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,*]

[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

[§]These authors contributed equally to this work

[*]To whom correspondence should be addressed

Contact: ak735@cornell.edu

Keywords

X chromosome, genetics, GWAS, software, association study, sexual dimorphism, complex diseases

Abstract

Background

The X chromosome plays an important role in complex human traits and diseases, especially those with sexually dimorphic characteristics. Special attention needs to be given to analysis of X, since unlike the non-sex chromosomes, males only carry one copy of the chromosome that they inherit from their mother, while in females, one of the two copies is transcriptionally silenced via X-inactivation. The different mode of inheritance leads to several analytical complications that have resulted in the majority of genome-wide association studies (GWAS) not considering the X chromosome or otherwise mishandling it by applying the same tools designed for non-sex chromosomes. Hence, there is a need for tools deploying association methods and quality control procedures that are specifically designed for the X chromosome and that account for its uniqueness.

Results

We present XWAS (chromosome X-Wide Analysis toolSet) – 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 procedures for X-linked data. We applied the analysis pipeline offered by this toolset to 16 GWAS datasets of immune-related disorders. We discovered several new associations on the X chromosome that have not been previously reported, one of which shows evidence of a pleiotropic affect across several diseases.

Conclusion

The XWAS toolset facilitates proper analysis of X-linked data from different types of association studies. This toolset, together with its use to discover genes underlying autoimmune disease risk, will provide the tools and incentive for others to incorporate X into GWAS, thereby enabling discoveries of novel X-linked loci implicated in many diseases and in their sexual dimorphism.

Background

Genome-wide association studies (GWAS) have successfully identified many loci underlying complex human diseases and other complex traits [1]. While very successful for the autosomes (non-sex chromosomes), most of these studies have either incorrectly analyzed or ignored the X chromosome (X) [2], due to analytical problems that follows its 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). 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 [3], cardiovascular diseases [4] and cancer [5,6], which suggests a potential contribution of X [7,8]. Recently, several studies have examined this issue and further demonstrated the value of analyzing X [9,10,11]. Autosomal methods cannot be directly applied to X without accounting for its unique patterns of genetic variation and mode of inheritance. Thus, while association methods and quality control (QC) procedures are well established for analysis of autosomes, association tests and QC pipelines for 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

Quality Control Steps

This 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 [12] and SMARTPCA [13]. These include the removal of both individual samples and SNPs (single nucleotide polymorphisms) according to multiple criteria. Samples are removed based on (i) relatedness, (ii) high genotype missingness rate, and (iii) differing genetic ancestry from the rest of the samples [13]. 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 Association Testing 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. 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 (FM02), corresponding to the female homozygotes. The toolset carries out the FM02 test by using the *--xchr-model* 2 option in PLINK [12]. 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 (FM01). Hence, the toolset further carries out an association analysis of a SNP by considering allele number as 0/1 in males by using the following options in PLINK: --*logistic* and --*linear* for binary and quantitative traits, respectively.

Single-Marker Sex-stratified Analysis on the X chromosome

The extended software, PLINK/XWAS, further provides a new sex-stratified test FMcomb 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. This functionality is accessible by the option --*stratsex*. It first carries out an association test separately in males and females and then combines the results of the two tests using Fisher's method [14] to obtain a final sex-stratified significance level. This test accommodates the possibility of differential effect size and direction between males and females. 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 as described above, thus making no assumptions regarding X-inactivation status. 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. Gene-based approaches may be better powered to discover associations than single-marker analysis 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 can be desirable to analyze whole-genes or all genes of a certain function combined based on their putatively differential effect between males and females, as illustrated in [10].

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 above) in and around each studied gene. To determine the significance, it follows the framework of [15] in comparing this observed statistic is to gene-level test statistics obtained from combining SNP-level statistics drawn from a multivariate normal distribution with the covariance determined by the empirical linkage disequilibrium between the SNPs in the tested gene. The significance level is then determined as the proportion, out of *x* drawings, for which the same gene-level statistic is more, or as, extreme compared to the empirical one. For computational efficiency, *x* is determined adaptively [15]. The gene-based test statistic can combined the individual SNP-level statistics in any of a number of ways. Specifically, instead of the statistic that considers the sum or minimum of *p*-values across SNPs in the gene, we implemented more powerful approaches based on truncated tail strength [16] and truncated product [17] methods [18]. 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 the SNPs

Additional Features

PLINK/XWAS can also output the allele frequencies for each SNP in males and females separately by specifying option --*freq-x*. Upcoming versions in the near future will offer additional features, including all options needed to conduct an extensive association study of quantitative traits, additional quality control that is different for males and females, and analysis of X-linked data from sequence-based association studies. They will also include implementation of new statistical methods that directly test for X-inactivation, for gene-gene interactions, and for differential effect size between males and females. We implemented but have not yet released many of these features, hence the XWAS toolset might include these by the time of publication of this manuscript.

Results and Discussion

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: ALS Finland [19] (phs000344), ALS Irish [20] (phs000127), Celiac disease CIDR [21] (phs000274), MS Case Control [22] (phs000171), Vitiligo GWAS1 [23] (phs000224), CD NIDDK [24] (phs000130), CASP [25] (phs000019), and T2D GENEVA [26] (phs000091). Similarly, we obtained the following datasets from the Wellcome Trust Case Control Consortium (WT): all WT1 [27] datasets, WT2 ankyolosing spondylitis [28], WT2 ulcerative colitis [29] and WT2 multiple sclerosis [30]. Finally, we also analyzed data from Vitiligo GWAS2 [31]. These datasets are described in more detail in [10].

We describe in the following the main results, and have included a more detailed description of the results in a separate paper [10]. We first applied the SNP-level FM02 and FMcomb 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-*HSPA8* with risk of viligo. The parent of this retrotransposed gene, *HSPA8* on chromosome 11, encodes a member of the heat shock protein family, which has been previously associated to vitiligo [32,33,34]. We discovered another association in WT2 ulcerative colitis of SNPs in an intron of *BCOR* contributing to ulcerative colitis disease risk. BCOR indirectly mediates apoptosis via co-repression of *BCL-6* [35]. Unfortunately, no SNPs in the same regions replicated in other inflammatory bowel disease datasets.

We next focused on a gene-based analysis of the X chromosome. We associated in Vitiligo GWAS1 and replicated in Vitiligo GWAS2 an association between the gene *FOXP3* and vitiligo disease risk (combined *P*-value = 9.5×10^{-6}). The same gene was associated to vitiligo in an earlier candidate gene study [36]. We also found a novel association of *ARHGEF6* to Crohn's disease and further replicated it in ulcerative colitis (combined *P*-value = 1.67×10^{-5}). ARHGEF6 binds to a surface protein of a gastric bacterium (*Helicobacter pylori*) that has been associated to inflammatory bowel disease [37,38]. Finally, we associated *CENPI* as contributing to the risk of three different diseases (amyotrophic lateral sclerosis, celiac disease and vitiligo) (*P*-value = 2.1×10^{-7}). Other, autosomal genes in the same family as *CENPI* have previously been associated to amyotrophic lateral sclerosis [39] as well as multiple sclerosis [40], supporting an involvement of *CENPI* with autoimmunity in general.

Conclusions

We have developed a toolset that includes both an extended version of PLINK [12] 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 a variety of X-adapted tests of association that are based on both single-marker and gene-based statistics. We applied this toolset to successfully discover and replicate a number of genes with autoimmune disease risk.

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, many researchers will potentially 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.

Availability and requirements

The XWAS software package, which includes (1) scripts, (2) the binary executable PLINK/XWAS, (3) all source code, and (4) a user manual, 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.

Authors' contributions

AK conceived and designed the project. DC and FG implemented the software and performed the analyses. DC, FG and AK wrote the paper. All authors read and approved the final manuscript.

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.

This work was supported by an NIH grant to AK (R01-HG006849), 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.

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L *et al*: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations**. *Nucleic acids research* 2014, **42**(Database issue):D1001-1006.

2. Wise AL, Gyi L, Manolio TA: **eXclusion: toward integrating the X chromosome in genome-wide association analyses**. *American journal of human genetics* 2013, **92**(5):643-647.

3. Voskuhl R: **Sex differences in autoimmune diseases**. *Biol Sex Differ* 2011, **2**(1):1.

4. Lerner DJ, Kannel WB: **Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population**. *Am Heart J* 1986, 111(2):383-390.

5. Matanoski G, Tao X, Almon L, Adade AA, Davies-Cole JO: **Demographics and tumor characteristics of colorectal cancers in the United States, 1998-2001**. *Cancer* 2006, 107(5 Suppl):1112-1120.

6. Muscat JE, Richie JP, Jr., Thompson S, Wynder EL: **Gender differences in smoking and risk for oral cancer**. *Cancer Res* 1996, 56(22):5192-5197.

7. Ober C, Loisel DA, Gilad Y: **Sex-specific genetic architecture of human disease**. *Nat Rev Genet* 2008, **9**(12):911-922.

8. Carrel L, Willard HF: **X-inactivation profile reveals extensive variability in X-linked gene expression in females**. *Nature* 2005, **434**(7031):400-404.

9. Tukiainen T, Pirinen M, Sarin AP, Ladenvall C, Kettunen J, Lehtimaki T, Lokki ML, Perola M, Sinisalo J, Vlachopoulou E *et al*: **Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation**. *PLoS genetics* 2014, **10**(2):e1004127.

10. Chang D, Gao F, Ma L, Sams A, Slavney A, Waldman YY, Billing-Ross P, Madar A, Spritz R, Keinan A: **Accounting for eXentricities: Analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases**. bioRxiv, doi: http://dx.doi.org/10.1101/009464.

11. Gilks WP, Abbott JK, Morrow EH: **Sex differences in disease genetics: evidence, evolution, and detection**. *Trends Genet* 2014, **30**(10):453-463.

12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *American journal of human genetics* 2007, **81**(3):559-575.

13. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nature genetics* 2006, **38**(8):904-909.

14. Fisher, R.A: *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh; 1925.

15. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Investigators A, Hayward NK, Montgomery GW, Visscher PM *et al*: **A versatile gene-based test for**

genome-wide association studies. *American journal of human genetics* 2010, **87**(1):139-145.

16. Jiang B, Zhang X, Zuo Y, Kang G: **A powerful truncated tail strength method for testing multiple null hypotheses in one dataset**. *Journal of theoretical biology* 2011, **277**(1):67-73.

17. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS: **Truncated product method for combining P-values**. *Genetic epidemiology* 2002, **22**(2):170-185.

18. Ma L, Clark AG, Keinan A: **Gene-based testing of interactions in association studies of quantitative traits**. *PLoS genetics* 2013, **9**(2):e1003321.

19. Laaksovirta H, Peuralinna T, Schymick JC, Scholz SW, Lai SL, Myllykangas L, Sulkava R, Jansson L, Hernandez DG, Gibbs JR *et al*: **Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study**. *Lancet Neurol* 2010, **9**(10):978-985.

20. Cronin S, Berger S, Ding J, Schymick JC, Washecka N, Hernandez DG, Greenway MJ, Bradley DG, Traynor BJ, Hardiman O: **A genome-wide association study of sporadic ALS in a homogenous Irish population**. *Hum Mol Genet* 2008, **17**(5):768-774.

21. Ahn R, Ding YC, Murray J, Fasano A, Green PH, Neuhausen SL, Garner C: **Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility loci**. *PLoS One* 2012, **7**(5):e36926.

22. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, Barkhof F, Radue EW, Lindberg RL, Uitdehaag BM, Johnson MR *et al*: **Genome-wide association analysis of**

susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet* 2009, **18**(4):767-778.

23. Jin Y, Birlea SA, Fain PR, Gowan K, Riccardi SL, Holland PJ, Mailloux CM, Sufit AJ, Hutton SM, Amadi-Myers A *et al*: **Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo**. *N Engl J Med* 2010, **362**(18):1686-1697.

24. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A *et al*: **A genome-wide association study identifies IL23R as an inflammatory bowel disease gene**. *Science* 2006, **314**(5804):1461-1463.

25. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJ *et al*: **Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways**. *Nature genetics* 2009, **41**(2):199-204.

26. Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Pare G *et al*: **Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes**. *Hum Mol Genet* 2010, **19**(13):2706-2715.

27. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.

28. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, Oppermann U, Dilthey A, Pirinen M, Stone MA *et al*: **Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility**. *Nature genetics* 2011, **43**(8):761-767.

29. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H *et al*: **Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region**. *Nature genetics* 2009, **41**(12):1330-1334.

30. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE *et al*: **Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis**. *Nature* 2011, **476**(7359):214-219.

31. Lerner DJ, Kannel WB: **Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population**. *Am Heart J* 1986, **111**(2):383-390.

32. Mosenson JA, Zloza A, Klarquist J, Barfuss AJ, Guevara-Patino JA, Poole IC: **HSP70i is a critical component of the immune response leading to vitiligo**. *Pigment Cell Melanoma Res* 2012, **25**(1):88-98.

33. Mosenson JA, Eby JM, Hernandez C, Le Poole IC: **A central role for inducible heat-shock protein 70 in autoimmune vitiligo**. *Exp Dermatol* 2013, **22**(9):566-569.

34. Abdou AG, Maraee AH, Reyad W: **Immunohistochemical expression of heat shock protein 70 in vitiligo**. *Ann Diagn Pathol* 2013, **17**(3):245-249.

35. Huynh KD, Fischle W, Verdin E, Bardwell VJ: **BCoR, a novel corepressor involved in BCL-6 repression**. *Genes Dev* 2000, **14**(14):1810-1823.

36. Birlea SA, Jin Y, Bennett DC, Herbstman DM, Wallace MR, McCormack WT, Kemp EH, Gawkrodger DJ, Weetman AP, Picardo M *et al*: **Comprehensive association**

analysis of candidate genes for generalized vitiligo supports XBP1, FOXP3, and TSLP. *J Invest Dermatol* 2011, **131**(2):371-381.

37. Luther J, Dave M, Higgins PD, Kao JY: **Association between Helicobacter pylori infection and inflammatory bowel disease: a meta-analysis and systematic review of the literature**. *Inflamm Bowel Dis* 2010, **16**(6):1077-1084.

38. Jin X, Chen YP, Chen SH, Xiang Z: **Association between Helicobacter Pylori infection and ulcerative colitis--a case control study from China**. *Int J Med Sci* 2013, **10**(11):1479-1484.

39. Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, Andersen PM, Armstrong J, Birve A, Blauw HM, Brown RH, Bruijn L, Chen W *et al*: **Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1**. *Neurobiol Aging* 2013, **34**(1):357 e357-319.

40. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, Barkhof F, Radue EW, Lindberg RL, Uitdehaag BM, Johnson MR *et al*: **Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis**. *Hum Mol Genet* 2009, **18**(4):767-778.