Computing effective sample size

The concept of effective sample size is widely used in statistical genetics to compare the relative power of two or more case/control genome-wide association studies (Newton-Cheh et al., 2009). Assuming that the effect sizes and allele frequencies are the same, the power of each study is determined by the total sample size and the case/control ratio. A balanced study with equal number of cases and control is the most powerful. But cases tend to be the limiting factor, so most studies are unbalanced. In order to compare the power of studies with different sample sizes and case/control ratios, the "effective sample size" (\(N_e\)) indicates the sample size of the balanced study with equivalent power. This statistic allows comparisons by putting all studies with different characteristics on the same scale. Moreover, intuition about effective sample size can guide study design to allocate resources to maximize power.

Designing gene expression studies to detect differential expression raises a corollary challenge: how to compare the relative power of studies with different numbers of total experiments and biological replicates, assuming all other factors are equivalent. Biological replicates from the same donor will be correlated because they measure the same underlying biological process. So a comparison of relative power must consider this degree of correlation.

Here we formalize the concept of effective sample size for studies with correlated samples by computing the sample size of a study of independent samples with equivalent power (Blainey et al., 2014). We start by assuming that all experiments have equal cost (in terms of labor, sequencing, etc.) and relax this assumption below.

Consider a study of \(k\) donors with \(m\) biological replicates per donor where \(\rho\) indicates the correlation between multiple experiments from the sample donor. This corresponds to \(mk\) total experiments. Following standard statistical theory of repeated measures study design (Diggle et al., 2002; Faes et al., 2009; Liu and Liang, 1997) the effective sample size is

\[
N_e = \frac{mk}{1 + \rho(m - 1)}
\]

Examination of this formula indicates key insights: 1) With 1 biological replicate per donor (\(m=1\)), the effective sample size equals the number of donors. 2) The increase in effective sample size obtained by increasing the number of biological replicates (i.e. \(m\)) is mediated by the correlation between biological replicates from the same donor (i.e. \(\rho\)).

Consider the contribution of each experiment to the power of the study as measured by effective sample size. Letting \(N_{total} = mk\) be the total number of
experiments, and $V$ be the contribution of each experiment to the effective sample size, then

$$V = \frac{N_e}{N_{total}}$$

$$= \frac{1}{1 + \rho(m - 1)}$$

Examination of this formula indicates two key insights: 1) $V$ represents the incremental impact of each successive experiment and is bounded between 0 and 1. 2) The incremental impact is highest when $\rho$ and $m$ are small. The latter point indicates that adding a biological replicate has a larger impact to increase power when there are few replicates or when the correlation between experiments from the same donor are small. When there are already, say, $m = 5$ replicates or $\rho$ is large then the contribution is minimal.

Computing effective sample size when costs are variable

In practice, there are substantial overhead costs for each donor in terms of recruitment, biopsy and hiPSC reprogramming. This overhead makes subsequent experiments from the same donor less expensive than the first experiment. When the total number of experiments in the study is fixed, then

$k = \frac{N_{total}}{m}$

is the number of biological replicates per donor. Consider that the cost per experiment varies so that the first experiment from a new donor costs $C_1$ units and all subsequent biological replicates cost $C_2$ units with $C_2 \leq C_1$. It follows that the first experiment costs $C_1$ units and the sum of all subsequent experiments from the same donor is

$C_2(m - 1)$

and the total cost per donor is

$C_1 + C_2(m - 1)$.

If the total cost of the study is fixed at $C$, it follows that the number of donors that can be afforded is

$k = \frac{C}{C_1 + C_2(m - 1)}$

A decreased cost of adding biological replicates changes computation of the effective sample size and pushes the calculation to favor increasing biological replicates when the total cost is fixed.

The companion website http://gabrielhoffman.shinyapps.io/design_ips_study/ creates interactive plots showing the effective sample size or the incremental
impact of each experiment when either the total cost or number of donors is fixed.


