PLOS GENETICS

RESEARCH ARTICLE

Why does the X chromosome lag behind autosomes in GWAS findings?

Ivan P. Gorlov*, Christopher I. Amos

Baylor College of Medicine, Institute for Clinical & Translational Research, One Baylor Plaza, Houston, Texas, United States of America

* ivan.gorlov@bcm.edu

Abstract

The X-chromosome is among the largest human chromosomes. It differs from autosomes by a number of important features including hemizygosity in males, an almost complete inactivation of one copy in females, and unique patterns of recombination. We used data from the Catalog of Published Genome Wide Association Studies to compare densities of the GWAS-detected SNPs on the X-chromosome and autosomes. The density of GWAS-detected SNPs on the X-chromosome is 6-fold lower compared to the density of the GWAS-detected SNPs on autosomes. Differences between the X-chromosome and autosomes cannot be explained by differences in the overall SNP density, lower X-chromosome coverage by genotyping platforms or low call rate of X-chromosomal SNPs. Similar differences in the density of GWAS-detected SNPs were found in female-only GWASs (e.g. ovarian cancer GWASs). We hypothesized that the lower density of GWAS-detected SNPs on the X-chromosome compared to autosomes is not a result of a methodological bias, e.g. differences in coverage or call rates, but has a real underlying biological reason–a lower density of functional SNPs on the X-chromosome versus autosomes. This hypothesis is supported by the observation that (i) the overall SNP density of X-chromosome is lower compared to the SNP density on autosomes and that (ii) the density of genic SNPs on the X-chromosome is lower compared to autosomes while densities of intergenic SNPs are similar.

Author summary

One of the most striking observations from the Genome Wide Association Studies (GWAS) is that the density of GWAS hits is much lower on X-chromosome compared to autosomes. This was initially explained by technical/analytical reasons such as lower coverage and lack of adequate methods to analyze X-chromosomal SNPs. Since then, a better coverage and better analytical methods to analyze X-chromosomal SNPs were developed. We recently revisited the issue and found that the density of GWAS hits on X-chromosome is at least 5-fold lower compared to autosomes. We demonstrated that the difference cannot be explained by technical or analytical reasons. We proposed a hypothesis of a real biological phenomenon underlying X versus autosomal differences in the density of GWAS-detected SNPs, namely that X-chromosome has a lower density of functional

Citation: Gorlov IP, Amos CI (2023) Why does the X chromosome lag behind autosomes in GWAS findings? PLoS Genet 19(2): e1010472. https://doi.org/10.1371/journal.pgen.1010472

Editor: Andreas Ziegler, Medizincampus Davos, SWITZERLAND

Received: October 12, 2022

Accepted: February 15, 2023

Published: February 27, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1010472

Data Availability Statement: The data analyzed herein was previously published and openly available. All data relevant to our analysis are shown in supplementary materials. We use GWAS-detected SNPs reported in the Catalog Of Published Genome Wide Association studies https://www.ebi.

ac.uk/gwas/. To count number of known SNPs on individual chromosomes we used data from dbSNP database https://www.ncbi.nlm.nih.gov/snp/, one thousand genome database https://www.genome.gov/27528684/1000-genomes-project#al-3, and TOPMed project https://topmed.nhlbi.nih.gov/. Sizes of intronic, exonic intergenic regions for individual chromosome were estimated using data from the consensus coding sequence (CCDS) database https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi.

Funding: Partial financial support was received from National Institutes of Health grants U19CA203654, U19CA203654S1 and Cancer Prevention and Research Institute of Texas grant RR170048. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

polymorphisms compared to autosomes because of a stronger selection against X-chromosomal mutations since X-chromosomal variants are more exposed to natural selection due to hemizygosity in males and X-chromosome inactivation in females. The hypothesis is supported by the analysis of the densities of intergenic, intronic and exonic SNPs on human chromosomes.

Introduction

The X-chromosome comprises about 156 million base pairs, which is comparable to the size of chromosome 7. For most of the X-chromosome, crossovers are limited to females only, resulting in a stronger linkage disequilibrium on the X-chromosome compared to autosomes [1]. The X-chromosome has two small pseudoautosomal regions (PARs) on Xp and Xq. While genes in PAR regions are expressed from both homologs in females, X-inactivation for most of the X chromosome that does not have a Y complement occurs during embryogenesis. In males, recombination between the X and Y chromosomes is limited to the pseudoautosomal regions which display an enormous excess of recombination compared to the genome average [2]. Both positive and negative (purifying) selection are stronger for the X chromosomal mutations because hemizygosity in males and X-chromosome inactivation in females make them more open to selective pressure compared to mutations occurring on diploid autosomes [3,4].

Genome-wide association studies (GWASs) have advanced the understanding of genetic control of human diseases and phenotypes [5,6]. Thousands of associations between Single Nucleotide Polymorphisms (SNPs) and phenotypic traits/diseases have been reported [7–10]. Since the X-chromosome has a number of unique characteristics compared to autosomes, it is important to understand why there are differences between the X-chromosome and autosomes in terms of GWAS findings.

A review of the distribution of GWAS hits by chromosomes found a substantial deficit of GWAS-detected SNPs on X-chromosome compared to autosomes (see study by Wise at al. [11]). The initial explanation of this deficit was underutilization of X-chromosomal SNPs by GWAS researchers because of lack of effective methods to analyze X chromosomal SNPs. However X-chromosome specific methods for GWAS were published 15 years ago [12,13] and since then it has been an active area of research [14,15]. Despite availability and ever improving analytical tools for X-chromosomal SNPs the shortage of GWAS SNPs located on X-chromosome persisted.

We reviewed summary statistics of thousands of published GWASs to identify proportions of SNPs associated with diseases/phenotypes for all chromosomes separately. We found that the proportion of disease/phenotype-associated SNPs is much lower for the X chromosome compared to autosomes. The persistent difference between X-chromosomal and autosomal SNPs in the densities of GWAS hits suggest that it is not a result of a "neglect" of X-chromosome by GWAS researchers [16] but that the deficit has a real underlying biological basis. We propose that a lower density of functional SNPs underlies the deficit of GWAS hits on X-chromosome compared to autosomes.

Results

Overall SNP densities on individual chromosomes

Fig 1 shows the overall density of SNPs computed as the number of SNPs per kilobase. SNPs from 3 sources are shown separately from the left to right: NCBI dbSNP database, 1000 Genome Project, and TOPMed project. Density of SNPs on X-chromosome is somewhat

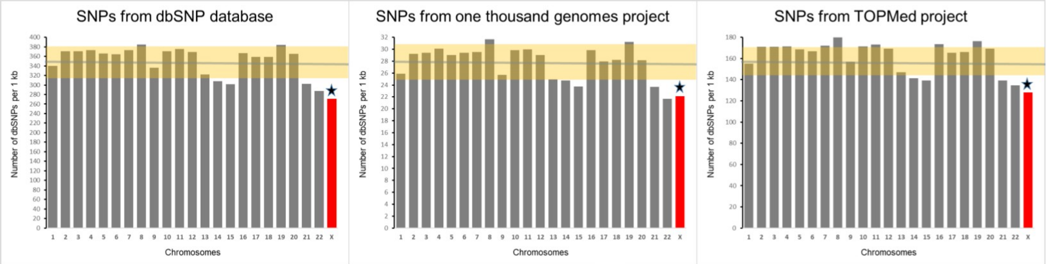

Fig 1. A comparison of X-chromosome and autosome by the overall SNP density. Green horizontal line depicts the mean for autosomes. Highlighted area represents SD for autosomes. Star marks the significant difference between X-chromosome and the mean for autosomes. The density was estimated as the number of SNPs per kilobase of chromosomal sequence.

https://doi.org/10.1371/journal.pgen.1010472.g001

lower compared to the average density on autosomes. In all 3 cases the differences between X-chromosome and autosomes in the density of SNPs were statistically significant with the corresponding p-values being: p = 0.003, 0.02 and 0.006.

The densities of the GWAS-detected SNPs on individual chromosomes

We found that the number of GWAS-detected SNPs per megabase is 6.8 times lower for X-chromosome compared to the average density for autosomes– 8.7 SNPs for X-chromosome compared to 59.4±2.9 for autosomes (Z-score = 3.8, p = 0.0001) (Fig 2, left panel). Accounting for the differences in the overall SNP density between X-chromosome and autosomes (see Fig 1) does not change the result materially (right panel of Fig 2). The average number of GWAS-detected SNPs per thousand SNPs reported by the NCBI dbSNP database is 0.17±0.01; however, the number of GWAS-detected SNPs per thousand X-chromosome SNPs is 0.03 (Z-score = 3.9, p = 0.00004). Therefore, there is a 5.7-fold difference between autosomes and X-chromosome in the proportion of GWAS-detected SNPs among all SNPs.

Coverage of X-chromosome and autosomes by common genotyping platforms

We analyzed chromosome coverage for 28 most often used genotyping platforms to test if the lower number of GWAS-detected SNPs on X-chromosome is a result of a lower coverage of X-

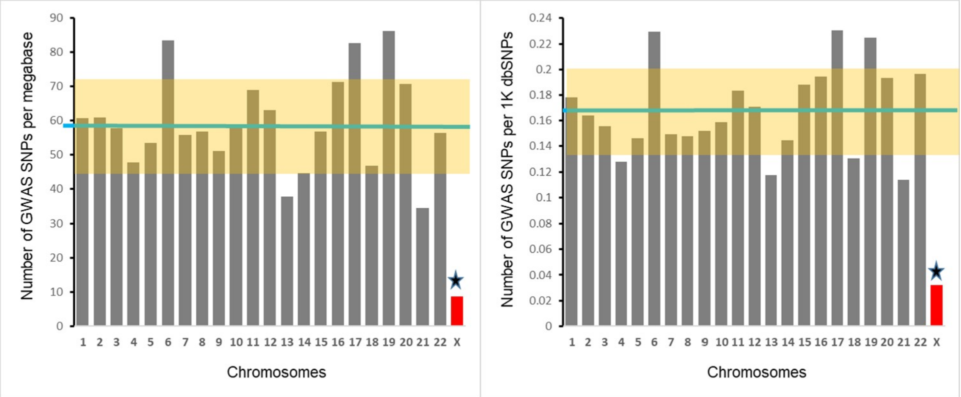

Fig 2. Chromosomal distributions of density (left panel) and fraction (right panel) of GWAS-detected SNPs. Left panel depicts the number of GWAS-detected SNPs per megabase, right panel shows the number of GWAS-detected SNPs per thousand SNPs. Each bar represents a chromosome. Green horizontal line represents the mean for all autosomes. Highlighted area represents SD for autosomes. Star marks the significant difference between X-chromosome and the mean for autosomes.

https://doi.org/10.1371/journal.pgen.1010472.g002

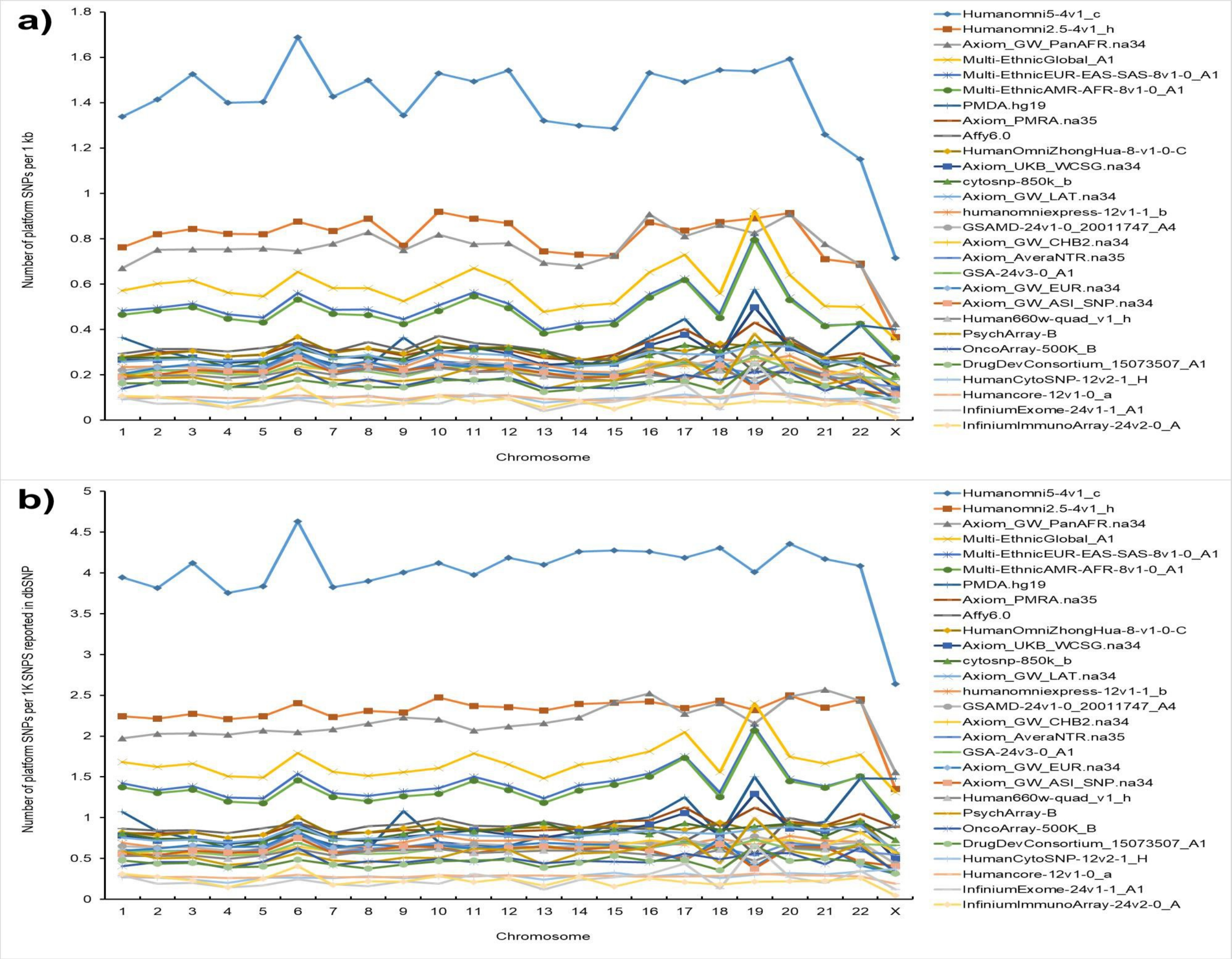

Fig 3. SNP coverage of individual human chromosomes on common genotyping platforms. (a) The number of SNPs on the platform per 1kb of the chromosomal sequence. **(b)** The number of SNPs on the genotyping platform per one thousand of the SNPs reported by dbSNP database for a given chromosome.

https://doi.org/10.1371/journal.pgen.1010472.g003

chromosome by genotyping arrays. The number of SNPs on individual chromosomes for each of the 28 platforms can be found in S2 Table. We used two metrics to test if all chromosomes are equally represented on genotyping platforms: (i) the number of platform SNPs per 1,000 nucleotides (chromosomal density); and (ii) the number of platform SNPs per 1,000 dbSNP SNPs reported for a given chromosome (fraction of genotyped SNPs). Fig 3 shows the results of the analysis. The results for these two metrics were similar. We found a lower density of X-chromosome SNPs compared to the SNPs on autosomes (upper panel of Fig 3): the mean ratio of the X-chromosome to autosome coverage across all platforms was 0.59±0.04, which significantly deviates from 1 expected under the assumption of equal coverage of X-chromosome and autosomes (t-test = 14.0, df = 28, p = $6.8x10^{-14}$). When the fraction of genotyped SNPs was used as a metrics for the coverage, the mean ratio of the X-chromosome to autosome coverage across all platforms was 0.75±0.05, which also significantly deviates from 1 expected under assumption of equal coverage of X-chromosome and autosomes (t-test = 15.0, df = 28, p = $3.8x10^{-9}$). Therefore, coverage of X-chromosomal SNPs is 25–40% lower compared to the coverage of autosomal SNPs.

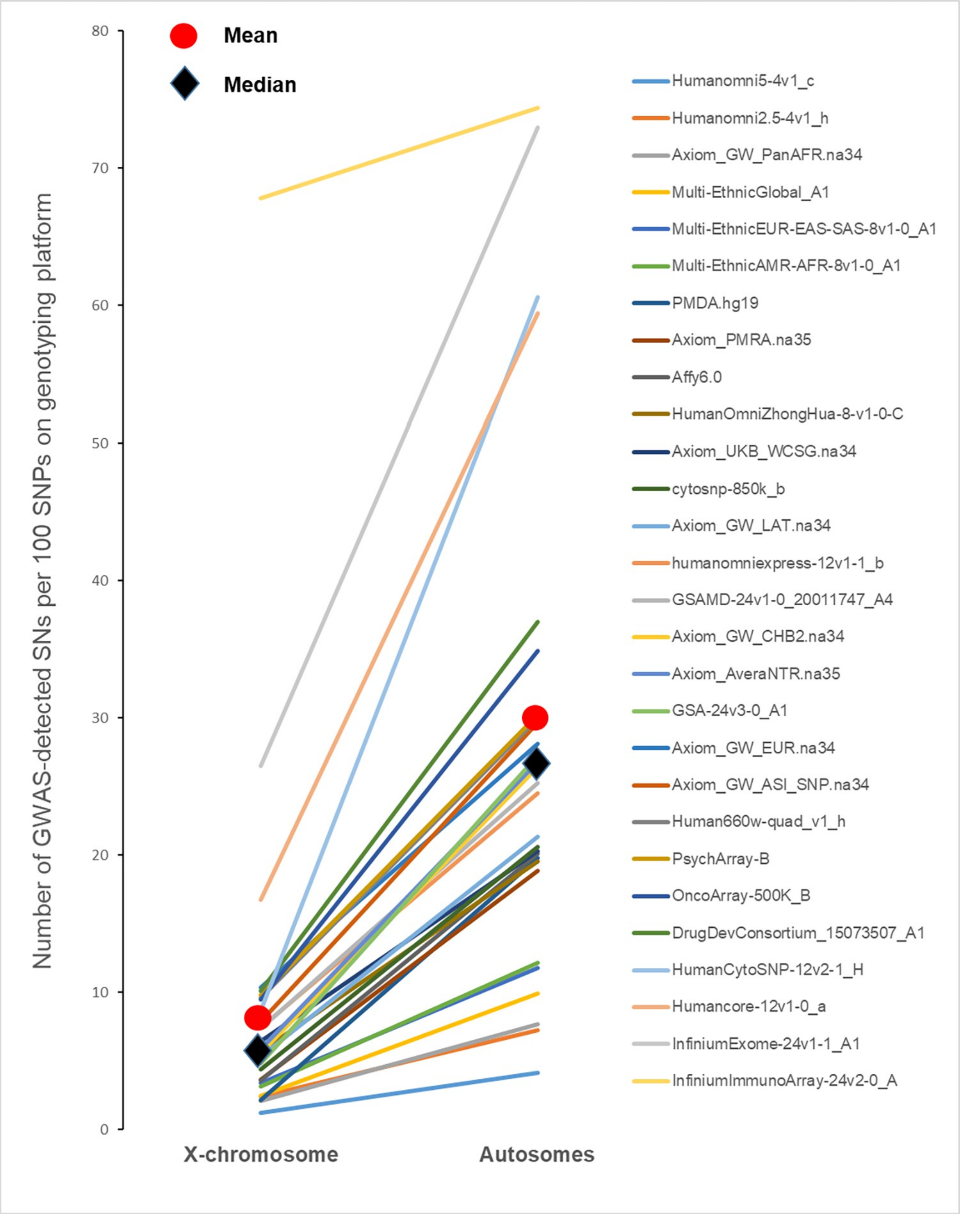

Fig 4. The number of GWAS-detected SNPs on X-chromosome and autosomes per 100 SNPs on genotyping platforms.

https://doi.org/10.1371/journal.pgen.1010472.g004

The number of GWAS-detected SNPs per hundred SNPs on genotyping platform

To test if the lower density GWAS hits can be explained by the lower coverage of SNPs located on X-chromosome we compared the densities of GWAS hits after adjustment for chromosome coverage. Our goal was to compare density of GWAS-detected SNPs among SNPs that were actually genotyped. We retrieved lists of SNPs for each genotyping platform and among them identified SNPs ever reported by GWASs. To make the results comparable across chromosomes we counted the number of GWAS-detected SNPs per 100 SNPs (Fig 4, S3 Table). The goal of the Fig 4 (S3 Table) was to demonstrate that the difference between autosomal and X-

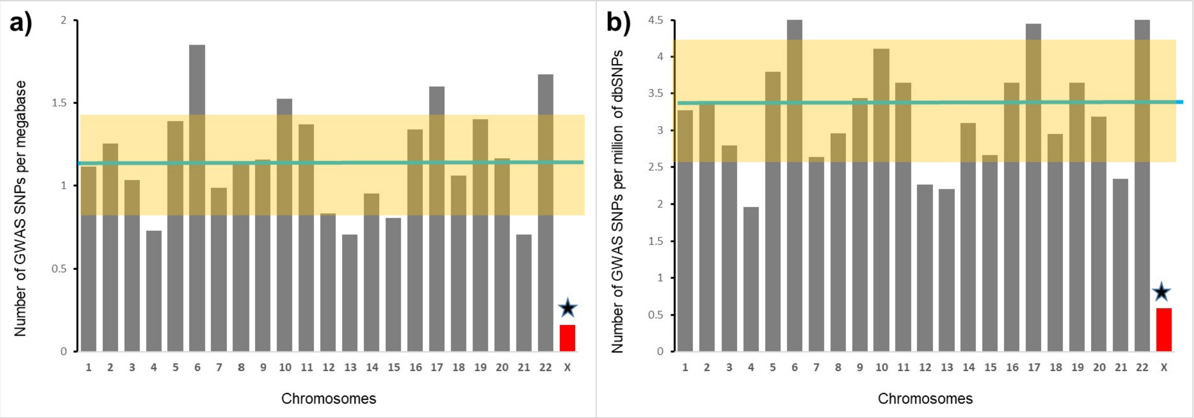

Fig 5. A comparison of X-chromosome and autosomes by the density of GWAS-detected SNPs in the female-only GWASs. The green horizontal line is the average among all autosomes; the highlighted area represents SD for autosomes; star marks a significant difference between X-chromosome and the average for autosomes. **(a)** The number of female-restricted GWAS-detected SNPs per megabase. **(b)** The number of female-restricted GWAS-detected SNPs per million of SNPs reported by dbSNP database for a given chromosome.

https://doi.org/10.1371/journal.pgen.1010472.g005

chromosome SNPs in the density of GWAS hits remains when we consider SNPs that **were actually included in GWASs**. Mean and median number of SNPs reported in at least one GWAS for X-chromosomal SNPs were 9.1% and 5.7% correspondingly, and 27.9% and 24.9% for autosomal SNPs. For each platform we also computed the ratio of the success rate for autosomal SNPs to the success rate for X-chromosomal SNPs. The mean ratio across 28 platforms was 4.03±0.28 (sign test = 5.1, p = 0.0000003). Therefore, autosomal SNPs have a much higher probability to be associated with a phenotype or disease compared to SNPs located on the X-chromosome after taking into account a lower X-chromosome coverage.

SNPs detected in female-restricted GWASs

Because of hemizygosity the number of genotyped X-chromosomes in males is half of the number of X-chromosomes in females compared to females while the number of genotyped autosomal SNPs is the same. This difference leads to a smaller effective sample size in males and, therefore, may affect the statistical power. Female-only GWASs, however, are expected to have a similar statistical power for X-chromosomal- and autosomal SNPs. The complete list of female-restricted diseases and phenotypes with the number of GWAS hits are shown in S4 Table. In total, 163 female-only phenotypes were used in the analysis, with the total number of GWAS hits of 4,581, comprising about 3% of the total number of SNPs reported in the Catalog of Published GWASs.

The results from the analysis are essentially identical to the results from all-GWAS-hits-together analysis (Fig 5). In female-restricted GWASs, the average number of GWAS-detected SNPs per megabase was 59.84±3.14 for autosomes, while the number of GWAS-detected SNPs per megabase of X-chromosomal sequence was much lower—8.69 (Z-score = 3.5, p = 0.0003). The average number of GWAS-detected SNPs per million of autosomal SNPs was 3.33±0.2 while the number of GWAS-detected SNPs per million of reported X-chromosomal SNPs was 0.59 (Z-score = 2.9, p = 0.002). For female-restricted phenotypes there is a 6.9-fold difference between autosomes and X-chromosome in the density of GWAS-detected SNPs and there is a 5.6-fold difference in the proportion of GWAS-detected SNPs among all reported SNPs. These differences are very similar to the all-GWAS-hits-together analysis: 6.8-fold and 5.7-fold differences, correspondingly (see Fig 2).

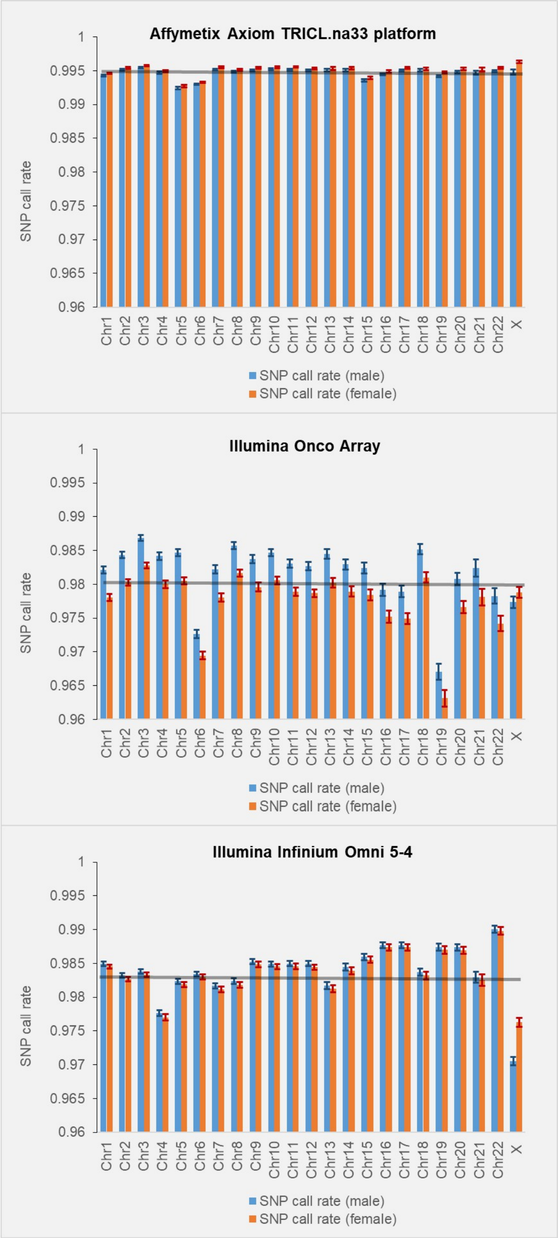

Fig 6. SNP call rates for the SNPs stratified by chromosome. Call rates in males and females are computed and shown separately. Error bars indicate standard errors (SEs).

https://doi.org/10.1371/journal.pgen.1010472.g006

Call rates for X-chromosomal and autosomal SNPs

One of the suggested "technical" explanations of the low density of GWAS hits on X-chromosome states that ". . .there might be problems with genotype calling for hemizygous males as a result of the lower intensity of some X chromosome variants, and so such males might cluster differently than females." [11]. To test this conjecture we revisited call rate data from several GWASs we have conducted or participated in [17–20]. Call rates were estimated for individual chromosomes and stratified by gender (Fig 6). For Affymetrix and OncoArray, call rates for X-chromosomal SNPs did not differ from the average rate call for autosomes. For Infinium

Omni array the call rate for X-chromosomal SNPs was slightly lower compared to average for autosomes: 0.971±0.002 versus 0.984±0.001. It is hard to say if this specifically relates to X-chromosome or it is a part of global inter-chromosomal differences in call rates, for example chromosome 4 also shows a slightly lower call rate compared to other autosomes. In any case, 1–2% differences in the SNP call rates cannot explain 6-fold differences in the density of GWAS-detected SNPs. The call rates were similar between autosomal and X-chromosomal SNPs both in males and females on other genotyping platforms also (personal report from Dr. Doheny (Johns Hopkins)).

Testing X-chromosomal SNPs for Hardy Weinberg Equilibrium

Testing SNPs for deviation from Hardy Weinberg Equilibrium (HWE) is a common QC test for genotyping quality. Including males in HWE test for X-chromosomal SNPs may lead to their preferential exclusion from the analysis because X-chromosomal SNPs are frequently deviate from HWE due to hemizygosity. The Initially recommended approach to deal with this issue is to run HWE test only in females [21–23]. More recently, an exact statistical test based on conditional distribution of the number of heterozygotes given the minor allele count has been proposed [24]. The test takes both male and female genotypes into account. For some SNPs the results from female-restricted HWE testing versus all-individuals-included testing are different, but overall the differences are not substantial (see Fig 7 in [24].) We randomly reviewed 20 GWAS studies published in the last 5 year and found none including males in HWE testing. The list of the GWASes randomly selected to test if they indicated including males in HWE testing can be found in S5 Table.

The density of genic SNPs is lower on X-chromosome compared to autosomes, while the density of intergenic SNPs is similar on X-chromosome and autosomes

We hypothesized that the density of functional SNPs is higher on autosomes compared to X-chromosome. To test this hypothesis we stratified dbSNP SNPs into those located in genic and intergenic regions and estimated their densities for each chromosome separately. SNPs located in genic regions are more likely to be functional compared to intergenic SNPs [25]. The latter is supported by the observation that SNPs in genic regions explain more phenotypic variation [26] and have a higher GWAS reproducibility rate compared to intergenic SNPs [25]. The density of SNPs located in genic regions of autosomes was higher than that of X-chromosome (Z score = 8.2, P = 8.7x10^{-16}), while the density of intergenic SNPs was similar for X-chromosome and autosomes (Z score = 1.2, P = 0.20).

Since the dbSNP database includes SNPs identified by targeted sequencing, this potentially can lead to a bias–a higher SNP density in more frequently targeted regions. Also because the data provided to the dbSNP database are submitted by individual users, in some cases this can lead to a bias against X-chromosome SNPs when the authors use inadequate variant calling methods [27]. To deal with this issue we have conducted an analysis using only SNPs detected by whole genome sequencing: SNPs from the 1000 Genomes Project (1KG) and SNPs from TopMED database. Fig 7 shows the densities of SNPs in intergenic, intronic and exonic regions separately for all autosomes and X-chromosome. Both 1KG and TOPMed SNPs' densities in intergenic regions were similar on X-chromosome and autosomes. For 1KG, SNP densities of intergenic SNPs were 2.7±0.1 and 2.2, respectively (p = 0.11). For intronic SNPs, the densities for autosomes and X-chromosome were 3.0±0.1 and 2.1 correspondingly (p = 1.2x10^{-6}). For exonic SNPs, the difference between autosomes and X-chromosome was

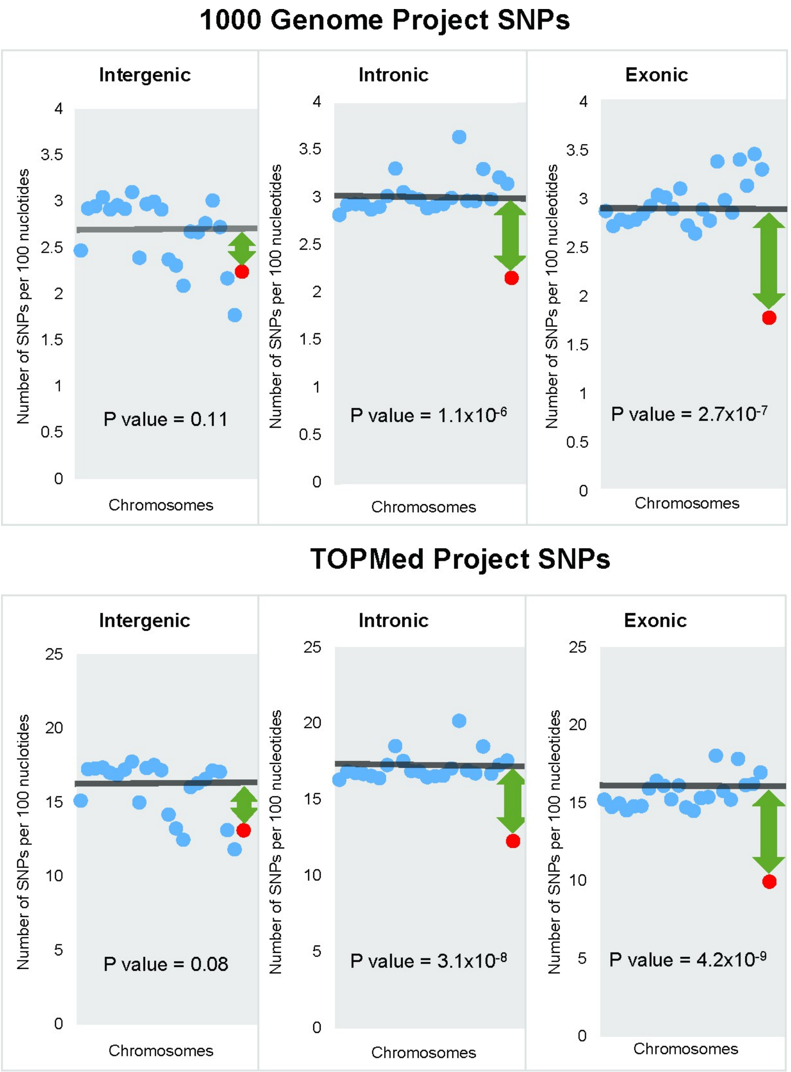

Fig 7. The densities of 1KG and TOPMed SNPs in intergenic, intronic and exonic regions. Each dot represents a chromosome: blue dots–autosomes and red dot–X- chromosome. Horizontal lines show the mean SNP density for all autosomes. Green arrows depict the difference between the mean autosomal and X-chromosome densities. Positions of chromosomes on the Fig from left to right correspond to chromosomal numbers.

https://doi.org/10.1371/journal.pgen.1010472.g007

slightly higher: 3.0±0.1 and 1.8, correspondingly ($p = 2.2x10^{-7}$). The results for TOPMed SNPs were similar to the results for 1KG SNPs (lower panel of Fig 7).

Proportions on segregating sites for missense mutations stratified by predicted effect on protein function

The ratio of segregating sites to the number of potential sites was used as a measure of purifying selection [28]. Sites under strong selection are expected to be less polymorphic. The results of the analysis stratified by chromosome are shown on Fig 8. Proportions of segregating sites increase with a higher Envision score. There are striking differences in the proportion of segregating sites between X-chromosome and autosomes: X-chromosome has a consistently lower proportion of segregating sites compare to autosomes. This is consistent with the idea that selection against X-chromosome missense mutations is stronger compared to autosomal missenses mutations.

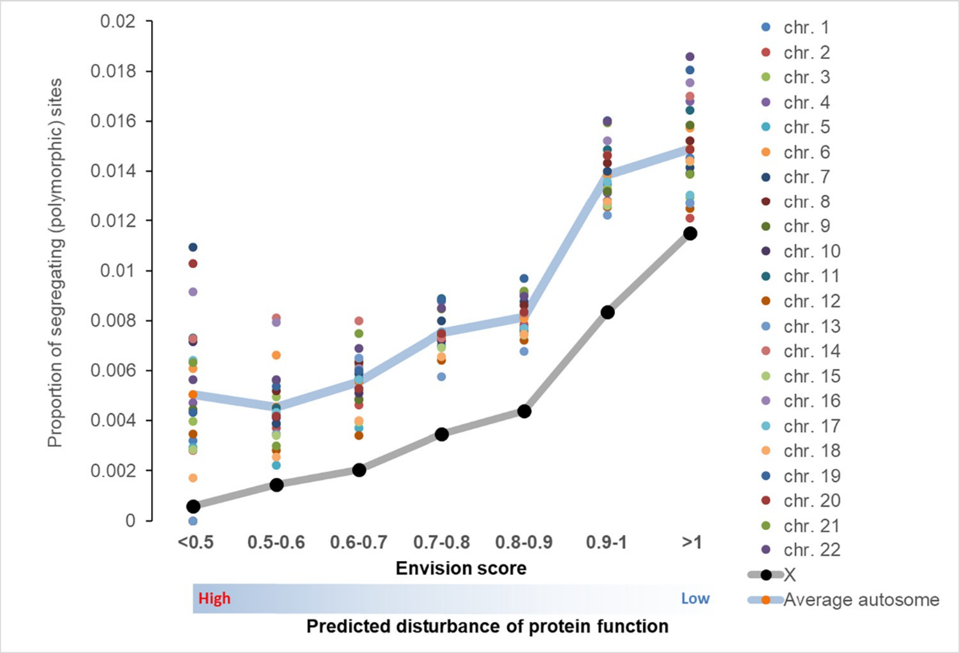

Fig 8. The proportion of segregating sites for missense mutations categorized by Envision score. The lower envision score is associated with a stronger effect on protein function. Of note, there is a higher inter chromosomal variation in the proportions of segregating sites on the tails of the distribution because of small sample sizes for those categories (see S6 Table).

https://doi.org/10.1371/journal.pgen.1010472.g008

Pseudoautosomal region 1 (PAR1)

We looked at the distribution of SNPs along X-chromosome (Fig 9). We binned the X-chromosome into 52 3-megabase segments. We chose this segment size because it is close to the size of the PAR1 region which is 2,781,479 bases. The total number of dbSNP SNPs in PAR1 is 63,848, which is almost 20 times lower compared to the average number of SNPs in other regions of X-chromosome– 828,112. A total of 15 GWAS-detected SNPs are located in PAR1,

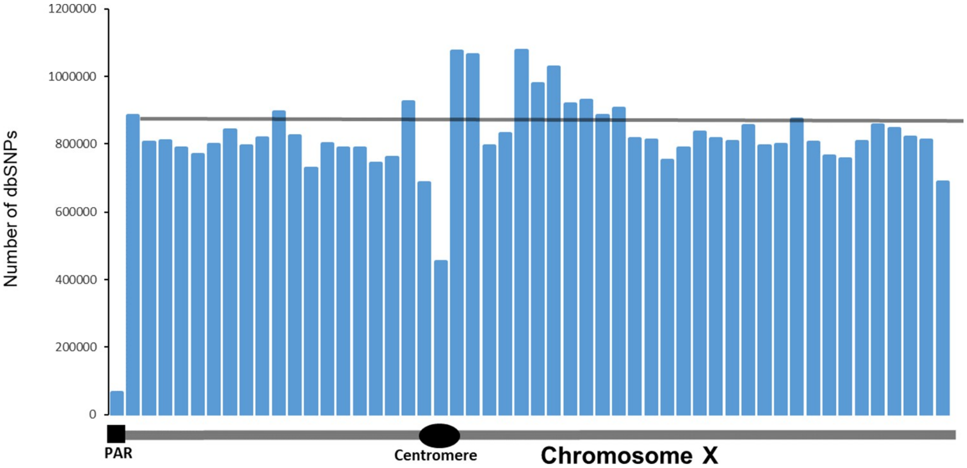

Fig 9. The distribution of the number of SNPs reported in dbSNP database. Each bar represents a 3-megabase fragment of X-chromosome. The first bar corresponds to the position and size of the pseudoautosomal region 1. Thin gray horizontal line represents the average number of SNPs in all bars except PAR1.

https://doi.org/10.1371/journal.pgen.1010472.g009

which translates into 6.7 SNPs per megabase. The estimated number of GWAS-detected SNPs per thousand of dbSNP variants in PAR1 is 0.21, which is comparable to the density of GWAS detected SNPs on autosomes– 0.17. Therefore, the fraction of GWAS-detected SNPs in the pseudoautosomal region 1 of X-chromosome is similar to that for autosomes. One cautionary note, however, is that the total number of SNPs in PAR1 region is small and evolutionary dynamics of PAR regions differ from the rest of X-chromosome [29]. These limitations call for revisited analysis of PAR regions when more data become available.

Discussion

The comprehensive analysis we conducted has found that the density of GWAS-detected SNPs on X-chromosome is 5-6-fold lower compared to that on autosomes. The difference cannot be explained by the overall lower density of SNPs on X-chromosome, a lower coverage of the X-chromosome by genotyping platforms, or an inferior call rate for X-chromosomal SNPs. Based on available publications [30,31] and our analysis (Fig 6), call rates for X-chromosome SNPs are similar to call rates for autosomal SNPs. The lower density of GWAS-detected SNPs on X-chromosome is not a result of a lower statistical power because female-restricted GWASs show the same level of the differences as all-phenotype GWASs.

Imputation is a commonly used approach to increase the number of SNPs analyzed in GWA studies [32,33]. Based on the results of a recent study [34], imputation accuracy is essentially the same for autosomal and X-chromosome SNPs: 89.3% versus 90.2. This suggests that imputation accuracy is unlikely to be the source of observed conspicuous differences in GWAS hit densities between X-chromosome and autosomes.

The other analytical technique frequently used in GWA studies is meta-analysis [35,36]. Meta-analysis improves the statistical power by combining statistics from several GWA studies. It is unlikely that this technique will contribute to X-chromosome versus autosomal differences in the GWAS hit densities since SNPs are included in meta-analysis regardless of their chromosomal location.

The other potential contributor to X-chromosome versus autosome differences in the density of GWAS findings is sex differences in allele frequencies. A recent study by Wang et al. [37] found that about 1% of SNPs on the X chromosome show sex differences in minor allele frequencies (MAFs). The study identified 2,039 SNPs with genome level differences in MAFs between males and females. We checked how many of those SNPs are reported by GWASs. We found that only one SNP from the list, rs4014653 was reported as a GWAS hit [38]. The frequency of GWAS-detected SNPs among SNPs with sex differences in MAF– 1/2039 is comparable with the fraction of GWAS-detected SNPs among all known SNPs (Fig 2). We estimated how many SNPs with sex differences in MAFs are on commonly used Illumina Humanomni5-4v1_c platform. We found that 21 SNPs are on the platform; therefore, the frequency of GWAS hits among directly genotyped SNPs is 1/21 = 0.05 which is similar to the overall frequency of GWAS hits on the platform, 0.08 (Fig 4). Thus it is unlikely that sex differences in MAFs is a major contributor to the deficit of X-chromosome SNPs in GWASs. X-chromosome SNPs do not show a higher deviation from HWE compared to autosomal SNPs. In fact, Graffelman et al. (2017) [39] found that the deviation from HWE is higher for autosomal SNPs. This finding, however, needs to be taken with caution because the study was based on the data from 1000 Genomes samples with a relatively low sequencing depth.

We hypothesized that the striking differences between X-chromosome and autosomes in the density of GWAS-detected SNPs is a result of a lower density of functional SNPs on the X-chromosome. Mutations on X-chromosome are more exposed to natural selection compared to autosomal mutations because of their hemizygosity in males and X-chromosome inactivation in females [40]. As a result, both positive and negative selection is stronger on X-chromosome–faster-X hypothesis [3]. A number of studies provide a strong support for faster-X hypothesis [3,4,41]. Consistent with the idea of stronger negative selection on X-chromosome that "weeds out" functional variants from it, the study by Kukurba et al. [42] reported a significant depletion of expression quantitative trait loci (eQTL) on X-chromosome. Lower genetic variation on X-chromosome compared to autosomes was first reported by the SNP consortium [43] and later confirmed by other studies [44,45].

The analysis of proportions of segregating sites for the 7 categories of missense mutations stratified by the predicted effect on the protein function demonstrated that (i) the proportion of segregating sites increases as the predicted functional effect of missense mutations decreases for all chromosomes, and that (ii) the proportion of segregating sites for X-chromosome mutations is lower than the proportion of segregating sites for missense mutations in autosomal genes. The results provide support for the hypothesis of lower density of functional polymorphisms on X-chromosome as a result of stronger purifying selection.

Since the majority of *de novo* mutations are deleterious [46], purifying selection is the most prevalent [47] pressure on new variants. Purifying selection against deleterious mutations may change frequencies of the linked SNPs. This form of the linked selection is called background selection [48–50]. Background selection is widespread and significantly influences evolutionary dynamics of the linked SNPs regardless of their functional significance [50,51]. Therefore, changes in frequency of functional polymorphisms (e.g. missense mutations) usually change frequencies of linked nonfunctional SNPs (e.g. synonymous SNPs). A stronger selection against X-chromosomal mutations compared to autosomal ones is expected to result in a lower density of functional polymorphisms on X-chromosome. Neutral mutations are not subject to selection and, therefore, their frequencies will be similar on X-chromosome and autosomes. The results of our analysis are consistent with this expectation. First, the stronger X-chromosomal selection is expected to result in a lower density of SNPs on X-chromosome. That is exactly what we observed: the overall densities of SNPs from dbSNP, 1KG, and TOPMed databases were correspondingly 23%, 20% and 21% lower on X-chromosome compared to autosomes. Second, we found that the density of X-chromosomal SNPs located in transcribed regions is significantly lower compared to autosomes, while the densities of intergenic SNPs are similar for X-chromosome and autosomes. SNPs located in genic regions are more likely to be functional compared to intergenic [25,26] even though some of the intergenic SNPs also can be functional [52,53]. We also noted that the densities of both exonic and intronic SNPs on autosomes were higher compared to the corresponding densities on X-chromosome.

The results of the analysis of PAR1 region are also in line with the hypothesis that the deficit of functional SNPs on X-chromosome results in a deficit of GWAS-detected SNPs on X-chromosome. We found that in the PAR1 the fraction of GWAS-detected SNPs among all dbSNP variants was similar to that of autosomes. The finding provides further support to the hypothesis that the lower density of GWAS-detected SNPs on X-chromosome is a result of a stronger selection against functional SNPs located on X-chromosome resulting from the hemizogous state of X-chromosomal SNPs in males and X-inactivation in females.

SNP density in the PAR1 is more than 10 times lower than the SNP density in other parts of X-chromosome. This difference is surprising because meiotic recombination has been shown to be associated with a higher mutation rate [54] and the recombination rate in PAR1 region is 20-fold higher compared to the genome-wide average [2,55]. GC-biased gene conversion (gBGC) may provide an explanation of the lower SNP density in the PAR1. gBDC is a recombination-associated conversion—unidirectional transfer of genetic information from a 'donor' to an 'acceptor' sequence [56,57]. gBGC accelerates the fixation of guanine or cytosine alleles and its effect is stronger than "mutagenic" effect of recombination [54]. gBGC is similar for functional and neutral polymorphisms because it acts regardless of the effects of the genetic polymorphisms on fitness.

In conclusion, we found that the density of GWAS-detected SNPs on X-chromosome is almost 6-fold lower compared to that on autosomes. Analysis stratified by SNP location in transcribed and intergenic regions supported the idea that the lower density of GWAS-detected SNPs on X-chromosome is a result of a stronger selection against functional SNPs on X-chromosome.

Methods

Data sources

Human Genome Assembly GRCh38.p13 was used to retrieve the sizes of individual chromosomes as well as the sizes of genic and intergenic regions. The dbSNP database Build 155 [58] was used to retrieve the number of SNPs in individual human chromosomes. To control for possible biases in the number of SNPs stemming from the targeted genotyping we used SNP data from two projects that used whole genome sequencing for SNP detection: 1000 Genomes Project phase 3 [59] and the TOPMed [60] project Freeze 5b. Phase 3 1000 Genomes Project has lower coverage of intergenic compared to genic regions [59]. Those differences are unlikely to influence the results of the analyses since proportions of exonic sequences are similar between X-chromosome and autosomes (S1 Table).

We used data from the Catalog of Published GWASs (CPGWAS) [61] to identify SNPs detected by GWAS studies. The catalog was accessed on June 12, 2022. For each chromosome we have identified the total number of SNPs reported by CPGWAS. SNPs detected and reported by multiple GWAS we counted as a single signal even when it was associated with different phenotypes or diseases. S1 Table shows the data on the chromosomal sizes including genic and intergenic regions, the number of SNPs on each chromosome and the number of GWAS-detected SNPs reported in CPGWAS for individual chromosomes. SNPs in the pseudoautosomal region one (PAR1) of the X-chromosome were analyzed separately because of its unique characteristics [29,62]. PAR2 region was not included in this analysis because it is too small and contains too few SNPs.

Stratification by intergenic, genic, intronic and exonic regions

Data from the consensus coding sequence (CCDS) database (release 22) [63] were used to estimate the sizes of exonic, intronic and intergenic regions for individual chromosomes. First we retrieved the start and end exon and intron positions of each coding gene transcript. The sum of the lengths of all exons on a given chromosome was used as a size of the exonic region. We excluded overlapping gene sequences to make sure that each nucleotide was counted only once. The similar approach was used to estimate the total size of intronic regions. Noncoding genes were counted as part of the genic region. Start and end positions of noncoding genes were retrieved from the NCBI Gene database [64]. The size of the intergenic region was computed as the total size of the chromosome minus the combined size of exons, introns and noncoding genes. Genes located on the Y-chromosome and mitochondrial genes were excluded from the analysis.

Possible explanations of the deficit of GWAS-detected SNPs on X-chromosome

We went through all technical/methodological explanations of the X-chromosome versus autosomal differences in the density of GWAS-detected SNPs of which we are aware. We tested whether the following factors could explain the lower density of GWAS-detected SNPs on X-chromosome:

i. *Chromosome sizes.* To account for size differences we computed the number of GWAS-detected SNPs per megabase or kilobase of sequence for each chromosome and used this metric to compare X-chromosome and autosomes.

ii. *Overall SNP density.* Differences in the number of GWAS hits could be a result of differences in the overall SNP density between X-chromosome and autosomes. We estimated the densities of all SNPs reported in dbSNP, 1000 Genomes Project, and TOPMed databases and compared chromosomes by the number of GWAS-associated SNPs per thousand of all SNPs on a given chromosome.

iii. *Coverage by genotyping platforms.* To estimate the coverage of chromosomes by genotyping platforms we analyzed 28 most commonly used genotyping arrays [65]. For each genotyping array we estimated the number of SNPs for individual chromosomes. To account for differences in chromosomal sizes we counted the number of array SNPs per thousand nucleotides and compared X-chromosome and autosomes by this measure.

iv. *Call rates.* Because the amount of DNA available for genotyping X-chromosomal SNPs is halfthe amount compared to autosomal SNPs in men, one could hypothesize that the success rate for genotype calls is lower for X-chromosome SNPs compared to autosomal SNPs. We tested this hypothesis by estimating call rates in several commonly used genotyping platforms.

v. *Effective sample size.* In males the number of analyzed alleles for X-chromosomal SNPs is only half of that for autosomes, therefore, the effective sample size is lower for X-chromosomal SNPs. This may contribute to X-chromosome versus autosomal differences in GWAS hit densities. To address this issue we conducted an analysis of the data from female-only GWASs where effective sample size is the same for X-chromosome and autosomes.

vi. *Density of functional SNPs.* Theoretical and observational studies show that selection on the X-chromosome is stronger than on autosomes [3]. Selection directly affects functional mutations while neutral variants are free from selection pressure. GWAS signals are driven by the presence of functional/causal SNPs. If the density of functional variants is lower on the X-chromosome, the density of GWAS-detected SNPs is also expected to be lower. To test this hypothesis we compared X-chromosome and autosomes by the densities of genic (likely functional) and intergenic (unlikely functional) SNPs.

Proportions of segregating sites for missense mutations with varying effect on protein function

We used Envision bioinformatics tool to predict the effect of missense mutations on protein function [66]. Envision combines experimental data on functional effects of missense mutations with a supervised, stochastic gradient boosting learning algorithm to quantify functional effects of missense mutations. Envision outperforms similar bioinformatics tools for prediction of functional effects [66]. Envision provides precomputed predictions for every possible single amino acid substitution in the human proteome. Among all possible amino acid substitutions (19 possible substations per amino acid) we identified the substitutions that can be produced by a single nucleotide substitution (SNS). For this we computationally mutated each single nucleotide in the coding sequence into 3 possible SNSs and identified those producing missense mutations (see [67] for details). Possible missense mutations were stratified into 7 categories based on the Envision score–the lower score is associated with a stronger functional disturbance. Therefore, for each gene we estimated the number of potential sites for each category. We used 1000 Genomes Project data to identify how many of the potential sites exits as SNPs (S6 Table).

Statistical analysis

We considered all autosomes as a group and tested the hypothesis that the X-chromosome belongs to this group. For each metric analyzed, e.g. the number of GWAS-detected SNPs per one million nucleotides, we computed the mean and standard deviation (σ) for autosomes. A standard Z-score was computed using the standard formula: $Z = \frac{x-\mu}{\sigma}$; where Z is a standard Z-score, x is the value of the metric for X-chromosome, μ –the mean value for autosomes, and σ is standard deviation for the metric in autosomes. P-values corresponding to the given Z-scores were used for comparisons between X-chromosome and autosomes.

Supporting information

S1 Table. Chromosomal sizes, number of GWAS detected SNPs and number of SNPs reported in dbSNP, 1K genomes and TOPMed databases for individual chromosomes.
(XLS)

S2 Table. Number of SNPs on chromosomes across 28 commonly used genotyping platforms.
(XLS)

S3 Table. Number of GWAS-detected SNPs per 100 SNPs on genotyping platform.
(XLS)

S4 Table. List of female only GWAS.
(XLS)

S5 Table. Genome wide association studies randomly selected to find if they mentioned including males in HWE testing.
(DOCX)

S6 Table. Number of potential and segregating sites and the fraction of segregating sites per potential site for missense mutations stratified by the predicted effect on protein function (Envision Score).
(XLS)

Acknowledgments

We thank Mr. Spyridon Tsavachidis for help with analysis and Dr. Gorlova for discussions of the results.

Author Contributions

Conceptualization: Christopher I. Amos.

Formal analysis: Ivan P. Gorlov.

Funding acquisition: Christopher I. Amos.

Methodology: Ivan P. Gorlov, Christopher I. Amos.

Writing – original draft: Ivan P. Gorlov.

Writing – review & editing: Ivan P. Gorlov, Christopher I. Amos.

References

1. Schaffner SF. The X chromosome in population genetics. Nat Rev Genet. 2004; 5(1):43–51. https://doi.org/10.1038/nrg1247 PMID: 14708015.
2. Hinch AG, Altemose N, Noor N, Donnelly P, Myers SR. Recombination in the human Pseudoautosomal region PAR1. PLoS Genet. 2014; 10(7):e1004503. Epub 2014/07/18. https://doi.org/10.1371/journal.pgen.1004503 PMID: 25033397; PubMed Central PMCID: PMC4102438.
3. Meisel RP, Connallon T. The faster-X effect: integrating theory and data. Trends Genet. 2013; 29 (9):537–44. Epub 2013/06/25. https://doi.org/10.1016/j.tig.2013.05.009 PMID: 23790324; PubMed Central PMCID: PMC3755111.
4. Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. Mol Biol Evol. 2014; 31(9):2267–82. Epub 2014/05/17. https://doi.org/10.1093/molbev/msu166 PMID: 24830675; PubMed Central PMCID: PMC4137703.
5. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019; 20(8):467–84. Epub 2019/05/10. https://doi.org/10.1038/s41576-019-0127-1 PMID: 31068683.
6. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012; 90(1):7–24. https://doi.org/10.1016/j.ajhg.2011.11.029 PMID: 22243964; PubMed Central PMCID: PMC3257326.
7. Bosse Y, Amos CI. A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol Biomarkers Prev. 2018; 27(4):363–79. https://doi.org/10.1158/1055-9965.EPI-16-0794 PMID: 28615365; PubMed Central PMCID: PMC6464125.
8. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet. 2018; 102(5):717–30. Epub 2018/05/05. https://doi.org/10.1016/j.ajhg.2018.04.002 PMID: 29727686; PubMed Central PMCID: PMC5986732.
9. Horwitz T, Lam K, Chen Y, Xia Y, Liu C. A decade in psychiatric GWAS research. Mol Psychiatry. 2019; 24(3):378–89. Epub 2018/06/27. https://doi.org/10.1038/s41380-018-0055-z PMID: 29942042; PubMed Central PMCID: PMC6372350.
10. Liang B, Ding H, Huang L, Luo H, Zhu X. GWAS in cancer: progress and challenges. Mol Genet Genomics. 2020; 295(3):537–61. Epub 2020/02/13. https://doi.org/10.1007/s00438-020-01647-z PMID: 32048005.
11. Wise AL, Gyi L, Manolio TA. eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am J Hum Genet. 2013; 92(5):643–7. Epub 2013/05/07. https://doi.org/10.1016/j.ajhg.2013.03.017 PMID: 23643377; PubMed Central PMCID: PMC3644627.
12. Clayton D. Testing for association on the X chromosome. Biostatistics. 2008; 9(4):593–600. Epub 2008/04/29. https://doi.org/10.1093/biostatistics/kxn007 PMID: 18441336; PubMed Central PMCID: PMC2536723.
13. Zheng G, Joo J, Zhang C, Geller NL. Testing association for markers on the X chromosome. Genet Epidemiol. 2007; 31(8):834–43. Epub 2007/06/06. https://doi.org/10.1002/gepi.20244 PMID: 17549761.
14. Accounting for sex in the genome. Nat Med. 2017; 23(11):1243. Epub 2017/11/09. https://doi.org/10.1038/nm.4445 PMID: 29117171.
15. Gao F, Chang D, Biddanda A, Ma L, Guo Y, Zhou Z, et al. XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. J Hered. 2015; 106(5):666–71. Epub 2015/08/14. https://doi.org/10.1093/jhered/esv059 PMID: 26268243; PubMed Central PMCID: PMC4567842.
16. Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, et al. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PLoS One. 2014; 9(12):e113684. Epub 2014/12/06. https://doi.org/10.1371/journal.pone.0113684 PMID: 25479423; PubMed Central PMCID: PMC4257614.
17. Landi MT, Bishop DT, MacGregor S, Machiela MJ, Stratigos AJ, Ghiorzo P, et al. Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. Nat Genet. 2020; 52(5):494–504. Epub 2020/04/29. https://doi.org/10.1038/s41588-020-0611-8 PMID: 32341527; PubMed Central PMCID: PMC7255059.
18. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across

histological subtypes. Nat Genet. 2017; 49(7):1126–32. https://doi.org/10.1038/ng.3892 PMID: 28604730; PubMed Central PMCID: PMC5510465.

19. Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. J Natl Cancer Inst. 2019; 111(2):146–57. Epub 2018/06/20. https://doi.org/10.1093/jnci/djy099 PMID: 29917119; PubMed Central PMCID: PMC6555904.

20. Zhou W, Liu G, Hung RJ, Haycock PC, Aldrich MC, Andrew AS, et al. Causal relationships between body mass index, smoking and lung cancer: Univariable and multivariable Mendelian randomization. Int J Cancer. 2021; 148(5):1077–86. Epub 2020/09/12. https://doi.org/10.1002/ijc.33292 PMID: 32914876; PubMed Central PMCID: PMC7845289.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–75. Epub 2007/08/19. https://doi.org/10.1086/519795 PMID: 17701901; PubMed Central PMCID: PMC1950838.

22. Konig IR, Loley C, Erdmann J, Ziegler A. How to include chromosome X in your genome-wide association study. Genet Epidemiol. 2014; 38(2):97–103. Epub 2014/01/11. https://doi.org/10.1002/gepi.21782 PMID: 24408308.

23. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics. 2012; 28(24):3329–31. Epub 2012/10/12. https://doi.org/10.1093/bioinformatics/bts610 PMID: 23052040; PubMed Central PMCID: PMC3519456.

24. Graffelman J, Weir BS. Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity (Edinb). 2016; 116(6):558–68. Epub 2016/04/14. https://doi.org/10.1038/hdy.2016.20 PMID: 27071844; PubMed Central PMCID: PMC4868269.

25. Smith EN, Koller DL, Panganiban C, Szelinger S, Zhang P, Badner JA, et al. Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. PLoS Genet. 2011; 7(6):e1002134. Epub 2011/07/09. https://doi.org/10.1371/journal.pgen.1002134 PMID: 21738484; PubMed Central PMCID: PMC3128104.

26. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43(6):519–25. Epub 2011/05/10. https://doi.org/10.1038/ng.823 PMID: 21552263; PubMed Central PMCID: PMC4295936.

27. Webster TH, Couse M, Grande BM, Karlins E, Phung TN, Richmond PA, et al. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. Gigascience. 2019; 8(7). Epub 2019/07/11. https://doi.org/10.1093/gigascience/giz074 PMID: 31289836; PubMed Central PMCID: PMC6615978.

28. Zivkovic D, Wiehe T. Second-order moments of segregating sites under variable population size. Genetics. 2008; 180(1):341–57. Epub 2008/08/22. https://doi.org/10.1534/genetics.108.091231 PMID: 18716326; PubMed Central PMCID: PMC2535686.

29. Monteiro B, Arenas M, Prata MJ, Amorim A. Evolutionary dynamics of the human pseudoautosomal regions. PLoS Genet. 2021; 17(4):e1009532. Epub 2021/04/20. https://doi.org/10.1371/journal.pgen.1009532 PMID: 33872316; PubMed Central PMCID: PMC8084340.

30. Huentelman MJ, Craig DW, Shieh AD, Corneveaux JJ, Hu-Lince D, Pearson JV, et al. SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. BMC Genomics. 2005; 6:149. Epub 2005/11/03. https://doi.org/10.1186/1471-2164-6-149 PMID: 16262895; PubMed Central PMCID: PMC1280925.

31. Bertolini F, Elbeltagy AR, Rothschild MF. Evaluation of the application of bovine, ovine and caprine SNP chips to dromedary genotyping. Livestock research for rural development. 2017; 29:31.

32. Naj AC. Genotype Imputation in Genome-Wide Association Studies. Curr Protoc Hum Genet. 2019; 102(1):e84. Epub 2019/06/20. https://doi.org/10.1002/cphg.84 PMID: 31216114.

33. Porcu E, Sanna S, Fuchsberger C, Fritsche LG. Genotype imputation in genome-wide association studies. Curr Protoc Hum Genet. 2013;Chapter 1:Unit 1 25. Epub 2013/07/16. https://doi.org/10.1002/0471142905.hg0125s78 PMID: 23853078.

34. Schurz H, Muller SJ, van Helden PD, Tromp G, Hoal EG, Kinnear CJ, et al. Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. Front Genet. 2019; 10:34. Epub 2019/02/26. https://doi.org/10.3389/fgene.2019.00034 PMID: 30804980; PubMed Central PMCID: PMC6370942.

35. Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JP. The power of meta-analysis in genome-wide association studies. Annu Rev Genomics Hum Genet. 2013; 14:441–65. Epub 2013/06/04. https://doi.org/10.1146/annurev-genom-091212-153520 PMID: 23724904; PubMed Central PMCID: PMC4040957.

36. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Pharmacogenomics. 2009; 10(2):191–201. Epub 2009/02/12. https://doi.org/10.2217/14622416.10.2.191 PMID: 19207020; PubMed Central PMCID: PMC2695132.

37. Wang Z, Sun L, Paterson AD. Major sex differences in allele frequencies for X chromosomal variants in both the 1000 Genomes Project and gnomAD. PLoS Genet. 2022; 18(5):e1010231. Epub 2022/06/01. https://doi.org/10.1371/journal.pgen.1010231 PMID: 35639794; PubMed Central PMCID: PMC9187127.

38. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell. 2020; 182(5):1214–31 e11. Epub 2020/09/06. https://doi.org/10.1016/j.cell.2020.08.008 PMID: 32888494; PubMed Central PMCID: PMC7482360.

39. Graffelman J, Jain D, Weir B. A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. Human genetics. 2017; 136(6):727–41. Epub 2017/04/05. https://doi.org/10.1007/s00439-017-1786-7 PMID: 28374190; PubMed Central PMCID: PMC5429372 conflict of interest.

40. Pereira G, Doria S. X-chromosome inactivation: implications in human disease. J Genet. 2021; 100. Epub 2021/09/24. PMID: 34553695.

41. Charlesworth B, Campos JL, Jackson BC. Faster-X evolution: Theory and evidence from Drosophila. Mol Ecol. 2018; 27(19):3753–71. Epub 2018/02/13. https://doi.org/10.1111/mec.14534 PMID: 29431881.

42. Kukurba KR, Parsana P, Balliu B, Smith KS, Zappala Z, Knowles DA, et al. Impact of the X Chromosome and sex on regulatory variation. Genome Res. 2016; 26(6):768–77. Epub 2016/05/20. https://doi.org/10.1101/gr.197897.115 PMID: 27197214; PubMed Central PMCID: PMC4889977.

43. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409(6822):928–33. Epub 2001/03/10. https://doi.org/10.1038/35057149 PMID: 11237013.

44. Wilson Sayres MA. Genetic Diversity on the Sex Chromosomes. Genome Biol Evol. 2018; 10(4):1064–78. Epub 2018/04/11. https://doi.org/10.1093/gbe/evy039 PMID: 29635328; PubMed Central PMCID: PMC5892150.

45. Arbiza L, Gottipati S, Siepel A, Keinan A. Contrasting X-linked and autosomal diversity across 14 human populations. Am J Hum Genet. 2014; 94(6):827–44. Epub 2014/05/20. https://doi.org/10.1016/j.ajhg.2014.04.011 PMID: 24836452; PubMed Central PMCID: PMC4121480.

46. Keightley PD. Rates and fitness consequences of new mutations in humans. Genetics. 2012; 190 (2):295–304. Epub 2012/02/22. https://doi.org/10.1534/genetics.111.134668 PMID: 22345605; PubMed Central PMCID: PMC3276617.

47. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A. 2003; 100(26):15754–7. Epub 2003/12/09. https://doi.org/10.1073/pnas.2536718100 PMID: 14660790; PubMed Central PMCID: PMC307640.

48. Stephan W. Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc Lond B Biol Sci. 2010; 365(1544):1245–53. Epub 2010/03/24. https://doi.org/10.1098/rstb.2009.0278 PMID: 20308100; PubMed Central PMCID: PMC2871815.

49. Hodgkinson A, Casals F, Idaghdour Y, Grenier JC, Hernandez RD, Awadalla P. Selective constraint, background selection, and mutation accumulation variability within and between human populations. BMC Genomics. 2013; 14:495. Epub 2013/07/24. https://doi.org/10.1186/1471-2164-14-495 PMID: 23875710; PubMed Central PMCID: PMC3727949.

50. Charlesworth B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J Hered. 2013; 104(2):161–71. Epub 2013/01/11. https://doi.org/10.1093/jhered/ess136 PMID: 23303522.

51. Wright SI. Charlesworth et al. on Background Selection and Neutral Diversity. Genetics. 2016; 204 (3):829–32. Epub 2017/01/24. https://doi.org/10.1534/genetics.116.196170 PMID: 28114095; PubMed Central PMCID: PMC5105860.

52. Chen J, Tian W. Explaining the disease phenotype of intergenic SNP through predicted long range regulation. Nucleic Acids Res. 2016; 44(18):8641–54. Epub 2016/06/10. https://doi.org/10.1093/nar/gkw519 PMID: 27280978; PubMed Central PMCID: PMC5062962.

53. Schierding W, Antony J, Cutfield WS, Horsfield JA, O'Sullivan JM. Intergenic GWAS SNPs are key components of the spatial and regulatory network for human growth. Hum Mol Genet. 2016; 25 (15):3372–82. Epub 2016/06/12. https://doi.org/10.1093/hmg/ddw165 PMID: 27288450.

54. Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc Natl Acad Sci U S A. 2015; 112(7):2109–14. Epub 2015/02/04. https://doi.org/10.1073/pnas.1416622112 PMID: 25646453; PubMed Central PMCID: PMC4343121.

55. Chen JF, Lu F, Chen SS, Tao SH. Significant positive correlation between the recombination rate and GC content in the human pseudoautosomal region. Genome. 2006; 49(5):413–9. Epub 2006/06/13. https://doi.org/10.1139/g05-124 PMID: 16767166.

56. Dutta R, Saha-Mandal A, Cheng X, Qiu S, Serpen J, Fedorova L, et al. 1000 human genomes carry widespread signatures of GC biased gene conversion. BMC Genomics. 2018; 19(1):256. Epub 2018/04/18. https://doi.org/10.1186/s12864-018-4593-1 PMID: 29661137; PubMed Central PMCID: PMC5902838.

57. Kostka D, Hubisz MJ, Siepel A, Pollard KS. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Mol Biol Evol. 2012; 29(3):1047–57. Epub 2011/11/15. https://doi.org/10.1093/molbev/msr279 PMID: 22075116; PubMed Central PMCID: PMC3278478.

58. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11. Epub 2000/01/11. https://doi.org/10.1093/nar/29.1.308 PMID: 11125122; PubMed Central PMCID: PMC29783.

59. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022; 185(18):3426–40 e19. Epub 2022/09/03. https://doi.org/10.1016/j.cell.2022.08.004 PMID: 36055201; PubMed Central PMCID: PMC9439720.

60. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021; 590(7845):290–9. Epub 2021/02/12. https://doi.org/10.1038/s41586-021-03205-y PMID: 33568819; PubMed Central PMCID: PMC7875770.

61. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019; 47(D1):D1005–D12. https://doi.org/10.1093/nar/gky1120 PMID: 30445434; PubMed Central PMCID: PMC6323933.

62. Charlesworth D. When and how do sex-linked regions become sex chromosomes? Evolution. 2021; 75(3):569–81. Epub 2021/02/17. https://doi.org/10.1111/evo.14196 PMID: 33592115.

63. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. 2018; 46(D1):D221–D8. Epub 2017/11/11. https://doi.org/10.1093/nar/gkx1031 PMID: 29126148; PubMed Central PMCID: PMC5753299.

64. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015; 43(Database issue):D36–42. Epub 2014/10/31. https://doi.org/10.1093/nar/gku1055 PMID: 25355515; PubMed Central PMCID: PMC4383897.

65. Verlouw JAM, Clemens E, de Vries JH, Zolk O, Verkerk A, Am Zehnhoff-Dinnesen A, et al. A comparison of genotyping arrays. Eur J Hum Genet. 2021; 29(11):1611–24. Epub 2021/06/19. https://doi.org/10.1038/s41431-021-00917-7 PMID: 34140649; PubMed Central PMCID: PMC8560858.

66. Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Syst. 2018; 6(1):116–24 e3. Epub 2017/12/12. https://doi.org/10.1016/j.cels.2017.11.003 PMID: 29226803; PubMed Central PMCID: PMC5799033.

67. Gorlova Olga KM, Tsavachidis Spiridon, Amos Christopher, Gorlov Ivan. Identification of lung cancer drivers by comparison of the observed and the expected numbers of missense and nonsense mutations in individual human genes. Oncotarget. 2022; V14(1):17–29. https://doi.org/10.18632/oncotarget.28231 PMID: 35634240