Published in final edited form as:

Genet Epidemiol. 2016 February ; 40(2): 123–132. doi:10.1002/gepi.21946.

Modeling X Chromosome Data using Random Forests: Conquering Sex Bias

Stacey J Winham¹, Gregory Jenkins¹, and Joanna M. Biernacka^{1,2}

¹Department of Health Sciences Research, Mayo Clinic Rochester, USA

²Department of Psychiatry and Psychology, Mayo Clinic Rochester, USA

Abstract

Machine learning methods, including Random Forests (RF), are increasingly used for genetic data analysis. However, the standard RF algorithm does not correctly model the effects of X-chromosome single nucleotide polymorphisms (SNPs), leading to biased estimates of variable importance. We propose extensions of RF to correctly model X SNPs, including a stratified approach and an approach based on the process of X chromosome inactivation. We applied the new and standard RF approaches to case-control alcohol dependence data from the Study of Addiction: Genes and Environment (SAGE), and compared the performance of the alternative approaches via a simulation study. Standard RF applied to a case-control study of alcohol dependence yielded inflated variable importance estimates for X SNPs, even when sex was included as a variable, but the results of the new RF methods were consistent with univariate regression-based approaches that correctly model X-chromosome data. Simulations showed that the new RF methods eliminate the bias in standard RF variable importance for X SNPs when sex is associated with the trait, and are able to detect causal autosomal and X SNPs. Even in the absence of sex effects, the new extensions perform similarly to standard RF. Thus, we provide a powerful multi-marker approach for genetic analysis that accommodates X chromosome data in an unbiased way. This method is implemented in the freely available R package ‘snpRF’ (<http://www.cran.r-project.org/web/packages/snpRF/>).

Keywords

Random Forest; X chromosome; sex differences; variable importance; bias

Introduction

Although the X chromosome is known to play a role in many Mendelian disorders, variants on the X chromosome are understudied for complex traits, and are routinely excluded from genome-wide association studies (GWAS) to simplify analyses [Wise, et al. 2013]. Because the number of copies of the X chromosome is confounded with sex, special analysis methods are required. Furthermore, the biology of the X chromosome differs from the

Corresponding Author: Stacey J Winham, Mayo Clinic, 200 First Street SW, Rochester MN 55905, Phone: (507) 266-3662, winham.stacey@mayo.edu.

The authors have no conflict of interest to report.

Author Manuscript

autosomes. For instance, transcription of one of the female copies of the X chromosome is randomly silenced during development during a process called X chromosome inactivation [Brown, et al. 1991], which equalizes levels of gene expression between males and females [Disteche 2012; Ross, et al. 2005]. This process is tissue- and cell-specific, and the complex biology of this process is not fully understood [Wu, et al. 2014]. Such phenomena are difficult to account for in statistical analyses, especially for studies including both males and females. However, it has recently been advocated that the X chromosome should not be excluded from genetic analyses [Chang, et al. 2014; Konig, et al. 2014], particularly for traits that exhibit clear sex differences [Winham, et al. 2015].

Author Manuscript

Because of the differences in genotype distributions between men and women for X chromosome single nucleotide polymorphisms (SNPs), special methods are required to avoid confounding of associations between X SNPs and phenotypes by sex. For univariate genome scans, a number of methods have been proposed to address this problem [Clayton 2008; Konig, et al. 2014; Purcell, et al. 2007; Thornton, et al. 2012; Wang, et al. 2014; Zheng, et al. 2007], and a method based on coding female genotypes as 0,1,2 copies and male genotypes as 0,2, has been shown to be a powerful approach. This approach essentially mimics X inactivation by reweighting male genotypes to balance the dosage across sex [Clayton 2008; Hickey and Bahlo 2011; Loley, et al. 2011]. The approach of Wang, et al expands on this idea, and maximizes the likelihood over alternative coding schemes that account for additional biological models, including random, non-random, and escape from X inactivation [Wang, et al. 2014].

Author Manuscript

Machine learning or data-mining techniques designed to build predictive models based on high-dimensional data are becoming popular in genetic association studies of common complex disease as an alternative to univariate regression-based genome scans [Szymczak, et al. 2009]. While univariate methods are often favored for GWAS due to computational simplicity, tools such as Random Forest (RF) have the advantage of building predictive models and ranking variables while simultaneously modeling the effects of a large number of genetic predictors. These approaches also have the potential to model risk factor interactions, although the ability of RF to capture interactions in high-dimensional data is limited [Winham, et al. 2012].

Author Manuscript

RF models non-linear SNP effects by constructing an ensemble of many classification or regression trees fit to bootstrap samples of subjects, with random subsets of predictors considered at each node during the tree-building process. Disease classifications are obtained based on an aggregate vote over all trees in the forest, and variable importance measures are computed based on the accuracy of these classifications in order to rank variables [Brieman 2001]. The standard RF algorithm does not allow for the proper adjustment for X chromosome SNPs. A naïve inclusion of X SNPs in the input dataset results in inflated variable importance measures for X chromosome SNPs if sex is associated with the trait, because the genotypes are confounded with sex—females have three possible values, whereas men have only two. Because RF models non-linear associations, traditional approaches to code X chromosome variants in males as 0,2 and females as 0,1,2 [Clayton 2009], which assume linearity, do not eliminate this confounding. In most prior studies X chromosome SNPs have been excluded from RF analyses [de Maturana, et al. 2013;

Goldstein, et al. 2010; Winham, et al. 2013]. However, in applications of tree-based approaches where X chromosome SNPs were not excluded, it is not clear if top-ranked X variants were observed simply because of potential confounding with sex [Ye, et al. 2005]. Recently, gene set analyses that examine multiple variants located within genes involved in a particular biological pathway—which may include SNPs on the X chromosome and the autosomes—have become popular [Fridley and Biernacka 2011]. For RF to provide correct results in such a setting, it is critical to ensure that the X chromosome variants and the autosomal variants can be modeled together in the same forest, without leading to spurious results.

In this paper, we propose extensions to RF that correctly incorporate X chromosome genotypes, including an approach that models the biological properties of X chromosome genetics. We illustrate the methods using a real data example from a case-control study of alcohol dependence, a trait that is more common in males. We also compare the variable importance measures based on the proposed extensions to those from traditional RF implementation using simulated data, and describe freely available software that implements our RF extensions (the R package ‘snpRF’).

Methods

Random Forests

RF is an ensemble method, where a large number of classification or regression trees (CART) [Breiman, et al. 1984] are fit to bootstrap samples of the data [Breiman 2001]. On average, one third of the samples are not included in a bootstrap sample that is used to generate a given tree, and these out-of-bag (OOB) samples are used to estimate prediction error (PE) [Breiman 1996]. In addition to estimates of prediction error, measures of variable importance are computed and used for ranking of variables in terms of their contribution to prediction of the outcome variable. One such measure is mean decrease in accuracy (MDA), or the reduction in prediction accuracy after the variable of interest is permuted within each tree [Schwarz, et al. 2010]. For a causal variable, permuting the variable disrupts the association and results in a large decrease in accuracy, while for a variable not associated with the outcome, permuting the variable should have no impact on prediction. However, in the case of X chromosome SNPs, genotypes are confounded with sex since females have three possible genotypes, but males have only two. Therefore, if the outcome variable (i.e. case/control status) is associated with sex, then X chromosome SNP genotypes are associated with the outcome (in a non-linear way) leading to inflated variable importance measures for X chromosome SNPs as compared to autosomal SNPs. We refer to this inflation in variable importance measures for X chromosome SNPs compared to autosomal SNPs under the null hypothesis of no genetic effects as X chromosome importance bias. In this manuscript, we describe this bias in X chromosome variable importance in standard RF analysis, and propose approaches to correct the bias.

We first consider two standard implementations of RF: 1) coding X SNPs as 0/1/2 copies of the minor allele in females (same as the autosomal SNPs) but 0/2 in males, as is routinely done in regression based analyses where additive effects are assumed [Chang, et al. 2015; Chang, et al. 2014; Clayton 2008]; and 2) coding the X SNPs as 0/1/2 in females and 0/2 in

males, but also including sex as a variable in the RF analysis. These approaches are both simple to apply, and can be implemented in standard RF software, such as the R package ‘randomForest’ [Liaw and Wiener 2002].

We also propose three alternative approaches to include X chromosome SNPs in RF models that aim to correct the confounding with sex. In all three approaches, the genotypes for X SNPs are coded in the same manner as for the standard RF approaches (0/1/2 for females and 0/2 for males). The first is a stratified approach, which fits RF models separately in each sex (RF-stratified). Combined variable importance estimates are calculated for each SNP as the average MDA across the male and female forests. As with the standard RF implementation above, this approach is also simple to compute and can be easily implemented with current software.

Rather than fitting two separate RFs on sex-specific samples, we also propose an alternative approach where both sexes are included in the same forest, but a split on sex is forced into the first node of every tree (RF-split). Therefore data from both sexes is modeled within the same forest, but the SNP effects are modeled within sex strata. Because this approach relies on constructing trees differently from standard RF, implementation of this method required modifications of current RF software.

The first two alternative approaches address potential sex-bias from the point of view of a confounding factor. The third approach that we developed differs by leveraging current knowledge of X chromosome biology into the construction of the forest, by mirroring the process of X chromosome inactivation (RF-XCI). Because in females, one chromosome is randomly silenced in every cell [Brown, et al. 1991; Carrel, et al. 1999; Lyon 1961], only one allele will be active at each X locus. Mimicking this idea, we randomly select one allele for females to include in each tree, so that in any given tree the effect of only one allele (coded 0/1) is modeled in both males and females. In order to preserve LD structure, we first phase the X chromosome SNPs in females to obtain two long range estimated haplotypes, and then sample these haplotypes before building each tree. Because the active chromosome is believed to be random for each cell in females, we randomly select a different haplotype for each tree to most closely mimic the natural biological process. Therefore female heterozygotes will have one of their alleles modeled in approximately 50% of trees, and their other allele in the remaining 50% of trees on average. The X alleles are then modeled along with the autosomal genotypes. Implementation of this approach requires pre-phasing female X chromosome data and additional data manipulation in the RF algorithm (random selection of female X haplotypes for each tree); we therefore modified current RF software to implement the approach in the R package ‘snpRF’ (<http://www.cran.r-project.org/web/packages/snpRF/>).

Example: Study of Alcohol Dependence

To illustrate the performance of RF when SNPs from the autosomes and X chromosomes are analyzed together, we utilize data from the Study of Addiction: Genetics and Environment (SAGE) [Bierut, et al. 2010], a genome-wide association study of alcohol dependence, a heritable disorder that is more common in males [Brady, et al. 1993; Goldman, et al. 2005].

The dataset is publically available and was obtained through dbGaP (study accession phs000092.v1.p1) [Mailman, et al. 2007].

Previously, using regression-based analyses of the SAGE data, it was shown that SNPs in the N-methyl-D-aspartate (NMDA)-dependent α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptor trafficking cascade pathway are associated with alcohol dependence [Karpyak, et al. 2012]. This gene set includes 988 SNPs from 12 autosomal genes (*RAP1GAP*, *GRIA2*, *CAMK2A*, *GRIA1*, *MAPK14*, *SYNGAP1*, *CAMK2B*, *MPDZ*, *GRIN1*, *GRIA4*, *GRIN2B*, *GRIN2A*) and one X chromosome gene (*GRIA3*). These univariate analyses were conducted assuming an additive genetic model, which would have limited power to detect non-linear effects. We also anticipated that a multivariate analysis could reveal different results. Because RF can model multiple variants simultaneously as well as capture nonlinear effects, analyses using standard RF and a weighted RF extension were subsequently performed, with SNPs on the X chromosome excluded due to the strong association between sex and alcoholism [Winham, et al. 2013].

We analyzed the full NMDA-dependent AMPA trafficking cascade pathway dataset including all 988 SNPs (including the X chromosome SNPs) in all 13 genes, and 1165 alcohol dependent cases (~70% male) and 1379 controls (~30% male) using RF, in order to illustrate how variable rankings for X chromosome SNPs compare to autosomal SNPs in a real data set where sex is strongly associated with the trait. Because RF requires complete data, a small number of missing SNP genotypes were imputed using the proximity matrix of the Random Forest algorithm [Stekhoven and Buhlmann 2012]. Autosomal SNP genotypes were coded as 0/1/2 copies of the minor allele, as were X genotypes in females; X SNP genotypes of males were coded 0/2. We utilized both standard RF, as well as the three alternative implementations to include X chromosome SNPs. All RF analyses were performed using 5000 trees and the total number of predictors considered at each node (mtry) of $\sqrt{988}$. We compared variable importance rankings across RF methods, and also compared results to the previously published univariate analyses.

Simulation Design

We performed a simulation study to better understand the results of the RF analysis of the SAGE data, and to compare the performance of the standard RF implementations with our alternative approaches to incorporate X SNP data. The simulated datasets were generated based on the SAGE dataset, with similar sample size, number of SNPs, sex distribution, and linkage disequilibrium (LD) patterns. Specifically, in order to preserve patterns of local LD, the NMDA-trafficking cascade pathway dataset was phased by chromosome using fastPhase [Scheet and Stephens 2006], and the most likely haplotypes were selected for each chromosome, for each individual. Haplotypes including 988 SNPs were then simulated from this haplotype distribution for $N=2000$ subjects using the program HapSim [Montana 2005]. For the autosomes, two haplotypes were simulated for each individual. For the X chromosome, two haplotypes were simulated for females, but only one haplotype was simulated for males; for males, the single haplotype was duplicated to make males appear to be homozygous for all X-chromosome SNPs (and to facilitate 0-2 genotype coding). For each individual and each chromosome, the phased haplotypes were converted to SNP

genotypes (with the exception of the X chromosome data for the RF-XCI method, which utilized the haplotype data). Furthermore, to simplify interpretations and to examine the impact of LD patterns and number of autosomal vs. X chromosome SNPs, genetic data were also simulated in the absence of LD for 500 autosomal and 500 X SNPs; for these datasets, genotypes were generated independently assuming Hardy-Weinberg Equilibrium with MAF=0.2 for all SNPs.

Binary phenotypes were simulated under the null distribution of no SNP effects, as well as under the alternative distribution with autosomal and X chromosome SNP effects. Under the null hypothesis, phenotypes were randomly generated from a Bernoulli distribution independently of the SNP genotypes, both when 1) SNP genotypes were independent, and 2) under the patterns of LD based on the SAGE data. To compare the performance of the RF methods for varying degrees of sample imbalance, sets of 2000 subjects were simulated under various ratios of cases to controls and males to females (Supplemental Table 1). Additionally, phenotypes were also simulated with and without strong sex effects (OR of males to females=0.25, 1, 4), in order to assess possible bias of variable importance measures for the RF methods in the presence of sex association with the trait.

To compare RF methods under the alternative hypothesis, genetic data was again simulated either with 1) independent SNPs, or 2) with LD patterns among SNPs based on the SAGE pathway data. Under the alternative hypothesis, phenotypes were randomly generated from a logistic model conditional on sex, an autosomal SNP (A , coded 0/1/2), and an X chromosome SNP (X , coded 0/1/2 for females and 0/2 for males) with additive genetic effects:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{sex}sex + \beta_A A + \beta_X X$$

Here $p = P(Y=1|sex, A, X)$, where Y is a binary indicator of case status and sex is a binary indicator of female sex. β_{sex} is the effect of sex, β_A is the effect of the autosomal SNP, and β_X is the effect of the X chromosome SNP, in terms of the log of the odds ratio. The intercept β_0 was chosen to achieve equal proportions of cases and controls on average. Effect sizes for sex and SNP effects were varied ($OR_{sex}=1.0, 1.5, 2.0, \text{ or } 2.5$; $OR_{SNP}=1.0, 1.1, \text{ or } 1.25$; Supplemental Tables 2 and 3). For the simulations with independent SNPs, $P=1000$ SNPs were simulated (2 causal and 998 non-causal); the causal X and autosomal SNPs were randomly selected. For the simulations with LD based on the SAGE pathway data, $P=988$ SNPs were simulated (2 causal and 986 non-causal); the causal X and autosomal SNPs were selected to have similar MAF and patterns of LD (Supplemental Figure 1).

Sensitivity analyses were also conducted to evaluate the performance of all of the methods under deviations from additivity, including nonlinear inheritance patterns and sex-specific effects. In particular, we considered a dominant genetic model, where A and X were coded 1 for heterozygotes and rare homozygotes (or males with the rare X allele) and 0 for common homozygotes (or males with the common X allele) for the purpose of generating the trait Y , using the above logistic model ($OR_{sex}=1.0 \text{ or } 2.0$; $OR_A = OR_X = 1.1 \text{ or } 1.25$); for analysis, A

and X were coded 0/1/2 as defined above. Furthermore, autosomal and X chromosome SNP*sex interactions were also considered under the following model, where the causal SNP effect (for A or X) is only present in females:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{sex}sex + \beta_{SNP,sex}SNP * sex$$

The coefficients were chosen such that $OR_A=OR_X=1.25$ in females and $OR_A=OR_X=1.0$ in males, and data was simulated with and without an independent effect of sex ($OR_{sex}=1.0$ or 2.0). Simulations for the dominant inheritance pattern and sex-specific effects were conducted assuming the LD patterns based on the SAGE data.

For each simulation scenario, 1000 datasets were generated with $N=2000$. Each simulated dataset was analyzed with the five RF approaches (standard RF without sex, standard RF with sex, RF-stratified, RF-split, and RF-XCI) using 5000 trees and $mtry=\sqrt{P}$, where P is the total number of SNPs in the dataset. The primary measure of interest evaluated by the simulations was variable importance, measured by mean decrease in accuracy (unscaled), a measure frequently used with RF to rank variables [Goldstein, et al. 2011]. Although RF is often used as a tool to develop predictive models, determining prediction error was not the goal of this study; rather, the goal was to examine the bias in variable importance measures that may be associated with including X SNPs in traditional RF analyses in comparison to autosomal SNPs, and whether the proposed alternatives can correct this X chromosome importance bias while preserving power in regards to variable rankings. Methods were compared based on the raw MDA, as well as variable rankings of the autosomal and X chromosome SNPs. All analyses were conducted in R statistical software version 2.14.0, and utilizing the R package ‘randomForest’ v4.6-7 [Liaw and Wiener 2002] as well as our new R package, ‘snprf’, that implements the newly proposed approaches.

Results

Study of Alcohol Dependence

Figure 1 shows the MDA variable importance measures for all SNPs in the NMDA-trafficking cascade pathway based on the analysis of the SAGE data using all 5 RF methods. For the standard RF approach that does not include sex as a variable (standard RF without sex), MDA is considerably higher for X chromosome SNPs than for autosomal SNPs, and this large difference remains even when sex is included as a predictor (standard RF with sex). On the other hand, for the three alternative RF approaches, the MDA for the X chromosome is similar to the MDA for the autosomes (Figure 1). In fact, when compared to the significance (p-value) rankings of the previously published univariate regression-based associations [Karpyak, et al. 2012], the standard RF methods clearly overestimate the importance of X SNPs; however, the new methods rank SNPs similarly to the univariate regression analyses. The RF-XCI approach ranks rs980272 in CAMK2A as the top SNP, which was the top SNP with univariate regression ($p=0.0006$) [Karpyak, et al. 2012], and the RF-stratified and RF-split methods also rank this SNP highly (Table 1); this SNP shows an additive pattern, where the minor allele increases risk for alcohol dependence (Supplemental

Figure 2A). Furthermore, rs8045712 in GRIN2A is also highly ranked by the alternative methods (Table 1). While associations with this gene were previously detected in the regression-based analyses, this particular SNP was not previously implicated in univariate logistic regression analyses ($p=0.13$); this SNP shows an over-dominance pattern, where the heterozygote has a reduced risk of alcohol dependence (Supplemental Figure 2B). The use of RF methods allowed for identification of such a nonlinear effect that was missed by univariate logistic regression. Furthermore, the use of a multivariate method such as RF estimates the importance of a particular variable while taking other variables into account, further explaining differences with univariate methods. Other multivariate methods, such as multiple logistic regression or penalized logistic regression, may yield similar results.

Simulation Results: Performance of Methods in the Absence of SNP Effects (Null Distribution)

As expected, when sex is not associated with the trait ($OR_{sex}=1$), all methods (standard and alternative RF) have variable importance centered at zero for the autosomes as well as the X chromosome SNPs, both when SNPs are independent and with LD patterns based on the SAGE data (Supplemental Figure 3). However, if sex is associated with the trait ($OR_{sex}=4$), then standard RF has inflated variable importance measures for the X SNPs—even when sex is included as a predictor—for both independent and correlated SNPs (Figure 2). The alternative methods of RF-stratified and RF-XCI remove this inflation; RF-split reduces the inflation, but still has higher variable importance for X SNPs, particularly when the SNPs are independent.

The degree of imbalance in the number of cases relative to the number of controls impacts the RF variable importance measures, dependent on the pattern of LD. When the number of cases and controls is not equal, the average MDA of X chromosome SNPs for the RF-stratified and RF-XCI methods is slightly lower than expected for SNPs simulated under LD. Conversely, the average MDA for the standard RF methods (with and without sex) is consistently higher for the X SNPs than the autosomal SNPs in all scenarios where sex is associated with the trait; but even when sex is not associated with the trait, the standard RF methods yield inflated X chromosome variable importance measures for SNPs simulated independently, if the number of cases differs from the number of controls (Supplemental Table 1).

RF-split (the method that splits on sex at the first node) gives unstable estimates of MDA across scenarios when comparing the X chromosome SNPs to the autosomes; often the average MDA is higher for the X SNPs, although in some situations it is lower.

Simulation Results: Performance of Methods in the Presence of SNP Effects (Alternative Distribution)

Because under the null hypothesis of no genetic effects we found that standard RF (with and without sex as a predictor) resulted in biased MDA for the X chromosome SNPs when sex was associated with the trait, the standard RF approach was only considered under the alternative scenarios when sex was not associated with the trait. Similarly, RF-split was also

biased under the null in many scenarios, so was also eliminated from consideration in simulations under the alternative hypothesis.

Simulation results demonstrated that when sex is associated with the trait ($OR_{sex} = 1$), the RF-XCI and RF-stratified methods perform similarly, although RF-XCI often ranks the causal autosomal SNP higher than RF-stratified, whereas RF-stratified often ranks the causal X SNP higher than RF-XCI (Supplemental Tables 2 and 3). However, when sex is not associated with the trait ($OR_{sex}=1$), RF-XCI performs better than RF-stratified, ranking the causal SNPs higher for both autosomal and X SNPs. Compared to standard RF, the RF-XCI obtains similar rankings for autosomal SNPs, although may rank the causal SNP lower for X SNPs (Supplemental Tables 2 and 3). Results were similar when SNPs were independent and when SNPs were generated under the SAGE LD patterns (for all choices of causal SNPs).

To illustrate these patterns, we present results for the low SNP effect size of $OR_{SNP}=1.10$ where SNP A=rs1806201 and SNP X=rs602086. When sex is associated with the trait ($OR_{sex}=2$), the RF-XCI method ranks the causal autosomal and X chromosome SNPs above the non-causal SNPs, whereas RF-stratified ranks the causal X chromosome SNP above the non-causal SNPs, but not the causal autosomal SNP (Figure 3A). When sex is not associated with the trait, RF-stratified and RF-XCI retain good performance as compared to standard RF, and continue to rank the causal SNPs highly (Figure 3B). Similar results were observed for all scenarios of the alternative hypothesis with additive SNP effects on the phenotype (Supplemental Tables 2 and 3). Similar results were also observed for the sensitivity analyses conducted under the dominant and SNP-sex interaction models; RF-stratified ranked the causal X chromosome SNP higher than RF-XCI, whereas RF-XCI ranked the causal autosomal SNP higher than RF-stratified (data not shown).

Discussion

In this study, we introduced three methods that allow for unbiased estimation of X chromosome SNP effects in the RF algorithm, and compared the variable importance measures based on these approaches to those obtained with naïve implementations of standard RF tools. By analyzing a candidate pathway for risk of alcohol dependence, in a dataset that includes more male than female cases but fewer male controls, we demonstrated that standard RF approaches ranked X chromosome SNPs highly as compared to the autosomal SNPs. Conversely, the newly proposed approaches produced rankings similar to univariate logistic regression analyses that properly account for X chromosome genotype differences between men and women, but also have the ability to identify nonlinear models missed by approaches that assume additivity while also adjusting for multiple variants simultaneously. The results of the alcohol dependence data analysis were confirmed through simulation studies, which showed that standard RF produces biased results for the X chromosome if sex is associated with the trait. The three proposed RF extensions corrected or reduced this bias.

It is notable that standard RF methods produce biased variable importance estimates for X chromosome variants, even if sex is included as a variable, and if X variants are coded 0,1,2,

in females and 0.2 in males, as is commonly suggested for regression-based approaches [Konig, et al. 2014]. Typically, this type of variable coding eliminates any sex-associated bias in X chromosome variants for regression based methods, because it assumes linearity in genetic effects across males and females, and aligns well with the common assumption of an additive genetic model. However, because the decision-tree structure of RF allows for the modeling of non-linear effects, the standard solution based on linearity fails to correct the problem if there is a strong association between sex and the phenotype. Furthermore, the standard RF method (with and without sex adjustment) also exhibits inflation of X chromosome importance under the null distribution—even when sex is not associated with the trait—when SNPs are independent for data with unequal numbers of cases and controls. This suggests that standard RF methods can lead to biased X chromosome rankings, even when X chromosome bias due to confounding with sex is not expected.

Surprisingly, the method based on splitting on sex in the first node of every tree in the forest (RF-split) resulted in poor performance, and gave variable importance results very different from RF-stratified. We hypothesize that this is because bootstrap samples for individual trees may result in substantial sex imbalances, resulting in an uneven distribution at the first node and unstable estimates of variable importance. This would be exacerbated if the sex distribution is unbalanced relative to the case/control distribution in the data as a whole, which explains why the performance of RF-split is more similar to standard RF. Conversely, the RF-stratified and RF-XCI methods are less sensitive to sample imbalances regarding sex. Of particular concern, RF-split produces upwardly biased X SNP importance measures under the null distribution of no genetic association, even for scenarios where sex is not associated with case-control status—for example, when there are more females than males, unequal numbers of cases and controls, and SNPs are independent.

When sex is associated with the phenotype of interest, both the RF-stratified and RF-XCI methods perform well. Potential bias in variable importance of X chromosome SNPs under the null distribution is eliminated. In the presence of SNP effects on the phenotype, both methods rank the true causal SNPs highly. For sex-associated traits, RF-XCI may perform better for autosomal SNPs, whereas RF-stratified may perform better for X SNPs. Although the RF-stratified method might be preferred because of ease of implementation using current software, for traits that are not associated with sex, RF-stratified is less powerful than RF-XCI, particularly for the autosomes. This is likely due to a reduction in sample size, since RF-stratified models data from each sex separately, whereas RF-XCI utilizes all samples in a single RF analysis. Therefore, although it requires haplotype phasing, the RF-XCI approach is preferred, because it retains similar performance to identify causal SNPs to standard RF when sex is not associated with the trait, but also removes potential X chromosome bias for traits that are sex-associated.

Additionally, the RF-XCI method not only addresses potential sex-bias due to confounding, but also leverages current knowledge of X chromosome biology. Because it attempts to model the effects of the biological process of X chromosome inactivation, it may more accurately capture the genetic etiology of the disease of interest. Furthermore, as more knowledge is gained regarding how XCI occurs, RF-XCI can be extended to model the biological process more accurately. For example, variants in the pseudo-autosomal region of

the X and Y chromosomes (which are inherited like autosomes)[Ross, et al. 2005] can be modeled with the autosomes; also, weights could be used in the random selection of haplotypes in females if it is known that one allele is preferentially expressed (e.g. via DNA methylation[Allen, et al. 1992] or RNA sequencing data[Wu, et al. 2014]).

This study also demonstrated that case/control sample imbalance in combination with the presence of LD also had an overall impact on the results under the null distribution of no genetic association for RF-stratified and RF-XCI. In particular, when SNPs are independent, RF-stratified and RF-XCI show no evidence of X chromosome importance bias for either data with case/control imbalance or for data with equal numbers of cases and controls; but when SNPs are in LD, the X chromosome importance estimates for RF-stratified and RF-XCI under the null distribution are lower than expected, implying that these methods are unlikely to yield inflated estimates of X chromosome variable importance under any scenario. In these scenarios it is likely that the differences in importance measures are driven by differences in LD between the X chromosome and the autosomes, in combination with case-control imbalance. Likely this is a problem of RF in general, not specific to the X chromosome; it has been previously shown that strong LD can result in biased importance measures [Meng, et al. 2009], although this has been previously unexplored regarding case-control imbalance.

Data-mining methods offer certain advantages for genetic studies over traditional univariate statistical approaches, such as the ability to model the effects of multiple variants simultaneously. Previously, there were no methods specifically designed to correctly model X chromosome data in these types of analyses. In this study, we have developed an extension of RF that allows for simultaneous unbiased modeling of the effects of autosomal and X-chromosome SNPs. We recommend the use of the RF-XCI approach for analysis of SNP data based on its good statistical performance and biological relevance. To facilitate ease of use, we have implemented the RF-XCI method in a freely-available R package ‘snpRF’, available on CRAN (<http://www.cran.r-project.org/web/packages/snpRF/>).

In this study, we utilized RF as a non-parametric multivariate tool to perform variable ranking in high-dimensional data using measures of variable importance [Janitza, et al. 2013; Lunetta, et al. 2004; Meng, et al. 2009; Strobl, et al. 2008]. However, RF is also often used for prediction. Because the focus of this paper was on assessing potential biases in the estimated variable importance measures of X chromosome SNPs compared to autosomal SNPs, we did not focus on comparisons of prediction error. Prediction error is a global measure that reflects the overall performance of all predictors; if sex is included as a variable in RF, then the prediction error represents the predictive ability of the SNPs and sex, jointly. Because in our study the prediction error estimates did not always represent the predictive ability of the SNPs alone, they were not comparable across standard RF and the alternative methods. Development of a measure of the genetic contribution to predictive ability is an avenue of future investigation for RF beyond the inclusion of X chromosome data.

This study represents important progress facilitating the analysis of X chromosome data with RF. Because various versions of RF were compared, other aspects important to analysis

with RF were not examined here, such as parameter tuning, which could be examined in future studies. Only the standard RF approach for classification, originally proposed by Breiman [Breiman 2001] and implemented in the R package 'randomForest', was utilized as a comparison approach; other variations of RF could also be considered, such as the 'cforest' function available in the R package 'party' [Strobl, et al. 2008; Strobl, et al. 2007]. Furthermore, the RF approaches were not compared to the many existing univariate approaches for analysis of X chromosome variants [Clayton 2008; Purcell, et al. 2007; Wang, et al. 2014] or multivariate methods such as multiple logistic regression or penalized logistic regression (with proper variant in coding in males and females). Such future comparisons will be critical, particularly regarding complex models with multiple SNPs and deviations from additivity, to determine the utility of the new methods.

Additionally, the RF-XCI method requires haplotype phasing, and the impact of haplotype uncertainty was not examined. Because the estimation of haplotypes is conducted for the purpose of preserving LD, but the RF analysis is performed on the genotypes rather than the haplotypes, the impact of phase uncertainty is expected to be minimal in regions of low LD (because there is little LD to preserve) and high LD (because haplotype uncertainty will be low). Although a thorough investigation of the impact of haplotype uncertainty is beyond the scope of this work, we note that the RF-XCI method could be easily extended to incorporate haplotype uncertainty by randomly sampling the haplotype for each tree from the haplotype probability distribution estimated for each female during phasing.

The assumptions of the alternative distribution were rather simplistic (e.g. additive SNP effects with a single X chromosome or autosomal SNP), and may not reflect the genetic etiology of complex human traits that manifest sex differences. Although performance was similar under models exhibiting nonlinear inheritance patterns and SNP-sex interactions, these were only considered as sensitivity analyses and were not a primary avenue of investigation. In particular, SNP*sex interaction effects were not thoroughly examined, which represents a critical avenue for future study, particularly regarding studies of traits with sex differences. In the scenarios examined here, RF-XCI and RF-stratified both performed well, although the stratified approach had less power to detect marginal effects due to reduced sample size, particularly for the autosomes; however, our results suggest that it may be a powerful approach to capture SNP*sex interactions.

The results of this study may have implications beyond the use of RF to analyze X chromosome SNP data. We demonstrated that simply including sex as a predictor in the standard RF algorithm did not eliminate confounding with sex, nor did requiring that sex be considered at the first node of each tree. These results suggest that there may be concerns with how covariates are incorporated into RF analyses in general. Furthermore, confounding with sex may be a concern with other tree-based data-mining approaches to genetic analysis beyond RF that were designed to capture non-linear effects. Similar extensions to those described here based on sex stratification or haplotype sampling to mimic X inactivation could also be incorporated in these other data-mining methods. Lastly, these research findings are relevant even for traits that are not associated with sex in the general population (i.e. no sex differences in disease prevalence), because sex proportions often differ between cases and controls due to study design; for instance, women are more likely to participate as

study controls. Therefore the use of standard RF may result in X chromosome importance bias even in situations where it is not expected.

In summary, in this study we have characterized the potential bias in variable importance measures that can occur if X chromosome data is naively included with autosome data in RF analyses. While X chromosome data is often excluded from genetic analyses to avoid such bias, we believe that it may represent an untapped resource for genetic epidemiology studies, particularly for studies of traits with sex differences in prevalence rates or clinical presentation. Therefore we have provided a powerful multi-marker approach for genetic analysis that accommodates X chromosome data in an unbiased way, and this method is freely available in the R package ‘snpRF’.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by funding from the Office of Women's Health Research (Building Interdisciplinary Careers in Women's Health award K12HD065987, PI Bahn) and the National Institute on Drug Abuse (R21 DA019570, PI Biernacka).

References

- Allen RC, Zoghbi HY, Moseley AB, Rosenblatt HM, Belmont JW. Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *Am J Hum Genet.* 1992; 51(6):1229–39. [PubMed: 1281384]
- Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S. A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci U S A.* 2010; 107(11):5082–7. others. [PubMed: 20202923]
- Brady KT, Grice DE, Dustan L, Randall C. GENDER DIFFERENCES IN SUBSTANCE USE DISORDERS. *American Journal of Psychiatry.* 1993; 150(11):1707–1711. [PubMed: 8214180]
- Breiman L. Bagging predictors. *Machine Learning.* 1996; 24(2):123–140.
- Breiman L. Random Forests. *Machine Learning.* 2001; 45:5–32.
- Breiman, L.; Friedman, J.; Stone, CJ.; Ohlsen, RA. *Classification and Regression Trees.* Chapman and Hall; 1984.
- Brieman L. Random Forests. *Machine Learning.* 2001; 45(1):27.
- Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature.* 1991; 349(6304):38–44. [PubMed: 1985261]
- Carrel L, Cottle AA, Goglin KC, Willard HF. A first-generation X-inactivation profile of the human X chromosome. *Proc Natl Acad Sci U S A.* 1999; 96(25):14440–4. [PubMed: 10588724]
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015; 4:7. [PubMed: 25722852]
- Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, 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. *PLoS One.* 2014; 9(12):e113684. [PubMed: 25479423]
- Clayton D. Testing for association on the X chromosome. *Biostatistics.* 2008; 9(4):593–600. [PubMed: 18441336]
- Clayton DG. Sex chromosomes and genetic association studies. *Genome Med.* 2009; 1(11):110. [PubMed: 19939292]

- de Maturana EL, Ye Y, Calle ML, Rothman N, Urrea V, Kogevinas M, Petrus S, Chanock SJ, Tardon A, Garcia-Closas M. Application of multi-SNP approaches Bayesian LASSO and AUC-RF to detect main effects of inflammatory-gene variants associated with bladder cancer risk. *PLoS One*. 2013; 8(12):e83745. others. [PubMed: 24391818]
- Disteche CM. Dosage Compensation of the Sex Chromosomes. *Annual Review of Genetics*, Vol 46. 2012; 46:537–560.
- Fridley BL, Biernacka JM. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet*. 2011; 19(8):837–43. [PubMed: 21487444]
- Goldman D, Oroszi G, Ducci F. The genetics of addictions: uncovering the genes. *Nat Rev Genet*. 2005; 6(7):521–32. [PubMed: 15995696]
- Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet*. 2010; 11:49. [PubMed: 20546594]
- Goldstein BA, Polley EC, Briggs FBS. Random Forests for Genetic Association Studies. *Statistical Applications in Genetics and Molecular Biology*. 2011; 10(1)
- Hickey PF, Bahlo M. X chromosome association testing in genome wide association studies. *Genet Epidemiol*. 2011; 35(7):664–70. [PubMed: 21818774]
- Janitz S, Strobl C, Boulesteix AL. An AUC-based permutation variable importance measure for random forests. *BMC bioinformatics*. 2013; 14:119. [PubMed: 23560875]
- Karpyak VM, Geske JR, Colby CL, Mrazek DA, Biernacka JM. Genetic variability in the NMDA-dependent AMPA trafficking cascade is associated with alcohol dependence. *Addict Biol*. 2012; 17(4):798–806. [PubMed: 21762291]
- Konig IR, Loley C, Erdmann J, Ziegler A. How to include chromosome X in your genome-wide association study. *Genetic epidemiology*. 2014; 38(2):97–103. [PubMed: 24408308]
- Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002; 2(3):18–22.
- Loley C, Ziegler A, König IR. Association tests for X-chromosomal markers--a comparison of different test statistics. *Human heredity*. 2011; 71(1):23–36. [PubMed: 21325864]
- Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet*. 2004; 5(1):32. [PubMed: 15588316]
- Lyon MF. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature*. 1961; 190:372–3. [PubMed: 13764598]
- Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*. 2007; 39(10):1181–6. others. [PubMed: 17898773]
- Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL. Performance of random forest when SNPs are in linkage disequilibrium. *BMC bioinformatics*. 2009; 10:78. [PubMed: 19265542]
- Montana G. HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics*. 2005; 21(23):4309–11. [PubMed: 16188927]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81(3):559–75. others. [PubMed: 17701901]
- Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP. The DNA sequence of the human X chromosome. *Nature*. 2005; 434(7031):325–37. others. [PubMed: 15772651]
- Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*. 2006; 78(4):629–644. [PubMed: 16532393]
- Schwarz DF, König IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*. 2010; 26(14):1752–8. [PubMed: 20505004]
- Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012; 28(1):112–8. [PubMed: 22039212]
- Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008; 9:307. [PubMed: 18620558]

- Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*. 2007; 8:25. [PubMed: 17254353]
- Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, Sun YV. Machine learning in genome-wide association studies. *Genetic epidemiology*. 2009; 33(Suppl 1):S51–7. [PubMed: 19924717]
- Thornton T, Zhang Q, Cai X, Ober C, McPeck MS. XM: association testing on the X-chromosome in case-control samples with related individuals. *Genet Epidemiol*. 2012; 36(5):438–50. [PubMed: 22552845]
- Wang J, Yu R, Shete S. X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genetic epidemiology*. 2014; 38(6):483–93. [PubMed: 25043884]
- Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, Biernacka JM. SNP interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*. 2012; 13(1):164. [PubMed: 22793366]
- Winham SJ, De Andrade M, Miller VM. Genetics of cardiovascular disease: Importance of sex and ethnicity. *Atherosclerosis*. 2015 In Press.
- Winham SJ, Freimuth RR, Biernacka JM. A Weighted Random Forests Approach to Improve Predictive Performance. *Stat Anal Data Min*. 2013; 6(6):496–505. [PubMed: 24501613]
- 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. [PubMed: 23643377]
- Wu H, Luo J, Yu H, Rattner A, Mo A, Wang Y, Smallwood PM, Erlanger B, Wheelan SJ, Nathans J. Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease. *Neuron*. 2014; 81(1):103–19. [PubMed: 24411735]
- Ye Y, Zhong X, Zhang H. A genome-wide tree- and forest-based association analysis of comorbidity of alcoholism and smoking. *BMC Genet*. 2005; 6(Suppl 1):S135. [PubMed: 16451594]
- Zheng G, Joo J, Zhang C, Geller NL. Testing association for markers on the X chromosome. *Genet Epidemiol*. 2007; 31(8):834–43. [PubMed: 17549761]

Figure 1.

Manhattan plot of MDA measures for the SAGE alcohol dependence and NMDA-dependent AMPA trafficking cascade pathway data for each of 5 RF methods: A) Standard RF without sex as a predictor, B) Standard RF with sex as a predictor, C) RF-Stratified, D) RF-Split, and E) RF-XCI.

Figure 2. MDA averaged across autosomal SNPs and X chromosome SNPs under the null distribution for 1000 simulations, when sex is associated with the trait ($OR_{sex}=4.0$), and equal numbers of cases and controls and males and females. (A) SNPs are independent. (B) SNPs were generated with LD based on real data.

Figure 3.

MDA variable importance rankings averaged across 1000 simulations for autosomal SNPs compared to X chromosome SNPs under the alternative distribution ($OR_A=OR_X=1.10$), when (A) sex is and (B) is not associated with the trait ($OR_{sex}=2.0$ vs. 1.0). A ranking of 1 indicates that the variable is least important, whereas a ranking 988 indicates that the variable is most important. SNPs were generated with LD based on real data, where causal SNP X=rs602086 and causal SNP A= rs1806201, denoted by '*'. *

Table 1

Top ranking SNPs based on variable importance for the 5 RF methods, compared to previously published univariate logistic regression p-value rankings (chromosome: rsID).

	Standard RF without sex	Standard RF with sex	RF-Stratified	RF-Split	RF-XCI	Logistic Regression
1	X: rs5911595	X: rs12557782	16: rs8045712	16: rs2267780	5: rs980272	5: rs980272
2	X: rs7058062	X: rs17325268	16: rs2267780	16: rs16966381	16: rs8045712	11: rs2508467
3	X: rs12557782	X: rs983007	16: rs9933624	5: rs980272	1: rs11591167	16: rs8050843
4	X: rs5911609	X: rs5911609	16: rs4628972	16: rs8045712	12: rs1806192	16: rs2077923
5	X: rs17325268	X: rs5911595	11: rs2508467	16: rs7205180	11: rs2508467	16: rs9933832