Bioinformatics, 31(16), 2015, 2691–2696
doi: 10.1093/bioinformatics/btv219
Advance Access Publication Date: 22 April 2015
Original Paper

OXFORD

Genetics and population analysis

ESPRESSO: taking into account assessment errors on outcome and exposures in power analysis for association studies

Amadou Gaye[1,*], Thomas W. Y. Burton[2] and Paul R. Burton[1]

[1]School of Social and Community Medicine, University of Bristol, UK and [2]School of Computing Science, Newcastle University, UK

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 18, 2014; revised on March 8, 2015; accepted on April 19, 2015

Abstract

Motivation: Very large studies are required to provide sufficiently big sample sizes for adequately powered association analyses. This can be an expensive undertaking and it is important that an accurate sample size is identified. For more realistic sample size calculation and power analysis, the impact of unmeasured aetiological determinants and the quality of measurement of both outcome and explanatory variables should be taken into account. Conventional methods to analyse power use closed-form solutions that are not flexible enough to cater for all of these elements easily. They often result in a potentially substantial overestimation of the actual power.

Results: In this article, we describe the Estimating Sample-size and Power in R by Exploring Simulated Study Outcomes tool that allows assessment errors in power calculation under various biomedical scenarios to be incorporated. We also report a real world analysis where we used this tool to answer an important strategic question for an existing cohort.

Availability and implementation: The software is available for online calculation and downloads at http://espresso-research.org. The code is freely available at https://github.com/ESPRESSO-research.

Contact: louqman@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A critical question to answer when designing or extending large scale studies and biobanks is what sample size is required to achieve adequate statistical power (i.e. the ability to detect the true determinants of the outcome). For a large study aimed at exploring weak effects, the answer can have major implications for funding and resources because the number of participants to be recruited (sample size) depends critically on the desired level of power; the higher the desired power, the larger the required sample size. Among other things, the answer also depends on the quality of the measurements of the variables of interest since assessment errors, in both outcome and explanatory variables, can substantially reduce the power of association studies (Wong *et al.*, 2003). Failure to take account of such assessment errors in power analyses at the design stage of a study may lead to a serious over-estimation of its true statistical power and result in a research platform that is critically underpowered when it comes to analysis.

Most conventional approaches to estimating the sample size required to achieve adequate statistical power fail to account rigorously for the sensitivity and specificity of the assessment of categorical outcome and explanatory variables or the reliability of quantitative variables (Burton *et al.*, 2009). Furthermore, proper incorporation of the impact of assessment error in multiple variables, and of the analytic disturbance that may arise from variables that have not been measured at all, generally demands non-trivial extension of the methods being used and so is not attempted. But given the

magnitude of the sample sizes that may well be required, the vast investment of time and resources that is needed, and the scientific and financial costs associated with significantly misjudging required study size, the failure to properly account for these factors can seriously undermine strategic planning. These effects may be substantial.

For example, having made entirely reasonable assumptions about issues such as the measurement error in outcome and explanatory variables and heterogeneity in disease risk it has been shown that, for a study with an intended statistical power of 80%, the required sample size can easily more than double (Burton *et al.*, 2009). But, if a study designed to have a power of 80% to detect a given effect at a (two-tailed) *P*-value of 0.01 is only half its required size, it is straightforward to calculate that the actual power of that study will be given by the cumulative standard normal standard distribution up to the quantile:

$$-0.159 = \left[\left((2.576 + 0.842) \times \sqrt{0.5}\right) - 2.576\right]$$

Where 2.576 is the normal standard deviate corresponding to a two-tailed *P*-value of 0.01, 0.842 is the deviate corresponding to a power of 80%, and $\sqrt{0.5}$ reflects the shrinkage in the standard error if sample size doubles. This quantile (−0.159) corresponds to a real power of 43.6% (rather than the intended 80%). Equivalent calculations for candidate gene studies using a (two-tailed) *P*-value of $P < 10^{-4}$ or genome wide association studies using $P < 10^{-8}$ generate corresponding powers of 29.3 and 13.9%, respectively.

ESPRESSO—Estimating Sample-size and Power in R by Exploring Simulated Study Outcomes—(Burton *et al.*, 2009) has been developed to provide a way to incorporate key elements and thereby to allow for more realistic power analysis and sample size calculation for large-scale epidemiological studies. It is a software tool, written in the R programming language (Ihaka and Gentleman, 1996), providing a simulation-based approach to power and sample size calculations for stand-alone studies, analyses nested in cohort studies and consortium-based meta-analyses.

ESPRESSO is aimed primarily at researchers involved in designing and setting up studies to investigate the genetic and/or environmental basis of complex traits. In particular, it enables those designing large cohorts and biobanks to better estimate the sample size required to achieve adequate power. ESPRESSO can also allow for reviewers and funding bodies to verify the statistical power calculations put forward by researchers in their grant applications, thereby helping to ensure that resources are not wasted on incorrectly powered studies.

In this article, we provide a concise introduction to the new ESPRESSO which is built as an R and web-based software. This new version was used in recent analysis by Gaye *et al.* (2014), and in this article, we illustrate its use by answering a question put through to us by the Canadian Partnership for Tomorrow (CPT) cohort.

2 The new ESPRESSO and its implementation

This section reports concisely the work undertaken to produce the new ESPRESSO version from the initial R script by Burton *et al.* The new version was built with the aim to (i) provide a more comprehensive and user friendly R tool accessible to a wider audience; (ii) allow for analyses with quantitative outcomes and quantitative environmental exposures and (iii) extend the range of biomedical scenarios that can be investigated, particularly enabling more interaction models.

2.1 From the precursor R script to fully fledged R libraries

The original version of ESPRESS0 (Burton *et al.*, 2009) consisted of a single R script which allowed for the fitting of a model with a binary outcome and two covariates (one single nucleotide polymorphism (SNP) and one binary environmental exposure) by calling a text file that held the input parameters for the calculations. This programming paradigm was fine for the initial project but a more flexible solution was required for the analyses of the wider range of biomedical scenarios sought under the new version of the tool. Hence, it was necessary to write functions to cope with the increased complexity. Whilst the initial R script allowed for the investigation of five biomedical scenarios, the newly build R packages (libraries) enable power and sample size calculations for a total of 14 scenarios reported in Table 1. Six R packages each with a dozen of functions were developed. The open source code of the packages, available freely from GitHub (https://github.com/ESPRESSO-research), allows for users proficient in the R programming language to download the tool and use it as is, or modify the code in ways that best suit their needs.

2.2 The new parameters

The initial parameters of ESPRESSO have already been detailed by Burton *et al.* (2009). Therefore, this section focuses on five new key parameters that were required in the new version along with some of the considerations for choosing how these parameters might be set.

In the new version of ESPRESSO, the outcome variable and the environmental determinants can be modelled as quantitative. Hence, the parameters '*pheno.reliability*' and '*env.reliability*' were introduced to set the level of uncertainty on the outcome and covariates measurements. These represent the reliability (test–retest reliability) of the assessment of a quantitative variable, which is a characteristic that reflects the consistency of the observed measurement across several repeats.

The new version of the tool allows for users to model two SNPs (rather than one in the initial script) as being in linkage disequilibrium (LD). To enable LD between the two SNPs, two new parameters (*LD* and *targetLD*) were added to the list of input parameters. The parameter '*LD*' is a binary indicator: set to 1 to introduce LD between the two SNPs, and to 0 to generate two independent SNPs. The correlated SNPs are generated using a multivariate normal distribution function and the method developed in the R package *HapSim* (Montana, 2005). *HapSim* models a haplotype as a multivariate random variable with known marginal distributions and pairwise correlation coefficients. The package allows for the simulation of a SNP haplotype of several biallelic loci. In our implementation of the method for ESPRESSO, we limited the number of loci to two because we generate only two SNPs in LD. Our implementation of the method consists of two main steps: (i) we compute the covariance matrix required to generate two correlated binary vectors of length *n* (each vector represents one SNP and *n* is the number of observations) and (ii) we use the covariance computed in step (1) to generate a matrix of data that follow a multivariate normal distribution. For two loci, there are four possible haplotypes; the sum of the frequencies of the four possible haplotypes across the *n* simulated individual is 1 and the Lewontin's *D* and Pearson's *r* correlation values calculated from the single frequencies of the four haplotypes is equal to the target level of correlation (desired level of LD) specified in the first step. If the number of individuals to simulate is large, the programme runs more slowly. Because this setting (i.e. LD between the two SNPs) could be time consuming, it is preferable to set the sample size and the number of simulations to low values for an initial explorative analysis.

Table 1. Overview of the models/scenarios that could be investigated with the initial ESPRESSO script (GA, GB, EB, EB × GA and EB × GB) versus those that are enabled under the new R libraries (all other nine models)

	Additive genetic variant (GA)	Binary genetic variant (GB)	Quantitative environmental exposure (EQ)	Binary environmental exposure (EB)
Additive genetic variant (GA)	GA × GA			
Binary genetic variant (GB)	GB × GA	GB × GB		
Quantitative environmental exposure (EQ)	EQ × GA	EQ × GB	EQ × EQ	
Binary environmental exposure (EB)	EB × GA	EB × GB	EB × EQ	EB × EB

The main effect scenarios that can be investigated are on the cells in the first column whilst the interactions scenarios are in the inner cells of the table.

The parameter '*target LD*' represents the desired level of LD between the two SNPs, if they are to be modelled as being LD. The user should consider the minor allele frequencies (MAFs) of the two SNPs when setting the desired level of LD. The minor allele frequencies of the SNPs should not be markedly different. This, simply because a prefect correlation (i.e. an absolute value of 1) cannot be obtained if the SNPs have markedly different minor allele frequencies as demonstrated mathematically in the Section 1.1 of the Supplementary Material.

2.3 The web-based version of ESPRESSO

The development of R packages improved greatly the usability of the tool and its maintenance (error tracking and debugging), but it was mainly *confined* to the R community. In order to widen the use of ESPRESSO it was important to make it accessible to non R users. We hence built a website based on Joomla content management system (Joomla, 2014) and embedded an online version of the tool (http://www.espresso-research.org/). The website contains extensive documentation about the tool and help information are available, upon a click, for each item on the graphical user interface of the calculator. The version of the ESPRESSO software running in the background of the page can be updated by simply installing the latest packages on the server where the page is hosted.

2.4 Overview of the ESPRESSO algorithm

An ESPRESSO simulation essentially comprises five steps as summarized graphically in Figure 1.

First (step 1 in Fig. 1), a series of input values required to set the simulation (e.g. number of runs), outcome, genetic and/or environmental determinant parameters are specified.

Then (step 2 Fig. 1); an error free dataset which contains the true outcome and determinant values for each simulated individual is generated. The word 'true' here refers not to the true value of some real individual in the real world, but rather the true values (without error) of each simulated variable in each simulated subject within ESPRESSO. If the outcome is binary, the 'true' outcome may be perturbed by heterogeneity in the base-line risk of disease arising from the impact of unmeasured determinants that do not themselves appear in the model.

At the next stage of the process (step 3 in Fig. 1) an error is generated and added to the true data to produce the 'observed' data. This error, in effect, disturbs (or may disturb) the observed values of the outcome variable and each of the covariates. The structure and magnitude of the error depends on the input parameters—for example, reflecting the presumed sensitivity and specificity of the assessments of binary variables or of the reliability of quantitative measures.

Then (step 4 in Fig. 1), the observed data generated in the simulation stage are analysed by generalized linear modelling (GLM).

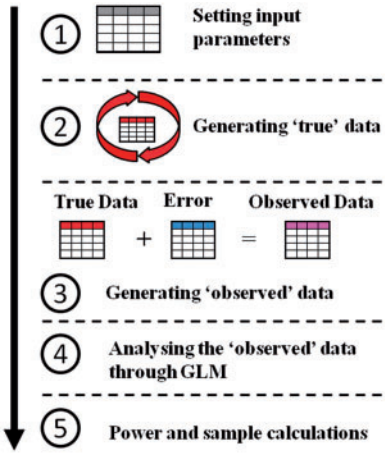

Fig. 1. Flowchart that shows the main steps in an ESPRESSO process

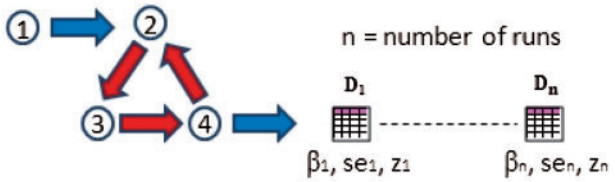

Fig. 2. Graphical view of the GLM analysis in ESPRESSO. After each simulation run a dataset of observed values is generated analysed and the beta coefficient, standard error and z-statistic stored

Steps 2, 3 and 4 are repeated for a number of times equal to the number of runs specified at step 1. After each run, as shown graphically in Figure 2, a matrix, D, of observed data is generated and analysed by GLM, and the estimates (beta, standard error and z-score), obtained from the GLM fit, are stored in three distinct vectors.

Finally (step 5 in Fig. 1), the sample size required to achieve the desired power specified at step 1 and the empirical and theoretical (modelled power achievable with the input sample size, also specified at step 1) are calculated.

The sample size required is the product of the relative change in standard error needed to achieve the desired power by the input sample size.

The empirical power is the proportion of runs in which the z-statistic for the parameter of interest exceeds the z-statistic for the desired level of statistical significance.

The theoretical power is the probability that the z-statistic obtained from the GLM fit takes any value less than or equal to the ratio of the mean beta coefficient to the mean standard error, i.e. it is the cumulative distribution function associated with the z-statistic obtained from the GLM fit.

The empirical power is not informative for extreme values of the standard error of the log odds ratio; in such cases one should consider the theoretical power.

3 Case study: analysis of the power of CPT cohort project to study quantitative traits

3.1 The CPT project

The CPT is a pan-Canadian initiative funded by the Canadian Partnership Against Cancer (CPAC). It aims to create a national biobank/bio-repository to provide a platform for future research on common chronic disease including cancer and cardiovascular disease (Borugian *et al.*, 2010). CPT is based on the integration of five large provincial cohorts each recruiting several tens of thousands of middle-aged participants. The planned (target), current and projected final recruitment numbers for the CPT project (at the time of this analysis) are outlined in Table 2 (Borugian *et al.*, 2010). The research team was aiming for a final sample size somewhere in the range 110 000–180 000 but was uncertain what inferential benefits would accrue from being towards the top of that range rather than towards the bottom.

3.2 Aim of the analysis and rational for using ESPRESSO

The aim of this analysis was to assess the statistical power of CPT as a platform for research projects exploring quantitative traits as outcomes, given its ultimate sample size. The analysis was requested by CPT to inform the primary (and immediate) strategic decisions to be made by CPAC on whether to continue recruitment at a rate that was likely to produce a total of approximately 110 000 participants by the end of March 2012, or alternatively to prioritise and step up recruitment with the aim of recruiting as many as 180 000. The ESPRESSO platform was used to carry out the calculations because unlike standard approaches it takes account of uncertainties around outcome and covariates measurements as mentioned in the introductory section of the article.

3.3 The outcome variables investigated

The exploration of the power profiles of CPT for quantitative outcomes was based on the estimates of participant-to-participant variation (standard deviation) for an extensive range of critical disease-related traits. These traits were originally drawn up for the power calculations of the CARTaGENE project (CARTaGENE, 2008) carried out under the direction of P. R. Burton. Forty-three outcome variables were investigated; they are either physical measures or biochemical and haematological parameters. An analysis of biochemical and haematological parameters on fresh blood is useful because: (i) it provides a series of valuable quantitative traits which are meaningful in their own right as complex traits that are worthy of aetiological study; (ii) it provides a series of quantitative traits that reflect intermediate traits that lie on the causal pathways leading to a number of complex binary traits that are of scientific interest; (iii) it includes a number of 'health screening' parameters that are of interest to potential recruits and therefore provide a tangible 'return' for agreeing to participate. The scientific rationale that justified the inclusion of each of the 43 variables in this study are the same as those that justified their inclusion in the CARTaGENE project (CARTaGENE, 2008). The rationale and the assumed distribution (mean and standard deviation) of each of the variables are available from earlier work (CARTaGENE, 2008; Gaye, 2012). This section serves mainly as an illustration of one of the possible uses of ESPRESSO, so the results are restricted to two outcome variables [systolic blood pressure (SBP) and a generic standardized variable]. The analyses of the other variables were conducted using the same strategy with full results reported elsewhere (Gaye, 2012).

Table 2. Configuration of the CPT cohort: sample size at the time of this analysis and target sample size

Name	Age-range at recruitment	Target sample size	Sample size by the time of this analysis
Atlantic cohort	40–69 years	30 000	15 000–25 000
British Columbia cohort	40–69 years	40 000	15 000–25 000
CARTaGENE (Quebec)	35–69 years	20 000	20 000
Ontario health survey (Ontario)	35–69 years	15 0 000	35 000–70 000
The tomorrow project (Alberta)	40–69 years	50 000	25 000–40 000
CPT project overall	Predominantly 35–69 years	250 000	110 000–180 000

3.4 Methods

3.4.1 The biomedical scenarios investigated and the strategy

The power profiles of CPT were analysed for the six scenarios summarized in Table 3. Because the fitted model is linear (i.e. continuous outcome), it is easy to mathematically calculate the minimum estimated effects from the first effect obtained empirically; however, we deliberately chose to obtain all the minimum estimated effects empirically. For each outcome and under each scenario, an iterative approach was used in ESPRESSO to determine the minimum estimated effect that can be detected with an empirical power of $80 \pm 2\%$ using the probable final samples sizes of CPT (110 000 and 180 000). The iterative approach consisted of looping through a range of effect sizes until reaching the smallest effect that ensures a power of 80%; these minimum effects were referred to as *minimum detectable effect sizes* (MDESs).

For each of the six scenarios in Table 3, the outcome, exposures and simulation parameters are described under Section 2 of the Supplementary Material and the values these parameters were set at are available under Section 4 of the Supplementary Material.

3.4.2 Analytic assumptions about the outcome and the genetic and environmental determinants

For each of the scenarios 1, 2 and 3 in Table 3, the GLM model fitted in ESPRESSO consists of one outcome (the quantitative trait being analysed) and one covariate; the three scenarios were analysed twice, once with a SNP as covariate and once with an environmental factor as covariate.

The genetic determinants were modelled as SNPs using an additive genetic model, as is now most commonly used (Wellcome Trust Case Control *et al.*, 2010): thus, the covariate was effectively modelled linearly as an ordinal variable taking a value 0, 1 or 2 indicating the number of minor alleles carried by each simulated subject. The genotyping error is generated as follows: we consider an observed marker that is not in complete LD with an unobserved causal variant ($r^2 < 1$); hence, the observed marker does not carry all of the information held by the unobserved causal variant. When the observed marker is typed, it is as if the unobserved causal variant has been typed with some error whose magnitude increases with decreasing LD. It is this error that we consider as the genotyping error. In other words, the genotyping error was taken as being

Table 3. The six scenarios explored to construct each power profile

Scenario	Minor allele frequency	Prevalence of 'at risk' environmental exposure	Mathematical model
1. Common determinants	0.30	0.50	Main effects only
2. Moderately common determinants	0.10	0.20	Main effects only
3. Uncommon determinants	0.05	0.10	Main effects only
4. Common determinants	0.30	0.50	Main effects + interaction
5. Moderately common determinants	0.10	0.20	Main effects + interaction
6. Uncommon determinants	0.05	0.10	Main effects + interaction

equivalent to the error that arises when the genotype at a locus of interest is inferred from the genotype of an observed marker (with the same allelic distribution) that is in LD with the unobserved causal variant at the same locus of interest with an r^2 value of 0.8. This corresponds to the weakest LD with HapMap 2 markers on the Affymetrix 500 K chip (Barrett and Cardon, 2006).

The environmental determinants were modelled as binary, and the measurement error was introduced by assuming an underlying latent variable with a reliability of 0.7. This reflects a moderate level of measurement error corresponding, for example, to blood pressure measurement in the Intersalt Study (Dyer *et al.*, 1994).

Gene-environment interactions were modelled using product terms again assuming an additive genetic model. Significance tests for genetic main effects and interactions were based on P-value < 0.0001 (i.e. assuming vague candidate genes) or P-value $< 10^{-7}$ (genome wide association studies); these P-values are considered conservative enough for genetic determinants (Pearson and Manolio, 2008; Pharoah *et al.*, 2004; Risch and Merikangas, 1996; Storey and Tibshirani, 2003). Non-genetic effects were tested at P-value < 0.01, a P-value more stringent than 0.05 the value widely used in association studies (Burton *et al.*, 2009; Ioannidis, 2005).

Unless otherwise specified, power estimation was based on the standard deviation and on the measurement reliability of the trait being considered as obtained from the analysis of the CARTaGENE optimization phase. When no firm evidence to the contrary was available to determine the likely measurement reliability of the quantitative trait being considered, it was taken to be 0.7.

3.5 Results

It is important to understand how the MDES should be interpreted. To illustrate the interpretation, consider Table 4 which provides an abstract of the power profile of CPT to investigate SBP as a quantitative outcome measured conventionally in a clinic setting (the results for all the scenarios on Table 3 are reported under Section 3 of the Supplementary Material). Conventional (peripheral) blood pressure is measured as the mean of three measurements. The device chosen (Colin Prodigy II Vital Signs Monitor OM-2200) is an automated device that uses the oscillometric method for assessing blood pressure.

Table 4. Minimal detectable effect sizes for SBP with 110 000 participants

	Genetic main effect	Environment main effect
P-value	10^{-7} (GWAS)	0.01
Moderately common determinants	1.0433	0.8123

The classification of the determinants as moderately common refers to the MAF of the genetic determinant (0.1) and the prevalence of the environmental (0.2), respectively, as reported on Table 3.

The population distribution of the variable reported in Table 4 (SBP) is: mean = 126 mm/Hg and SD = 18.2. In the body of the table, the MDES for the environmental main effect for the moderately common exposure was reported as 0.8123 mm/Hg. This scenario (Table 3) invokes a binary environmental exposure with a prevalence of 0.2 (20%). The reported results therefore imply that if the final sample size of CPT was 110 000 participants, if conventional clinic blood pressure was measured using the standard operating protocol (SOP) outlined in the first paragraph above, and if scientific interest focused on the impact of a binary environmental exposure which had realistic characteristics corresponding to those outlined in Section 3.4.2, the power calculations would indicate that there was an 80% *chance* of detecting, at P-value < 0.01, a real effect of that environmental determinant corresponding to an increase (or decrease) in SBP of 0.8123 mm/Hg, on average. Similarly, if interest were focused on a moderately common SNP in a genome wide association study (GWAS) there would be an 80% *chance* of detecting the effect of a SNP with a minor allele frequency of 0.10 (10%) at P-value $< 10^{-7}$ (for genome-wide inference) if that SNP really increased or decreased SBP by at least 1.0433 mm/Hg.

3.6 Conclusions

Given the magnitudes of the effect sizes that could be detected using the entire data of the CPT project, the cohort has a good potential to study the aetiological architecture of quantitative traits. Scrutiny of the power profiles of the quantitative variables tabulated individually demonstrates that given a sample size of 110 000 or 180 000, genetic and environmental main effects associated with any quantitative variables that are collected across the whole CPT project can potentially be studied with substantial power—effect sizes as small as 1/12th of a standard deviation will be reliably detectable even under the most challenging scenario (uncommon genotype, i.e. MAF = 0.05, with testing at P-value $< 10^{-7}$ under GWAS) (Bansal *et al.*, 2010; Bodmer and Bonilla, 2008; Burton *et al.*, 2009). But, as would be anticipated (Burton *et al.*, 2009; Wong *et al.*, 2003) the power to detect gene-environment interactions is considerably less strong. Given the central relevance of such interactions, it is important to note that a sample size of 180 000 rather than 110 000 would markedly enhance the capacity to study gene-environment interactions when the interacting determinants are both other than common. The larger sample size will also allow additional scope for data sub-setting.

4 Discussions

The ESPRESSO software allows for elements that are not taken into account in conventional power calculators to be included more readily in the power and sample size calculations for stand-alone case-control and cohort studies as well as for case-control analyses nested in cohort

studies. The new version of ESPRESSO is implemented as open source R packages to allow researchers proficient in the R programming language to use it in a flexible way and give them the ability to access and alter the code to answer further scientific questions that require some modification to the downloadable version. This new version comes also with an online, menu-driven, interface for non R users or for those who just prefer a Graphic User Interface.

With the current version of ESPRESSO, it is not possible to carry out power calculations, for genetic association studies, that precisely represent reality, i.e. when an analysis might choose to model a substantive number of genetic variants at the same time. This is because the current version of the tool currently allows for the modelling of no more than two genetic exposures and two environment/life style exposures at one time. However, given Mendelian randomization which ensures that SNPs must be closely located to be correlated because of LD, and given the large number of participants in most studies in which ESPRESSO might be used, which means that the residual error structure will be affected very little by the degrees of freedom used up by including even tens of genetic covariates, we feel this is not a serious problem. There would only be a problem if the inferences based on using just two SNPs in isolation were to differ substantially from those derived if all SNPs were to be considered together and in many settings there is no substantial difference at all. For example, if a GWAS analysis deals sequentially with one million separate variant-disease associations, it is perfectly acceptable to model one of those associations in isolation.

4.1 Limitations and future work

The version of ESPRESSO presented here does not allow for power and sample size calculation where the genetic component is a complex haplotypes (i.e. several loci in LD). With the current version at most only two loci in LD can be modelled and this is not very realistic representation of haplotype blocks. So the current ESPRESSO approach is potentially restrictive in this regard. It is therefore desirable to extend the software by implementing additional methods that do enable the joint consideration of a larger number of SNPs in LD. This could potentially be done by adapting the method developed by Montana in the R package *Hapsim* (Montana, 2005) which we mentioned in Section 2.2.

In the current version of the tool, the process consists of simulating a dataset with a number of specific characteristics and seeing in what proportion of the simulations the effect of interest is detected. As part of further work, we consider exploiting this feature to build in a function that allows for the estimation of false discovery rate (FDR) by first determining the proportion of false discoveries among all the discoveries and then calculating the FDR as defined by Benjamini and Hochberg (Benjamini and Hochberg, 1995).

Acknowlegements

Public Population Project in Genomics and Society (P^3G), Biobank Standardization and Harmonization for Research Excellence in the European Union (BioSHaRE-EU), CPT project, Prof Nuala Sheehan, Dr Andrew J. Turner.

Funding

The initial development of the ESPRESSO tool used in this work was supported by the Canadian Institutes of Health Research through the Public Population Project in Genomics and Society (P^3G) [grant number 64070822]. The current development of ESPRESSO is funded by the European Union through the Biobank Standardization and Harmonization for Research Excellence in the European Union (BioSHaRE-EU) project [FP7-HEALTH-F4-201433].

Conflict of Interest: none declared.

References

Bansal,V. *et al.* (2010) Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.*, **11**, 773–785.

Barrett,J.C. and Cardon,L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.

Borugian,M.J. *et al.* (2010) The Canadian partnership for tomorrow project: building a pan-Canadian research platform for disease prevention. *CMAJ*, **182**, 1197–1201.

Burton,P.R. *et al.* (2009) Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int. J. Epidemiol.*, **38**, 263–273.

CARTaGENE. (2008) Cartagene, the world within you, http://www.cartagene.qc.ca (5 May 2015, date last accessed).

Dyer,A.R. *et al.* (1994) Urinary electrolyte excretion in 24 hours and blood pressure in the INTERSALT study. II. Estimates of electrolyte-blood pressure associations corrected for regression dilution bias. The INTERSALT cooperative research group. *Am. J. Epidemiol.*, **139**, 940–951.

Gaye,A. (2012) Key determinants of statistical power in large scale genetic association studies, https://lra.le.ac.uk/handle/2381/27882 (5 May 2015, date last accessed).

Gaye,A. *et al.* (2014) Understanding the impact of pre-analytic variation in haematological and clinical chemistry analytes on the power of association studies. *Int J Epidemiol*, **43**, 1633–1644.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Ioannidis,J.P. (2005) Why most published research findings are false. *PLoS Med.*, **2**, e124.

Joomla (2014) *The Platform Millions of Website Are Built On.* Open Source Matters, http://www.joomla.org/ (5 May 2015, date last accessed).

Montana,G. (2005) HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics*, **21**, 4309–4311.

Pearson,T.A. and Manolio,T.A. (2008) How to interpret a genome-wide association study. *J. Am. Med. Assoc. (JAMA)*, **299**, 1335–1344.

Pharoah,P.D. *et al.* (2004) Association studies for finding cancer-susceptibility genetic variants. *Nat. Rev. Cancer*, **4**, 850–860.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 9440–9445.

Wellcome Trust Case Control, C. *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.

Wong,M.Y. *et al.* (2003) The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int. J. Epidemiol.*, **32**, 51–57.