Biochemical and Biophysical Research Communications 422 (2012) 363–368

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

IPGWAS: An integrated pipeline for rational quality control and association analysis of genome-wide genetic studies

Yan-Hui Fan [a], You-Qiang Song [a,b,*]

[a] Department of Biochemistry, The University of Hong Kong, Pokfulam, Hong Kong
[b] The Centre for Reproduction, Development and Growth, The University of Hong Kong, Pokfulam, Hong Kong

ARTICLE INFO

Article history:
Received 14 April 2012
Available online 30 April 2012

Keywords:
Genome-wide association study
Quality control
Association analysis
Manhattan plot
QQ plot

ABSTRACT

Large numbers of samples and marker loci were tested for association in genome-wide association studies (GWAS). Hence, quality control (QC) by removing individuals or markers with low genotyping quality is of utmost importance to minimize potential false positive associations. IPGWAS was developed to facilitate the identification of the rational thresholds in QC of GWAS datasets, association analysis, Manhattan plot, quantile–quantile (QQ) plot, and format conversion for genetic analyses, such as meta-analysis, genotype phasing, and imputation. IPGWAS is a multiplatform application written in Perl with a graphical user interface (GUI) and available for free at http://sourceforge.net/projects/ipgwas/.

1. Introduction

The advances in high throughput genotyping techniques have resulted in an explosion of genome-wide association studies (GWAS). There are 1449 published GWAS with *p*-values less than 5×10^{-8} for 237 traits by the second quarter of 2011 [1]. GWAS have considered common genetic variations across the entire human genome to identify susceptibility and/or resistance alleles and/or genotypes associated with a disease trait. A large number of samples, as well as hundreds and thousands of single nucleotide polymorphisms (SNP), were tested for association in GWAS. Hence, even a small systematic error or bias can result in detrimental false positive and/or false negative association signals. Genotype calling assigns a subject either homozygous of allele A or B, or heterozygous genotype for each SNP according to the signal intensities, where allele A and B is the minor and the major allele, respectively [2–4]. However, due to DNA quality and/or specific features during the hybridization process, the signal intensities vary among individuals and some individuals may be mistyped in genotyping calls [2,4–6]. SNP mistyping is an important cause of spurious association results [7].

Quality control filters are applied to remove potentially problematic individuals and/or makers to minimize potential false findings [3,8]. Discordance of genetic gender and phenotypic gender indicates plating errors or sample mix-ups [3,9,10]. It is advisable to exclude these individuals from further analysis. Both missing genotypes and heterozygosity rate are measures of DNA quality [3,4,9,11]. Individuals with high missing genotypes and anomalously heterozygosity rate should be removed from further analysis. Case-control study compares the differences between unrelated individuals, so individuals with cryptic relatedness should be excluded from further analysis. Population structure has the potential to cause confounding and biases, statistical software such as EIGENSTRAT [12] has been successfully developed to detect and correct for it in genome-wide association studies. SNPs with lower call rate are more likely present as false positive associations [3,13], SNPs with call rates $\geqslant$95% had few spurious associations [13]. SNPs show extensively departure from Hardy–Weinberg equilibrium (HWE) can indicate problems with genotyping and/or genotype calling, or a strong signal of association [3,9,14]. Therefore, HWE test should be only applied to controls. SNPs with HWE of $P \geqslant 0.001$ did not result in spurious associations [13]. SNPs with lower minor allele frequency (MAF) should be removed because both the genotyping quality and the power to detect an association signal tends to decrease with decreasing MAF [9,15]. SNPs with MAF $\geqslant$ 5% did not exhibit spurious associations [13].

Inconsistence between SNP arrays, genotype calling algorithms, batch size and batch composition are potential sources for false positive and false negative errors [16–18]. When cases and controls have been genotyped at different genotyping centers and/or at different times, recalling genotypes after detecting and removing poor quality samples is the most effective way to eliminate the plate biases [10,19]. SAQC has been developed to detect poor-quality SNP arrays and/or DNA samples [20]. The linkage disequilibrium

* Corresponding author. Address: 21 Sassoon Road, 3/F, Laboratory Block, Department of Biochemistry, Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong.
E-mail addresses: nolanfyh@gmail.com (Y.-H. Fan), songy@hkucc.hku.hk (Y.-Q. Song).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.04.117

(LD) based post-association cleaning approach detects spurious association by checking the neighboring markers in LD with 84% detection rate [21].

However, no filtering scheme is perfect, the ultimate step of quality control is manual inspection of the SNP calls using cluster graph, which provides a visual inspection and aids in identifying problematic SNPs. After association analysis, it is essential to check cluster graph of these SNPs which will be replicated to prevent wasteful replication efforts [2–4,15]. IPGWAS systematically integrates several quality control (QC) protocols [3,9] to identify and remove individuals, as well as SNPs, that may introduce bias. It is not reasonable to apply the same QC criteria to different studies because of differences in overall call rates between studies [13]. IPGWAS implements intelligent approaches using plots in each QC step to decide the thresholds, rather than arbitrarily deciding the thresholds.

Association analysis by comparing the allele or genotype frequency difference in unrelated individuals is a powerful method for identifying disease susceptibility loci or genotypes. The null hypothesis of no association between rows and columns of the 2×2 matrix of the counts of the two alleles among cases and controls are usually tested using Pearson's χ^2 test, Fisher exact test, linear regression for quantitative traits and logistic regression for disease traits. The null hypothesis of no association between rows and columns of the 2×3 matrix of the counts of the three genotypes among cases and controls are usually tested using Pearson's χ^2 test, Cochran-Armitage (CA) trend test, or Fisher exact test. The CA trend test is used to determine association in a $2 \times k$ contingency table, where k is the number of genotype classes in the genetic studies [3,22,23]. This test modifies the χ^2 test, is more conservative, and does not require Hardy–Weinberg equilibrium (HWE) holds [24]. A particular set of scores is assigned to genotypes to maximize the power of this test. The CA trend test is widely used in GWAS [25]. IPGWAS implements the CA trend test in addition to the Pearson's χ^2 test.

2. Methods

This section describes functions that were implemented in IPGWAS. These functions include identification of the rational thresholds in QC of GWAS datasets, combination of GWAS datasets, association analysis, Manhattan plot, quantile–quantile (QQ) plot, format conversion for genetic analyses, such as meta-analysis, genotype phasing, imputation and other functions.

2.1. Quality control

2.1.1. Individual Quality Control

Individual QC of GWAS consists of several steps. IPGWAS identifies individuals with discordant sex information because of labeling errors or sample contamination using the plot of the inbreeding coefficient F and the labeled gender of each individual. IPGWAS filters out individuals with an outlying call rate using the plot of ordered one minus missingness rate. Individual departures from expected heterozygosity caused by DNA contamination or problems with the correlation of samples would then be removed. Population-based case-control studies assume that all samples are unrelated. IPGWAS employs two methods to identify samples that have cryptical relatedness. The first method re-moves the individual with a lower call rate from each pair with an identical-by-descent (IBD) value larger than a threshold, such as 0.1875, which is halfway between the expected IBD value of second-degree and third-degree relatives. The second method plots the mean and variance of identical-by-state (IBS). Individual pairs with similar relationship will then be clustered together. Then, one individual with lower call rate from each pair that cannot be clustered with most individual pairs is removed.

2.1.2. SNP Quality Control

IPGWAS filters out SNPs with outlying call rates, similar with individual QC, using the ordered one minus missingness rate plot. SNPs departure from the HWE can be indicative of problems with genotyping or genotype calling. Hence, SNPs that strongly deviate from HWE in the control data are customarily excluded. IPGWAS explores the quantile quantile (QQ) plot of HWE p-value to identify SNPs that deviate from HWE. Missingness among SNPs that are not random with respect to phenotype can lead to false positive association results. IPGWAS can identify and remove these SNPs as well as SNPs with low minor allele frequency (MAF). A MAF threshold of 1–2% is usually applied, but a higher such as 5% is also used in small sample size studies.

2.2. Combination of GWAS datasets

It is useful to combine GWAS datasets when the genotype datasets were genotyped by different collaborators or genotyped in different batches. When carry out population stratification analysis, it is also useful to combine HapMap reference genotypes to investigate the population structure. It is very important to make sure that the two sets of SNPs are concordant in terms of positive or negative strand when combine two datasets. IPGWAS employs two methods to ensure the two sets of SNPs are concordant in terms of positive or negative strand. The first method removes all symmetric SNPs first, then checks the remaining SNPs to find out which SNPs in the two datasets were not on the same strand, finally, flips the strand of these SNPs in one dataset. This method works well for Illumina datasets. However, this method will remove too much SNPs for Affymetrix datasets, which may reduce the power of the study. The second method uses the Affymetrix annotation files to decide whether one SNP is on the positive or negative strand, and then flip all SNPs to the same strand for the two datasets.

2.3. Association analysis

Association analysis compares the allele or genotype frequency difference in unrelated cases and controls to identify disease susceptibility loci or genotypes. IPGWAS use the logistic regression and linear regression function of PLINK [26] directly to do the regression analysis. Besides logistic regression and linear regression, IPGWAS also provides Cochran-Armitage trend test with additive, dominant, and recessive models, and Pearson's χ^2 test with more detail statistical information than PLINK [26].

2.3.1. Pearson's χ^2 Test

Assuming that we have N individuals, R cases, and S controls, then there are total n_0, n_1, and n_2 individuals with 0, 1, and 2 risk alleles, respectively. There are r_0, r_1, and r_2 cases, as well as s_0, s_1, and s_2 controls with 0, 1, and 2 risk alleles, respectively (Table 1). The genotypic test provides a general test of association in the 2×3 table of disease-by-genotype, and the χ^2 with 2 degrees of freedom is calculated from the following formula:

Table 1
Distribution of genotypes.

	AA[a]	AB	BB	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

[a] Allele B is the risk allele and allele A is the other allele.

Table 2
Distribution of alleles.

	A	B	Total
Cases	a	b	$a+b$
Controls	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

In allelic model test: a is the count of allele A in all cases and $a = 2r_0 + r_1$; b is the count of allele B in all cases and $b = 2r_2 + r_1$; c is the count of allele A in all controls and $c = 2s_0 + s_1$; d is the count of allele B in all controls and $d = 2s_2 + s_1$.
In dominant and recessive model test: a, b, c, d are not the count of alleles but the count of genotype categories. Suppose allele A is dominant to allele B. In the dominant model test: $a = r_0 + r_1$, $b = r_2$, $c = s_0 + s_1$, and $d = s_2$. In the recessive model test: $a = r_0$, $b = r_2 + r_1$, $c = s_0$, and $d = s_2 + s_1$.

$$\chi^2 = \sum_{i=0}^{2} \frac{(r_i - E(r_i))^2}{E(r_i)} + \sum_{i=0}^{2} \frac{(s_i - E(s_i))^2}{E(s_i)} \tag{1}$$

$$E(r_i) = \frac{Rn_i}{N} \tag{2}$$

$$E(s_i) = \frac{Sn_i}{N} \tag{3}$$

where $E(r_i)$ and $E(s_i)$ are the expected number of genotypes in cases and controls, respectively.

To test the allele frequency difference (Table 2) between cases and controls, the χ^2 with 1 degree of freedom is calculated from the following formula:

$$\chi^2 = \frac{(ad - bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} \tag{4}$$

The χ^2 with 1 degree of freedom of dominant and recessive model can also be calculated using formula 4 with different methods to calculated a, b, c and d (Table 2).

The output of Pearson's χ^2 test includes SNP identifier (rs ID); chromosome; physical position; minor allele; major allele; counts of three genotypes in cases and controls separately; model of the test; number of individuals with non-missing genotypes; minor allele frequencies in cases, controls and both cases and controls; χ^2 statistic; asymptotic p value; estimated odds ratio; standard error of odds ratio; and 95% confidence interval for odds ratio. The comprehensive information provides details of the association analysis and can also be used as input file for Manhattan plot, QQ plot, and meta-analysis directly.

2.3.2. Cochran-Armitage trend test

Assuming the genotype distribution is as Table 1 shown. Score vectors (0, 1, 2), (0, 1, 1), and (1, 1, 0) are used for additive, dominant, and recessive models, respectively. The test statistics can then be written as follows:

$$\chi^2_{add} = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{(N-R)R[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]} \sim \chi^2_1 \tag{5}$$

$$\chi^2_{dom} = \frac{N[N(r_1 + r_2) - R(n_1 + n_2)]^2}{R(N-R)n_0(n_1 + n_2)} \sim \chi^2_1 \tag{6}$$

$$\chi^2_{rec} = \frac{N[N(r_0 + r_1) - R(n_0 + n_1)]^2}{R(N-R)n_2(n_0 + n_1)} \sim \chi^2_1 \tag{7}$$

The output file contains the same information as Pearson's χ^2 test except that the allelic odds ratio, standard error and 95% confidence interval was used for all three models.

2.4. Plot the results

2.4.1. QQ plot

QQ plot is a useful tool for interpreting the results of an association test. *P*-values should follow a uniform distribution between zero and one under null hypothesis. Hence, if a set of m *p-values* is ordered from lowest to highest, then the observed quantile of the jth ordered item should, on average, be equal to $j/(m+1)$. The log QQ p-value plot is usually used, the negative logarithm of the jth smallest p-value is plotted against -log $(j/(m+1))$, where m is the number of SNPs. Loci that deviate from the null line correspond to SNPs that deviate from the null hypothesis.

2.4.2. Manhattan plot

A Manhattan plot is a type of scatter plot, Manhattan plot of GWAS graphically displays the results of association analysis to chromosomal locations. The relative genomic coordinates for each chromosome are displayed along the *X* axis, and the negative logarithm of the association P values for each single nucleotide polymorphism displayed on the *Y* axis.

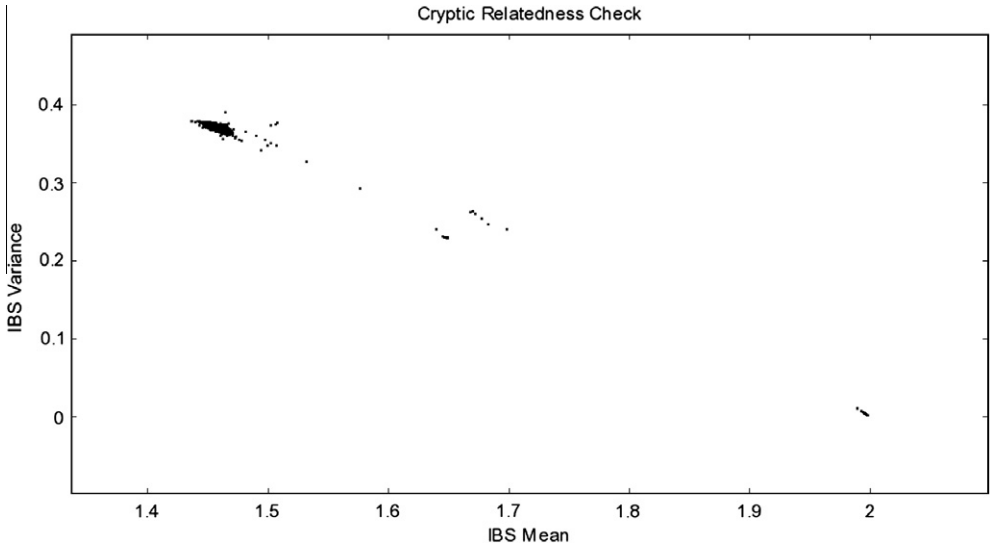

Fig. 1. Plot of the IBS mean versus IBS variance. This plot suggests 1.5 should be used as the IBS mean threshold to remove individuals that have cryptic relatedness with one or more other individuals.

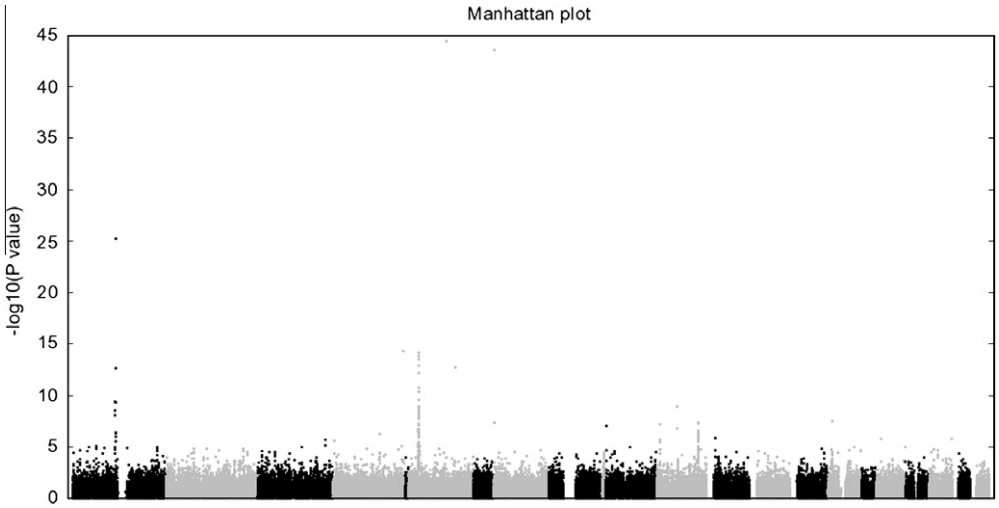

Fig. 2. Manhattan plot of the genome-wide association results. The genome-wide distribution of –log10 P values of Cochran-Armitage trend test is shown across the chromosomes (1–22).

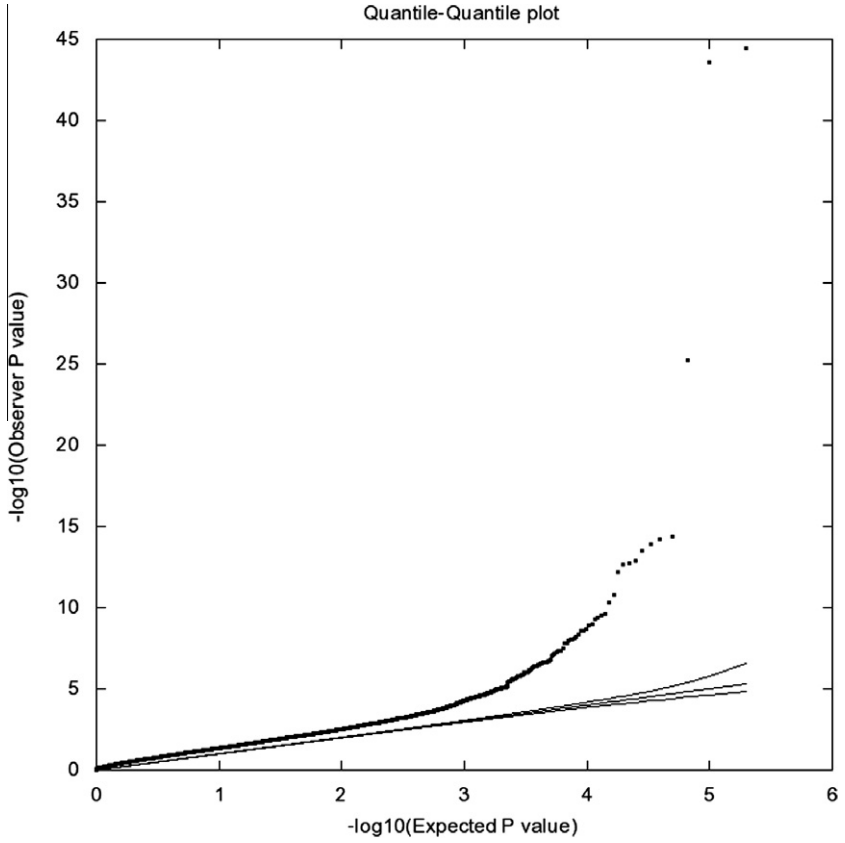

Fig. 3. Quantile–quantile plot of the genome-wide association results. The region between the line below and the line above the null line is the 95% confidence interval.

2.5. Data manipulation

IPGWAS contains a lot of useful functions for data manipulation. Such as change the affection status of individuals; filter subjects on affection status and/or gender status; split GWAS dataset by chromosome; filter GWAS results based on *P* value or remove singleton significant SNPs; and convert file format for other genetic analysis. IPGWAS can convert the EIGENSTRAT [12] adjusted χ^2 results to the PLINK [26] association results format, which then can be used for Manhattan plot, QQ plot and meta-analysis. IPGWAS can convert the PLINK [26] format genotype to PHASE [27,28] format for haplotype reconstruction and to BEAGLE [29,30] format for imputing genotypes, inferring haplotype phase and detecting identity-by-descent. IPGWAS can convert the association results of PLINK [26] and SNPTEST [31] to GWAMA [32] format for meta-analysis. IPGWAS can also convert the imputation results of MACH [33] into PLINK [26] or SNPTEST [31] format for further association analysis.

Table 3
Comparisons of IPGWAS results and the original results.

P-value[a]	WTCCC	IPGWAS	Common[b]
$<10^{-5}$	55/760 (0.07)	15/131 (0.11)	14/78 (0.18)
$<10^{-6}$	46/635 (0.07)	10/93 (0.11)	9/52 (0.17)
$<10^{-7}$	44/533 (0.08)	8/68 (0.12)	7/33 (0.21)
$<10^{-8}$	38/454 (0.08)	7/54 (0.13)	7/22 (0.32)

The values in plain font are the numbers of SNPs pass the *P*-value thresholds (after the slash) and the numbers of SNPs were reported by other genome-wide association studies (before the slash). The values in the parenthesis are the results of the numbers before the slash divide the numbers after the slash.

[a] *P*-value thresholds, the frequentist association test with additive model results were used for WTCCC, and Cochran–Armitage trend test with additive model results were used for IPGWAS.

[b] The number of SNPs that passed the *P*-value threshold in both WTCCC results and IPGWAS results.

2.6. Pathway analysis

Pathway analysis groups SNPs by the pathways they are involved in instead of analyzing one single SNP. It reduces complexity and increases explanatory power. Since most disease associated SNPs have been found by GWAS have very small effect size, if these SNPs are from the same biological pathway, then pathway analysis can increase the power to detect association between the pathways and a disease. It is also easy to interpret the pathways association results than the SNPs association results, especially when the SNPs are not located in any functional regions.

IPGWAS integrated the SNP ratio test (SRT) [34], which assesses the enrichment of significant associations in a pathway and computes an empirical *P*-value for each pathway based on permuting case and control status. The gene ontology (GO) based pathway analysis is under developing.

3. Results

The WTCCC dataset consists of about 2000 cases for each of seven major diseases and a shared set of about 3000 controls [35]. The subjects in the WTCCC study were genotyped with the GeneChip 500 K Mapping Array Set (Affymetrix chip). In this report, we used the genotype data of type 1 diabetes (T1D) and common controls to demonstrate how IPGWAS works.

Three cases and eleven controls were removed from the following analysis due to gender mismatch. Five percent was used as call rate threshold to exclude individuals and SNPs with low call rate. Five percent was also used as minor allele frequency threshold to exclude SNPs with low minor allele frequency. According the plot (Fig. 1) of mean and variance of IBS, 38 individuals were removed due to they have cryptic relatedness with one or more other individuals (1.5 was used as the IBS mean threshold). The result of this method is consistent with the result using 0.1875 as IBD estimates threshold, which removes 35 individuals, 34 of them were included in the 38 individuals. Furthermore, 92 population outliers were removed by EIGENSTRAT [12].

There are total 366658 SNPs in 4853 cases and controls passed the quality control. Cochran–Armitage trend test implemented in IPGWAS was used to calculate the association *p*-values. The Manhattan plot and QQ plot were shown in Fig. 2 and Fig. 3, respectively. Then we compared the IPGWAS results and the original WTCCC results and searched the GWAS catalog [1]. The results (Table 3) suggested that IPGWAS is more stringent, which will reduce the false positive association results and increase the success rate of the following replication studies. It is also cheaper and easier to conduct the replication studies.

4. Discussions

IPGWAS is designed to select the rational thresholds for QC, to combine and split GWAS datasets, to perform and plot the results of an association analysis, and to convert the format of different software for downstream analysis. IPGWAS uses the summary results of PLINK [26] and provides graphical output to assist the selection of rational thresholds for QC, which then will be used by PLINK [26] to complement to PLINK [26]. IPGWAS tests different models for association and generates more detailed association results, which then can be used for Manhattan plot, QQ plot and meta-analysis. IPGWAS and the popular genetic analysis tool such as PLINK [26], IGG3 [36] and GenABEL [37] complement each other. IPGWAS is a multiplatform application written in Perl with a user-friendly graphical interface, which is convenient to use, especially for genetic scientists who know little about programming. The module-based design is easy to maintain. Additional details regarding the function of IPGWAS can be found in its user manual.

Acknowledgments

We thank Patrick Kao for the comments and suggestions. We thank the users of IPGWAS for their valuable feedbacks. We also thank Wellcome Trust Case Control Consortium (WTCCC) and dbGAP for making their data available for our analysis (project #1488). This work was supported by grants from the Research Grant Council (HKU775208 M) and from the Research Fund for the Control of Infectious Disease (RFCID: No. 11101032) to YQS.

References

[1] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proc. Natl. Acad. Sci. USA 106 (2009) 9362–9367.

[2] A. Ziegler, I.R. König, J.R. Thompson, Biostatistical aspects of genome-wide association studies, Biochem. J. 50 (2008) 8–28.

[3] C.A. Anderson, F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris, K.T. Zondervan, Data quality control in genetic case-control association studies, Nat. Protocols 5 (2010) 1564–1573.

[4] Y.Y. Teo, Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure, Curr.Opin. Lipidol. 19 (2008) 133–143.

[5] R.J.L. Anney, E. Kenny, C.T. O'Dushlaine, J. Lasky-Su, B. Franke, D.W. Morris, B.M. Neale, P. Asherson, S.V. Faraone, M. Gill, Non-random error in genotype calling procedures: Implications for family-based and case–control genome-wide association studies, Am. J. Med. Genet. 147B (2008) 1379–1386.

[6] D.G. Clayton, N.M. Walker, D.J. Smyth, R. Pask, J.D. Cooper, L.M. Maier, L.J. Smink, A.C. Lam, N.R. Ovington, H.E. Stevens, S. Nutland, J.M.M. Howson, M. Faham, M. Moorhead, H.B. Jones, M. Falkowski, P. Hardenbol, T.D. Willis, J.A. Todd, Population structure, differential bias and genomic control in a large-scale, case-control association study, Nat. Genet. 37 (2005) 1243–1246.

[7] D. Mitry, H. Campbell, D.G. Charteris, B.W. Fleck, A. Tenesa, M.G. Dunlop, C. Hayward, A.F. Wright, V. Vitart, SNP mistyping in genotyping arrays—an important cause of spurious association in case-control studies, Genet. Epidemiol. 35 (2011) 423–426.

[8] M. Pongpanich, P.F. Sullivan, J.-Y. Tzeng, A quality control algorithm for filtering SNPs in genome-wide association studies, Bioinformatics 26 (2010) 1731–1737.

[9] M.E. Weale, Quality control for genome-wide association studies, Methods Mol. Biol. 628 (2010) 341–372.

[10] S. Turner, L.L. Armstrong, Y. Bradford, C.S. Carlson, D.C. Crawford, A.T. Crenshaw, M. de Andrade, K.F. Doheny, J.L. Haines, G. Hayes, G. Jarvik, L. Jiang, I.J. Kullo, R. Li, H. Ling, T.A. Manolio, M. Matsumoto, C.A. McCarty, A.N. McDavid, D.B. Mirel, J.E. Paschall, E.W. Pugh, L.V. Rasmussen, R.A. Wilke, R.L. Zuvich, M.D. Ritchie, Quality control procedures for genome-wide association studies, Curr Protoc Hum Genet, John Wiley & Sons, Inc., 2001, pp. Unit1.19.

[11] C.C. Laurie, K.F. Doheny, D.B. Mirel, E.W. Pugh, L.J. Bierut, T. Bhangale, F. Boehm, N.E. Caporaso, M.C. Cornelis, H.J. Edenberg, S.B. Gabriel, E.L. Harris, F.B. Hu, K.B. Jacobs, P. Kraft, M.T. Landi, T. Lumley, T.A. Manolio, C. McHugh, I. Painter, J. Paschall, J.P. Rice, K.M. Rice, X. Zheng, B.S. Weir, G.I. for the, Quality control and quality assurance in genotypic data for genome-wide association studies, Genet. Epidemiol. 34 (2010) 591–602.

[12] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, Nat. Genet. 38 (2006) 904–909.
[13] T. Miyagawa, N. Nishida, J. Ohashi, R. Kimura, A. Fujimoto, M. Kawashima, A. Koike, T. Sasaki, H. Tanii, T. Otowa, Y. Momose, Y. Nakahara, J. Gotoh, Y. Okazaki, S. Tsuji, K. Tokunaga, Appropriate data cleaning methods for genome-wide association study, J. Hum. Genet. 53 (2008) 886–893.
[14] A. Ziegler, I.R. König, A statistical Approach to Genetic Epidemiology: Concepts and Applications, Wiley-VCH, Weinheim, 2006.
[15] A. Ziegler, Genome-wide association studies: quality control and population-based measures, Genet. Epidemiol. 33 (2009) S45–S50.
[16] H. Hong, L. Shi, Z. Su, W. Ge, W.D. Jones, W. Czika, K. Miclaus, C.G. Lambert, S.C. Vega, J. Zhang, B. Ning, J. Liu, B. Green, L. Xu, H. Fang, R. Perkins, S.M. Lin, N. Jafari, K. Park, T. Ahn, M. Chierici, C. Furlanello, L. Zhang, R.D. Wolfinger, F. Goodsaid, W. Tong, Assessing sources of inconsistencies in genotypes and their effects on genome-wide association studies with HapMap samples, Pharmacogenomics J. 10 (2010) 364–374.
[17] H. Hong, Z. Su, W. Ge, L. Shi, R. Perkins, H. Fang, D. Mendrick, W. Tong, Evaluating variations of genotype calling: a potential source of spurious associations in genome-wide association studies, J. Genet. 89 (2010) 55–64.
[18] H. Hong, Z. Su, W. Ge, L. Shi, R. Perkins, H. Fang, J. Xu, J. Chen, T. Han, J. Kaput, J. Fuscoe, W. Tong, Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples, BMC Bioinformatics 9 (2008) S17.
[19] A. Pluzhnikov, J.E. Below, A. Konkashbaev, A. Tikhomirov, E. Kistner-Griffin, C.A. Roe, D.L. Nicolae, N.J. Cox, Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping, Am. J. Hum. Genet. 87 (2010) 123–128.
[20] H.-C. Yang, H.-C. Lin, M. Kang, C.-H. Chen, C.-W. Lin, L.-H. Li, J.-Y. Wu, Y.-T. Chen, W.-H. Pan, SAQC: SNP array quality control, BMC Bioinformatics 12 (2011) 100.
[21] B. Han, B.M. Hackel, E. Eskin, Postassociation cleaning using linkage disequilibrium information, Genet. Epidemiol. 35 (2011) 1–10.
[22] W.G. Cochran, Some methods for strengthening the common χ^2 Tests, Biometrics 10 (1954) 417–451.
[23] P. Armitage, Tests for linear trends in proportions and frequencies, Biometrics 11 (1955) 375–386.
[24] D.J. Balding, A tutorial on statistical methods for population association studies, Nat. Rev. Genet. 7 (2006) 781–791.
[25] B. Freidlin, G. Zheng, Z. Li, J.L. Gastwirth, Trend tests for case-control studies of genetic markers: power, sample Size and robustness, Hum. Hered. 53 (2002) 146–152.
[26] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.
[27] M. Stephens, P. Donnelly, A comparison of bayesian methods for haplotype reconstruction from population genotype data, Am. J. Hum. Genet. 73 (2003) 1162–1169.
[28] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978–989.
[29] B.L. Browning, S.R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals, Am. J. Hum. Genet. 84 (2009) 210–223.
[30] B.L. Browning, S.R. Browning, A fast, powerful method for detecting identity by descent, Am. J. Hum. Genet. 88 (2011) 173–182.
[31] J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes, Nat. Genet. 39 (2007) 906–913.
[32] Reedik Mägi, A.P. Morris, GWAMA: software for genome-wide association meta-analysis, BMC Bioinformatics 11 (2010) 288.
[33] Y. Li, C.J. Willer, J. Ding, P. Scheet, G.R. Abecasis, MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes, Genet. Epidemiol. 34 (2010) 816–834.
[34] C. O'Dushlaine, E. Kenny, E.A. Heron, R. Segurado, M. Gill, D.W. Morris, A. Corvin, The SNP ratio test: pathway analysis of genome-wide association datasets, Bioinformatics 25 (2009) 2762–2763.
[35] Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls, Nature 447 (2007) 661–678.
[36] M.-X. Li, L. Jiang, P.Y.-P. Kao, P.-C. Sham, Y.-Q. Song, IGG3: a tool to rapidly integrate large genotype datasets for whole-genome imputation and individual-level meta-analysis, Bioinformatics 25 (2009) 1449–1450.
[37] Y.S. Aulchenko, S. Ripke, A. Isaacs, C.M. van Duijn, GenABEL: an R library for genome-wide association analysis, Bioinformatics 23 (2007) 1294–1296.