**XWAS: a software toolset for genetic data analysis and association studies of the X chromosome**

Feng Gao$^{1,5}$, Diana Chang$^{1,2,5}$, Arjun Biddanda$^{1,5}$, Li Ma$^{1,3}$, Yingjie Guo$^{1,4}$, Zilu Zhou$^1$, and Alon Keinan$^{1,2,*}$

$^1$Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

$^2$Program in Computational Biology and Medicine, Cornell University, Ithaca, NY 14853, USA

$^3$Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20740, USA

$^4$School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

$^5$These authors contributed equally to this work

*To whom correspondence should be addressed: ak735@cornell.edu*
Abstract

XWAS is a new software for the analysis of the X chromosome in association studies and similar studies. The X chromosome plays an important role in human disease, especially those with sexually dimorphic characteristics. Special attention needs to be given to its analysis due to the unique inheritance pattern, leading to analytical complications that have resulted in the majority of genome-wide association studies (GWAS) either not considering X or mishandling it with GWAS toolsets that have been designed for non-sex chromosomes. Hence, XWAS fills the need for tools that are specially designed for analysis of X. Following extensive, stringent, and X-specific quality control, XWAS offers an array of statistical tests of association, including: (1) the standard test between a SNP (single nucleotide polymorphism) and disease risk, including after first stratifying individuals by sex, (2) a test for a differential effect of a SNP on disease between males and females, (3) motivated by X-inactivation, a test for higher variance of a trait in heterozygous females as compared to homozygous females, and (4) for all tests, a version that allows for combining evidence across all SNPs in a whole gene. We applied the toolset analysis pipeline to 16 GWAS datasets of immune-related disorders and to 7 risk factors of coronary artery disease, and discovered several new X-linked genetic associations. XWAS will provide the tools and incentive for others to incorporate the X chromosome into GWAS, hence enabling discoveries of novel loci implicated in many diseases and in their sexual dimorphism.

Key words: GWAS, software, genetic association study, sexual dimorphism, complex diseases
Introduction

Genome-wide association studies (GWAS) have identified thousands of loci underlying complex human diseases and other complex traits (Welter, et al., 2014). While very successful for the autosomes (non-sex chromosomes), the vast majority of these studies have either incorrectly analyzed or completely ignored the X chromosome (X) (Wise, Gyi and Manolio, 2013). In most studies, all variants on the X chromosomes have been removed as a consequence of the quality control (QC) procedures (Chang, et al., 2014; Mailman, et al., 2007; Tryka, et al., 2014; Wise, et al., 2013). Many other studies that did analyze the X chromosome incorrectly applied methods that have been designed for the autosomes, without accounting for the analytical problems arising from X’s unique mode of inheritance (i.e. males only bear one copy of X, while one of the two copies in females is transcriptionally silenced via X-inactivation) and consequentially from X’s population genetics patterns and evolution forces that differs from that of autosome’s (Emery, Felsenstein and Akey, 2010; Lambert, et al., 2010; Wilson and Makova, 2009). As a result, the role X plays in complex diseases and traits remains largely unknown.

Many human diseases commonly studied in GWAS show sexual dimorphism, including autoimmune diseases (Voskuhl, 2011), cardiovascular diseases (Lerner and Kannel, 1986) and cancer (Matanoski, et al., 2006; Muscat, et al., 1996), which suggests a potential contribution of X (Carrel and Willard, 2005; Ober, Loisel and Gilad, 2008). Several recent studies have examined this issue and further demonstrated the value of analyzing X (Chang, et al., 2014; Gilks, Abbott and Morrow, 2014; Li, et al.; Ma, Hoffman and Keinan, 2015; Tukiainen, et al., 2014). However, while association methods, QC and analysis pipelines are well established for analysis of autosomes, respective pipelines for X-linked data are not widely available. Hence, in
this paper, we introduce the software package XWAS (chromosome X-Wide Analysis toolSet), which is tailored for analysis of genetic variation on X. It implements extensive functionality that carries out QC specially-designed for the X chromosome, statistical tests of single-marker association that account for the unique mode of inheritance, gene-based tests of association and additional distinct tests that capitalize on the mode of inheritance that are only applicable to X. In implementing these features, the toolset builds on--and complements--the commonly-used PLINK (Purcell, et al., 2007) software. Combined, the XWAS toolset integrates X into GWAS as well as into the next generation of sequence-based association studies.
Features and Functionality

Quality Control Procedures

The XWAS toolset implements a whole pipeline for performing QC on genotype data for the X chromosome. The pipeline first follows standard GWAS QC steps as implemented in PLINK (Purcell, et al., 2007) and SMARTPCA (Price, et al., 2006) by running these tools. These include the removal of both individual samples and SNPs (single nucleotide polymorphisms) according to multiple criteria. Specifically, samples are removed based on (i) relatedness, (ii) high genotype missingness rate, and (iii) genetic ancestry differing from the majority of the samples (Price, et al., 2006). SNPs are removed based on criteria such as their missingness rate, their minor allele frequency (MAF), and deviation from Hardy-Weinberg Equilibrium (HWE). While the toolset is currently focused on case-control GWAS (binary traits), the entire QC pipeline is also applicable to GWAS of quantitative traits. One filter applied only for binary traits is the removal of SNPs which missingness correlates with trait, i.e. with case- or control-status (--test-missing).

To consider differences in genotyping between hemizygous males and diploid females, XWAS applies all the aforementioned QC steps of samples separately for males and females. Consequently, a unified dataset is generated for subsequent analyses that include all SNPs and individuals passing the above filtering criteria in both the male and female QC groups.

The pipeline then applies X-specific QC steps, which are exclusively built into XWAS, to the unified dataset. These include (i) removing SNPs with significantly different MAF between male and female samples in the control group (--freqdiff-x), (ii) removing SNPs with significantly different missingness rates between male and female controls (--missdiff-x), and (iii) removal of
SNPs in the pseudoautosomal regions (PARs). The first two of these steps capture problems in genotype calling when plates include both males and females (Korn, et al., 2008). Further details regarding specific QC procedures can be found in the user manual that is available with the toolset.

**Single-Marker Association Testing on the X chromosome**

For an X-linked SNP, while females have 0, 1, or 2 copies of an allele, hemizygous males have at most 1 copy. If X-inactivation is complete, it produces monoallelic expression of X-linked protein-coding genes in females. Therefore, when considering loci that undergo complete X-inactivation, it may be apt to consider males as having 0/2 alleles, corresponding to the female homozygotes (the FM02 test). The toolset carries out this test for association between a SNP and disease risk by using the --xchr-model 2 option in PLINK (Purcell, et al., 2007). For other scenarios though, including where some genes on the X escape X-inactivation or different genes are inactivated in different cells, it can be more indicative to code males as having 0/1 alleles. Hence, the toolset further carries out such an association test (FM01 test) of a SNP by using the following options in PLINK (Purcell, et al., 2007): --logistic and --linear for binary and quantitative traits, respectively.

All tests, including tests described in following sections, allow for covariates such as population structure, sex, and traits that are correlated with the disease, as commonly considered in GWAS. We suggest calculating principal components by using EIGENSTART (Price, et al., 2006) and include them as covariates to control for population structure. Ten such principal components are considered by default, unless otherwise specified. Any other user-defined covariates can also be incorporated.
Single-Marker Sex-stratified Analysis on the X chromosome

The XWAS software further includes new tests that are not a part of PLINK. First, we implemented a new sex-stratified test $\text{FM}_\text{comb}$ that is particularly relevant for X analyses since SNPs and loci on the sex chromosomes are potentially more likely to exhibit different effects on disease risk between males and females. In such scenarios, a sex-stratified test as described in the following can be better powered, as well as in scenarios where the effect is only observed in one sex. This functionality is accessible by the option --strat-sex. The $\text{FM}_\text{comb}$ tests first carry out an association test separately in males and females and then combines the results of the two tests to obtain a final sex-stratified significance level. The combination of the two test statistics is implemented using both Fisher’s method (--fishers) (Fisher, 1925) (in the $\text{FM}_{F,\text{comb}}$ test) and Stouffer’s method (--stouffers) (Stouffer, 1949) (in the $\text{FM}_{S,\text{comb}}$ test).

Each of these two tests is more powerful in different scenarios (Chang, et al., 2014), e.g. $\text{FM}_{F,\text{comb}}$ allows the SNP tested to have different, even an opposite, effect on disease risk in males and females. $\text{FM}_{F,\text{comb}}$ is also insensitive to whether males are coded as 0/2 (as in the $\text{FM}_{02}$ test) or as 0/1 (as in the $\text{FM}_{01}$ test), thus making no assumptions regarding X-inactivation status. Alternatively, $\text{FM}_{S,\text{comb}}$ directly accounts for the potentially differing sample sizes between males and females to maximize power. For this latter test, XWAS weighs by the sample size in males and females in cases and controls following the approach of (Willer, Li and Abecasis, 2010).

Single-Marker Sex-differentiated Effect Size Test on the X chromosome

We described above sex-stratified tests that accommodate associations that have different effect size between males in females. In another type of test ($\text{FM}_{\text{diff}}$), we directly test whether the effect
size is different between the sexes, which we find especially plausible for the X chromosome. This test, applied for each SNP, runs a t-test to test for difference between the odds ratio (OR) in males alone and the OR in females alone, while accounting for hemizygosity in males. This test is implemented under the --sex-diff option and is further described in (Chang, et al., 2014). For this test and the sex-stratified test introduced in the previous section, both odds ratios and regression coefficients in each sex can be provided as output for further examination.

**Single-Marker variance-based testing informed by X-inactivation in Females**

During X-inactivation, the expression of one copy of the X chromosome in females is randomly silenced, thereby increasing variation in the expression of X-linked quantitative trait loci (QTL). Specifically, females that are heterozygotes for a QTL might exhibit higher phenotypic variance than homozygous females since one or the other allele might be more dominantly affecting the phenotype in each given female heterozygote, such that for some individuals the QTL expression is more similar to one type of female homozygous, while to the other type in other individuals. We developed a test aimed at capturing this increased variance as a means for detecting X-linked QTL in females. An overview of the test and its implementation follows, while we refer readers to (Ma, et al., 2015) for a full description of the test. This test (F_var) is currently implemented under the --var-het option.

The null hypothesis of the F_var test is that phenotypic variances of the three genotypic groups of a SNP with 0, 1, or 2 copies of a reference allele are all equal. The alternative hypothesis is that female heterozygotes show a higher phenotypic variance than others. Hypothesis testing is carried out using a modified Brown-Forsythe test of variances (Brown and Forsythe, 1974). We first normalize the phenotypic value and remove the effects of possible covariates by a linear
regression as conventionally done, namely $y = \mu + XB + e$, where $y$ is a vector of quantitative trait levels, $\mu$ is the population mean, $X$ is the matrix of possible covariates, and $e$ is a vector of residuals. Assume $y_{i|g=j}$ is the phenotypic value of the $i$th individual in the $j$th genotypic group and $z_{i|g=j} = |e_{i|g=j}|$ is the absolute residual value of the $i$th individual in the $j$th genotypic group ($j = 0, 1, \text{or} 2$ copies of an allele of a SNP). A test statistic is derived as

$$T_{\text{var}} = \frac{\frac{\bar{z}_1 - \bar{z}_{0/2}}{n_1} + \frac{s_{0/2}^2/n_1}{n_0 + n_2}}{\sqrt{s_1^2/n_1 + s_{0/2}^2/n_0 + n_2}}$$

where $\bar{z}_1$ is the sample mean of $z_{i|g=1}$ over $i$, $\bar{z}_{0/2}$ is the sample mean of $z_{i|g=0}$ and $z_{i|g=2}$ combined, $s_1^2$ and $s_{0/2}^2$ are the sample variances respectively, and $n_j$ is the sample size of the $j$th genotypic group. Under the null hypothesis, the statistic follows a $t$-distribution with degrees of freedom given by

$$df = \frac{\left(\frac{s_1^2/n_1 + s_{0/2}^2/(n_0 + n_2)}{s_1^2/n_1 + s_{0/2}^2/(n_0 + n_2)}\right)^2}{\left(\frac{s_1^2/n_1}{n_1 - 1} + \left(\frac{s_{0/2}^2/(n_0 + n_2)}{(n_0 + n_2)}\right)^2/(n_0 + n_2 - 1)\right)}.$$

This variance-based test captures a novel signal of X-linked associations by directly testing for higher phenotypic variance in heterozygous females than homozygotes. As a test of variance it is generally less powerful than standard tests of association that consider means; however, it provides an independent and complementary test to the standard association test for QTLs on X (Ma, et al., 2015).

**X-linked Gene-based Analysis**

XWAS also includes unique features for carrying out gene-based association analysis on the X chromosome. Gene-based approaches may be better powered to discover associations than single-marker analyses in cases of a gene with multiple causal variants of small effect size, or of
multiple markers that are each in incomplete linkage disequilibrium with underlying causal variant/s. Furthermore, in studying the effect of X on sexual dimorphism in complex disease susceptibility, it may be desirable to analyze whole-genes or all genes of a certain function combined based on their unique function or putatively differential effect between males and females, as illustrated in (Chang, et al., 2014).

The XWAS package determines the significance of association between each gene as a whole and disease risk by implementing a gene-level test statistic that combines individual SNP-level test statistics for all SNPs in and around each studied gene. This gene-level approach is applicable to any of the different tests described above. For instance, beyond tests of association, it can be applied to the sex-differentiated tests. In this case the gene-based test captures any scenario whereby SNPs within the gene display different effects in males and females, without restricting such differential effects to be of a similar nature across SNPs. By default, genes are considered from the UCSC browser “knownCanonical” transcript ID track. SNPs were mapped to a gene if they are in the gene or within 15 kb of the gene’s start or end positions. The user can also provide a different set of gene definitions or alternate regions of interest and a different window length around them in which SNPs are also to be considered.

Combining SNP statistics across a gene is implemented in the general framework of (Liu, et al., 2010). Specifically, the significance of an observed gene-based test statistic is assessed from the distribution of test statistics that is expected given the linkage disequilibrium between the SNPs in the gene. In (Liu, et al., 2010), the test statistics for all SNPs in the gene are summed. Here, we have implemented a slight modification to this procedure, whereby we combined SNP-based p-values with either the truncated tail strength (Jiang, et al., 2011) or the truncated product...
(Zaykin, et al., 2002) method, which have been suggested to be more powerful in some scenarios (Ma, Clark and Keinan, 2013; Zaykin, et al., 2002).

To determine significance, XWAS follows the procedure in (Liu, et al., 2010). The observed statistic is compared to gene-level test statistics obtained when SNP-level statistics are randomly drawn from a multivariate normal distribution with a covariance determined by the empirical linkage disequilibrium between the SNPs in the tested gene. The significance level is then the proportion, out of many such drawings, for which this sampled gene-level statistic is more, or as extreme compared to the empirical one. For computational efficiency, the number of drawings is determined adaptively, as in (Liu, et al., 2010). By combining the truncated tail measures with this procedure, our new gene-based method combines the test statistics from multiple SNPs that show relatively low p-values, while also accounting for the dependency between these p-values due to linkage disequilibrium between the SNPs. Such a p-value is estimated for each gene and for each of the X-linked tests described above.
Examples of Use

In this section, we summarize several sets of results obtained using the XWAS software and publicly-available GWAS datasets. For many of the results, we include herein a brief description of the main results, with the full description appearing in separate papers (Chang, et al., 2014; Ma, et al., 2015). All associations presented herein are significant and details of the variety of p-values can be found in the respective papers.

Association of X-linked SNPs with autoimmune diseases

We applied the XWAS software described above to 16 GWAS datasets of autoimmune disease and other disorders with a potential autoimmune-related component. These include the following datasets that we obtained from dbGaP (Mailman, et al., 2007; Tryka, et al., 2014): ALS Finland (Laaksovirta, et al., 2010) (phs000344), ALS Irish (Cronin, et al., 2008) (phs000127), Celiac disease CIDR (Ahn, et al., 2012) (phs000274), MS Case Control (Baranzini, et al., 2009) (phs000171), Vitiligo GWAS1 (Jin, et al., 2010) (phs000224), CD NIDDK (Duerr, et al., 2006) (phs000130), CASP (Nair, et al., 2009) (phs000019), and T2D GENEVA (Qi, et al., 2010) (phs000091). Similarly, we obtained the following datasets from the Wellcome Trust Case Control Consortium (WT): all WT1 (The Wellcome Trust Case Control Consortium, 2007) datasets, WT2 ankylosing spondylitis (Evans, et al., 2011), WT2 ulcerative colitis (UK IBD Genetics Consortium, et al., 2009) and WT2 multiple sclerosis (International Multiple Sclerosis Genetics Consortium, et al., 2011). Finally, we also analyzed data from Vitiligo GWAS2 (Jin, et al., 2012). These datasets are described in more detail in (Chang, et al., 2014).

Following application of the QC pipeline as described above, we applied the SNP-level FM02, FMF.comb, and FMS.comb tests to all SNPs in each of the 16 datasets. Based on the Vitiligo GWAS1
datasets, we associated SNPs in a region 17 kilobases (kb) away from the retrotransposed gene retro-\textit{HSPA8} with risk of vitiligo. The parent of this retrotransposed gene, \textit{HSPA8} on chromosome 11, encodes a member of the heat shock protein family, which has been previously associated to vitiligo (Abdou, Maraee and Reyad, 2013; Mosenson, \textit{et al.}, 2013; Mosenson, \textit{et al.}, 2012). We discovered another association in WT2 ulcerative colitis of SNPs in an intron of \textit{BCOR} contributing to ulcerative colitis disease risk. BCOR indirectly mediates apoptosis via co-repression of \textit{BCL-6} (Huynh, et al., 2000).

**Association of X-linked genes with autoimmune diseases**

We next focused on a gene-based analysis of the X chromosome by using the SNP-level results of all the three tests in the above results as a basis for gene-based tests in the same 16 datasets. This analysis led to the discovery of the first X-linked gene-based associations with any disease or trait, which supports the utility of the XWAS package in facilitating such analyses. We associated in Vitiligo GWAS1 and replicated in Vitiligo GWAS2 an association between the gene \textit{FOXP3} and vitiligo disease risk, in support of an earlier candidate gene study (Birlea, \textit{et al.}, 2011). We also found a novel association of \textit{ARHGEF6} to Crohn’s disease and further replicated it in ulcerative colitis, another inflammatory bowel disorder (IBD). ARHGEF6 binds to a surface protein of a gastric bacterium (\textit{Helicobacter pylori}) that has been associated to IBD (Jin, \textit{et al.}, 2013; Luther, \textit{et al.}, 2010). Finally, we associated \textit{CENPI} as contributing to the risk of three different diseases (amyotrophic lateral sclerosis, celiac disease and vitiligo). Other, autosomal genes in the same family as \textit{CENPI} have previously been associated to amyotrophic lateral sclerosis (Ahmeti, \textit{et al.}, 2013) as well as multiple sclerosis (Baranzini, \textit{et al.}, 2009), supporting an involvement of \textit{CENPI} with autoimmunity in general.
X-linked SNPs showing sex-differentiated effect size with autoimmune disease

As a final analysis on the 16 autoimmune datasets, we applied the FM\textsubscript{diff} test and its gene-based version. Based on this test, we discovered and replicated the gene \textit{C1GALT1C1} (also known as \textit{Cosmc}) as exhibiting sex-differentiated effect size in risk of IBD. \textit{C1GALT1C1} is necessary for the synthesis of many O-glycan proteins (Ju and Cummings, 2005), which are components of antigens. We further found \textit{CENPI}, which we previously associated with several diseases, to show significantly different effects in males and females in the same diseases as in the association analysis.

Increased variance of systolic blood pressure in heterozygous females for an X-linked SNP

As an example of application of the variance-based testing informed by X-inactivation, we considered data on 7 quantitative traits from the Atherosclerosis Risk in Communities (ARIC) study (Williams, 1989) along with Affymetrix 6.0 data from the participating individuals, which included 34,527 X-linked SNPs. Note that for an example use of this test we turned to a different dataset since this test is currently only applicable to quantitative traits. First, we applied the entire set of QC procedures implemented in XWAS for quantitative traits. Then, we applied our single-marker variance-based testing and compared with application of standard testing for a QTL. Across the 7 traits, we found one SNP with a significant association based on the variance test (Ma, \textit{et al.}, 2015). Importantly, the signals of this test are not in the same loci as those of the standard test, in line with them capturing different types of signals. Specifically, the significant SNP, rs4427330, which is associated with systolic blood pressure based on the variance test, is not associated with any trait based on the standard test. It is located upstream of \textit{AFF2}, which might regulate \textit{ATRX}. \textit{ATRX} is associated with alpha-thalassemia, a disease that can cause anemia and has been associated with hypertension (Bowie, Reddy and Beck, 1997).
Implementation and Availability

The XWAS software package is implemented in C++ and includes functions from PLINK (Purcell, et al., 2007). Additional features are also implemented in scripts, including in shell (for QC), Perl (for converting file formats and using SMARTPCA), and R (gene-based testing), as appropriate. The entire package is freely available for download from http://keinanlab.cb.bscb.cornell.edu and includes (1) scripts, (2) the binary executable XWAS, (3) all source code, (4) a user manual, and (5) example data and run. Additional help is provided via the --xhelp option. The XWAS has been initially designed and optimized for Linux systems, hence exhibits best performance in such systems. However, the binary executable (but not scripts) is also available for Windows and MAC OS. The Makefile is written for Linux systems but it can be easily adjusted for Windows and MAC OS by changing several indicated lines therein.
Conclusions

We have developed an extensive toolset that facilitate the inclusion of the X chromosome in genome-wide association studies. It offers X-specific QC procedures, a variety of X-adapted tests of association, and an X-specific test of variance testing, available for both single-marker and gene-based statistics. We applied this toolset to successfully discover and replicate a number of genes with autoimmune disease risk and blood pressure.

We are continually developing the software and upcoming versions in the near future will offer additional features, including all features needed to conduct an extensive association study of quantitative traits (many features for quantitative traits are already functional in the current version). Similarly, while imputation of unobserved SNPs is presently performed as a preprocessing step using IMPUTE2 (Howie, et al., 2012), we will incorporate X-specific imputation as part of the pipeline. Additional features will include analysis of X-linked data from sequence-based association studies (including burden tests), statistical methods that have been previously designed for the X chromosome (Clayton, 2008; Clayton, 2009; Loley, Ziegler and Konig, 2011; Thornton, et al., 2012; Zheng, et al., 2007), tests for gene-gene interactions, and additional tests we previously proposed based on the workings of X-inactivation (Ma, et al., 2015). Finally, we will directly account for whether or how often any gene undergo or escape X-inactivation (Carrel and Willard, 2005; Cotton, et al., 2011; Disteche, 2012; Slavney, et al.). For computational efficiency, we will also upgrade to the most recent version of PLINK.

Considering the availability of unutilized data for the X chromosome from thousands of GWAS, and the additional X-linked data that is being generated as a part of ongoing GWAS, many researchers will find extensive utility in the XWAS toolset. It will facilitate the proper analysis of
these data, incorporate X into GWAS and enable discoveries of novel X-linked loci as implicated in many diseases and in their sexual dimorphism.
Funding

This work was supported by a NIH grants to AK (R01HG006849 and R01GM108805), as well as by an award from The Ellison Medical Foundation to AK, and an award by The Edward Mallinckrodt, Jr. Foundation to AK. FG is a Howard Hughes Medical Institute (HHMI) International Student Research fellow.

Acknowledgements

We thank Paul Billing-Ross, Aviv Madar, Aaron Sams, Andrea Slavney, Richard Spritz, Yedael Y. Waldman and Liang Zhang for helpful comments on the software and previous versions of this manuscript.
References


THE WELLCOME TRUST CASE CONTROL CONSORTIUM. 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447(7145):661-678.


