QpowR: Interactive power calculator for two-stage genetic association studies of quantitative traits.

Juan P. Steibel and Gonçalo Abecasis.

November 2008.

Contents

Introduction

This document describes the installation and use of QpowR, an interactive power calculator for two stage association studies in quantitative traits. An appendix gives the technical details on the methods used to compute power and to optimize experimental designs.

The calculator assumes that a sample of N unrelated individuals is genotyped following a two-stage design. In a two-stage design (Figure 1) the sample is partitioned into two subsets. On one subset, comprising a proportion π_s of the individuals, all (M) available markers are typed (stage one) and tested for association with a quantitative trait. A fraction π_m of the markers with stronger evidence of association to the quantitative trait are typed in the rest of the samples (stage two).

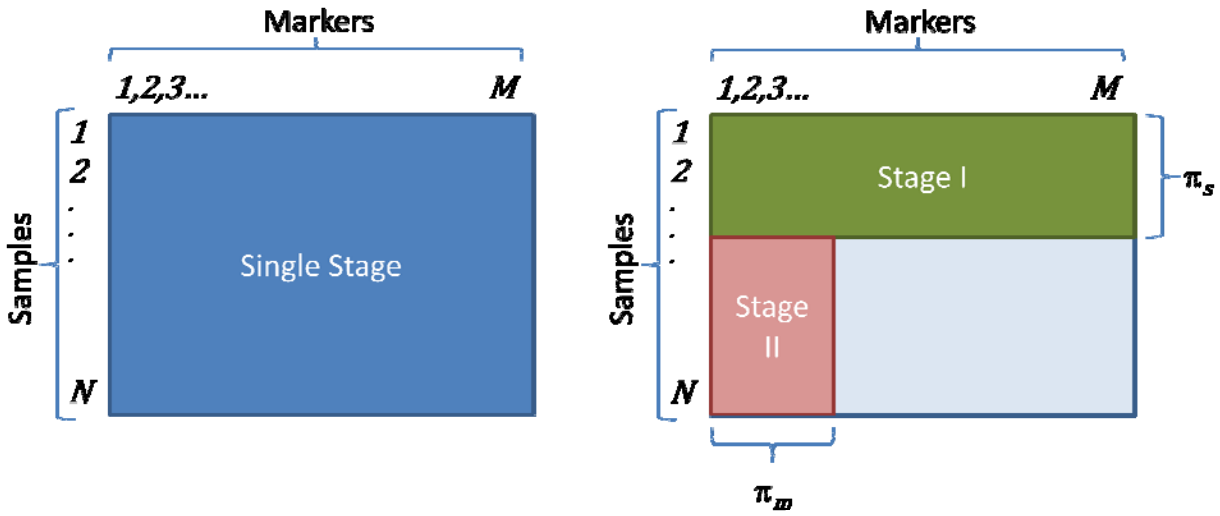

Figure 1. Single stage and two stage designs. In a single stage design all M markers are typed in all N individuals (blue rectangle). In a two stage design, all M markers are typed in a fraction π_s of the individuals (Stage I, green rectangle) and a fraction π_m of the markers are typed in the rest of the samples (Stage II, red rectangle).

If the effect sizes in both stages are the same, the two stage design will be less powerful than the single stage design with equal total sample size. However, depending on the sample allocation and the proportion of markers typed in the second stage, the loss in power may be minimal. On the other side, two stage designs may result in a significant reduction in cost. QpowR compares the power of these designs at a fixed sample size and at a fixed cost. It also allows optimizing a two stage design to yield either maximum power or minimum cost.

A very efficient and robust way of analyzing the results from a two stage design consists in pooling the two test statistic into a single joint test. This strategy increases the power compared to the traditional "replication" analysis where the results from the two stages are analyzed separately. Additionally, the joint test statistic is robust to data heterogeneity that can seriously affect a test on the combined data. The power calculator considers both joint and

replication analyses and provides the power and the test statistic threshold needed to implement both tests.

Download and run the program.

The power calculator requires that R (http://www.r-project.org/) be installed in the system. The QpowR script file can be downloaded from http://www.msu.edu/~steibelj/JP_files/qpowr.html. The script is portable to any system with a basic R installation.

To start the calculator:

1) Unix systems
    a. Copy the script file to the working directory.
    b. Open the console and Type: R <ENTER>.
    c. At the R-prompt, TYPE: `source("QpowR.r")` <ENTER>
2) Windows[1] and Mac systems
    a. Copy the script file to the working directory
    b. Start R
    c. Go to the R-menu: `FILE>CHANGE DIR>BROWSE`
    d. Browse to the working directory and click OK.
    e. At the R-prompt, TYPE: `source("QpowR.r")` <ENTER>

Using the interactive calculator.

Figure 2 presents the interface of QpowR. The input variables for power analysis are: total sample size (N), % of the phenotypic variance explained by the marker (H^2 *or* h^2), proportion of samples typed in the first stage (π_s), proportion of markers re-typed in the second stage (π_m) and experiment-wise error rate (α_m).

[1] When using R windows, the focus of the QpowR interface may be lost and the windows sent to the background. To avoid this problem, run R in SDI mode. Open R. Go to MENU>EDIT>GUIpreferences and Select the SDI option in the "single or multiple windows" line. Save the preferences and select OK. Close and re-start R.

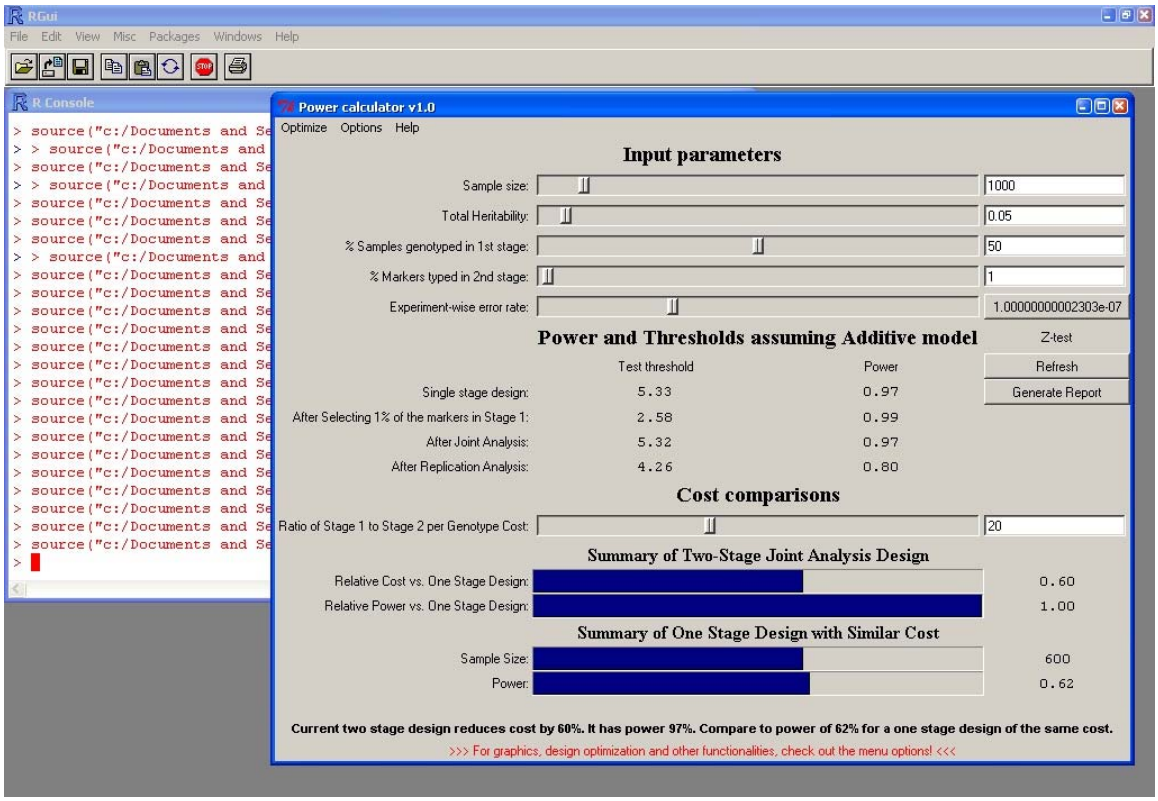

Figure 2. Main window of QpowR. Most of the functionality can be accessed from slider bars and data entry slots.

Values for these variables can be input using the sliding scale or the data entry slot in the top-panel of the window. To modify a parameter, click on the slider tab, hold the button and move the tab right or left. The actual value is displayed in the data entry slot or label to the right of the slider. Pointing on the scale to the right or left of the slider tab and left-clicking will change the value of the parameter by the minimum allowed increment. A right click will bring the slider to the place indicated by the mouse pointer. The calculator updates automatically when the slider is moved.

A precise value may be entered manually through the entry slot to the right of each slider. Simply click and type the value (erase old value if needed) and press <ENTER> to update the calculator.

The Input of α_m has two particularities. First, the sliding scale is logarithmic, to allow better resolution for small error rates. Second, the data entry has been replaced by a click button. A click on the α_m result button will open a simple dialog (Figure 3) to compute the experiment-wise error rate. Enter the total number of tests and the comparison-wise error rate in the dialog box and click OK to update the value in the calculator. A Bonferroni corrected p-value is computed and updated to the main window.

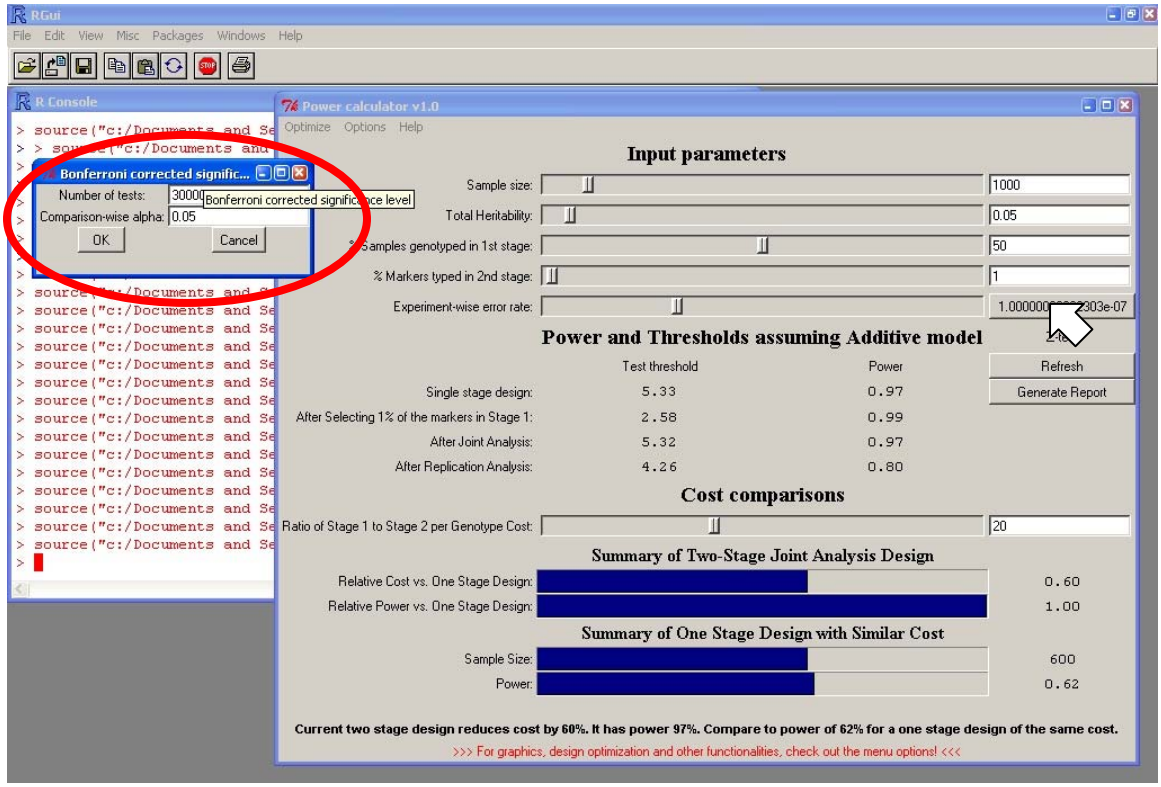

Figure 3. The dialog box to calculate the experiment-wise error rate (red oval) is opened when clicking on the result label (white arrow). The user enters the number of tests and comparison-wise error rate and clicks <OK> to update the main calculator.

The results of the power and test statistic threshold are displayed in the middle panel. The output rows are discussed below.

One Stage experiment: This is an experiment where all N samples are genotyped in the first stage. This provides the maximum power possible.

First Stage selection: This is the first step of the two-stage experiment. The test statistic threshold is the expected cutoff value to select π_m markers to carry onto the next stage.

Joint Analysis: Presents the power and test-statistic cutoff value if a joint statistic is computed (see details in the appendix).

Replication Analysis: The power and test-statistic cutoff value of a replication (independent) second stage statistic.

Relative cost comparisons are presented in the bottom panel of the calculator. A sliding scale or a data entry slot is used to set the relative genotyping cost of the second stage to the first stage (on a per-marker basis).

The results are represented graphically (through a slider) and numerically.

Relative cost: It is the cost of the two stage experiment, relative to the cost of a single stage experiment using the same N samples. Note: this can be larger than one, but the slider will grow until reaching the end of the scale (value 1.0) and it will stay at that point. The numerical output, on the other hand, will display the full value (>1.0).

N of single stage at same cost: The number of samples that could be typed at the same cost of the two-stage experiment.

Power of single stage at same cost: The power of the experiment mentioned in the previous paragraph.

Optimal designs

The *Optimize* menu entry allows improving the design selected by the user. Three options are available.

Cost: The power of the joint analysis is kept fixed and π_m and π_s are selected such that the minimum possible cost is attained.

Power: Optimal values of π_m and π_s are found such that a two stage experiment with the same cost will produce maximum power.

Plots: Power and relative cost are represented in a contour plot, as function of π_m and π_s. A small dialog is presented to add the current design point to the plot. Also, minimum cost and maximum power designs are presented. The optimization is only approximate as it is performed over a coarse grid of π_m and π_s values.

The computations may take a few seconds in some systems, and there is a message box indicating this. The results are displayed in a dialog window (figure 3). Press UPDATE to copy these values to the calculator. Press CANCEL to reject the proposed design.

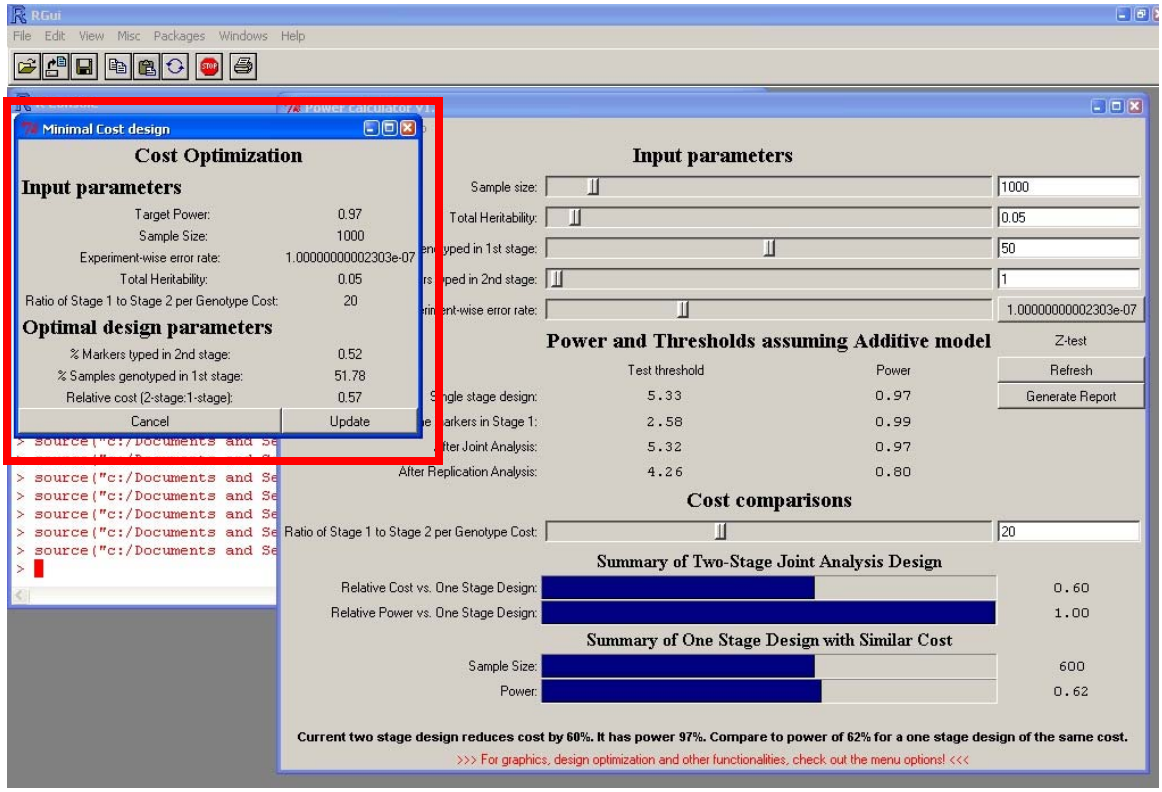

Figure 4. Cost optimization invoked from the Optimize menu (red square). Click on Update to copy the values to the main calculator window. A similar dialog is produced for Power optimization.

Option panel

In this version of the calculator, there are two options. By default the calculator assumes an additive model and an asymptotic test, resulting in a z-statistic. There are four combinations of these two options (Table 1). The selected model and test is indicated in the result panel. See ***technical details*** for further information on this.

Model	Type of Test	Statistic	Optimization
Additive	Asymptotic	z	Available
Additive	Exact	t	Unavailable
Dominance	Asymptotic	χ^2	Unavailable
Dominance	Exact	F	Unavailable

Table 1. Four test can be selected through the two option switches.

Alternative output

To facilitate exporting the results, the calculator includes a click-button labeled *Generate Report*. Click to this button to open a text window with a summary of the current state of the power calculator. Text can be highlighted by clicking at the first character and dragging the

mouse while keeping the mouse button pressed. The keystroke <CTRL>+<C> will copy the highlighted test to the clipboard. Figure 4 presents a dump these results pasted from the clipboard.

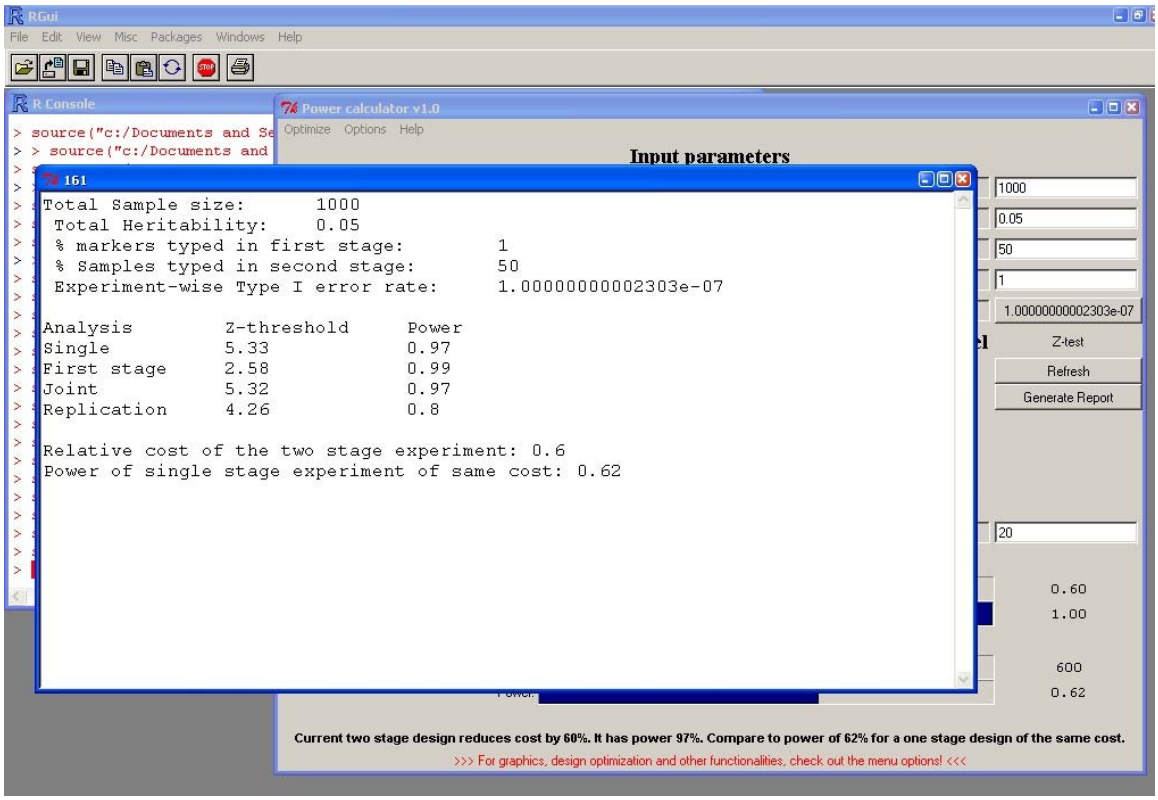

Figure 5. Report generated by QpowR. The text in the window can be cut and pasted to other applications.

Appendix: Technical details.

The model:

A linear fixed effects model with additive of marker (SNP) is assumed.

$$y_i = \mu + b \cdot x_i + e_i, \quad [1]$$

where x_i takes the value -1,0 or 1 for genotypes aa, Aa, AA respectively. μ is the overall mean, b is the additive effect and $e_i \; i.i.d \; N(0, \sigma_e^2)$ is the residual value.

The heritability, h^2, can be expressed in terms of the additive effect, the residual variance and the allelic frequency (p).

$$h^2 = \frac{V_b}{V_b + \sigma_e^2};$$
$$V_b = 2p(1-p)(b)^2$$

The user is asked to enter the expected heritability in the population.

Test statistic: The z-test can be used to test the overall hypothesis of association between the SNP genotype and the phenotype. Under the Null hypothesis, $z \sim \phi_{0,1}$, a standard normal distribution.

Distribution under the alternative hypothesis: Under the alternative hypothesis, the test statistics follows a normal distribution with mean $\boldsymbol{\lambda}$:

$$\lambda = \sqrt{N \frac{H^2}{1-H^2}}. \quad [2]$$

Critical value and Power:

The critical value (C) and the power $(1-\beta)$ of the z test can be computed as follows.

$$C = \pm \Phi^{-1}\left(1 - \frac{\alpha_m}{2}\right)). \quad [3]$$
$$1 - \beta = 1 - \Phi(C - \lambda) + \Phi(-C - \lambda), \quad [4]$$

where Φ is the normal cumulative density function(c.d.f) with mean m and variance v, Φ^{-1} is the inverse-normal c.d.f and α_m is the experiment-wise error rate.

Two stage tests: The sample is divided into two fractions: $N_1 = \pi_s N$ and $N_2 = (1-\pi_s)N$, a first stage test statistic z_1 is computed for every marker, and a proportion π_m of those SNPs is selected to re-test in the second stage. The second stage yields statistics z_2. Two possible tests are considered from those statistics.

Replication based test statistic: Reject the null hypothesis if both tests statistics are rejected individually.

Critical values:

$$C_1 = \Phi^{-1}\left(1 - \frac{\pi_m}{2}\right), \quad [5]$$

$$C_{rep} = \Phi^{-1}\left(1 - \frac{\alpha_m}{\pi_m}\right), \quad [6]$$

Power:

$$1 - \beta_{rep} = \left(1 - \Phi(C_1 - \lambda_1)\right)\left(1 - \Phi\left(C_{rep} - \lambda_2\right)\right) + \left(\Phi(-C_1 - \lambda_1)\right)\left(\Phi\left(-C_{rep} - \lambda_2\right)\right), \quad [7]$$

Joint test statistic:

Combine the statistics from each stage into a single test and test for significance:

$$z_j = \sqrt{\pi_s} z_1 + \sqrt{(1 - \pi_s)} z_2, \quad [8]$$

The distribution of this statistic conditional on the first stage statistic is a Gaussian distribution. The tail area probabilities can be calculated using the following formula.

$$\begin{aligned} P\left(|z_j| > C_j; |z_1| > C_1\right) &= \int_{-\infty}^{-C_j} \left(\left(1 - \Phi_{z_2}\left(\frac{C_j - \sqrt{\pi_s} x_1}{\sqrt{1 - \pi_s}} \right) + \Phi_{z_2}\left(\frac{-C_j - \sqrt{\pi_s} x_1}{\sqrt{1 - \pi_s}} \right) \right) \varphi_{z_1}(x_1) dx_1 \right) \\ &+ \int_{C_j}^{\infty} \left(\left(1 - \Phi_{z_2}\left(\frac{C_j - \sqrt{\pi_s} x_1}{\sqrt{1 - \pi_s}} \right) + \Phi_{z_2}\left(\frac{-C_j - \sqrt{\pi_s} x_1}{\sqrt{1 - \pi_s}} \right) \right) \varphi_{z_1}(x_1) dx_1 \right), \end{aligned} \quad [9]$$

To implement the joint test, equation [9] is first equated to α_m and solved for C_j assuming the null distribution for both test statistics. Power is obtained by computing the integrals in equation [9], replacing the C_j value obtained in the previous step, and assuming the distributions under the alternative hypothesis.

The implementation of the non-asymptotic test is straightforward. The same formulas presented above can be used replacing the Gaussian p.d.f and c.d.f with appropriate student-t distributions.

Dominance model

$$y_i = \mu + b \cdot x_i + d \cdot w_i + e_i, \quad [10]$$

Critical values:

$$C_1 = F^{-1}_{2,N_1-3}(1-\pi_m), \quad [11]$$

$$C_{rep} = F^{-1}_{2,N_2-3}\left(1-\frac{\alpha_m}{\pi_m}\right), \quad [12]$$

Power:

$$1-\beta_1 = 1 - F_{2,N_1-3,\lambda_1}(C_1), \quad [13]$$

$$1-\beta_2 = 1 - F_{2,N_2-3,\lambda_2}(C_{rep}), \quad [14]$$

$$1-\beta_{rep} = (1-\beta_1)(1-\beta_2). \quad [15]$$

Joint test statistic:

Combine the two test statistics into a single test[2] and test for significance:

$$f_j = \pi_s f_1 + (1-\pi_s) f_2. \quad [16]$$

The distribution of this test statistic is not a standard one, but the significance threshold and power can be computed numerically from the following expression.

$$P(f_j > C_j; f_1 > C_1) = \int_{C_1}^{\frac{C_j}{\pi_s}} \left(1 - F_{f_2}\left(\frac{C_j - \pi_s x_1}{1-\pi_s}\right) \mathrm{f}_{f_1}(x_1) dx_1\right) + \left(1 - F_{f_2}\left(\frac{C_j}{\pi_s}\right)\right) \quad [17]$$

where F_f is the cumulative density function of statistic f and f_f is the probability density function of statistic f.

A similar procedure is used for the chi-square test (assuming know residual variance or large *N*). The tests under the dominance model are included in the calculator as an experimental feature, as they are not optimal.

Relative cost:

The cost of a two stage experiment relative to a single stage experiment using the same number of samples N will be:

$$RC = \pi_s + (1-\pi_s)\pi_m\pi_c, \quad [18]$$

where $\pi_c \geq 1$ is the relative per genotype cost in stage two compared to stage one.

[2] The weights used in these tests are not optimal.

Number of samples in one stage experiment at the same cost:

If we fix the total cost $RC \leq 1$, we can compute the sample size of a single stage experiment of equivalent cost:

$$N_{1e} = N(\pi_s + (1 - \pi_s)\pi_m\pi_c), \qquad [19]$$

Power of one stage experiment with the same cost:

Using the sample size obtained from equation [19], the power of a single stage experiment of the same cost can be computed using equation [4].

Design Optimization (only available for z-test statistics).

Two types of optimizations are performed by QpowR. The optimal design consists of a pair of π_s and π_m values, such that the cost is minimized or the power is maximized. The cost minimization may be particularly difficult and sometimes the optimization is not successful. In those cases, a warning message will appear. The R system `optimize` function is used in both optimizations.

Power maximization:

For a given cost level $RC \leq 1$ and a value of π_s, the proportion of markers to be assayed in the second step can be obtained from equation [18] as:

$$\pi_m = \frac{RC - \pi_s}{(1 - \pi_s)\pi_c}, \qquad [20]$$

This expression can be replaced in equation [9] that is then minimized for π_s

Cost minimization:

To optimize cost, equation [18] is minimized over $\log(\pi_m)$, with π_s constrained to ensure the desired power level (this is done by solving equation [9] for π_s given π_m). Direct minimization of [18] over π_m or minimizing [18] over π_s, after solving [9] for π_m did not produce successful optimization in many scenarios.

Contour plots of cost and power:

The contour plot is computer over a grid of π_m and π_s values. The maximum value of π_s, is arbitrarily fixed at 0.8. The maximum value of π_m is $1/\pi_c$, because it represents the break-even point where the two stage design costs more than the one stage design.

After generating the plot, a widget is activated to add points to the plot.

Maximum power point corresponds to the design with maximum power among those with cost equal or smaller than the current design.

Minimum cost point corresponds to the parameter combination that yields the minimum cost while the power is equal or larger than the power of the current design.

The minimization or maximization performed in this way is vey imprecise because a coarse grid is used in the search. But the functionality has been included for those cases where the numerical optimization fails. A good strategy could be using the results from the contour plot as a starting point for a new round of numerical optimizations.