HHS Public Access

Author manuscript

Hum Mutat. Author manuscript; available in PMC 2018 May 25.

Published in final edited form as:

Hum Mutat. 2012 December ; 33(12): 1708–1718. doi:10.1002/humu.22161.

Use of Support Vector Machines for Disease Risk Prediction in Genome-Wide Association Studies: Concerns and Opportunities

Florian Mittag[1], **Finja Büchel**[1], **Mohamad Saad**[2,3], **Andreas Jahn**[1], **Claudia Schulte**[4], **Zoltan Bochdanovits**[5], **Javier Simón-Sánchez**[5], **Mike A. Nalls**[6], **Margaux Keller**[6,7], **Dena G. Hernandez**[6,8], **J. Raphael Gibbs**[6,8], **Suzanne Lesage**[9,11,12], **Alexis Brice**[9,10,11,12], **Peter Heutink**[5], **Maria Martinez**[2,3], **Nicholas W Wood**[8], **John Hardy**[8], **Andrew B. Singleton**[6], **Andreas Zell**[1], **Thomas Gasser**[4], and **Manu Sharma**[4] **for the International Parkinson's Disease Genomics Consortium (IPDGC)**

[1]Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, Tubingen, Germany [2]Institut National de la Sante et de la Recherche Medicale, UMR 1043, Centre de Physiopathologie de Toulouse-Purpan, Toulouse, France [3]Département des Sciences du Vivant, Paul Sabatier University, Toulouse, France [4]Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Centre for Neurodegenerative Diseases, Tübingen, Germany [5]Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre, Amsterdam, The Netherlands [6]Laboratory of Neurogenetics. National Institute on Aging, National Institutes of Health, Bethesda, Maryland [7]Department of Biological Anthropology. Temple University, Philadelphia, Pennsylvania [8]Department of Molecular Neuroscience, Institute of Neurology, University College London, London, UK [9]Université Pierre et Marie Curie-Paris, Centre de Recherche de l'Institut du Cerveau et de laMoelle Epinière, UMR-S975, Paris, France [10]AP-HP, Hôpital de la Salpêtrière, Département de Génétique et Cytogénétique, Paris, France [11]Institut National de la Sante et de la Recherche Medicale, UMR_S975 CRicm, Paris, France [12]Centre National de la Recherche Scientifique, UMR 7225, Paris, France

Abstract

*Correspondence to: Manu Sharma, Hertie-Institute of Clinical Brain Research, Department of Neurology, University of Tuebingen, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany. manu.sharma@uni-tuebingen.de.

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsors: Intramural Research Programs of the National Institute on Aging; National Institute of Neurological Disorders and Stroke; National Institute of Environmental Health Sciences; National Human Genome Research Institute of the National Institutes of Health; Department of Health and Human Services (project numbers Z01-AG000949-02 and Z01-ES101986); Human subjects protocol 2003-077; US Department of Defense (award number W81XWH-09-2-0128); National Institutes of Health (grants NS057105 and RR024992); American Parkinson disease Association (APDA); Barnes Jewish Hospital Foundation; Greater St Louis Chapter of the APDA; Hersenstichting Nederland; Neuroscience Campus Amsterdam; the section of medical genomics, the Prinses Beatrix Fonds; and the Michael J. Fox foundation; the KORA (Cooperative Research in the Region of Augsburg) research platform was started and nanced by the Forschungszentrum für Umwelt und Gesundheit, which is funded by the German Federal Ministry of Education, Science, Research, and Technology and by the State of Bavaria; German National Genome Network (NGFNplus number 01GS08134, German Ministry for Education and Research); German Federal Ministry of Education and Research (NGFN 01GR0468, PopGen); Initiative and Networking Fund of the Helmholtz Association (01EW0908 in the frame of ERA-NET NEURON); French National Agency of Research (ANR-08-MNP-012).

The success of genome-wide association studies (GWAS) in deciphering the genetic architecture of complex diseases has fueled the expectations whether the individual risk can also be quantified based on the genetic architecture. So far, disease risk prediction based on top-validated single-nucleotide polymorphisms (SNPs) showed little predictive value. Here, we applied a support vector machine (SVM) to Parkinson disease (PD) and type 1 diabetes (T1D), to show that apart from magnitude of effect size of risk variants, heritability of the disease also plays an important role in disease risk prediction. Furthermore, we performed a simulation study to show the role of uncommon (frequency 1–5%) as well as rare variants (frequency <1%) in disease etiology of complex diseases. Using a cross-validation model, we were able to achieve predictions with an area under the receiver operating characteristic curve (AUC) of ~0.88 for T1D, highlighting the strong heritable component (~90%). This is in contrast to PD, where we were unable to achieve a satisfactory prediction (AUC ~0.56; heritability ~38%). Our simulations showed that simultaneous inclusion of uncommon and rare variants in GWAS would eventually lead to feasible disease risk prediction for complex diseases such as PD. The used software is available at http://www.ra.cs.uni tuebingen.de/software/MACLEAPS/.

Keywords

genome-wide association studies; disease risk prediction; machine learning; support vector machines; Parkinson disease

Introduction

Apart from confirming previously reported genetic associations, genome-wide association studies (GWAS) also identified new genetic loci that led to a new mechanistic insight into the genetic architecture of complex diseases [Hirschhorn and Daly, 2005; Ioannidis et al., 2009; Zeggini et al., 2008; Zondervan and Cardon, 2007]. Although the first wave of GWAS unequivocally confirmed a number of susceptible variants in different disorders in different populations, they explain only a small proportion of heritability, that is, the proportion of phenotypic variability in a population due to additive genetic factors. With the release of the 1000 Genomes data (www.1000genomes.org), it is anticipated that future genetic studies will fill the gap between known and yet-to-be defined heritability in complex disorders [Altshuler et al., 2010; Olsen et al., 2007]. For example, a recent study on Parkinson disease using 1000 Genomes data revealed new putative loci that were overlooked in previous genome scans [Nalls et al., 2011]. Apart from identifying risk factors for complex diseases, geneticists are also interested in disease risk prediction for multifactorial disorders, such as Diabetes, Crohn's disease, and neurodegenerative disorders, to provide patients with an early diagnosis that is based on their genetic architecture, as determined by array-based genotyping technologies.

Compared to complex disorders, genetic risk profiling already exists for monogenic disorders such as phenylketonuria (PKU) and Huntington disease's (HD) [Janssens et al., 2006]. Monogenic disorders are caused by mutation(s) within a single gene; the risk of a disease to carriers is substantially higher than to noncarriers. Unlike monogenic disorders, multiple genes either act alone or synergistically along with environmental factors in

complex disorders, each having a minor effect in influencing the disease susceptibility and progression, which makes genetic risk prediction difficult in complex disorders [Janssens and van Duijn, 2008; Manolio, 2010].

A number of studies have performed *genomic profiling* using most significant single-nucleotide polymorphisms (SNPs) nominated by either a single GWAS or a meta-analysis of GWAS [Janssens and van Duijn, 2006; Janssens et al., 2006; Nalls et al., 2011]. Genomic profiling is carried out based on the assumption that top-validated markers could also be effective classifiers and eventually be used in clinical decision-making to improve treatment. Thus far, risk-profiling analyses performed on top-validated SNPs have not been very encouraging [Jakobsdottir et al., 2009; Janssens et al., 2006]. The results showed only a marginal increase in area under the receiver operating characteristic (ROC) curve (area under the receiver operating characteristic curve [AUC]) performance, thus making it difficult to put these findings into clinical practice [Jakobsdottir et al., 2009]. These results, nevertheless, were not unexpected because most GWAS showed that the effect sizes of uncommon variants on complex disorders are smaller than anticipated and heritability ismoderate [So et al., 2011;Visscher et al., 2008]. Therefore, the clinical utility of genetic predictions based on a small number of SNPs is likely to be negligible.

The field of machine learning provides a variety of methods to approach these challenges, such as linear or logistic regression techniques, decision trees, random forests, and Bayesian approaches, which can provide an improvement over standard regression approaches by including a statistical analysis in the context of Bayesian inference. Although it is possible to directly apply numerical regression methods to classification tasks, these approaches have well-known deficiencies compared to methods designed specifically for classification. Further, standard regression techniques are not well suited for genome-wide analyses because they assume independence of markers, which is usually not true due do linkage disequilibrium (LD) [Szymczak et al., 2009].

Recently, the support vector machine (SVM) was proposed to performgenome-wide disease risk predictions based on GWAS data [Szymczak et al., 2009; Wei et al., 2009] and was shown to outperform logistic regression on type 1 diabetes (T1D) dataset. Both of these techniques find linear boundaries in feature space: an intuitive separation surface that leads to efficient learning methods.

Unlike logistic regression or number-of-risk-allele-based approaches, which depend upon a predetermined model to determine the probability of an event by fitting the data to a logistic curve, SVMs discriminate between two classes (here, “case,” and “control”) by finding a separating hyperplane for the data points, potentially by transforming input data (SNP genotypes) into a higher-dimensional feature space [Ben-Hur et al., 2008; Evans et al., 2009; Szymczak et al., 2009]. Most importantly, while many methods find a linear decision boundary, the SVM finds the one with the largest margin between both classes (see section “Materials and Methods”).

This leads to better generalization performance and deterministic solution. Because of these advantages and its prominence in classification, we focus on SVMs in this study. SVMs have

been used successfully for disease risk prediction [Wei et al., 2009, 2011] and, particularly, they have been successfully applied to type 1 diabetes (heritability is high ~90%). However, the utility of SVMs has not been tested on neurodegenerative disorders, such as PD, where the heritability component is moderate (~38%).

We applied this unbiased approach on four independent PD-GWAS data sets to assess the disease risk prediction. The advantage of using this approach is that it does not require external prior knowledge about known risk variants of the underlying disease.

Furthermore, we performed a simulation study using different scenarios to determine whether uncommon variants alone or multiple uncommon/rare variants are helpful in clinical settings to make better diagnostic assessments for common diseases.

In addition, we developed a stand-alone user-friendly Java-based software, Machine Learning Analysis Pipeline for SNP data (MACLEAPS), to perform genome-wide disease risk prediction for complex diseases.

Material and Methods

Study Cohorts

For PD, we used four-stage 1 GWAS datasets (French, Dutch, CIDR, and German/US) for our study (see Table 1). Those datasets have already been subject to extensive quality control procedures, whose details are described elsewhere [Pankratz et al., 2009; Saad et al., 2011; Simón-Sánchez et al., 2009, 2011]. All four datasets have nonoverlapping samples. In addition, we filtered all datasets to exclude samples or SNPs with more than 5% missing values, variants with less than 5% minor allele frequency (MAF), and samples deviating from the Hardy–Weinberg equilibrium using the PLINK command line tool [Purcell et al., 2007].

It should also be noted that in the German/US GWAS dataset [Simón-Sánchez et al., 2009], individuals with “three or more relatives with parkinsonism or with an apparent Mendelian inheritance of PD were excluded” from the US cohort and “cases showing clear evidence of dominant inheritance were excluded” from the German cohort, giving it focus on sporadic cases of PD, whereas some of the datasets used for external validation have a focus on familiar PD.

In brief, when using a simple cross-validation model, we considered the German/US GWAS dataset as a single dataset, as it is also done in the study from which it originates [Simón-Sánchez et al., 2009] because analyses of the pairwise Identity by State (IBS) showed that the cohort clearly shared common Caucasian ancestry. When using an independent GWAS dataset approach, we also evaluated a model where we treated German and US GWAS datasets as separate studies. Apart from that, we also accessed type 1 diabetes (T1D), type 2 diabetes (T2D), and bipolar disorder (BD) data from WTCCC as described in The Wellcome Trust Case Control Consortium [The Wellcome Trust Case Control Consortium, 2007]. For T1D, T2D, and BD, roughly 2,000 cases and 1,500 controls from the 1958 British Birth

Control Cohort were used (Table 2). The WTCCC data was genotyped using Affymetrix 500K chips.

Nomenclature

Throughout this document, we use the following terms to denote variants with a certain minor allele frequency: “rare” for SNPs with an MAF < 1%, “uncommon” for SNPs with an MAF between 1% and 5%, and “common” for variants with an MAF > 5%, as it was also done in Cirulli and Goldstein (2010).

Overview of the Analysis Pipeline

We applied the support vector machine (SVM) algorithm to train models based on GWAS SNP data that act as binary classifiers for new datasets. We then evaluated the ability of these models to correctly predict the phenotypes of a new set of samples with their given genotypes. We performed two different types of evaluation for our models: (1) within-study cross-validation, where one dataset is split into a training and validation set, and (2) between-study validation, where the data from one study is used to train a classifier, which is then validated using the data from a second study. Finally, we evaluated the predictive performance of the models using the AUC.

Feature Encoding

A GWAS dataset consisting of n individuals and p SNPs can be represented by an $n \times p$ matrix $G = (g_{i,j})$, where $g_{i,j}$ denotes the genotype of SNP j for individual i. A straightforward approach for the numerical encoding of a genotype is to represent it as the number of minor alleles that are present, that is, as 0, 1 or 2. Hence, each individual i is represented through his/her genotype vector $x_i = (g_{i1}, \ldots, g_{ip})$ and his/her disease status $y_i \in \{-1, +1\}$, where −1 denotes unaffected (control) and +1 denotes affected (case).

Support Vector Machines

The basic principle of SVMs is to find a hyperplane that separates a set of labeled data points into two classes, with a large gap (or margin) between them. Their main difference compared to regression methods is that this hyperplane is only defined by the data points that lie on or violate the margin, which are called support vectors, and enables SVM models to perform well on unknown data. Further, using kernel functions, SVMs are also able to learn nonlinear decision boundaries (as discussed below).

As described above, the data is represented as vectors $x_i \in \Re^p$ for each of the n data points (genotype vectors) with their respective class label y_i. The goal of training is to find a decision function $f(x)$ that assigns the correct class label to each data point:

$$f(x_i) = y_i \; \forall 1 \leq i \leq n .$$

SVMs use a decision function of the form $f(x) = \text{sgn}(\langle w, x \rangle + b)$, where w is the hyperplane’s normal vector, $\frac{b}{\|w\|}$ the offset from the origin, and sgn is the sign function which yields +1, if its input is greater than zero and −1 otherwise. The hyperplane can be

represented as a linear combination of the input vectors, so that $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ with $\alpha_i \geq 0$, making the decision function equivalent to

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x\rangle + b\right). \quad (1)$$

Consequently, only those input vectors with $\alpha_i > 0$ define the decision boundary and are thus called support vectors. In addition, w is constrained so that the distance of the hyperplane to these support vectors point is $\frac{1}{\|w\|}$, the margin of the hyperplane. Therefore, the optimization problem of maximizing the separating hyperplane's margin is equivalent to minimizing $\|w\|^2$. Thus, the SVM optimization problem is

$$\min_{w,b}\left\{\frac{1}{2}\|w\|^2\right\} \text{ subject to: } y_i(\langle w, xi\rangle + b) \geq 1. \quad (2)$$

However, perfect linear separation of the data into cases and controls is often not possible. One way to circumvent this problem is to introduce slack variables ξ_i, which allow data to violate the margin. The penalty for this violation is controlled by the cost parameter C. The soft-margin SVM optimization can be then defined as

$$\min_{w,b,\xi}\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \text{ subject to } y_i(\langle w, xi\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \quad (3)$$

The choice of C has a large impact on the classification performance because it directly affects the generalization ability of the learned models. Selecting a large C value can lead to overfitting of data resulting in scenarios in which the model is accurate on training data, but fails to perform well on new data [Guyon and Elisseefi, 2003]. Alternatively, selecting a small C threshold may allow for too many misclassifications and thus leads to failure in learning.

There still remains the problem that the data may not be linearly separable. To address this issue, a standard modification to the linear SVM was made by substituting the dot product $\langle x, x_i\rangle$ in Equation (1) for a kernel function $K(x, x_i)$ as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right). \quad (4)$$

This enables the SVM to learn a model that incorporates nonlinear interactions between features, which in this case can be SNP–SNP interactions. A commonly used kernel function is the Gaussian radial basis function [Ben-Hur et al., 2008], which introduces the γ parameter controlling the width of the Gaussian "bell" that determines the influence of a support vector on its surroundings:

$$K(x, x_i) = e^{-\gamma \|x - x_i\|^2}. \quad (5)$$

Smaller values of gamma widen the bell, increasing the area of influence of each support vector. This limits the overall number of support vectors and yields more linear decision boundaries [Ben-Hur et al., 2008]. Large values for gamma restrict the influence of each support vector, which results in a k-nearest neighbor-like behavior and can lead to overfitting. The optimal choice of gamma has to be evaluated for each training set and also depends on the choice for the parameter C (see below).

For the radial basis function (RBF) kernel, we used the LIBSVM implementation [Chang and Lin, 2011] and a modified LIBLINEAR implementation for linear models. Our version of LIBLINEAR is based on the work of Fan et al. (2008).

Parameter Optimization of SVM

We performed an extensive grid search, where C and γ span a grid of parameter combinations. Usually, the step size of this grid is exponentially growing, for example, 2^{-5}, 2^{-3}, …, 2^{15} for C and 2^{-15}, 2^{-13}, …, 2^{3} for γ (for more details on grid search, please refer to Guyon and Elisseefi, 2003). In brief, each combination of C and γ is evaluated using a k-fold cross-validation. This process is repeated m times. For each fold, $k-1$ parts of the dataset are used to train a model, which is then used to predict the phenotypes of the remaining part. The results of the m cross-validation runs with k-folds each are then averaged and the parameters that exhibit the best performance are chosen to train a model using the entire dataset. Our parameter optimization performed 10 twofold cross-validations ($k = 2, m = 10$) for each parameter combination in the above stated range.

Filtering and Cross-Validation

To increase the optimization of SVM, we applied standard quality control measures on all our datasets. The details are described elsewhere [Nalls et al., 2011; Pankratz et al., 2009; Saad et al., 2011; Simón-Sánchez et al., 2009, 2011]. In brief, we included SNPs that have a call rate greater than 95%, minor allele frequency greater than 5%, and pass a Hardy–Weinberg equilibrium test with P value of 0.01 or better. No external knowledge about known risk variants or similar was used to maintain an unbiased approach. Moreover, we selected only those SNPs for model building that pass a certain significance level in a standard case/control association analysis, a basic allelic trend test, in the training data using PLINK and we used different P value thresholds ranging $P < 1 \times 10^{-8}$ to $P < 1 \times 10^{-3}$ to build our disease risk prediction models. We used twofold and fivefold cross-validation

models to test our data, where 50% and 80% of the data were used to train a model and the rest of the data was used to assess its performance for disease prediction.

For the WTCCC-T1D dataset, we also compared the results to an analysis where we removed all SNPs in the MHC region (chromosome 6, from 25–34 Mbps, see [Wei et al., 2009]) prior to model training to assess the influence of this region.

Genotype Imputation

We used a multimarker tagging-based approach for imputation as implemented in PLINK [Purcell et al., 2007]. For SNPs to be imputed, we used the HapMap phased haplotype data (release 22) on CEU subjects, as downloaded from HapMap website (www.hapmap.org). A reference panel was used to select a small set of flanking SNPs, which when phased together lead to a haplotype background with high LD with these SNPs. In the second step, genotype data from the study sample and reference panel were jointly phased with these SNPs and the missing genotypes in the study samples were imputed during phasing. In addition to that, we used the Markov chain haplotyper (MACH) for imputation on markers which are present on our training datasets but not on our prediction model. We used low coverage sequencing of 112 samples of Caucasian ancestry in the 1000 genomes project (August 2009) as a reference panel. A default two-stage procedure is used for imputation. In brief, 200 randomly selected individuals are used to estimate error rates and crossover rates for parameter estimation for imputation. In the second step, default parameters were used to impute genotypes on more than 7 million SNPs for disease risk prediction.

Performance Evaluation

We evaluated the predictive performance of our models using the AUC value, which is the area under the ROC curve. The ROC curve is obtained by plotting the true positive rate (TPR) versus the false positive rate (FPR) for varying decision thresholds. Removing the sgn-function from the decision function of the SVM leads to a continuous decision value for each sample that reflects the predicted distance to the hyperplane, thus making it possible to rank the predictions from lowest to highest:

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\,. \quad (6)$$

The AUC can be interpreted as the probability that, given a pair of random samples from the positive and the negative class, the positive sample will be ranked higher than the negative one [Wray et al., 2010]. Consequently, classifiers with an AUC of 0.5 are as good as random predictions, whereas an AUC of 1.0 represents a perfect classification.

We chose the AUC as our performance measure because of its robustness against class skew and for comparability to other disease risk assessment studies.

Role of Uncommon and/or Multiple Rare Variants

We simulated two different scenarios (model A and model B) to investigate the discriminative accuracy of genomic profiling, which is the extent to which genomic profiles can distinguish between those who will or will not develop disease. We simulated all genomic profiles consisting of 500,000 SNPs with only 100 of them being associated to the phenotype using PLINK.

In the first scenario, we generated genomic profiles for 1,000, 3,000 and 5,000 subjects, where the odds ratio (OR) of the associated SNPs varies from 2 to 5 and risk allele frequencies vary from 1% to 5%. We designed model A to assess the performance of MACLEAPS when uncommon variants with strong effect size are involved in disease susceptibility. As is shown in recently published GWAS, effect sizes for most common diseases are more moderate than previously anticipated. On the other hand, a recent GWAS [Do et al., 2011] and a meta-analyses of GWAS of PD [Nalls et al., 2011] already reach sample sizes of the next order of magnitude with ~33,000 and ~17,000 samples total. Therefore, in model B (mixed model), genomic profiles were designed for 1,000, 3,000, 5,000, and 20,000 subjects, where we simulated 10 rare variants (this includes LRRK2 and GBA mutations) and 90 uncommon risk variants, where the OR varies from 3 to 9 and from 0.7 to 1.5 for rare variants and uncommon variants, respectively. Furthermore, risk allele frequency varies from <1% to 5% for the mixed model. Because neurodegenerative disorders are age dependent, their prevalence increases linearly with age. We, therefore, varied the prevalence of the disease from 1% to 3% to assess the age-dependent genomic profile for neurodegenerative disorders such as PD.

Results

Evaluation of MACLEAPS Using PD-GWAS Data

We used predefined P value thresholds (see section “Methods”) to filter SNPs for disease prediction. Compared to recently published studies, which only focus on top-validated SNPs to determine genetic prediction, we conducted genomic profiling in an unbiased and systematic manner. We incorporated SNPs that, based on a standard association analysis on the respective dataset, passed a genome-wide threshold ($P < 1 \times 10^{-8}$) to SNPs that showed suggestive evidence of association for PD ($P < 1 \times 10^{-3}$). We observed that when using a liberal threshold for SNP selection, those SNPs are scattered across the genome, whereas we found only a few SNPs, when using genome-wide threshold (this mainly covers a few SNPs in the α-synuclein [SNCA] and microtubule-associated protein tau genes [MAPT]). The SNP selection process does not take LD into account, but the results show that the predictive performance drops noticeably when the datasets are preprocessed using LD pruning (Supp. Fig. S1), which is also in agreement with the results in Wei et al. (2009), and does not improve significantly when additional nongenotyped SNPs were imputed (Supp. Table S1).

We followedtwo different approaches to evaluate the performance of MACLEAPS. First, we evaluated the risk assessment by performing a within-study cross-validation approach using the dataset of German and United States as a single study. AUC values range from 0.49 to 0.56, which suggests that our model did not increase the disease prediction beyond chance

alone (Table 3). Using radial kernel functions, which can implicitly assess gene–gene interaction, showed no significant difference in the predictive power compared to a linear kernel (data not shown).

One of the major concerns in GWAS is that observed associations could be false positives. This could be caused by population structuring, the use of different genotyping platforms, or a lack of statistical power due to small sample sizes. Within-study cross-validation strategies cannot overcome such biases. Therefore, we evaluated the performance of MACLEAPS using three independent PD-GWAS datasets (Tables 4 and 5, and Fig. 1). Because these PD datasets were genotyped using different SNP arrays, we used imputation strategies (see "Methods") to match SNPs with our training model. The highest AUC values (0.55–0.58) could be observed for *P* value thresholds of 1×10^{-7} and 1×10^{-6} with only 3 and 13 SNPs, respectively. We observed that, using independent PD datasets, AUC values did not increase beyond chance, highlighting the importance of large sample sizes to detect clinically relevant markers for PD.

When comparing the results of the linear SVM models with the ones from RBF SVMs, we observe that if only a small number of SNPs were selected (four or less), the performance of RBF models drops to the level of chance or even below. For larger numbers of selected SNPs, the performance of both types of SVM models is comparable, without an obvious superiority of one over the other.

Evaluation of T1D, T2D, and BD Using within Study Cross-Validation Model

As a control analysis, we evaluated the performance of MACLEAPS on the WTCCC type 1 diabetes (T1D) dataset using a fivefold cross validation; that is, as described in the section "Methods," 80% of the data are used to train a disease model and the rest of the data is used for disease prediction. Using different *P* value threshold, MACLEAPS was able to achieve a maximum performance (AUC = 0.88) when using a threshold of $P < 1 \times 10^{-5}$, and an only slightly worse performance (AUC = 0.81–0.87) when using the other thresholds, for the within-study cross validation approach for T1D (see Table 6). These results are in agreement with a recently published study [Wei et al., 2009] using the same approach as in our study. It is important to mention here that, unlike other complex diseases, the majority of the genetic architecture of T1D can be explained by the major histocompatibility complex (MHC) region because the AUC drops to 0.64 when all SNPs from this region are removed (see Table 6). Therefore, simultaneous evaluation of two extreme traits (heritability component in T1D ~90% compared to ~38% in PD) provided a better perspective on the underlying genetic architecture of PD.

In addition, the results of our analyses of the WTCCC type 2 diabetes (T2D), and bipolar disorder (BD) datasets show similar predictive power (Table 7), with AUCs ranging from 0.58 to 0.62 for T2D and from 0.55 to 0.61 for BD, two diseases having epidemiological characteristics related to PD.

Understanding the Role of Uncommon/Multiple Rare Variants for Disease Prediction

To understand the lack of predictability for PD, we simulated different scenarios by simultaneously accounting for the role of multiple uncommon and/or rare variants. Using

model A (see “Methods”), which assessed the role of uncommon risk variants (MAF 1% to 5%) with large effects (OR > 2), we observed that even for a moderate sample size (N = 1,000) MACLEAPS is successfully able to classify the subjects (AUC ~ 0.78). Likewise, by simultaneously increasing the frequency and effect size of risk variants, we were able to achieve a very good classification (AUC ~ 0.87) for diseases in which uncommon variants with strong effects have been shown to play an important role in disease susceptibility (Supp. Tables S2–S5). These results are in agreement with recent studies, which showed high predictive values for a few complex diseases such as age macular degeneration and hypertriglyceridemia. Similarly, using model B, we observed that our AUC scores increased from 0.56 to 0.84 when we incrementally increased the threshold of our parameters (number of samples from 1,000 to 20,000; MAF from 1% to 5%; OR from 0.7 to 1.5) (Fig. 2, Supp. Tables S6–S20). As expected, MACLEAPS achieves the best performance when using the largest sample size (N = 20,000).

Our observations argue in favor of selecting SNPs that are in the lower tail of distribution statistics for disease prediction models, because risk variants with a small effect size will most likely not reach stringent thresholds of $P < 1 \times 10^{-7}$ or better. This is probably more helpful for those disorders for which the heritability component is low (e.g., bipolar disorder, type 2 diabetes, and neurodegenerative disorders) arguing in favor of the role of multiple loci’s with small effects in disease pathogenesis. The role of multiple uncommon and/or rare variants for complex diseases is not surprising because GWAS studies have already suggested the role of uncommon as well as rare variants in influencing the disease susceptibility [Altshuler et al., 2010].

Discussion

We developed MACLEAPS to perform genome-wide risk prediction in an unbiased manner. We tested our algorithm to perform genome-wide risk profiling for PD and T1D. Compared to PD, MACLEAPS achieved satisfactory performance when applied to T1D (AUC scores range from 0.81 to 0.88). The lack of finding clinical utility markers for PD may be due to several reasons: (1) We observed that using strict threshold criteria few SNPs were used to train our model. For example, using P value threshold ranges from 1×10^{-6} to 1×10^{-8}, we obtained only 13 SNPs to train our model. Most of these SNPs were clustered in the SNCA or MAPT gene, the most common risk factors for the sporadic form of PD identified to date [Singleton et al., 2010]. Therefore, it was not unexpected to have poor prediction for our datasets (Table 5). This is in contrast to a recently published study for type 1 diabetes, which used approximately 240 SNPs for the same threshold for disease prediction [Wei et al., 2009]. (2) The genome-wide risk profiling also depends on the heritability component of the disease. By comparing the genomic profiling estimates between T1D and PD, we showed that in case of T1D, the majority of the genetic component could be explained by the MHC region alone [Wei et al., 2009] (see Table 6). Therefore, it is not surprising to observe that genome-wide risk profiling of T1D outperforms PD. (3) The magnitude of effect size also influenced the ability of disease prediction for most complex diseases. The effect size of most complex diseases varies from 0.6 to 1.5 and heritability estimates are very moderate [So et al., 2011]. Therefore, for most of these diseases, inclusion of low-frequency variants in the GWAS (<5%) along with liberal P value thresholds will lead to an improved disease

risk prediction, as is shown by our simulations. (4) Finally, for complex diseases such as PD, apart from considering only genetic factors, other nongenetic factors should be taken into consideration for disease risk profiling. Although the role of strong genetic components cannot be underestimated in early onset cases of PD, for late onset cases other nongenetic factors also contribute to disease pathogenesis. Therefore, inclusion of such factors into the disease prediction model will eventually increase the disease prediction estimates. Note that MACLEAPS has built-in functions to incorporate such additional variables into the analysis.

The lack of high AUC scores for PD might also reflect the fact that our estimates are conservative. As described above, we used fivefold and fivefold cross validation (within-study validation) and independent datasets for disease prediction. As the effect size of PD risk variants is low, splitting the data further reduces the statistical power to detect associated SNPs, and thus the ability of the trained model to accurately predict the rest of the datasets. As expected, we observed approximately 6% of the total genetic variance explained by our training datasets using a genome-wide complex trait analysis [Yang et al., 2011], further underscoring the need for large sample sizes for disease risk prediction where effect size is moderate. Likewise, the exclusion of cases with apparent Mendelian inheritance of PD from the cohort [Simón-Sánchez et al., 2009], leading to an underrepresentation of known risk variants with strong effect sizes such as LRRK2 and GBA mutations may have also influenced our AUC estimates. Indeed, our simulation study also suggests an increase in AUC (see “Results”) when we simultaneously increased the sample size from 1,000 to 20,000 and included the low frequency variants (Tables 8 and 9, and Supp. Tables S6–S20). The use of 1000 Genomes data along with next-generation sequencing technologies will help to identify more rare variants, which will eventually be used in GWAS settings. This will probably help in improving the disease risk assessment for complex diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov). We used DNA panels, samples, and clinical data from the National Institute of Neurological Disorders and Stroke Human Genetics Resource Center DNA and Cell Line Repository. People who contributed samples are acknowledged in descriptions of every panel on the repository website. We thank the French Parkinson’s Disease Genetics Study Group: Y Agid, M Anheim, A-M Bonnet, M Borg, A Brice, E Broussolle, J-C Corvol, P Damier, A Destée, A Dürr, F Durif, S Klebe, E Lohmann, M Martinez, P Pollak, O Rascol, F Tison, C Tranchant, M Vérin, F Viallet, and M Vidailhet. We also thank the members of the French 3C Consortium: A Alpérovitch, C Berr, C Tzourio, and P Amouyel for allowing us to use part of the 3C cohort, and D Zelenika for support in generating the genome-wide molecular data. This studymakes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

References

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008; 4:e1000173. [PubMed: 18974822]

Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:1–27.

Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. 2011; 7:e1002141. [PubMed: 21738487]

Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 2009; 18:3525–3531. [PubMed: 19553258]

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: a library for large linear classification. J Mach Learn Res. 2008; 9:1871–1874.

Guyon I, Elisseefi A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3:1157–1182.

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005; 6:95–108. [PubMed: 15716906]

Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet. 2009; 10:318–329. [PubMed: 19373277]

Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet. 2009; 5:e1000337. [PubMed: 19197355]

Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med. 2006; 8:395–400. [PubMed: 16845271]

Janssens AC, van Duijn CM. Towards predictive genetic testing of complex diseases. Eur J Epidemiol. 2006; 21:869–870. [PubMed: 17160427]

Janssens AC, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet. 2008; 17:R166–R173. [PubMed: 18852206]

Manolio T. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; 363:166–176. [PubMed: 20647212]

Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simon-Sanchez J, Schulte C, Lesage S, Sveinbjornsdottir S, Stefansson K, Martinez M, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet. 2011; 377:641–649. [PubMed: 21292315]

Olsen RJ, Curry CV, Vanguri AL, Dunphy CH, Chang CC. The JAK2 V617F mutation is not identified in hematological malignancies associated with bone marrow fibrosis other than chronic myeloproliferative disorders. Lab Invest. 2007; 87:254a–254a.

Pankratz N, Wilk JB, Latourelle JC, DeStefano AL, Halter C, Pugh EW, Doheny KF, Gusella JF, Nichols WC, Foroud T, Myers RH. Genomewide association study for susceptibility genes contributing to familial Parkinson disease. Hum Genet. 2009; 124:593–605. [PubMed: 18985386]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker Paul IW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

Saad M, Lesage S, Saint-Pierre A, Corvol JC, Zelenika D, Lambert JC, Vidailhet M, Mellick GD, Lohmann E, Durif F, Pollak P, Damier P, et al. Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population. Hum Mol Genet. 2011; 20:615–627. [PubMed: 21084426]

Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Krüger R, Federoff M, et al. Genomewide association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 2009; 41:1308–1312. [PubMed: 19915575]

Simón-Sánchez J, van Hilten JJ, van de Warrenburg B, Post B, Berendse HW, Arepalli S, Hernandez DG, de Bie RM, Velseboer D, Scheffer H, Bloem B, van Dijk KD, et al. Genome-wide association

study confirms extant PD risk loci among the Dutch. Eur J Hum Genet. 2011; 19:655–661. [PubMed: 21248740]

Singleton AB, Hardy J, Traynor BJ, Houlden H. Towards a complete resolution of the genetic architecture of disease. Trends Genet. 2010; 26:438–442. [PubMed: 20813421]

So H-C, Gui AHS, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol. 2011; 35:310–317. [PubMed: 21374718]

Szymczak S, Biernacka J, Cordell HO. Machine learning in genome-wide association studies. Genetic. 2009; 33:51–57.

Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet. 2008; 9:255–266. [PubMed: 18319743]

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; 447:661–678. [PubMed: 17554300]

Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 2009; 5:e1000678. [PubMed: 19816555]

Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010; 6:e1000864. [PubMed: 20195508]

Wu C, Walsh KM, Dewan AT, Hoh J, Wang Z. Disease risk prediction with rare and common variants. BMC Proc. 2011; 5(Suppl 9):S61. [PubMed: 22373337]

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008; 40:638–645. [PubMed: 18372903]

Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. Nat Protocols. 2007; 2:2492–2501. [PubMed: 17947991]

International Parkinson Disease Genomics Consortium Members

Michael A. Nalls (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA), Vincent Plagnol (UCL Genetics Institute, London, UK), Dena G. Hernandez (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Manu Sharma (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Una-Marie Sheerin (Department of Molecular Neuroscience, UCL Institute of Neurology), Mohamad Saad (INSERM UMR 1043 CPTP, Toulouse, France; and Paul Sabatier University, Toulouse, France), Javier Simón-Sánchez (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre, Amsterdam, Netherlands), Claudia Schulte (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Suzanne Lesage (INSERM, UMR_S975 [formerly UMR_S679], Paris, France; Université Pierre et Marie Curie-Paris, Centre de Recherche de l’Institut du Cerveau et de la Moelle épinière, Paris, France; and CNRS, Paris, France), Sigurlaug Sveinbjörnsd óttir (Department of Neurology, Landspítali University Hospital, Reykjavík, Iceland; Department of Neurology, MEHT Broomfield Hospital, Chelmsford, Essex, UK; and Queen Mary College, University of

London, London, UK), Sampath Arepalli (Laboratory of Neurogenetics, National Institute on Aging), Roger Barker (Department of Neurology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK), Yoav Ben-Shlomo (Department of Social Medicine, Bristol University, UK), Henk W Berendse (Department of Neurology, VU University Medical Center), Daniela Berg (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Kailash Bhatia (Department of Motor Neuroscience, UCL Institute of Neurology), Rob M. A. de Bie (Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands), Alessandro Biffi (Center for Human Genetic Research and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA), Bas Bloem (Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Zoltan Bochdanovits (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Michael Bonin (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen, Tübingen, Germany), Jose M. Bras (Department of Molecular Neuroscience, UCL Institute of Neurology), Kathrin Brockmann (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Janet Brooks (Laboratory of Neurogenetics, National Institute on Aging), David J. Burn (Newcastle University Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne, UK), Gavin Charlesworth (Department of Molecular Neuroscience, UCL Institute of Neurology), Honglei Chen (Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, NC, USA), Patrick F. Chinnery (Neurology M4104, The Medical School, Framlington Place, Newcastle upon Tyne, UK), Sean Chong (Laboratory of Neurogenetics, National Institute on Aging), Carl E. Clarke (School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK; and Department of Neurology, City Hospital, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK), Mark R. Cookson (Laboratory of Neurogenetics, National Institute on Aging), J. Mark Cooper (Department of Clinical Neurosciences, UCL Institute of Neurology), Jean Christophe Corvol (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and INSERM CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France), Carl Counsell (University of Aberdeen, Division of Applied Health Sciences, Population Health Section, Aberdeen, UK), Philippe Damier (CHU Nantes, CIC0004, Service de Neurologie, Nantes, France), Jean-François Dartigues (INSERM U897, Université Victor Segalen, Bordeaux, France), Panos Deloukas (Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK), Günther Deuschl (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein, Campus Kiel, Christian-Albrechts-Universität Kiel, Kiel, Germany), David T. Dexter (Parkinson's Disease Research Group, Faculty ofMedicine, Imperial College London, London, UK), Karin Dvan Dijk (Department ofNeurology, VU University Medical Center), Allissa Dillman (Laboratory of Neurogenetics, National Institute on Aging), Frank Durif (Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France), Alexandra Dürr (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and AP-HP, Pitié-Salpêtrière Hospital), Sarah Edkins (Wellcome Trust Sanger Institute), Jonathan R Evans (Cambridge Centre for Brain Repair, Cambridge, UK), Thomas Foltynie (UCL Institute of Neurology), Jianjun Gao (Epidemiology Branch, National Institute of Environmental Health Sciences), Michelle

Gardner (Department of Molecular Neuroscience, UCL Institute of Neurology), J. Raphael Gibbs (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology), Alison Goate (Department of Psychiatry, Department of Neurology, Washington University School of Medicine, MI, USA), Emma Gray (Wellcome Trust Sanger Institute), Rita Guerreiro (Department of Molecular Neuroscience, UCL Institute of Neurology), Ómar Gústafsson (deCODE genetics and Department of Psychiatry, Oslo University Hospital, N-0407 Oslo, Norway), Clare Harris (University of Aberdeen), Jacobus J. van Hilten (Department of Neurology, Leiden University Medical Center, Leiden, Netherlands), Albert Hofman (Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands), Albert Hollenbeck (AARP, Washington DC, USA), Janice Holton (Queen Square Brain Bank for Neurological Disorders, UCL Institute of Neurology), Michele Hu (Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Xuemei Huang (Departments of Neurology, Radiology, Neurosurgery, Pharmacology, Kinesiology, and Bioengineering, Pennsylvania State University—Milton S Hershey Medical Center, Hershey, PA, USA), Heiko Huber (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Gavin Hudson (Neurology M4104, The Medical School, Newcastle upon Tyne, UK), Sarah E. Hunt (Wellcome Trust Sanger Institute), Johanna Huttenlocher (deCODE genetics), Thomas Illig (Institute of Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Pálmi V. J ónsson (Department of Geriatrics, Landspítali University Hospital, Reykjavík, Iceland), Jean-Charles Lambert (INSERM U744, Lille, France; and Institut Pasteur de Lille, Université de Lille Nord, Lille, France), Cordelia Langford (Cambridge Centre for Brain Repair), Andrew Lees (Queen Square Brain Bank for Neurological Disorders), Peter Lichtner (Institute of Human Genetics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Patricia Limousin (Institute of Neurology, Sobell Department, Unit of Functional Neurosurgery, London, UK), Grisel Lopez (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI, National Institutes of Health), Delia Lorenz (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein), Alisdair McNeill (Department of Clinical Neurosciences, UCL Institute of Neurology), Catriona Moorby (School of Clinical and Experimental Medicine, University of Birmingham), Matthew Moore (Laboratory of Neurogenetics, National Institute on Aging), Huw R. Morris (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK), Karen E. Morrison (School of Clinical and Experimental Medicine, University of Birmingham; and Neurosciences Department, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK), Ese Mudanohwo (Neurogenetics Unit, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery), Sean S. O'Sullivan (Queen Square Brain Bank for Neurological Disorders), Justin Pearson (MRC Centre for Neuropsychiatric Genetics and Genomics), Joel S. Perlmutter (Department of Neurology, Radiology, and Neurobiology at Washington University, St Louis, MO, USA), Hjörvar Pétursson (deCODE genetics; and Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Pierre Pollak (Service de Neurologie, CHU de Grenoble, Grenoble, France), Bart Post (Department of Neurology, Radboud University Nijmegen Medical Centre), Simon Potter (Wellcome Trust Sanger Institute), Bernard Ravina (Translational Neurology, Biogen Idec,

MA, USA), Tamas Revesz (Queen Square Brain Bank for Neurological Disorders), Olaf Riess (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Fernando Rivadeneira (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Patrizia Rizzu (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Mina Ryten (Department of Molecular Neuroscience, UCL Institute of Neurology), Stephen Sawcer (University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's hospital, Cambridge, UK), Anthony Schapira (Department of Clinical Neurosciences, UCL Institute of Neurology), Hans Scheffer (Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Karen Shaw (Queen Square Brain Bank for Neurological Disorders), Ira Shoulson (Department of Neurology, University of Rochester, Rochester, NY, USA), Ellen Sidransky (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI), Colin Smith (Department of Pathology, University of Edinburgh, Edinburgh, UK), Chris C. A. Spencer (Wellcome Trust Centre for Human Genetics, Oxford, UK), Hreinn Stefánsson (deCODE genetics), Stacy Steinberg (deCODE genetics), Joanna D. Stockton (School of Clinical and Experimental Medicine), Amy Strange (Wellcome Trust Centre for Human Genetics), Kevin Talbot (University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Carlie M. Tanner (Clinical Research Department, The Parkinson's Institute and Clinical Center, Sunnyvale, CA, USA), Avazeh Tashakkori-Ghanbaria (Wellcome Trust Sanger Institute), François Tison (Service de Neurologie, Hôpital Haut-Lévêque, Pessac, France), Daniah Trabzuni (Department of Molecular Neuroscience, UCL Institute of Neurology), Bryan J. Traynor (Laboratory of Neurogenetics, National Institute on Aging), André G. Uitterlinden (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Daan Velseboer (Department of Neurology, Academic Medical Center), Marie Vidailhet (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225), Robert Walker (Department of Pathology, University of Edinburgh), Bart van de Warrenburg (Department of Neurology, Radboud University Nijmegen Medical Centre), Mirdhu Wickremaratchi (Department of Neurology, Cardiff University, Cardiff, UK), Nigel Williams (MRC Centre for Neuropsychiatric Genetics and Genomics), Caroline H Williams-Gray (Department of Neurology, Addenbrooke's Hospital), Sophie Winder-Rhodes (Department of Psychiatry and Medical Research Council and Wellcome Trust Behavioural and Clinical Neurosciences Institute, University of Cambridge), Kári Stefánsson (deCODE genetics), Maria Martinez (INSERM UMR 1043; and Paul Sabatier University), John Hardy (Department of Molecular Neuroscience, UCL Institute of Neurology), Peter Heutink (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Alexis Brice (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225, AP-HP, Pitié-Salpêtrière Hospital), Wellcome Trust Case-Control Consortium 2 (webappendix p 13), Thomas Gasser (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, and DZNE, German Center for Neurodegenerative Diseases), Andrew B. Singleton (Laboratory of Neurogenetics, National Institute on Aging), and Nicholas W. Wood (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology).

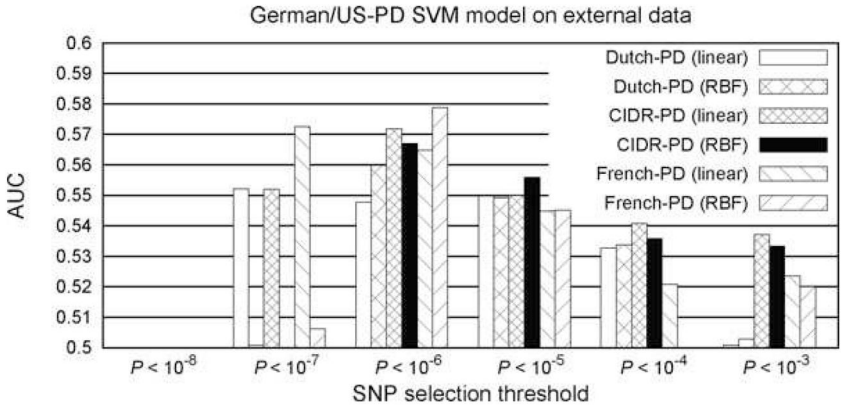

Figure 1.

Performance of risk assessment models trained on the German/US-PD dataset. For all three external datasets (Dutch-PD, CIDR-PD, and French-PD) SVM models achieved the best performance for a SNP selection threshold of $P < 1 \times 10^{-7}$ and $P < 1 \times 10^{-6}$.

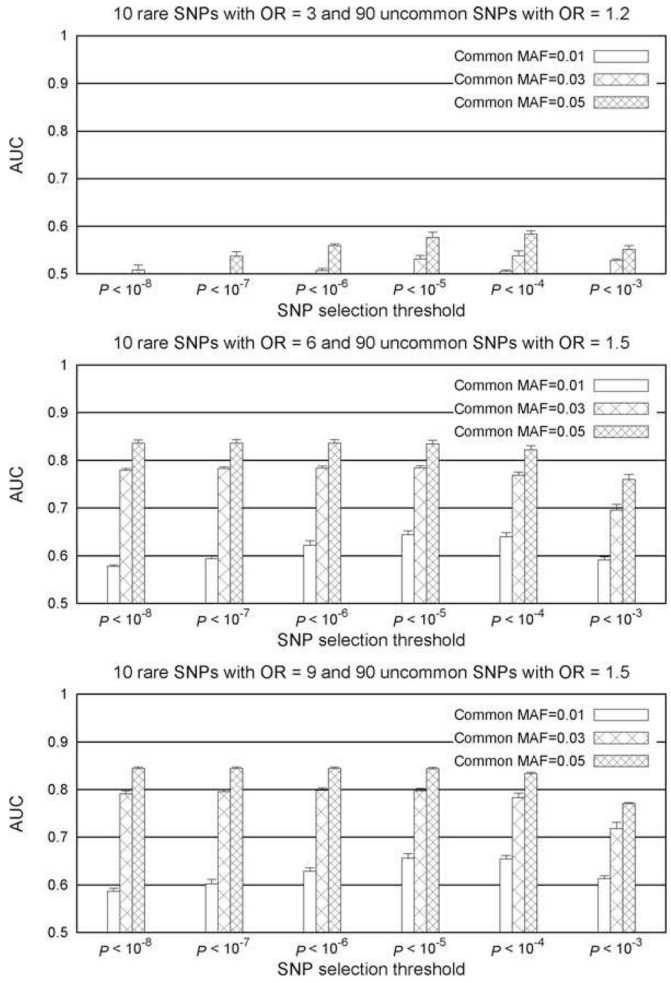

Figure 2.

Performance of risk assessment models using 20,000 simulated samples and mixed risk model (model **B**). In the mixed model **B**, 100 SNPs (out of 500,000 total) were generated having an effect on the phenotype, consisting of 10 rare SNPs (MAF 0.1%) with strong effect and 90 uncommon SNPs with moderate effect size.

Table 1

Description of the Parkinson Disease (PD) GWAS Datasets Used in the Study

Dataset	Number of cases	Number of controls	Array platform	Reference
GWAS-PD (German cohort)	742	944	Illumina HumanHap550 v3	[Simón-Sánchez et al., 2009]
GWAS-PD (US cohort)	971	3,034	Illumina HumanHap550 v3/HumanHap300+HumanHap240S	[Simón-Sánchez et al., 2009]
GWAS-PD (German + US)	1,713	3,978	(See cohorts)	[Simón-Sánchez et al., 2009]
Dutch-PD GWAS	772	2,024	Illumina Human 660W-Quad	[Simón-Sánchez et al., 2011]
CIDR-PD GWAS	857	876	Illumina HumanCNV370 v1_C	[Pankratz et al., 2009]
French-PD GWAS	1,039	1,984	Illumina Human610-Quad	[Saad et al., 2011]

Table 2

Description of the WTCCC GWAS Datasets Used in the Study

Dataset	Number of cases	Number of controls	Array platform
WTCCC-T1D (cases only)	2,000	–	Affymetrix GeneChip 500K
WTCCC-T2D (cases only)	1,999	–	Affymetrix GeneChip 500K
WTCCC-BD (cases only)	1,998	–	Affymetrix GeneChip 500K
WTCCC-58C (shared controls)	–	1,504	Affymetrix GeneChip 500K

Table 3

Results of a Fivefold Cross-Validation Using SVM on the German/US-PD and the WTCCC-T1D Datasets

	Parkinson disease (GWAS-PD)					Type 1 diabetes (WTCCC-T1D)				
	AUC (SD)		# SNP			AUC (SD)		# SNP		
SNP selection threshold	Lin	RBF	min	max	(SD)	Lin	RBF	min	max	(SD)
$P < 1 \times 10^{-8}$	0.52 (0.000)	0.49 (0.000)	0	1	0.45	0.88 (0.009)	0.88 (0.007)	151	201	20.25
$P < 1 \times 10^{-7}$	0.52 (0.009)	0.50 (0.017)	0	5	2.07	0.88 (0.009)	0.88 (0.005)	183	239	21.7
$P < 1 \times 10^{-6}$	0.53 (0.023)	0.53 (0.021)	5	10	2.30	0.88 (0.006)	0.88 (0.006)	224	281	24.11
$P < 1 \times 10^{-5}$	0.55 (0.022)	0.55 (0.022)	19	39	8.41	0.88 (0.005)	0.88 (0.005)	295	350	20.94
$P < 1 \times 10^{-4}$	0.56 (0.010)	0.55 (0.011)	113	195	32.69	0.87 (0.009)	0.87 (0.011)	430	527	39.97
$P < 1 \times 10^{-3}$	0.56 (0.019)	0.56 (0.017)	848	1054	77.64	0.81 (0.006)	0.81 (0.007)	1027	1123	36.64

Lin, linear SVM; RBF, SVM with radial basis function kernel; AUC, area under the receiver operating characteristic curve; SD, standard deviation;

#SNPs, number of SNPs.

Table 4

Result of a Cross-Study Validation of the German/US SVM Model on Dutch and CIDR Datasets

	Model trained on German/US-PD, validated on					Model trained on Dutch-PD			Model trained on CIDR-PD		
	Dutch-PD		CIDR-PD			Validated on German/US-PD					
SNP selection threshold	Lin AUC	RBF AUC	Lin AUC	RBF AUC	#SNPs	Lin AUC	RBF AUC	#SNPs	Lin AUC	RBF AUC	#SNPs
$P < 1 \times 10^{-8}$	–	–	–	–	0	–	–	0	–	–	0
$P < 1 \times 10^{-7}$	0.552	0.501	0.552	0.462	3	–	–	0	–	–	0
$P < 1 \times 10^{-6}$	0.548	0.560	0.572	0.567	13	0.495	0.501	1	–	–	0
$P < 1 \times 10^{-5}$	0.550	0.549	0.550	0.556	36	0.524	0.531	6	0.523	0.526	5
$P < 1 \times 10^{-4}$	0.533	0.534	0.541	0.536	171	0.509	0.510	79	0.520	0.522	49
$P < 1 \times 10^{-3}$	0.501	0.503	0.537	0.533	1111	0.514	0.517	643	0.530	0.533	494

Lin, linear SVM; RBF, SVM with radial basis function kernel; AUC, area under the receiver operating characteristic curve; #SNPs, number of SNPs.

Table 5

Results of a Cross-Study Validation of the German/US Linear SVM Model on CIDR, Dutch, and French Datasets

	Model trained on GWAS-PD, validated on						
	CIDR-PD AUC		Dutch-PD AUC		French-PD AUC		
SNP selection threshold	Lin	RBF	Lin	RBF	Lin	RBF	#SNPs
$P < 1 \times 10^{-8}$	–	–	–	–	–	–	0
$P < 1 \times 10^{-7}$	0.552	0.462	0.552	0.501	0.573	0.506	3
$P < 1 \times 10^{-6}$	0.572	0.567	0.548	0.560	0.565	0.579	13
$P < 1 \times 10^{-5}$	0.550	0.556	0.550	0.549	0.545	0.545	36
$P < 1 \times 10^{-4}$	0.541	0.536	0.533	0.534	0.521	0.483	171
$P < 1 \times 10^{-3}$	0.537	0.533	0.501	0.503	0.524	0.520	1111

Lin, linear SVM; RBF, SVM with radial basis function kernel; AUC, area under the receiver operating characteristic curve; #SNPs, number of SNPs.

Table 6

Results of a Fivefold Cross-Validation Using SVM on the WTCCC-T1D Dataset with and without the MHC Region

	Type 1 diabetes (WTCCC-T1D)					Type 1 diabetes (WTCCC-T1D) without MHC region				
	AUC (SD)		#SNP			AUC (SD)		#SNP		
SNP selection threshold	Lin	RBF	min	max	(SD)	Lin	RBF	min	max	(SD)
$P < 1 \times 10^{-8}$	0.88 (0.009)	0.88 (0.007)	151	201	20.25	0.69 (0.010)	0.69 (0.014)	7	11	1.79
$P < 1 \times 10^{-7}$	0.88 (0.009)	0.88 (0.005)	183	239	21.7	0.70 (0.009)	0.70 (0.016)	10	12	0.84
$P < 1 \times 10^{-6}$	0.88 (0.006)	0.88 (0.006)	224	281	24.11	0.70 (0.010)	0.70 (0.014)	13	19	2.79
$P < 1 \times 10^{-5}$	0.88 (0.005)	0.88 (0.005)	295	350	20.94	0.71 (0.008)	0.70 (0.006)	28	42	6.19
$P < 1 \times 10^{-4}$	0.87 (0.009)	0.87 (0.011)	430	527	39.97	0.69 (0.010)	0.68 (0.005)	86	116	12.1
$P < 1 \times 10^{-3}$	0.81 (0.006)	0.81 (0.007)	1027	1123	36.64	0.64 (0.009)	0.64 (0.006)	546	691	57.33

Lin, linear SVM; RBF, SVM with radial basis function kernel; AUC, area under the receiver operating characteristic curve; SD, standard deviation; #SNPs, number of SNPs.

Table 7

Results of a Fivefold Cross-Validation Using SVM on the WTCCC-T2D and the WTCCC-BD Datasets

	Type 2 Diabetes (WTCCC-T2D)					Bipolar Disorder (WTCCC-BD)				
	AUC (SD)		#SNP			AUC (SD)		#SNP		
SNP selection threshold	Lin	RBF	min	max	(SD)	Lin	RBF	min	max	(SD)
$P < 1 \times 10^{-8}$	0.59 (0.021)	0.59 (0.021)	1	5	1.52	0.55 (0.008)	0.55 (0.008)	0	3	1.3
$P < 1 \times 10^{-7}$	0.60 (0.022)	0.58 (0.040)	2	6	1.52	0.57 (0.012)	0.56 (0.012)	1	5	1.79
$P < 1 \times 10^{-6}$	0.61 (0.017)	0.62 (0.012)	5	19	5.54	0.56 (0.018)	0.57 (0.020)	4	16	4.67
$P < 1 \times 10^{-5}$	0.61 (0.018)	0.62 (0.025)	20	32	4.42	0.56 (0.019)	0.57 (0.017)	17	34	7.01
$P < 1 \times 10^{-4}$	0.60 (0.016)	0.60 (0.018)	86	113	9.81	0.56 (0.027)	0.56 (0.019)	108	119	5.15
$P < 1 \times 10^{-3}$	0.59 (0.011)	0.59 (0.008)	589	652	24.59	0.61 (0.021)	0.61 (0.018)	672	774	40.41

Lin, linear SVM; RBF, SVM with radial basis function kernel; AUC, area under the receiver operating characteristic curve; SD, standard deviation; #SNPs, number of SNPs.

Hum Mutat. Author manuscript; available in PMC 2018 May 25.

Table 8

Results of a Fivefold Cross-Validation on Simulation Data with a Mixed Model of Common and Rare Variants for 5,000 Samples

OR_{rare}	6			6			9		
$OR_{uncommon}$	1.2			1.5			1.5		
MAF	**0.01**	**0.03**	**0.05**	**0.01**	**0.03**	**0.05**	**0.01**	**0.03**	**0.05**
SNP selection threshold									
$P < 1 \times 10^{-8}$	–	–	–	–	0.51	0.58	0.52	0.52	0.56
$P < 1 \times 10^{-7}$	0.50	–	–	0.50	0.51	0.62	0.53	0.54	0.62
$P < 1 \times 10^{-6}$	0.50	0.50	–	0.51	0.54	0.67	0.54	0.59	0.70
$P < 1 \times 10^{-5}$	0.51	0.51	0.49	0.54	0.58	0.74	0.54	0.63	0.74
$P < 1 \times 10^{-4}$	0.52	0.54	0.50	0.54	0.61	0.74	0.56	0.63	0.74
$P < 1 \times 10^{-3}$	0.49	0.52	0.50	0.53	0.54	0.60	0.50	0.53	0.59

OR, odds ratio of rare/uncommon SNPs; MAF, minor allele frequency of uncommon SNPs.

Table 9

Results of a Fivefold Cross-Validation on Simulation Data with a Mixed Model of Common and Rare Variants for 20,000 Samples

OR_{rare}	6			6			9		
$OR_{uncommon}$	1.2			1.5			1.5		
MAF	**0.01**	**0.03**	**0.05**	**0.01**	**0.03**	**0.05**	**0.01**	**0.03**	**0.05**
SNP selection threshold									
$P < 1 \times 10^{-8}$	0.54	0.54	0.55	0.58	0.78	0.84	0.59	0.79	0.84
$P < 1 \times 10^{-7}$	0.54	0.55	0.57	0.59	0.78	0.84	0.60	0.80	0.84
$P < 1 \times 10^{-6}$	0.54	0.55	0.59	0.62	0.78	0.84	0.63	0.80	0.84
$P < 1 \times 10^{-5}$	0.55	0.56	0.60	0.64	0.78	0.83	0.66	0.80	0.84
$P < 1 \times 10^{-4}$	0.55	0.56	0.60	0.64	0.77	0.82	0.65	0.78	0.83
$P < 1 \times 10^{-3}$	0.54	0.54	0.58	0.59	0.70	0.76	0.61	0.72	0.77

OR, odds ratio of rare/uncommon SNPs; MAF, minor allele frequency of uncommon SNPs.