

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

Diana Chang^{1,2,§}, Feng Gao^{1,§} and Alon Keinan^{1,2,*}

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

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

[§]These authors contributed equally to this work

^{*}To whom correspondence should be addressed

ABSTRACT

Summary: We present XWAS (chromosome X-Wide Analysis tool-Set) – a toolset specially designed for analysis of the X chromosome in association studies, both on the level of single markers and the level of entire genes. It further offers other X-specific analysis tools, including quality control (QC) procedures for X-linked data. We have applied and tested this software by carrying out several X-wide association studies of autoimmune diseases.

Availability and Implementation: The XWAS software package, which includes scripts, the binary executable PLINK/XWAS and all source code is freely available for download from <http://keinanlab.cb.bscb.cornell.edu/content/tools-data>. PLINK/XWAS is implemented in C++ and other features in shell scripts and Perl. This software package is designed for Linux systems.

Contact: ak735@cornell.edu

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified many loci underlying complex human diseases and other complex traits (Welter et al. 2014). However, the majority of these GWAS have either omitted or incorrectly analyzed the X chromosome (X) (Wise et al. 2013). Recently, several studies have examined the role that X plays in GWAS (Tukiainen et al. 2014 and Chang et al. 2014) and demonstrated the value of its analysis. Special attention needs to be given to analysis of X since unlike the non-sex chromosomes, males only carry one copy of X that they inherit from their mother. Additionally, for most X-linked genes, one of the two copies in females is transcriptionally silenced via X-inactivation. While association methods and quality control (QC) procedures are well established for analysis of autosomes, association tests and QC pipe-lines that account for the differences in X-linked data are not widely available. In this paper, we introduce the software toolset XWAS (chromosome X-Wide Analysis toolSet), which is tailored for analysis of genetic variation on X. This toolset integrates X into GWAS as well as into the next generation of sequence-based association studies.

IMPLEMENTATION

PLINK (Purcell et al. 2007) is a widely-used software toolset for GWAS. Hence, we implemented the core of our XWAS package as an extended version of PLINK, termed PLINK/XWAS. It adds new functions on top of those in PLINK that facilitate X-tailored QC and analyses. Beyond these extensions of PLINK, XWAS includes several shell and Perl scripts that bring together the different functionalities into a pipeline that carries out an inte-grated analysis.

Quality Control (QC) Steps

This toolset implements a whole pipeline for performing QC on X data. The pipeline first follows standard GWAS QC steps as implemented in PLINK (Purcell et al. 2007) and SMARTPCA (Price et al. 2006). These include the removal of both individual samples and SNPs (single nucleotide polymorphisms) according to multiple criteria. Samples are removed due to (i) relatedness, (ii) high genotype missingness rate, and (iii) differing genetic ancestry from the rest of the samples (Price et al. 2006). SNPs are removed based on criteria such as their missingness rate and their minor allele frequency (MAF). The pipeline then applies X-specific QC steps, including the removal of SNPs with significantly different MAF between males and females across individuals in the control group (option `--freqdiff-x`), and SNPs in the pseudoautosomal regions (PARs). Further details regarding specific QC procedures can be found in the user manual that is available with the toolset.

Single-Marker Analysis on the X Chromosome

For an X-linked SNP, while females have 0, 1, or 2 copies of an allele, males only carry 0 or 1 copies. Complete X-chromosome inactivation produces monoallelic expression of X-linked protein-coding genes in females. Therefore, it may be apt to code males as having 0/2 alleles,

corresponding to the female homo-zygotes, when considering loci that undergo complete X-inactivation. The toolset uses this coding in carrying out a test of association between a single SNP and disease risk by using the `--xchr-model 2` option in PLINK (Purcell et al. 2007). For other scenarios, 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. The toolset carries out an association analysis of a single SNP coded as 0/1 in males by using the following options in PLINK: `--logistic` and `--linear` for binary and quantitative traits, respectively.

The extended software, PLINK/XWAS, further provides a new sex-stratified test that is particularly relevant for X analyses since SNPs and loci that are associated with disease risk are presumably more likely to exhibit different effects on risk between males and females when they are on the sex chromosomes. This functionality is accessible by the option `--stratsex`. It first carries out an association test in each of males and females separately and then combines the results of the two tests using the Fisher's method (Fisher, 1925) to obtain a final sex-stratified significance level. Since the test in males is separate, this sex-stratified test is independent of whether 0/1 or 0/2 allele coding is considered in males, thus making no assumptions regarding X-inactivation status. Furthermore, the sex-stratified test may be better powered in situations where an allele has opposing effects between males and females in association to the tested trait, as well as when the effect is only observed in one sex.

X-linked Gene-based Analysis

XWAS also includes an R script for carrying out gene-based association analysis in the framework established by VEGAS (Liu et al. 2010). Gene-based approaches may be better powered to discover associations than single-marker analysis in cases such of a gene with multiple causal variants of small effect size, or of multiple markers that are each in incomplete linkage disequilibrium with one underlying causal variant. Furthermore, in studying the effect of X on sexual dimorphism in complex disease susceptibility, it can be desirable to analyze whole-genes, either chromosome-wide or by focusing on certain gene functions that differ between males and females, as illustrated in Chang et al. 2014.

An R script we provide as part of XWAS determines the significance of association between a gene and disease risk. It implements a gene-level test statistic that combines all individual SNP-level test statistics (any of the different statistics described in section 2.2) in and around each studied gene. To determine the significance, this observed statistic is compared to gene-level test statistics drawing from a multivariate normal distribution with the covariance determined by the linkage disequilibrium between the SNPs considered in the tested gene. Instead of the gene-based statistic that considers the sum or minimum p-value across SNPs in the gene, we implemented more powerful approaches based on truncated tail strength (Jiang et al. 2011) and truncated product (Zaykin et al. 2002) methods (Li et al. 2013). Thus, the 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 these SNPs. The significance level of the resulting gene-level statistic is calculated as the proportion, out of x drawings from the multivariate normal distribution, for which the same

gene-level statistic is more, or as, extreme compared to the empirical one. As in the original implementation of VEGAS, x is determined adaptively (Liu et al. 2010).

Additional Features

PLINK/XWAS can output the allele frequencies for each SNP in males and females separately by specifying option `--freq-x`. Upcoming versions will offer additional features, including all options needed to conduct an extensive association study of quantitative traits, tests of X-inactivation, tests for gene-gene interactions and analysis of sequence-based data.

CONCLUSION

We have developed a toolset that includes both an extended version of PLINK (Purcell et al. 2007) and additional scripts that, combined, facilitate including the X chromosome as part of a genome-wide association study. It offers X-specific QC procedures, as well as tests of association based on both single-marker and gene-based statistics. We have applied this software to X-linked data from many GWAS of several autoimmune diseases, with the detailed results presented in Chang et al. 2014. Briefly, we discovered and independently replicated several X-linked genes associated with disease risk. These results provide a proof-of-principle of our XWAS toolset and the type of discoveries it can advance.

Considering the availability of unutilized data for the X chromosome from hundreds of GWAS, and the additional X-linked data that is being generated as a part of many ongoing GWAS, 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.

ACKNOWLEDGEMENTS

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

Funding: This work was supported by the National Institutes of Health [to A.K.; R01-HG006849]; The Ellison Medical Foundation [to A.K.]; and the Edward Mallinckrodt, Jr. Foundation [to A.K.].

REFERENCES

- Chang,D. et al. No eXceptions: Accounting for the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. In revision.
<http://www.biorxiv.org/content/early/2014/09/21/009464>
- Fisher,R.A. (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
- Jiang,B. et al. (2011) A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. J. Theor. Biol., 277, 67-73.
- Liu,J.Z. et al. (2010) A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet., 87, 139-145.
- Price,AL. et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet., 38, 904-909.
- Purcell,S. et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet., 81, 559-575.
- Tukiainen,T. et al. (2014) Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. PLoS Genet., 10, e1004127.
- Welter,D. et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic. Acids. Res., 42(Database issue), D1001-D1006.
- Wise,AL. et al. (2013) eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am. J. Hum. Genet., 92, 643-647.
- Zaykin,DV. et al. (2002) Truncated product method for combining P-values. Genet. Epidemiol., 22, 170-185.