to the PARs, within the XTR. The XTR makes up ~3.04% of the X chromosome, yet the numbers of called SNPs increase substantially when using the appropriate SCC-aware reference compared to the Default for both genic (2.2–2.4%) and indiscriminate SNPs (4.0–4.3%).

Our expression analyses of RNAseq data may be the first published RNAseq analyses using the CHM13_v2.0 assembly. Our

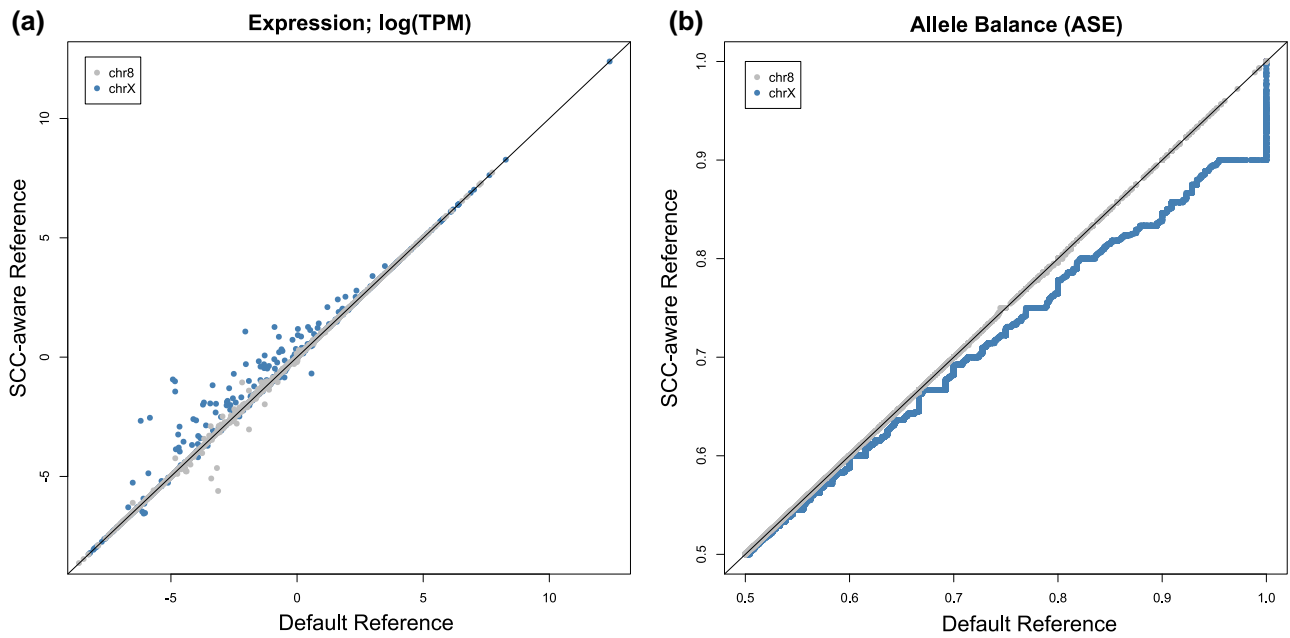


Fig. 2. Effects of the SCC-aware reference genome on common RNAseq analyses: a) gene expression (TPM) and b) allele balance (ASE). For allele balance, we used the SCC-aware reference called VCF as a measure to increase the total number of transcripts included (see Table 3). Both analyses use the T2T-CHM13v2 genome sequence for mapping.

Table 3. Efficacy of ASE analysis across differing modes of variant calling and RNAseq alignment strategies.

ASE mode	chr8 #	chrX #
Default VCF, Default RNAseq	885	819
Default VCF, SCC-aware RNAseq	885	823
SCC-aware VCF, Default RNAseq	885	834
SCC-aware VCF, SCC-aware RNAseq	885	835

comparative expression analysis suggests that a notable amount of gene expression differences can be found throughout the X chromosome but are most notable in PAR1 (Supplementary Fig. 3a). Interestingly, we note that ASE analysis especially suffers from a 2-fold increase in error when using an inappropriate reference genome. The first introduction of error, as mentioned previously, is the substantial number of false negatives introduced during variant calling via mapping WGS reads (Tables 1 and 2). The second error is introduced during mapping RNAseq reads to the Default reference, whereby correcting for either factor (called SNPs or RNAseq mapping) can partially recover some of the potentially missed transcripts in an ASE experiment (Table 3). However, to take full advantage of ASE analyses on the X chromosome, it is essential to include both correctly called variants and correctly mapped RNAseq reads (Table 3; Supplementary Table 4).

In line with previous conclusions (e.g. Wise et al. 2013), the general absence of the X chromosome in many analyses may be due, in part, to an increase in technical effort/ability to prepare the reference genome prior to analysis (Webster et al. 2019). The X chromosome makes up 5% of the haploid genome size of the typical XX human individual. Therefore, the “scorched earth” error rate of not including the X chromosome in genomics analyses of XX individuals is at least 5%. The introduction of read mapping errors on the X chromosome only affects 5% of the total length of the X chromosome, which equates to only 0.25% of the variants called become unreliable when not accounting for SCC and using a Default reference genome (Fig. 1; Table 2). Thus, the common

practice of purposefully introducing an error rate of 5% (excluding the X chromosome) to potentially avoid an error rate of 0.25% (including the X chromosome) is excessive and, technically speaking, precludes the use of the term “genome-wide” in most association studies in humans (Wise et al. 2013; Sun et al. 2023). At a minimum, using a reference genome with Y PARs masked would provide a substantial improvement to the total variants called (Table 2). However, it is a relatively trivial task to inform the reference genome with the SCC when mapping samples and accommodate changes in ploidy across different regions; thus ensuring that reliable variant calls across, even within the PARs and XTR (Webster et al. 2019; Carey et al. 2022). We expect the broader utilization of the SCC-aware reference genome for alignment could be catalyzed by it being made available alongside the Default on repositories such as NCBI’s GenBank, where the main hurdle to its inclusion may be low (Carey et al. 2022).

In conclusion, we conducted a pilot study of replicating a series of commonly used genomics tools/analyses across a subset of the GTEx data available on the cloud. We showed that technical artifacts introduced by using the Default reference genome affect about 5% across the X chromosome but are most extensive in the PARs and XTR, ranging upwards of 700% in some regions. In line with prior work, we provided additional evidence that technical artifacts of including the sex chromosomes in genomics analyses can be negated with available information and tools (Olney et al. 2020; Webster et al. 2019). We are aware that, though the “eXclusion” of the X chromosome is widespread (Wise et al. 2013), the exclusion of the Y is even more extensive in empirical and clinical genomics (Sun et al. 2023). SCC-aware reference genomes can effectively negate the effects of homology on the sex chromosomes in XX individuals and reduce this mismapping in XY individuals, allowing for their accurate inclusion in human genomics studies (Oill 2022). We are hopeful that research groups will make the inclusion of SCC-aware references a staple part of their future projects, not only to better reflect the original intent behind the National Institutes of Health of the USA’s policy on the consideration of sex as a biological variable (<https://orwh>).

od.nih.gov/sex-gender/nih-policy-sex-biological-variable) but also to bring to humanity a better understanding of how sex chromosomes affect human health and disease states across the world.

Data availability

The data used in this study are available as follows: reference genome T2T-CHM13v2.0, GenBank: GCA_009914755.4. The GTE Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx data are described and available through dbGaP under accession phs000424.v8.p1. We received approval to access this data under dbGaP accession #8834 and code to replicate results on the Terra cloud computing environment on GitHub/Dockstore (https://github.com/DrPintoThe2nd/XYalign_AC3).

Supplemental material available at G3 online.

Acknowledgments

We acknowledge Research Computing at Arizona State University for providing high-performance computing and storage resources that have contributed to the research results reported in this paper (<https://cores.research.asu.edu/research-computing>). We would also like to thank Wilson lab members for helpful feedback on the research and manuscript.

Funding

This work was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health grant R35GM124827 to M.A.W. Funding for the computational aspects of this research was provided to M.A.W by the AnVIL Cloud Credits (AC3) pilot program for computation using the Terra platform.

Conflicts of interest

The author(s) declare no conflict of interest.

Literature cited

Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, Graubert A, Hadley K, Handsaker RE, Huang KH, Kashin S, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318–1330. doi:10.1126/science.aaz1776.

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(12):1691–1692. doi:10.1093/bioinformatics/btr174.

Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Berkeley (CA): Lawrence Berkeley National Laboratory; 2014. No. LBNL-7065E.

Carey SB, Lovell JT, Jenkins J, Leebens-Mack J, Schmutz J, Wilson MA, Harkess A. Representing sex chromosomes in genome assemblies. *Cell Genom*. 2022;2(5):100132. doi:10.1016/j.xgen.2022.100132.

Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Francisco M. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv*. 2015. doi:10.1101/023754.

Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. 2014;30(17):2503–2505. doi:10.1093/bioinformatics/btu314.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;117(17):9451–9457. doi:10.1073/pnas.1921046117.

Graves JAM. Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annu Rev Genet*. 2008;42(1):565–586. doi:10.1146/annurev.genet.42.110807.091714.

Inkster AM, Wong MT, Matthews AM, Brown CJ, Robinson WP. Who's afraid of the X? Incorporating the X and Y chromosomes into the analysis of DNA methylation array data. *Epigenetics Chromatin*. 2023;16(1):1. doi:10.1186/s13072-022-00477-0.

Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. *Nat Rev Genet*. 2019;20(3):173–190. doi:10.1038/s41576-018-0083-1.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37(8):907–915. doi:10.1038/s41587-019-0201-4.

Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med*. 2020;12(1):91. doi:10.1186/s13073-020-00791-w.

Köferle A, Schlattl A, Hörmann A, Thatikonda V, Popa A, Spreitzer F, Ravichandran MC, Supper V, Oberndorfer S, Puchner T, et al. Interrogation of cancer gene dependencies reveals paralog interactions of autosome and sex chromosome-encoded genes. *Cell Rep*. 2022;39(2):110636. doi:10.1016/j.celrep.2022.110636.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–2993. doi:10.1093/bioinformatics/btr509.

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, Shi L. A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq. *Sci Rep*. 2017;7(1):1–10. doi:10.1038/s41598-017-04520-z.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17(1):10. doi:10.14806/ej.17.1.200.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–1303. doi:10.1101/gr.107524.110.

McKinney W. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, No. 1) p. 51–56.

Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J*. 2014;239:2. doi:10.5555/2600239.2600241.

Natri H, Garcia AR, Buetow KH, Trumble BC, Wilson MA. The pregnancy pickle: evolved immune compensation due to pregnancy underlies sex differences in human diseases. *Trends Genet*. 2019;35(7):478–488. doi:10.1016/j.tig.2019.04.008.

Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44–53. doi:10.1126/science.abj6987.

Oill AM. 2022. Assessment of genetic variation in globally diverse human populations and its implications for human health and disease. Unpublished dissertation. Arizona State University.

Olney KC, Brotman SM, Andrews JP, Valverde-Vesling VA, Wilson MA. 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. *Biol Sex Differ*. 2020;11(1):42. doi:[10.1186/s13293-020-00312-9](https://doi.org/10.1186/s13293-020-00312-9).
- OpenSSL Project. 2003. OpenSSL: the Open Source toolkit for SSL/TLS.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417–419. doi:[10.1038/nmeth.4197](https://doi.org/10.1038/nmeth.4197).
- Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Res*. 2020;9:304. doi:[10.12688/f1000research.23297.1](https://doi.org/10.12688/f1000research.23297.1).
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842. doi:[10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).
- Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. The complete sequence of a human Y chromosome. *bioRxiv*. 2022. doi:[10.1101/2022.12.01.518724](https://doi.org/10.1101/2022.12.01.518724).
- Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, Culotti A, Ellrott K, Goecks J, Grossman RL, et al. Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics lab-space. *Cell Genomics*. 2022;2(1):100085. doi:[10.1016/j.xgen.2021.100085](https://doi.org/10.1016/j.xgen.2021.100085).
- Sun L, Wang Z, Lu T, Manolio TA, Paterson AD. Exclusionary: ten years later, where are the sex chromosomes in GWAS? *Am J Hum Genet*. 2023;110(6):903–912. doi:[10.1016/j.ajhg.2023.04.009](https://doi.org/10.1016/j.ajhg.2023.04.009).
- Webster TH, Couse M, Grande BM, Karlins E, Phung TN, Richmond PA, Wilson MA. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *Gigascience*. 2019;8(7):giz074. doi:[10.1093/gigascience/giz074](https://doi.org/10.1093/gigascience/giz074).
- Wise AL, Gyi L, Manolio TA. Exclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet*. 2013;92(5):643–647. doi:[10.1016/j.ajhg.2013.03.017](https://doi.org/10.1016/j.ajhg.2013.03.017).
- Zverinova S, Guryev V. Variant calling: considerations, practices, and developments. *Hum Mutat*. 2021;43(8):976–985. doi:[10.1002/humu.24311](https://doi.org/10.1002/humu.24311).

Editor: M. Hahn