Cross validation of pooling/resampling GWAS using the WTCCC data

Jorge I. Vélez1,2, Cameron A. Jack1, Aaron Chuah1, Bob Buckley3, Juan C. Correa4, Simon Easteal5 and Mauricio Arcos-Burgos1,2*

Abstract

Recently, we presented a new method of pooling/resampling genome-wide association study (prGWAS) that uncovered new and known loci associated to Alzheimer’s disease. In here, we contrast this method with the Welcome Trust Case Control Consortium (WTCCC) data, a well-known GWAS on seven human complex diseases. Our results suggest that prGWAS can be considered an efficient, specific, and accurate alternative to the conventional GWAS approach at a fraction of the genotyping cost, and provide insights into other potential applications such as next generation sequencing.

Keywords: Case/control data, GWAS, DNA pooling, prGWAS, subsampling

Introduction

In a recent manuscript we presented a new strategy for genome wide association studies (GWAS) that uses resampling and DNA pooling, and that we denominated pooling/bootstrap-based GWAS (prGWAS) [1]. This methodology is well suited to identify disease-associated genetic variants using limited and relatively small samples at a fraction of the individual genotyping cost [1]. We applied the prGWAS strategy to a unique cohort of patients with an autosomal-dominant form of Alzheimer’s disease (AD) segregating a mutation in the presenilin-1 gene (PSEN1), [1-4] and identified new and previously reported loci underpinning the susceptibility to AD and/or modifying the age of onset of this dementia [1].

Here we cross-validate the prGWAS methodology using the Welcome Trust Case Control Consortium (WTCCC) data, a collection of publically available genetic information obtained by a network of ~50 research groups across the United Kingdom (http://www.wtccc.org.uk/). Based on performance measures, [5] we found that the prGWAS provides efficient, specific and accurate parameters comparable to those obtained from the application of traditional GWAS.

Furthermore, prGWAS detects 98% of all regions of the genome showing strongest or moderate evidence of association [6] whilst using a reasonably number of DNA pools. Overall, these findings demonstrate that prGWAS might be considered a feasible alternative to individual genotyping, and provide insights of its use into other potential whole genome designs.

Methods

WTCCC data

Sample description

Briefly, the WTCCC aimed to understand patterns of human genome sequence variation, and to explore their utility for the design and analyses of GWAS. The WTCCC analyzed n~2,000 cases of each of seven complex human diseases and n~3,000 control individuals, all of them from England, Scotland and Wales. Genotyping of ~1.8 million single nucleotide polymorphism (SNPs) in the n~17,000 samples was performed using high-throughput technologies [6].

The two control collections, assumed to be samples from the general population, are n~3,000 individuals from the 1958 British Birth Cohort (58C, n~1,500) and from blood donors...
recruited as part of this project (NBS cohort, \( n \sim 1,500 \)). Both cohorts have approximately the same number of males and females. Individuals in the 58C cohort, aged 44-45 years, were born in the same week in 1958; NBS samples are from blood donors with 18-69 years of age. Previous comparison of SNP allele frequencies between the 58C and NBS cohorts showed few significant differences [6] and therefore treated as a unique cohort in this study.

Approximately 14,000 individuals were phenotyped for seven human complex diseases (bipolar disorder [BD], Crohn’s disease [CD], coronary artery disease [CAD], hypertension [HT], rheumatoid arthritis [RA], type 1 diabetes [T1D] and type 2 diabetes [T2D]) and comprised the case cohorts, each of which includes \( n \sim 2000 \) samples [6].

**Genotype data**

Control cohorts were genotyped using the Illumina 1.2M and Affymetrix V6.0 platforms, and cases were genotyped using the Affymetrix 500K SNP-chip. Complete information for the genotyped SNPs was obtained from Illumina (www.illumina.com) and Affymetrix (www.affymetrix.com). Non-autosomal SNPs and those with no available coordinate information were excluded. Genotypes for all samples were obtained after application to the WTCCC, and processed in R [7] and Python (https://www.python.org/) (Supplementary material).

**Genetic association analysis**

Genetic association analysis were performed in PLINK [8] version 1.07. SNPs with a minimal allele frequency (MAF) of at least 1%, a \( p \)-value greater than \( 10^{-8} \) for the Hardy-Weinberg equilibrium test, and a genotype frequency of at least 95% across all samples were included for analysis. To some extent, these parameters mimic those previously used by the WTCCC [6]. After completion, results of the association analyses for all chromosomes in a specific case cohort were combined in a single file.

**prGWAS method**

**Construction of DNA pools**

Let \( N_n \) be the total number of controls, \( N_x \) the total number of cases, \( n \) and \( r \) the total number of DNA samples in the master plate for cases and controls, respectively, and, \( n' = \lfloor x/2 \rfloor \) with \( x \) the integer part of \( x \). Note that, by design, \( n_r \leq N_x \) and \( n_r \leq N_r \) and that \( N_r = 3,000 \) and \( N_r = 2,000 \) for each disease in the WTCCC data under this notation.

DNA pools were constructed as previously described [1]. First, a random sample of size \( n \) is drawn from \( N_n \) to form the master plates; and secondly, \( k \) samples of size \( n' \) are randomly selected without replacement from the master plate (Figure 1). During the random selection process within master plates, it is unlikely that an individual’s DNA sample would be selected more than once. However, it is possible that some individuals belong to several of the \( k \) DNA pools but not to the same DNA pool. In the context of prGWAS, \( k \) represents the number of DNA pool pairs to be generated from the master plates [1].

**Allele frequency estimation**

After constructing the pools of DNA pools, the genotypes for all \( m \) genotyped SNPs were retrieved from the ped files previously constructed (Supplementary material) by matching the IDs of the individuals in the DNA pools with those of the WTCCC samples. The allele frequency for the \( j \)th SNP in the \( i \)th group was estimated as \( \hat{p}_{ij} = \frac{2y_{ij}}{2x} \), where \( y_{ij} \) and \( y_{ij} \) are, respectively, the number of individuals with one and two alleles in the \( j \)th SNP (i=1,2; j=1,2,…, m). The number of alleles was determined based on the ordered two-string genotypes for each SNP.

**Detecting associated SNPs**

As described in Vélz et al., [1] for each SNP it is of interest to test

\[
H_0: p_{1j} = p_{2j}, \quad j = 1,2, \ldots, m
\]

against a suitable alternative hypothesis, say \( H_1 \). Because
a total of $k$ DNA pools are generated for cases and controls, (1) is to be tested $k$ times for each SNP. For $k$ fixed, the test statistic for this purpose is given by [1]

$$T_j = \frac{2n_1 n_2 (p_{1,j} - p_{2,j})^2}{n_2 p_{1,j} (1 - p_{1,j}) + n_2 p_{1,j} (1 - p_{2,j})}$$

(2)

with $n_1$ and $p_1$ as previously defined ($i=1,2; j=1,2,\ldots,m$). Under the null hypothesis of no difference between the allele frequency of cases and controls for the $j$th SNP, $T_j$ follows with one degree of freedom; the $P$-value for the $j$th SNP can be calculated as $P_j = Pr(T_j > x)$, $j=1,2,\ldots,m$.

Let $\alpha$ be the type I error probability of the test. Because $m>1$ statistical tests are being performed (and usually $m \rightarrow \infty$), it is imperative to control by multiple testing [9] the main problem of rejecting $H_0$, when $\rho_1 < \alpha$ is that actual type I error probability of the $m$ tests is $\alpha_0 = (1 - (1 - \alpha)^m)$, which leads to an increased number of false positives as $m \rightarrow \infty$. Although several criteria have been proposed to control by multiple testing in GWAS, [10-16] the FDR, [10] a relatively new approach, suitable for exploratory analysis, [14] less stringent and more powerful than Bonferroni’s, [10,14,15] is preferable in prGWAS [1]. We shall say that the $j$th SNP is disease-associated in the $l$th pair of DNA pools if $p_j^0 < \alpha$, with $p_j^0$ the FDR-corrected $P$-value for the $j$th SNP when the $l$th pair of DNA pools is generated ($j=1,2,\ldots,m; l=1,2,\ldots,k$).

### Combination of $P$-values

Based on the $k$ pairs of DNA pools being generated and subsequently compared using (2), a total of $k$ $P$-values are calculated for each of $m$ SNPs in the SNP-chip. Further, $P$-values for each SNP are combined using the Stouffer’s $Z$-transformed method [17,18] after introducing some degree of dependence [19]. The test statistic for the $j$th SNP is given by [19]

$$Z_j = \begin{cases} k \sqrt{k_1} & \text{if } \rho \text{ is known} \\ k_1 \sqrt{k_2} & \text{if } \rho \text{ is not known} \end{cases}$$

(3)

where \(k=1-(1-p^0)A + pB)^{1/2}\), $k_2=\sum_{l=1}^m h(l) Z(l)^2$, $k_1=\left(A(1-C) + BC \right)^{1/2}$, $A=\sum_{l=1}^m h(l)$, $B=\left(\sum_{l=1}^m h(l)^2\right)^2$, $C = \rho + \sqrt{2k} \cdot (1 - \rho) \cdot (k + 1)^{-1/2}$.

When combining the $k$ $P$-values for each SNP, the null hypothesis of interest when the is whether the allele frequency between cases and controls is the same across all pairs of DNA pools, that is

$$H_0: p_{1,l} = p_{2,l} \quad l=1,2,\ldots,k$$

(4)

with the alternative hypothesis being $H_1: p_{1,l} > p_{2,l}$ for some $l$. Under $H_0$ in (4), the statistic $Z_j$ follows a N(0,1) distribution.

Hartung (1999) [19] showed that

$$\rho = \max \left\{ \frac{1}{k-1}, \rho_j \right\}$$



Thus, the calculation of one-and two-tailed $P$-values follows [18].

### Comparison of prGWAS and GWAS

Let $S_{j}^0$ and $S_{j}^1$ be binary variables constructed after applying the prGWAS and traditional GWAS on the WCC data, such that $S_{j}^0=1$ if $p_j^0 < \alpha$ and zero otherwise, and $S_{j}^1=1$ if $p_j < \alpha$ and equal to zero otherwise. Here, $p_j^0$ the Bonferroni-corrected $P$-value of the $j$th SNP.

Results from the prGWAS method were compared to those obtained in GWAS using several performance measures [5] after constructing 2×2 contingency tables (Table 1a) for every combination of disease, chromosome, $k$ and sample sizes of the master plates. Initially, the results from the association analysis using GWAS and those using prGWAS were merged by SNP after the $P$-values from the GWAS analysis were corrected by multiple testing using Bonferroni’s criterion. For the $l$th pair of DNA pools and $n_1$ and $n_2$ fixed, the entries of the 2×2 contingency table are

$$a^{(l)} = \sum_{j=1}^m I\{S_{j,l}^0 = 1, S_{j,l}^1 = 1\}, \quad b^{(l)} = \sum_{j=1}^m I\{S_{j,l}^0 = 1, S_{j,l}^1 = 0\}$$

$$c^{(l)} = \sum_{j=1}^m I\{S_{j,l}^0 = 0, S_{j,l}^1 = 1\} \quad \text{and} \quad d^{(l)} = \sum_{j=1}^m I\{S_{j,l}^0 = 0, S_{j,l}^1 = 0\}$$

where $I_{ij}$ is an indicator. For $l$ fixed, $a$ corresponds to the number of statistically significant SNPs (i.e., positive markers) found by both methods, $b$ is the number of positive SNPs according to prGWAS but not positive in GWAS, $c$ is the number of not significant markers using prGWAS but positive using GWAS, and $d$ is the number of SNPs not statistically significant by neither method. In this context, markers found to be positive using the complete data set are said to be true positives (Table 1a).

In addition to the classical measures, we also calculated the lift, a performance measure initially introduced as ‘interest’ by Brin et al., [20] and recently used in data mining and marketing [21] to evaluate the relative performance of alternative classification models (Table 1b) [22]. If $A$ is the event ‘detecting a marker as being statistically significant using prGWAS’ and $B$ the event ‘detecting a marker as being statistically significant using GWAS’, lift measures how many times more often $A$ and $B$ occur together than expected if they were statistically independent [22]. In other words, this is equivalent to quantify how much more successful the prGWAS method is likely to be than if no predictive model (i.e., random selection) was used to detect statistically significant markers.

### Results

**Previously reported SNPs**

Table 2a presents the association signals at previously replicated loci. Using prGWAS, we were able to detect 11 of...
Table 1a. Results and expressions for calculating the performance measures used to quantitatively compare prGWAS and GWAS.

| Measure       | Expression                                      |
|---------------|-------------------------------------------------|
| Sensitivity   | \( a/(a+c) \)                                   |
| Specificity   | \( d/(c+d) \)                                   |
| Positive PPV  | \( a/(a+b) \)                                   |
| Negative PPV  | \( d/(c+d) \)                                   |
| FDR           | \( 1-PPV \)                                     |
| False negatives | \( 1-NPV \)                                   |
| Lift          | \( (a+d)/(a+b+c+d) \)                           |

(a) Possible results when the GWAS and prGWAS are compared. Here, \( a \) is the number of significant markers found by both methods, \( b \) is the number of positive SNPs according to prGWAS but not positive according to GWAS, \( c \) corresponds to those markers not found to be significant by prGWAS but that are true positives according to GWAS, and \( d \) is the number of SNPs labeled as not significant by both methods; (b) Expressions for calculating the performance measures used to quantitatively compare prGWAS and GWAS (see the Methods).

the 13 associations previously reported; [6] marker rs4420638 in CAD did not pass quality control (see Methods) and marker rs3087243 in T1D was not prGWAS significant at 5%. As a function of the sample size of the master plates, no difference was found in the number of detected markers (trend \( P=0.2421 \), Figure 2a).

In Table 2b regions of the genome with the strongest association signals are presented. Twenty of the 21 signals previously reported in a standard analysis [6] were detected by prGWAS, with 43% (9/21) of these signals being statistically significant for at least one value of \( n \). Overall, the number of markers detected increases with \( n \) (trend \( P=0.00343 \), Figure 2c). An extensive and more detailed description of the associated variants for each disease can be found in the Supplementary Material.

Number of DNA pools

Figure 2 depicts the number of DNA pools to be generated \( k \) as a function of the sample size of the master plates \( n \) to obtain, using prGWAS, similar \( P \)-values to those reported previously (see Table 2) when the full genotype data was used.

Panel (a) depicts the results to detect markers in previously replicated loci (Table 2a). As a function of \( n \), regression analysis shows that \( k \) decreases as a function of \( n \) (trend \( P=7.88 \times 10^{-6} \)). Furthermore, analysis of variance (ANOVA) discloses statistically significance difference in the average value of \( k \) as a function of \( n \) (\( F_{2,60}=9.197, P=0.00225 \)). The highest Tukey’s honest significance difference was obtained after comparing \( k \) for \( n=1000 \) and \( n=250 \) (\( d=-4.25, P_{adj}=-0.0237 \)). These results suggest that, in prGWAS, the larger the value of \( n \) the lower the number of DNA pools to be generated in order to obtain comparable \( P \)-values to those obtained at previously robustly replicated loci.

As a function of \( n \) (trend \( P=0.0252 \)) for detecting regions with the strongest association signals (Table 2b). However, no statistically significance difference was found between the average value of \( k \) across sample sizes (\( k_{250}=5.63, k_{250}=7.11, k_{250}=5.93, k_{250}=4.68, F_{2,40}=2.723, p=0.0778 \)). Furthermore, no linear relationship between \( n \) and \( k \) (trend \( P=0.189 \)) nor in the average value of \( k \) for all sample sizes (\( k_{250}=5.59, k_{250}=5.85, k_{250}=5.76, k_{1000}=4.92, F_{2,61}=0.898, p=0.413 \)) was found when detecting regions of the genome with moderate evidence of association (Table 2c). Altogether, these results indicate that prGWAS detects those regions of the genome showing the strongest or moderate evidence of association previously reported by randomly generating a reasonably number of DNA pools regardless of the sample size of the master plates.

Table 2a. Classification rate (CR) and related measures for detecting regions with the strongest association signals.

| Measure       | Expression                                      |
|---------------|-------------------------------------------------|
| CR            | \( (a+d)/(a+b+c+d) \)                           |
| Sensitivity   | \( a/(a+c) \)                                   |
| Specificity   | \( d/(c+d) \)                                   |
| Positive PPV  | \( a/(a+b) \)                                   |
| Negative PPV  | \( d/(c+d) \)                                   |
| FDR           | \( 1-PPV \)                                     |
| False negatives | \( 1-NPV \)                                   |
| Lift          | \( (a+d)/(a+b+c+d) \)                           |

Panel (a) depicts the results to detect markers in previously replicated loci (Table 2a). As a function of \( n \), regression analysis shows that \( k \) decreases as a function of \( n \) (trend \( P=7.88 \times 10^{-6} \)). Furthermore, analysis of variance (ANOVA) discloses statistically significance difference in the average value of \( k \) as a function of \( n \) (\( F_{2,60}=9.197, P=0.00225 \)). The highest Tukey’s honest significance difference was obtained after comparing \( k \) for \( n=1000 \) and \( n=250 \) (\( d=-4.25, P_{adj}=-0.0237 \)). These results suggest that, in prGWAS, the larger the value of \( n \) the lower the number of DNA pools to be generated in order to obtain comparable \( P \)-values to those obtained at previously robustly replicated loci.

The number of DNA pools to be generated in prGWAS decreases as a function of \( n \) (trend \( P=0.0252 \)) for detecting regions with the strongest association signals (Table 2b). However, no statistically significance difference was found between the average value of \( k \) across sample sizes (\( k_{250}=5.63, k_{250}=7.11, k_{250}=5.93, k_{250}=4.68, F_{2,40}=2.723, p=0.0778 \)). Furthermore, no linear relationship between \( n \) and \( k \) (trend \( P=0.189 \)) nor in the average value of \( k \) for all sample sizes (\( k_{250}=5.59, k_{250}=5.85, k_{250}=5.76, k_{1000}=4.92, F_{2,61}=0.898, p=0.413 \)) was found when detecting regions of the genome with moderate evidence of association (Table 2c). Altogether, these results indicate that prGWAS detects those regions of the genome showing the strongest or moderate evidence of association previously reported by randomly generating a reasonably number of DNA pools regardless of the sample size of the master plates.

Performance measures

A total of 4,159 (7 diseases 22 chromosomes×3 master plates’ sample sizes×9 combined \( P \)-values) 2×2 contingency tables were constructed in all in silico experiments. The main results are presented in Figure 3.

Across all seven diseases, the sensitivity ranges from 0.15% to 1.4%. Regardless of \( k \), the lowest sensitivity values are obtained when the sample size of the master plate is \( n=500 \) and the highest, except for CAD, RA and T1D, when \( n=500 \). Furthermore, no linear relationship between \( n \) and \( k \) (trend \( P=0.189 \)) nor in the average value of \( k \) for all sample sizes (\( k_{250}=5.59, k_{250}=5.85, k_{250}=5.76, k_{1000}=4.92, F_{2,61}=0.898, p=0.413 \)) was found when detecting regions of the genome with moderate evidence of association (Table 2c). Altogether, these results indicate that prGWAS detects those regions of the genome showing the strongest or moderate evidence of association previously reported by randomly generating a reasonably number of DNA pools regardless of the sample size of the master plates.

In Figure 3b the results for the specificity are depicted. It can be seen that specificity increases slightly as a function of \( k \) when \( n \) is fixed, and is slightly higher for \( k \) fixed when \( n \) is large. In practical terms this implies that applying prGWAS using master plates of relatively large size would result in higher specificity values (i.e., ~99%) regardless of \( k \). However, as per
### Table 2a

Comparison of previously replicated loci, and genomic regions with the strongest and moderate association signals between prGWAS and GWAS.

| Collection | SNP              | prGWAS P-value (k) | Genotypic P-value |
|------------|------------------|---------------------|-------------------|
|            |                  | n=250               | n=500             | n=1000            |
| CAD        | rs4420638        | ND                  | ND                | ND                |
|            |                  |                     |                   | 1.7x10^1          |
| CD         | rs17231417       | 1.33x10^-13 (3)     | 9.46x10^-9 (2)    | 1.49x10^-13 (3)   |
|            | rs11805303       | 1.66x10^-10 (6)     | 5.38x10^-13 (5)   | 5.9x10^-11        |
| RA         | rs615672         | 3.99x10^-10 (7)     | 7.24x10^-15 (9)   | 6.97x10^-14 (3)   |
|            | rs6679677        | 1.09x10^-4 (9)      | 2.38x10^-25 (8)   | 2.09x10^-26 (2)   |
| T1D        | rs3087243        | *                   | *                 | 1.8x10^-5         |
|            | rs2104286        | 2.25x10^-2 (8)      | 8.43x10^-7 (2)    | 6.26x10^-10 (2)   |
|            | rs3788964        | *                   | *                 | 7.6x10^-3         |
|            | rs6679677        | 1.089x10^-6 (9)     | 7.33x10^-2 (9)    | 2.04x10^-30 (5)   |
|            | rs9270986        | 3.74x10^-13 (4)     | 6.31x10^-16 (2)   | 3.18x10^-4 (2)    |
| T2D        | rs1801282        | *                   | *                 | 5.4x10^-3         |
|            | rs4506565        | 1.41x10^-3 (9)      | 7.32x10^-4 (9)    | 3.24x10^-14 (3)   |
|            | rs5215           | *                   | *                 | 5.6x10^-3         |

### Table 2b

| Collection | SNP              | prGWAS P-value (k) | Genotypic P-value |
|------------|------------------|---------------------|-------------------|
|            |                  | n=250               | n=500             | n=1000            |
| BD         | rs420259         | *                   | *                 | 3.82x10^-8 (10)   |
|            |                  |                     |                   | 6.29x10^-8        |
| CAD        | rs13303049       | 8.39x10^-13 (3)     | 1.58x10^-12 (4)   | 1.51x10^-13 (3)   |
|            | rs11805303       | 1.66x10^-10 (9)     | 3.19x10^-12 (7)   | 6.38x10^-13 (5)   |
|            | rs10210302       | 3.76x10^-14 (10)    | 3.17x10^-15 (4)   | 6.50x10^-17 (3)   |
|            | rs9858542        | 2.53x10^-9 (9)      | 1.02x10^-7 (10)   | 1.94x10^-4 (7)    |
| RA         | rs17234657       | 1.34x10^-29 (3)     | 8.01x10^-13 (4)   | 3.73x10^-13 (2)   |
|            | rs1000113        | *                   | *                 | 2.03x10^-7 (7)    |
|            | rs10761659       | *                   | *                 | 5.60x10^-3 (8)    |
|            | rs10883365       | *                   | *                 | 1.75x10^-6        |
|            | rs17221417       | *                   | *                 | 5.58x10^-8        |
|            | rs2542151        | *                   | *                 | 2.03x10^-7        |
| T1D        | rs6679677        | *                   | *                 | 2.38x10^-26 (4)   |
|            | rs6457617        | 3.46x10^-34 (10)    | 3.21x10^-28 (5)   | 1.21x10^-29 (3)   |
| T2D        | rs6679677        | *                   | *                 | 7.33x10^-13 (10)  |
|            | rs9272346        | 1.30x10^-12 (8)     | 3.92x10^-16 (4)   | 1.19x10^-12 (2)   |
|            | rs11171739       | *                   | *                 | 4.71x10^-9 (9)    |
|            | rs17696736       | *                   | *                 | 6.07x10^-14 (4)   |
|            | rs12708716       | 4.90x10^-7 (8)      | *                 | 4.26x10^-7 (10)   |
|            | rs9465871        | *                   | *                 | 3.34x10^-7        |
|            | rs4506565        | *                   | *                 | 3.24x10^-14 (3)   |
|            | rs9939609        | 1.14x10^-7 (4)      | 1.23x10^-4 (4)    | 3.40x10^-11 (3)   |
| Collection | SNP         | prGWAS P-value (k)          | Genotypic P-value |
|------------|-------------|----------------------------|-------------------|
|            |             | n=250                      | n=500             | n=1000          |
| BD         | rs4027132   | 8.33x10^{-6} (8)           | *                 | 9.88 x 10^{-4} (4) | 9.68x10^{-4} |
|            | rs7570682   | *                         | 7.58x10^{-8} (9)  | 1.64 x 10^{-5}  |
|            | rs1375144   | 7.7x10^{-6} (4)            | 5.93x10^{-8} (5)  | 9.64x10^{-8} (7) | 1.31x10^{-8} |
|            | rs2953145   | *                         | 3.91x10^{-6} (3)  | 6.57x10^{-6}    |
|            | rs4276227   | *                         | 5.61x10^{-8} (2)  | *               | 2.62x10^{-8} |
|            | rs683395    | ND                        | ND                | ND              | 5.11x10^{-4} |
|            | rs6458307   | ND                        | ND                | ND              | 4.35x10^{-4} |
|            | rs2609653   | ND                        | ND                | ND              | b              |
| CAD        | rs17672135  | *                         | 6.07x10^{-10} (9) | 2.51x10^{-8} (5) | 2.35x10^{-8} |
|            | rs383830    | *                         | 7.92x10^{-8} (4)  | 1.34x10^{-8}    |
|            | rs6922269   | *                         | 1.91x10^{-6} (6)  | 1.50x10^{-6}    |
|            | rs8055236   | *                         | 2.09x10^{-8} (8)  | 1.26x10^{-8} (2) | 5.60x10^{-8} |
|            | rs7250581   | *                         | 2.48x10^{-8} (3)  | 2.50x10^{-8}    |
|            | rs688034    | *                         | 3.77x10^{-8} (5)  | 3.75x10^{-8}    |
| CD         | rs12037606  | 1.08x10^{-6} (6)          | 9.54x10^{-8} (5)  | *               | 1.09x10^{-5} |
|            | rs6596075   | -                         | 1.04x10^{-4} (3)  | *               | 3.19x10^{-4} |
|            | rs6908425   | *                         | *                 | *               | 1.10x10^{-5} |
|            | rs9469220   | *                         | 2.99x10^{-8} (3)  | 1.82x10^{-8} (2) | 2.28x10^{-8} |
|            | rs7753394   | *                         | 4.34x10^{-8} (2)  | 6.06x10^{-8} (2) | 2.59x10^{-8} |
|            | rs7807268   | *                         | 1.60x10^{-8} (7)  | 2.36x10^{-8} (2) | 4.42x10^{-8} |
|            | rs6601764   | *                         | 9.84x10^{-10} (8) | 4.64x10^{-10} (8) | 8.95x10^{-10} |
|            | rs8111071   | *                         | 4.88x10^{-9} (10) | 1.45x10^{-9} (6) | 1.75x10^{-9} |
| HT         | rs2820037   | *                         | *                 | *               | 7.66x10^{-5} |
|            | rs6997709   | *                         | 3.58x10^{-5} (5)  | 5.70x10^{-5} (4) | 4.36x10^{-5} |
|            | rs7961152   | *                         | 3.65x10^{-8} (9)  | 3.03x10^{-8}    |
|            | rs11110912  | *                         | 1.45x10^{-8} (6)  | 1.94x10^{-8}    |
|            | rs1937506   | *                         | *                 | *               | 4.53x10^{-8} |
|            | rs2398162   | *                         | 7.89x10^{-10} (10)| 9.06x10^{-10} (8) | 5.67x10^{-10} |
| RA         | rs6684865   | a                         | -                 | 9.87x10^{-8} (6) | 3.14x10^{-9} |
|            | rs11162922  | a                         | 9.50x10^{-6} (5)  | 2.13x10^{-6} (4) | 5.70x10^{-6} |
|            | rs3816587   | a                         | *                 | 9.25x10^{-6}    |
|            | rs6920220   | a                         | *                 | 6.44x10^{-5} (5) | 1.58x10^{-5} |
|            | rs11761231  | a                         | *                 | 4.33x10^{-5} (3) | 2.65x10^{-5} |
|            | rs2104286   | *                         | 2.44x10^{-5} (5)  | 2.85x10^{-11} (2) | 2.52x10^{-5} |
|            | rs9550642   | 3.24 x 10^{-16} (2)       | 5.19x10^{-29} (2) | 2.31x10^{-101} (3) | 3.90x10^{-9} |
Continuation of Table 2c.

| Collection | SNP      | prGWAS P-value (k) | Genotypic P-value |
|------------|----------|--------------------|-------------------|
|            |          | n=250              | n=500             | n=1000            |       |
| T1D        | rs2837960* | *                  | -                 | 1.68x10⁻⁶         |       |
|           | rs743777  | *                  | -                 | 1.15x10⁻⁶         |       |
|            | rs2639703* | -                  | 2.67x10⁻⁸ (8)     | 1.74x10⁻⁵         |       |
| rs17388568 | *         | 3.68x10⁻⁴ (4)      | 3.27x10⁻⁶         |                   |       |
| rs2544677  | *         | *                  | 3.11x10⁻⁵         | 4.43x10⁻⁵         |       |
| rs17166496 | *         | *                  | 5.20x10⁻⁵         |                   |       |
| rs2104286  | *         | 8.43x10⁻² (2)      | 4.32x10⁻⁵         |                   |       |
| rs11052552 | 1.80x10⁻⁷ (10) | 1.40x10⁻⁴ (9)   | 7.24x10⁻⁷         |                   |       |
| rs2542151  | *         | *                  | 2.47x10⁻⁴ (4)     | 1.16x10⁻⁵         |       |
| T2D        | rs4655595* | *                  | 1.49x10⁻⁴ (5)     | 5.24x10⁻⁴ (2)     | 1.33x10⁻⁵ |
| rs6718526  | 1.39x10⁻⁵ (7) | 2.31x10⁻⁷ (6)     | 3.05x10⁻⁵         |                   |       |
| rs358806   | ND        | ND                 | ND                |                   |       |
| rs7659604  | *         | *                  | 9.42x10⁻⁵         |                   |       |
| rs9326506  | *         | *                  | 2.99x10⁻⁵         |                   |       |
| rs12304921 | ND        | ND                 | ND                | 7.07x10⁻⁶         |       |
| rs1495377  | *         | *                  | 2.51x10⁻⁷ (5)     | 6.52x10⁻⁶         |       |
| rs2930291  | *         | 7.15x10⁻⁴ (7)      | *                 | 4.40x10⁻⁵         |       |
| rs2903265  | *         | *                  | 1.89x10⁻⁴ (9)     | 4.98x10⁻⁵         |       |

(a) Comparison of association at previously replicated loci between prGWAS and the GWAS using the full data. Reported P-values in prGWAS are those closer to the P-values found with the full data.

(b) Comparison of regions with the strongest association signals between prGWAS and Burton et al., [6] as a function of the master plate’s sample size. Reported P-values in prGWAS are those closer to the P-values found with the full data.

(c) Comparison of regions of the genome showing moderate association in prGWAS and the GWAS using the full WTCCC data. Reported P-values in prGWAS are those closer to the P-values found with the full data.

*Not statistically significant.

ND=Not determined as a consequence of not passing quality control (see Methods).

SNP=Single nucleotide polymorphism; Chr=Chromosome; n=sample size of the master plate (see Methods); k=number of DNA pools being combined. BD=Bipolar disorder; CAD=Coronary artery disease; CD=Conduct disorder; RA=Rheumatoid arthritis; HT=Hypertension; T1D=Type 1 diabetes; T2D=Type 2 diabetes.

*Statistically significant in prGWAS but the resulting combined P-value was higher than that obtained using the full data set.

1As reported in Table 2 of Burton et al [6].

the results when the GWAS approach is used, it also seems that the change in specificity is not particularly high for large n. Figure 3c presents the results for the classification rate (CR). Obtained values range between 80% and >90%, with master plates of larger size and higher values k producing lower CRs. Conversely, when n is small (i.e., n=250), prGWAS produces the higher CRs for k>2. Furthermore, the CR tends to increase slightly when k>5 for this sample size.

Results for the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are presented in Figures 3d and e, respectively. In the former, master plates of size n=1000 provide PPV of >70% regardless of k, with PPVs stabilizing when k>3. Despite not being the highest, master plates of ~16% of controls and ~50% of cases (i.e., n=500) produce PPVs between 50% and 80%. On the other hand, the NPVs range from 80% to 90%, with the highest values found when n=250 regardless of k, followed by n=500 and n=1000. As a function of k, the NPV decreases for n=1000 and n=500, and slightly increases when k>3 for n=250 (panel e, blue line).

Figure 3f depicts the FDR. The results indicate that the FDR
Figure 2. Number of pairs of DNA pools to be randomly generated in \textit{prGWAS} to obtain comparable \textit{P}-values to those (a) previously replicated loci (Table 2a); (b) regions with the strongest association signals (Table 2b), and (c) regions of the genome showing moderate association when the full genotype data is used (Table 2c). Here, \( n \) is the sample size of the master plates across all seven diseases. Number of markers with similar \textit{P}-values to those reported in Burton et al., [6] are presented in parenthesis. Individual data points were jittered to facilitate interpretation.

Figure 3. Results of (a) sensitivity, (b) specificity, (c) classification rate, (d) positive predictive value, (e) negative predictive value, (f) false discovery and (g) lift across all chromosomes as a function of \( k \) and the sample size of the master plates (\( n=250 \) in blue, \( n=500 \) in pink and \( n=1000 \) in green). Measures are presented in %. BD=Bipolar disorder; CD=Crohn’s disease; CAD=Coronary artery disease; HT=Hypertension; RA=Rheumatoid arthritis; T1D=Type 1 diabetes; T2D=Type 2 diabetes.

does not hold for \( n=250 \) regardless of the disease, and the lowest values when \( n=250 \) regardless of \( k \). On the other hand, the lift for predicting negative markers is slightly greater than one for all diseases (data not shown). Altogether, these results suggest that the \textit{prGWAS} method is potentially useful for predicting positive markers in future data sets.

Discussion

By varying the number of pairs of DNA pools and the sample size of the master plates, it was possible to demonstrate that \textit{prGWAS} is efficient, specific and accurate compared to conventional GWAS [1].

In the GWAS approach, a total of \( \sim 17,000 \) DNA samples...
were genotyped, whereas using prGWAS this number would have considerably been reduced to at least 80 DNA pools (up to 10 DNA pools per disease + 10 DNA pool from the controls) regardless of how many individuals’ DNA samples are present in the master plates. As in DNA pooling genotyping every pool is treated as a single DNA sample, the reduction in genotyping costs by using prGWAS is substantial.

Despite the encouraging results of the prGWAS method when compared to GWAS using the WTCCC data, some limitations need to be acknowledged. First, in contrast with the reported GWAS results [6]. In the analyses reported here all cases and controls individual genotypes were used for the genetic association analysis; secondly, no control by population structure was performed, and thirdly, no SNPs were removed after visual inspection.

Potential applications of prGWAS include whole-exome and whole-genome sequencing for cases and controls, and GWAS when families of (not necessarily) identical structure with affected and unaffected siblings are recruited [23].

Acknowledgement and funding

This research is supported in part by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship and The Australian National University (ANU) High Degree Research scholarships granted to JIV. JIV is a doctoral student at ANU. Some of this work is to be presented in partial fulfillment of the PhD degree requirements. The first author thanks Ms. Lindsay Nailer from ANU and Dr. Fernando Marmolejo-Ramos from Stockholm University, Sweden for critical reading of an earlier version of this document.

Publication history

EIC: Kenneth Maiase, Wayne State University, USA.
Received: 03-Nov-2014 Final Revised: 09-Dec-2014
Accepted: 14-Jan-2015 Published: 20-Jan-2015

References

1. Velez JI, Chandrasekharappa SC, Henao E, Martinez AF, Harper U, Jones M, Solomon BD, Lopez L, Garcia G, Aguirre-Acevedo DC, Acosta-Baena N, Correa JC, Lopera-Gomez CM, Jaramillo-Elorza MC, Rivera D, Kosik KS, Schork NJ, Swanson JM, Lopera F and Arcos-Burgos M. Pooling/bootstrap-based GWAS (pGWAS) identifies new loci modifying the age of onset in PSEN1 p.Glu280 Ala Alzheimer’s disease. Mol Psychiatry. 2013; 18:568-75. | Article | PubMed Abstract | PubMed Full Text

2. Lalli MA, Cox HC, Arcila ML, Cadavid L, Moreno S, Garcia G, Madrigal L, Reiman EM, Arcos-Burgos M, Bedoya G, Brunkow ME, Glusman G, Roach JC, Hood L, Kosik KS and Lopera F. Origin of the PSEN1 E280A mutation causing early-onset Alzheimer’s disease. Alzheimers Dement. 2014; 10:5277-5283 e10. | Article | PubMed

3. Londono AC, Castellanos FX, Arbelaez A, Ruiz A, Aguierre-Acevedo DC, Richardson AM, Easteal S, Lidbury BA, Arcos-Burgos M and Lopera F. An 1H-MRS framework predicts the onset of Alzheimer’s disease symptoms in PSEN1 mutation carriers. Alzheimers Dement. 2014; 10:552-61. | Article | PubMed

4. Lopera F, Ardilla A, Martinez A, Madrigal L, Arango-Viana JC, Lemere CA, Arango-Lasprilla JC, Hincapie L, Arcos-Burgos M, Ossa JE, Behrens IM, Norton J, Lendon C, Goate AM, Ruiz-Linares A, Rosselli M and Kosik KS. Clinical features of early-onset Alzheimer disease in a large kindred with an E280A presenilin-1 mutation. JAMA. 1997; 277:793-9. | Article | PubMed

5. Parikh R, Mathai A, Parikh S, Chandra Sekhar G and Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008; 56:45-50. | Article | PubMed Abstract | PubMed Full Text

6. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661-78. | Article | PubMed Abstract | PubMed Full Text

7. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. | Website

8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559-75. | Article | PubMed Abstract | PubMed Full Text

9. Shaffer JP. Multiple Hypothesis Testing. Ann Rev Psychol. 1995; 46:561-84. | Article

10. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57:289-300. | Article

11. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. In Studi in Onore del Professor Salvatore Ortu Carboni. 1935; 13-60. | Article

12. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG and Macgregor S. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010; 87:139-45. | Article | PubMed Abstract | PubMed Full Text

13. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004; 74:765-9. | Article | PubMed Abstract | PubMed Full Text

14. Storey JD and Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003; 100:9440-5. | Article | PubMed Abstract | PubMed Full Text

15. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B (Methodological). 2002; 64:479-98. | Article

16. Velez JI, Correa JC and Arcos-Burgos M. A new method for detecting significant p-values with applications to genetic data. Revista Colombiana de Estadistica. 2014; 37:67-76. | Article

17. Stouffer SA, Suchman EA, DeVinney LC, Star SA and Williams RMJ. Adjustment During Army Life. Princeton University Press. 1949. | Pdf

18. Whitlock MC. Combining probability from independent tests: the weighted z-method is superior to Fisher’s approach. J Evol Biol. 2005; 18:1368-73. | Article | PubMed

19. Hartung J. A Note on Combining Dependent Tests of Significance.
20. Brin S, Motwani R, Ullman JD and Tsur S. Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD International Conference on Management Data. 1997; 265-76. | Article

21. Piatetsky-Shapiro G and Masand B. Estimating Campaign Benefits and Modeling Lift. San Diego, CA, USA. 1999. | Article

22. McNicholas PD, Murphy TB and O’Regan M. Standardising the Lift of an Association Rule: Department of Statistics, Trinity College Dublin, Ireland. 2007. | Pdf

23. Risch N and Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res. 1998; 8:1273-88. | Article | PubMed

Citation:
Vélez JI, Jack CA, Chuah A, Buckley B, Correa JC, Eastal S and Arcos-Burgos M. Cross validation of pooling/resampling GWAS using the WTCCC data. Mol Biol Genet Eng. 2015; 3:1. http://dx.doi.org/10.7243/2053-5767-3-1