Mamm Genome (2010) 21:467–476
DOI 10.1007/s00335-010-9279-1

Optimising experimental design for high-throughput phenotyping in mice: a case study

Natasha A. Karp · Lauren A. Baker · Anna-Karin B. Gerdin · Niels C. Adams · Ramiro Ramírez-Solis · Jacqueline K. White

Received: 8 June 2010 / Accepted: 26 July 2010 / Published online: 27 August 2010

Abstract To further the functional annotation of the mammalian genome, the Sanger Mouse Genetics Programme aims to generate and characterise knockout mice in a high-throughput manner. Annually, approximately 200 lines of knockout mice will be characterised using a standardised battery of phenotyping tests covering key disease indications ranging from obesity to sensory acuity. From these findings secondary centres will select putative mutants of interest for more in-depth, confirmatory experiments. Optimising experimental design and data analysis is essential to maximise output using the resources with greatest efficiency, thereby attaining our biological objective of understanding the role of genes in normal development and disease. This study uses the example of the noninvasive blood pressure test to demonstrate how statistical investigation is important for generating meaningful, reliable results and assessing the design for the defined research objectives. The analysis adjusts for the multiple-testing problem by applying the false discovery rate, which controls the number of false calls within those highlighted as significant. A variance analysis finds that the variation between mice dominates this assay. These variance measures were used to examine the interplay between days, readings, and number of mice on power, the ability to detect change. If an experiment is underpowered, we cannot conclude whether failure to detect a biological difference arises from low power or lack of a distinct phenotype, hence the mice are subjected to testing without gain. Consequently, in confirmatory studies, a power analysis along with the 3Rs can provide justification to increase the number of mice used.

Introduction

The mouse is the model organism of choice for studying the role of genes in normal development and disease, not least because advances in genetic engineering have made the genome highly tractable (Oliver et al. 2007; Zambrowicz and Sands 2003). There is a coordinated, international effort to produce (International Mouse Knockout Consortium 2007; Pettitt et al. 2009) and phenotype (Brown et al. 2006) knockouts for all mouse genes and release the resulting resource to the scientific community. Multiple phenotyping centres are performing high-throughput, systematic, primary phenotypic screening of mutant mouse strains to identify and highlight potential phenotypes of interest associated with mutant strains (Brown et al. 2006; Justice 2008). Hypotheses can be developed from these data to explain the role of the gene under investigation. The complementary role of in-depth, follow-up phenotyping aims to confirm and extend the primary observations into

Electronic supplementary material The online version of this article (doi:10.1007/s00335-010-9279-1) contains supplementary material, which is available to authorized users.

N. A. Karp · L. A. Baker · A.-K. B. Gerdin · N. C. Adams · R. Ramírez-Solis · J. K. White (✉)
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
e-mail: jkw@sanger.ac.uk

Present Address:
L. A. Baker
Division of Cardiovascular Medicine, Level 6, Addenbrooke's Centre for Clinical Investigation (ACCI), Addenbrooke's Hospital, University of Cambridge, Hills Road, Box 110, Cambridge CB2 0QQ, UK

N. C. Adams
MRC Harwell, Harwell Science and Innovation Campus, Oxfordshire OX11 0RD, UK

specialised fields of research. The pressures and objectives associated with primary and secondary phenotyping are therefore distinct. Within the primary centres there is a need for high throughput, optimising the use of resources whilst maximising sensitivity. In contrast, secondary centres need to ensure sufficient sensitivity is obtained to interrogate confidently the phenotype of interest. Different statistical power is required to fulfil these distinct objectives.

Typically, differences are highlighted using univariate statistical methods such as Student's *t* test (Crawley 2005). These tests calculate the probability (p) that the populations under comparison have the same mean and any difference arises from sampling variation. A change is deemed significant if the calculated p value falls below a prescribed level, typically 0.05 (the "nominal significance level"). Two types of error are possible: type I (α) and type II (β). Type I represents a false-positive error which occurs when a difference is declared to be significant erroneously. Type II represents a false-negative error which occurs when the test fails to detect a true biological difference. Power ($1 - \beta$) is the ability of a test to detect change and it depends on the variance (noise), effect size (magnitude of the treatment effect), number of replicates, and nominal significance that the researcher sets. To increase the power for a given technique, the researcher has most control over the number of replicates, but increasing the number of replicates beyond a certain point has little impact on the power. An undersized study will not have the capacity to detect some changes as statistically significant, whilst an oversized study will use more resources than necessary. Typically, for primary, screening experiments, a target power of 0.8 is used to ensure that the majority of times a biological difference is detected, whilst for secondary, confirmation experiments, a power of 0.95 ensures that a difference is not missed if it exists (Cohen 1988). Previously in the literature, power analyses in animal experiments have focused on only simple experimental designs (Festing 2003; Meyer et al. 2007), and yet a review of toxicological experiments involving animals suggests that a third of the experiments might be unnecessarily large (Festing 1996).

The high-throughput nature of primary phenotyping introduces the statistical problem of multiple testing, where false positives accumulate. For example, at the 0.05 confidence level, 5% of sample differences will be statistically significant even though no biological difference exists. It has been argued that in the context of exploratory experiments, where confirmatory investigations are performed, allowing a low frequency of false leads would not present a serious problem if the majority of significant findings were correctly chosen (Cui and Churchill 2003; Draghici 2002; Qian and Huang 2005). This has led to the development of methodologies to control the false discovery rate (FDR), where the focus is on achieving an acceptable ratio of true and false positives. These methods maintain sensitivity whilst addressing the multiple-testing problem.

We have performed a case study to optimise experimental design and ensure robust analysis using results from noninvasive blood pressure (NIBP) testing. This method uses an inflatable cuff secured at the base of the tail to measure heart rate and blood pressure. Indirect methods, such as this tail-cuff approach, have been recommended for experimental studies with animals (Kurtz et al. 2005). The protocol is automated and has been shown to give reproducible results with conscious mice (Feng et al. 2008; Krege et al. 1995) and to detect phenotypic differences in models of cardiovascular disease (Cho et al. 2008; Roncon-Albuquerque et al. 2008).

Raw NIBP data have a nested structure as measurements are frequently taken multiple times within a day and can also be taken across multiple days (Fig. 1) (Cho et al. 2008; Feng et al. 2008; Krege et al. 1995; Roncon-Albuquerque et al. 2008; Whitesall et al. 2004). These repeated measures are not independent and this is an example of temporal pseudoreplication where multiple time series measurements are made that are of short duration such that no time-dependent effects occur. Unless the statistical approach considers this nesting, the correlation will lead to an inflated estimate of statistical significance and thus type I errors (Hulbert 1984). Current users have addressed this by using a standard Student's *t* test on the mean of means (Cho et al. 2008; Krege et al. 1995; Roncon-Albuquerque et al. 2008; Whitesall et al. 2004). An alternative approach is to use a nested ANOVA, which models the covariance structure introduced by the grouping of the data. With this nested design, the effects of day and mouse are classed as random effects because they influence only the variance, not the overall mean, of the dependent variable. A major benefit of the nested design is that it economises on the number of degrees of freedom used by the factor levels and thus maintains power (Crawley 2005).

To ensure that robust results were obtained, the assumptions underlying the three-level nested ANOVA were investigated. In addition, the variance sources within the assay were considered. With the findings, a statistical power analysis assessing the sensitivity of the NIBP protocol was performed. The power analysis output directed the experimental design to optimise throughput, minimising animal testing whilst retaining sufficient power to detect phenotypes of interest. Finally, the nested ANOVA was applied to 46 mutant–control data sets for the three parameters monitored. This comparison allowed us to assess how the FDR can be used to address the risk of false positives arising from the multiple-testing problem. There are publications on various aspects of how experimental design and statistical analysis are essential for experiments

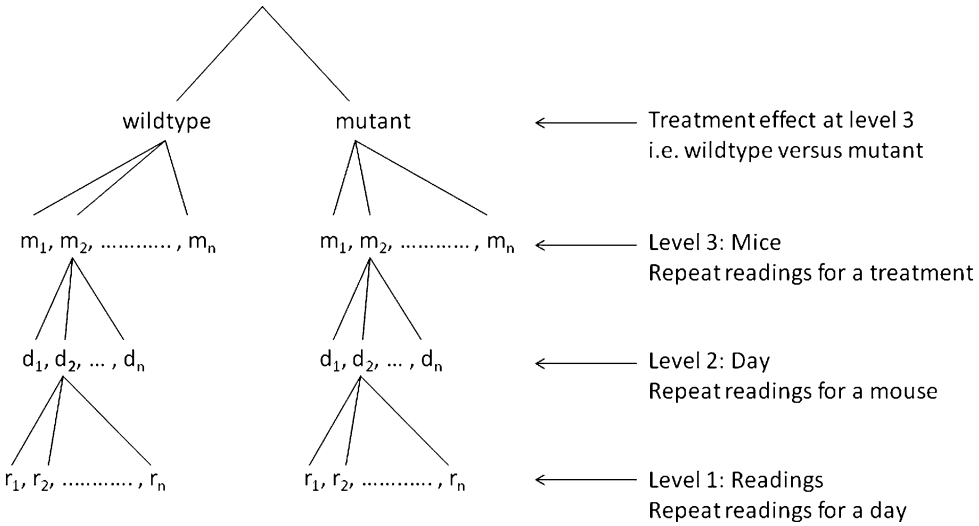

Fig. 1 The NIBP experiment can have a three-level random-effect nested design, where readings from days are nested within mice and the replicate readings are nested within day. The *n* represents the number of readings taken in-house at each level. Mice (m) are the level 3 unit, days (d) are the level 2 unit, and readings (r) are the level 1 unit. The mutation contrast is defined at level 3

in animal research (Festing 1994, 1996, 1997, 2003; Gaines Das 2002; Kilkenny et al. 2009). This report provides a comprehensive case study applied in an animal research setting and pulls together all the individual components to demonstrate how statistical techniques can help optimise experimental design and ensure that findings are robust.

Materials and methods

Mice

The care and use of all mice in this study were in accordance with UK Home Office regulations, UK Animals (Scientific Procedures) Act of 1986. Mice were created from either the NIH-funded Knock Out Mouse Project (KOMP) (Collins and Consortium 2007; Pettitt et al. 2009) or the EU-funded European Conditional Mouse Mutagenesis program (EUCOMM) (Firebaugh and Gibbs 1985) targeted ES cells by blastocyst injection (Pettitt et al. 2009) or were generously donated to the Mouse Genetics Programme by Wellcome Trust Sanger Institute Faculty groups. Details of all lines of mice used in this study are provided in Supplementary Data 1.

Mice were maintained in a specific pathogen-free unit on a 12-h light:12-h dark cycle with lights off at 7:30 p.m. and no twilight period. The ambient temperature was 21 ± 2°C and the humidity was 60 ± 10%. Mice were housed at a stocking density of 3–5 mice per cage [overall dimensions of caging (L × W × H): 365 × 207 × 140 mm, floor area = 530 cm^2] in individually ventilated cages (Tecniplast Seal Safe1284L) receiving 60 air changes per hour. In addition to Aspen bedding substrate, standard environmental enrichment of two nestlets, a cardboard Fun Tunnel, and three wooden chew blocks were provided. Mice were given water and diet ad libitum. At 4 weeks of age, mice were transferred from Mouse Breeders Diet (Lab Diets 5021-3) to a high-fat (21.4% fat by crude content) dietary challenge (Special Diet Services Western RD-829100).

Blood pressure and heart rate analyses

The noninvasive blood pressure assay was performed on approximately 11-week-old conscious mice using the automated tail-cuff MC4000 Blood Pressure Analysis System (Hatteras Instruments, Inc., Cary, NC, USA). Equipment was calibrated weekly and pressure tests performed daily following the manufacturer's recommendation. To facilitate acclimatisation, thereby reducing stress effects, mice were transferred to the measurement room at least 1 day prior to the start of the procedure and remained there for the entire data collection cycle. To address circadian variation, readings were collected between 08:30 and 12:30. To avoid introduction of bias, the experimenters were blinded to the genotype during the procedure. Furthermore, cages were processed randomly, and different genotypes could be housed together, hence there was no pattern to the order in which animals were processed. The procedure was spread over 5 days, including one training (thus these data were discarded) and four measurement days, each consisting of 5 acclimatisation and 15 measurement cycles. Within one measurement cycle, 70 consecutive waveforms were collected to provide a heart rate measure, and then the tail cuff was inflated occluding blood flow to the tail. Systolic and diastolic blood pressures were recorded as the pressure required to decrease the intensity of the original waveform by 20 and 50%, respectively. For each mutant–control comparison, we aimed for ten mice in each study group; however, the number of mice did vary between 5 and 23 for operational reasons. For example, some lines were subviable so a full set of mice could not be

breed, or welfare-related issues arose during the NIBP procedure resulting in the termination of the experiment.

Data analysis

Data analysis was completed using the freeware statistical program R (Ihaka and Gentleman 1996; http://www.r-project.org/) unless otherwise stated. To estimate statistical power, the freeware Optimal Design program was used (Raudenbush 1997; http://www.wtgrantfdn.org/resources/overview/research_tools). To estimate q values, the freeware Q-value program, which generates a graphical user interface with R, was used (Storey 2002; http://genomics.princeton.edu/storeylab/qvalue/).

Results

NIBP data capture

The data used for the following analysis were collected from a total of 1086 mice as an integral part of the high-throughput primary phenotyping programme ongoing at the Sanger Institute. Each mouse was assessed using a standardized battery of phenotypic tests, including NIBP which was performed over a 5-day period around 11 weeks of age, as described above. Prior to NIBP, the mice were exposed to a high-fat dietary challenge and weekly body weight measurement (week 4 onwards), and had a simple dysmorphology screen performed on them (week 9). Data from a total of 23 unique alleles were split by gender creating 46 unique mutant–control combinations (data from mice heterozygous and homozygous for the targeted allele were available for a subset of colonies) for analysis (Supplementary data 1). The following analysis was completed on all the raw data and hence omits the user-review stage (visual inspection of waveforms to ensure typical structure is obtained), which is frequently used to try to improve the data quality.

Selecting the appropriate statistical test

The selection of the statistical test depends on the research objectives, the experimental design, and the data properties. The importance of considering these was highlighted in 2009 when it was found that 60% of animal-based research articles reviewed had issues with the transparency and robustness of the statistical analysis (Kilkenny et al. 2009). The NIBP procedure is designed in such a way that it gives data with a hierarchical structure. A three-level random-effect nested ANOVA is a superior method of statistical analysis for hierarchical data than the alternative Student's t test on the mean of means, both of which are tests to identify differences in the variable mean (Crawley 2005). For knockout lines, where we have data for both genders, we could use a two-way version of the above techniques where the data are considered simultaneously. This approach has the advantage of assessing whether a statistically significant interaction is occurring where the effect of the genotype is not the same for the two genders. When the interaction is not significant, then a two-way ANOVA would be more sensitive in detecting change than an analysis that considers the genders independently. Interpreting a two-way ANOVA can be tricky and is more involved. In this article we have focused on a nested ANOVA that considers genders independently as the findings on this will be equally applicable to the two-way nested ANOVA.

Like all statistical tests, a nested ANOVA has a number of assumptions regarding the data under analysis. If these assumptions are not met, the test becomes unreliable. A nested ANOVA assumes that the observations within each subgroup are (1) normally distributed, (2) have equal variance, and (3) are independent. These assumptions were tested on the raw data as described below.

Normality was assessed for 38 randomly selected groups of readings, where a group comprises those measured from a mouse for a day, with the Anderson–Darling test for normality and a Q–Q normality plot. The data were found to be unimodal. For systolic and diastolic blood pressure data, the majority of groups of readings passed the test of normality (76 and 75%, respectively), whilst for heart rate data only 47% passed. The majority of failure arose from outliers that were included in the data due to omission of the user-review step which is frequently used to try and improve the data quality.

To assess the equal variance assumption, residual diagnostic plots were examined after fitting a linear model equivalent to a three-way nested ANOVA to mutant–control data sets (Supplementary data 2). Here, the residues (the difference between the actual value and the value predicted by the model) were plotted as a function of the independent variable to assess whether any systematic behaviour was present. The residues were found to be random in their distribution and not dependent on the signal strength of the independent variable. The presence of outliers could be seen; however, these spanned the entire signal strength range. Thus, the assumption of equal variance was met.

In assuming independence, we are, in effect, assuming that across the time span of the measurements no significant time-dependent effect exists. To assess for a day effect, wild-type C57BL/6NTac (Taconic Denmark) mice were examined both individually and as a gender group. Individually for each mouse, the mean readings were plotted against day and no pattern with time was found

(Supplementary data 3). For each gender, the mean of 20 wild-type mice was plotted with time (Supplementary data 4) and the readings between days compared with a two-tailed paired Student's *t* test. For both genders, no visual pattern with time was apparent and no statistically significant effect was seen for the three parameters across the 4 days of measurement.

As an ANOVA approach is considered robust and can tolerate departures from the normality assumption, our findings support the use of a three-level nested ANOVA as an appropriate tool for studying NIBP data.

Assessing the variance at each level

The variance at each level of the data (number of mice, days, and readings) was estimated by examining 46 mutant–control comparison data sets (Table 1; Fig. 2; Supplementary data 5). No statistically significant difference was seen between the estimated variances for the two genders when assessed with an independent or paired Student's *t* test (data not shown). For blood pressure measurements, on average 69% of the variation lay between mice, 30% of the variation lay between days, and 1% lay between readings for a given mouse on a given day. For heart rate, a higher proportion of the variance lay between mice (84%), suggesting that either the variation between days was much lower or that the variation between mice was higher. These data sets were prepared without user filtering, which is a common procedure (Whitesall et al. 2004). Omitting this stage did increase the number of outliers seen during the testing of normality; however, even with these outliers, only 1–3% of the variation arose from the readings taken within a day.

Power analysis for optimisation of experimental design

In a nested design, the variance and number of readings at each level influence the statistical power, with the factors at higher levels having more influence (Raudenbush 1997; Raudenbush and Xiao-Feng 2001). Power was calculated using the three-level-model module of the freeware program Optimal Design (Raudenbush 1997; Raudenbush and Xiao-Feng 2001) for a 0.95 confidence. To allow comparison across the different parameters, which have different units, a standardised effect-size measure, Cohen's *d*, was used (Cohen 1988). Here the effect size of interest is standardised for the variability in the data, hence a *d* of 1 means that the difference in the mean is equivalent to 1 standard deviation unit. These values can also be related to percentage overlap between distributions where a Cohen's *d* of 0.8 is equivalent to 50% overlap and as *d* increases the overlap decreases (Cohen 1988). For an effect size of

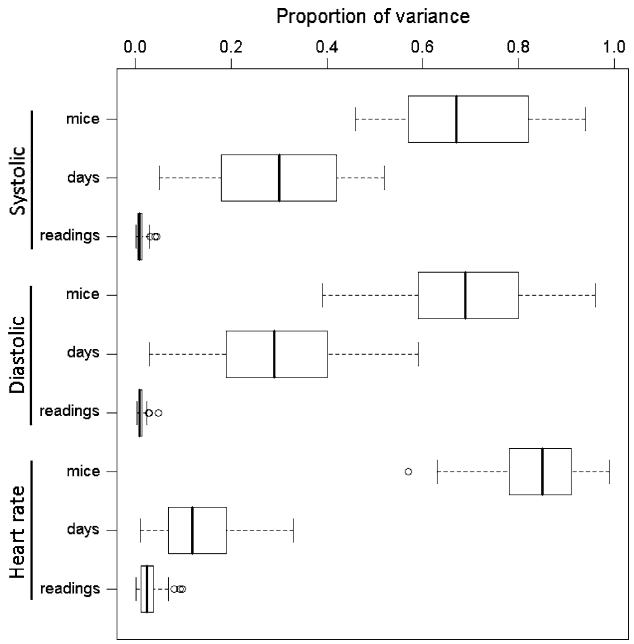

Fig. 2 Boxplot comparison showing the distribution of the variance between mice, days, and readings for each of the three parameters monitored in the NIBP procedure from the 46 mutant–control data set comparisons. The whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range. Points beyond this are classed as outliers and are shown as individual circles

Table 1 Variability at each experimental level

Parameter	Proportion of variance					
	Level 3—between mice		Level 2—between days		Level 1—between readings	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Diastolic BP	0.69	0.13	0.30	0.13	0.01	0.01
Systolic BP	0.69	0.14	0.30	0.14	0.01	0.01
Heart rate	0.84	0.09	0.13	0.07	0.03	0.02

BP blood pressure

The variability between clusters was estimated as a proportion of the variance at each level for 46 mutant–control comparison data sets for each of the parameters studied in the NIBP procedure. The mean was calculated across all 46 comparisons and thus pools data from both genders and from different genetic backgrounds

interest, the statistical power can be calculated for various designs to investigate the influence of altering the number of mice, days, or readings. Our typical experiment would result in a data set from 20 mice (ten per group) with readings from four consecutive days and 15 readings per day. Consequently, the effect of the number of readings, number of days, and number of mice on power was analysed by altering each variable one at a time whilst keeping all other aspects of the design constant (Fig. 3).

For all effect sizes studied, the power analysis found that once one reading per day was obtained no further increase in power was realised with additional readings (Fig. 3a, b). This is unsurprising and arises because the variation between readings is so low, and as a level 3 factor, it is least influential on the power obtained. Increasing the number of days gave rise to a more typical power curve, where after an initial increase, further increases in the number of days results in diminishing returns (Fig. 3c, d). Changes in the number of mice was the most influential parameter, which arose from variation in mice dominating but also because number of mice is a level 1 factor (Fig. 3e, f). With the design of four days of readings for ten mice per genotype (Fig. 3a, b), a Cohen's *d* of 0.91, where the distributions overlap by 48%, will reproducibly be detected for the blood pressure measurements (power = 0.80), whilst a Cohen's *d* of 0.77 (54% overlap) will reproducibly

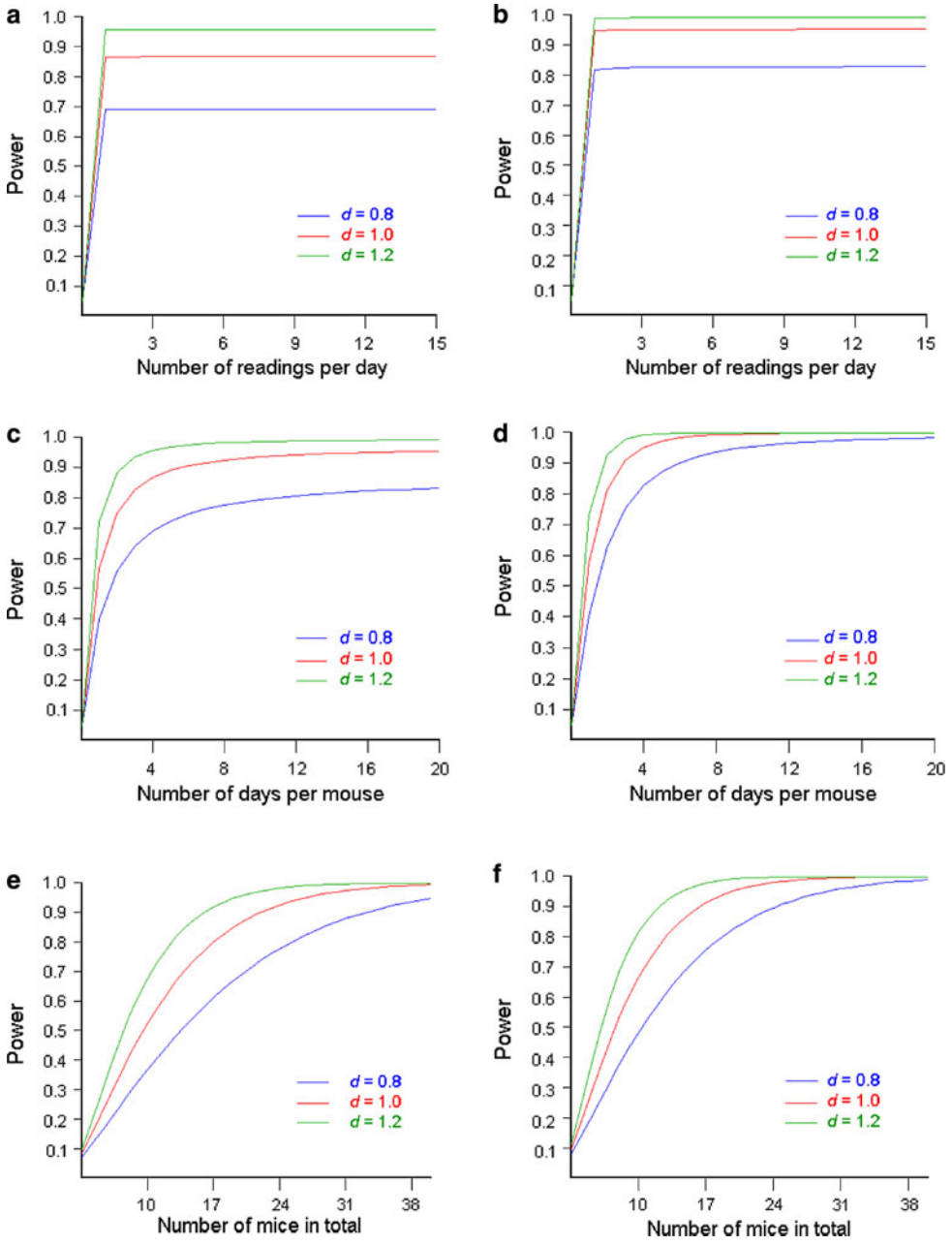

Fig. 3 Power curves to examine the effect on blood pressure (**a**, **c**, **e**) and heart rate (**b**, **d**, **f**) of changing one aspect of the experimental design at a time [per day (**a**, **b**), number of days (**c**, **d**), and number of mice (**e**, **f**)] whilst keeping all other aspects of the experimental design constant. When they are kept constant they default to 10 mice per group, 4 days of readings, and 15 readings per day. The power was calculated for three Cohen's *d* effect sizes (0.8, 1.0, and 1.2)

be detected for the heart rate parameter (power = 0.80). In comparison, we can detect only larger effect sizes [Cohen's d of 1.20 (38% overlap) in the blood pressure parameters and Cohen's d of 1.00 (45% overlap)] for heart rate with this experimental design if the target power is set at 0.95, as required in confirmatory experiments.

Mutant–control comparisons: identifying statistically significant change whilst addressing the multiple-testing problem

In this study, 46 mutant and control data set comparisons were completed for the three parameters monitored (Supplementary data 6). Twenty-eight significant findings were identified from the resulting 138 statistical tests performed. However, with a large number of statistical tests such as this, false positives (type I errors) can accumulate such that if no biological differences were present, then seven false positives would be expected if $p = 0.05$ significance threshold was used. Storey's q-value method addresses the multiple-testing problem by allowing control of the FDR, which is the proportion of false calls of those classified as significant (Storey 2002) (Table 2). Storey's method found that with a p value threshold of 0.05, the FDR is estimated at 22%. This means that of the 28 statistically significant findings, six are estimated to be false discoveries. Alternatively, a p value threshold of 0.025 leads to an estimate of three false calls in 21 statistically significant finds, whilst a p value threshold of 0.005 identifies seven as statistically significant with no false positives predicted. These results demonstrate how allowing a low number of false calls increases the sensitivity. For each statistical test completed in this study, the p and q values are reported and the relationship between p value and q value is shown (Supplementary data 7).

Table 2 Estimated number of false discoveries for various p value thresholds

p Value threshold	q Value threshold	Number tests classified as statistically significant	Estimated number of false positives
0.05	0.219	28	6
0.025	0.147	21	3
0.01	0.078	13	1
0.005	0.062	7	0

The p value is a measure of significance in terms of the false-positive rate and focuses on the test in isolation. The q value is a measure in terms of the false discovery rate across all the statistical tests within one experimental family

Assessing biological significance by calculating effect-size measures

For the 28 statistically significant findings ($p < 0.05$), the biological significance of each difference was assessed by calculating the associated effect size (Table 3). The proportion of the total variance that is attributed to the genotype difference was calculated (η^2) and is equivalent to the coefficient of determination (r^2). It was related to Cohen's d using Eq. 1 (Rosnow and Rosenthal 1996). This allows the findings to be related to the Cohen's d effect-size measure used in the power calculations. For reference, the absolute difference in mean was calculated for these putative biologically significant changes in blood pressure and heart rate.

$$\left[d = 2r/\sqrt{(1 - r^2)}\right] \quad (1)$$

Discussion

This study provides an example of how statistical investigation is essential to ensure that the experiments deliver meaningful results. This research finds that the three-level nested ANOVA is a statistically appropriate method to apply to NIBP data when multiple readings are collected for a mouse, as the assumptions are met. Use of appropriate statistical tools is essential to ensure correct leads are identified for future studies, with the false-positive rate controlled to the level selected by the researcher. This study also demonstrates how optimisation of the experimental design is essential to achieve the research objectives in question but also to reduce work (and therefore cost) and enhance welfare (refine). As these are significant issues in animal research, it is critical therefore to complete these analyses before embarking on experiments, particularly in a high-throughput scenario.

To optimise the design of the experiment, the variation sources in the data were investigated and used in a power analysis. For heart rate measurements in NIBP, the variation between mice dominated such that on average 83% of the variation arose from variation between mice, 13% between days, and 3% between readings for a mouse from a given day. For blood pressure measurements, 69% of the variation arose from variation between mice, 30% between days, and only 1% between readings for a mouse from a given day. These results arose from data sets where no user review occurred. This suggests that there is little value to the user review process, omission of which saves a considerable amount of time. With so little variation arising between readings, a power analysis confirms that once one reading is obtained there is little benefit from additional readings. The number of days was influential in the

Table 3 Effect-size measures for the mutant–control comparisons that were identified as statistically significant ($p < 0.05$) when assessed with a three-level random-effect nested ANOVA

Allele	Genotype comparison	Gender	Parameter	p Value	q Value	η^2	Cohen's d	Difference in mean (control–mutant) [HR (bpm) or BP (mmHg)]
Tpm1^{tm1aWtsi}	HETvWT	F	Heart rate	0.0366	0.1881	0.211	1.03	33.1
Mta1^{tm1aWtsi}	HETvWT	F	Heart rate	0.0021	0.0432	0.267	1.21	41.7
	HETvWT	M	Heart rate	0.0328	0.1759	0.148	0.83	24.6
	HETvWT	F	Diastolic BP	0.0178	0.1292	0.168	0.90	10.9
	HOMvWT	M	Heart rate	0.0009	0.0278	0.322	1.38	51.6
	HOMvWT	F	Heart rate	0.0035	0.0617	0.304	1.32	56.4
Akt2^{tm1Wcs}	HOMvWT	M	Heart rate	0.025	0.1468	0.133	0.78	−5.1
	HOMvWT	F	Diastolic BP	0.0193	0.1323	0.065	0.53	8.0
	HOMvWT	F	Systolic BP	0.0223	0.1448	0.060	0.51	8.13
Herc3$^{tm1a(EUCOMM)Wtsi}$	HOMvWT	M	Heart rate	0.025	0.1468	0.103	0.68	−39.5
	HOMvWT	F	Heart rate	0.04	0.1955	0.097	0.65	−32.2
Epc1^{tm1aWtsi}	HETvWT	M	Systolic BP	0.0114	0.0946	0.132	0.78	−21.1
	HETvWT	M	Diastolic BP	0.0083	0.0788	0.142	0.81	−20.1
	HETvWT	M	Heart rate	0.0051	0.0650	0.180	0.94	−46.5
	HETvWT	F	Heart rate	0.0052	0.0650	0.164	0.89	−35.1
Baz1b$^{tm1a(KOMP)Wtsi}$	HETvWT	M	Heart rate	0.0412	0.1955	0.040	0.41	−25.4
	HETvWT	F	Heart rate	0.0235	0.1449	0.062	0.51	−28.6
Mysm1$^{tm1a(KOMP)Wtsi}$	HETvWT	M	Systolic BP	0.0064	0.0658	0.111	0.71	−10.9
	HETvWT	M	Diastolic BP	0.0054	0.0650	0.107	0.69	−10.5
Tmc1^{dn}	HOMvHET	F	Heart rate	0.0007	0.0278	0.055	0.48	−40.1
Cadm1$^{tm1.2Brd}$	HOMvWT	F	Heart rate	0.0003	0.0278	0.236	1.11	−49.1
Cdh23^{v}	HOMvHET	F	Systolic BP	0.0498	0.2194	0.045	0.43	−7.4
	HOMvHET	F	Heart rate	0.0013	0.0321	0.175	0.92	−127.2
Magi2^{tm1Grnt}	HETvWT	M	Systolic BP	0.0115	0.0946	0.089	0.62	−10.2
	HETvWT	M	Diastolic BP	0.0469	0.2143	0.048	0.45	−7.2
Mta3$^{tm1a(KOMP)Wtsi}$	HOMvWT	M	Diastolic BP	0.0281	0.1576	0.024	0.32	35.7
	HOMvWT	F	Heart rate	0.0127	0.0321	0.066	0.53	18.7
Brd7^{tm1aWtsi}	HETvWT	M	Heart rate	0.0058	0.0650	0.050	0.46	16.1

BP blood pressure, *HR* heart rate

Within the genotype comparison column, WT indicates that mice were wild type for the gene of interest. HET and HOM indicate that mice were heterozygous and homozygous, respectively, for the targeted allele

sensitivity obtained, but most significant was the number of mice used in the analysis. With this information, the optimal design, which balances the cost with the available resources and experimental objective, can be chosen. Specifically, the current design in our facility achieves the target power of 0.8 to detect large changes (80% of a SD unit), which we feel is an appropriate goal for primary-screening, hypothesis-generating research. Therefore, we do not need to alter the number of mice or days in the current experimental design. However, the number of readings per day (up to 15) is excessive, yielding little added value, and can confidently be reduced without loss of power. We settled on five readings per day to allow for missing values that can arise from mouse movement during the procedure. This is a refinement from a welfare perspective as it reduces the number of measurement cycles and, hence, the experimental duration.

For both practical and ethical reasons there is a drive to reduce the number of animals used in a study, as reiterated by the mantra of the three Rs (Burch and Russell 1959). If an experiment is underpowered, the findings are inconclusive and hence a power analysis, along with the three Rs, can be used to justify an increase in the number of mice. However, with an overpowered study, the additional readings are not necessary and the number of mice should be reduced.

Across the 46 mutant–control comparisons, a number of statistically significant findings could be identified depending on the significant threshold (p value) used. The lower the p value threshold used, the lower the risk of a false positive, which is a particular issue with a multiple-testing scenario. However, protecting against a false positive in this manner increases the risk of a false negative, where biologically significant differences are missed. To address the multiple-testing problem but maintain sensitivity, the FDR was estimated for various thresholds of significance. This data set demonstrates that allowing some false calls increases sensitivity whilst giving a measure of the associated risk.

The most robust hit was found for metastasis associated 1 (*Mta1*). Homozygous null mice of both genders displayed an increase in heart rate of approximately 50–60 bpm ($p < 0.05$). This increase was detected to a lesser degree (20–40 bpm) in heterozygous mice of both genders ($p < 0.05$), indicating a gene-dosage effect. *Mta1* is a broadly expressed gene [(Simpson et al. 2001); in-house observation from *lacZ* reporter gene study] known to be a component of the Mi-2/nucleosome remodeling and histone deacetylase (NuRD) complex and therefore plays a key role in regulation of gene expression. There are no prior publications linking *Mta1* with cardiac function, although an alternative transcript was detected in the heart (Simpson et al. 2001).

This case study demonstrates the value of using statistical analysis to direct experimental design, thus allowing an informed decision to ensure that the three Rs are being met. Additional statistical analysis with effect size and false discovery measures can ensure that the findings are robust and that future downstream work is efficient. This is essential for minimising the experiments whilst maximizing the potential benefit to scientific knowledge.

Acknowledgment We thank the staff of the Sanger Institute's Research Support Facility, Mouse Genetics Programme, and Mouse Informatics Group for their excellent support. This work was funded by the Wellcome Trust (grant no. WT077157/Z/05/Z) and from the EUMODIC project (funded by the European Commission under contract number LSHG-CT-2006-037188).

References

Brown SD, Hancock JM, Gates H (2006) Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. PLoS Genet 2:e118

Burch W, Russell R (1959) The principles of humane experimental technique. Methuen & Co, London

Cho H, Park C, Hwang IY, Han SB, Schimel D et al (2008) Rgs5 targeting leads to chronic low blood pressure and a lean body habitus. Mol Cell Biol 28:2590–2597

Cohen J (1988) Statistical power analysis for the behaviour sciences. Lawrence Earlbaum Associates, Mahwah, NJ

Collins F, Rossant J, Wurst W, International Mouse Knockout Consortium (2007) A mouse for all reasons. Cell 128:9-13

Crawley M (2005) Statistics: an introduction using R. Wiley, New York

Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 4:210

Draghici S (2002) Statistical intelligence: effective analysis of high-density microarray data. Drug Discov Today 7:S55–S63

Feng M, Whitesall S, Zhang Y, Beibel M, D'Alecy L et al (2008) Validation of volume-pressure recording tail-cuff blood pressure measurements. Am J Hypertens 21:1288–1291

Festing MF (1994) Reduction of animal use: experimental design and quality of experiments. Lab Anim 28:212–221

Festing MF (1996) Are animal experiments in toxicological research the 'right' size? In: Morgan BJT (ed) Statistics in toxicology. Clarendon Press, Oxford, pp 3–11

Festing MF (1997) Experimental design and husbandry. Exp Gerontol 32:39–47

Festing MF (2003) Principles: the need for better experimental design. Trends Pharmacol Sci 24:341–345

Firebaugh G, Gibbs J (1985) User's guide to ratio variables. Am Sociol Rev 50:713–722

Gaines Das RE (2002) Role of ancillary variables in the design, analysis, and interpretation of animal experiments. ILAR J 43:214–222

Hulbert S (1984) Pseudoreplication and the design of ecological field experiments. Ecol Mongr 54:187–211

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

Justice MJ (2008) Removing the cloak of invisibility: phenotyping the mouse. Dis Model Mech 1:109–112

Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC et al (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PloS ONE 4:e7824

Krege JH, Hodgin JB, Hagaman JR, Smithies O (1995) A noninvasive computerized tail-cuff system for measuring blood pressure in mice. Hypertension 25:1111–1115

Kurtz TW, Griffin KA, Bidani AK, Davisson RL, Hall JE (2005) Recommendations for blood pressure measurement in humans and experimental animals. Part 2. Blood pressure measurement in experimental animals: a statement for professionals from the subcommittee of professional and public education of the American Heart Association council on high blood pressure research. Hypertension 45:299–310

Meyer CW, Elvert R, Scherag A, Ehrhardt N, Gailus-Durner V et al (2007) Power matters in closing the phenotyping gap. Naturwissenschaften 94:401–406

Oliver PL, Bitoun E, Davies KE (2007) Comparative genetic analysis: the utility of mouse genetic systems for studying human monogenic disease. Mamm Genome 18:412–424

Pettitt SJ, Liang Q, Rairdan XY, Moran JL, Prosser HM et al (2009) Agouti C57BL/6N embryonic stem cells for mouse genetic resources. Nat Methods 6:493–495

Qian HR, Huang S (2005) Comparison of false discovery rate methods in identifying genes with differential expression. Genomics 86:495–503

Raudenbush SW (1997) Statistical analysis and optimal design for cluster randomized trials. Psychol Methods 2:173–185

Raudenbush SW, Xiao-Feng L (2001) Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. Psychol Methods 6:387–401

Roncon-Albuquerque R Jr, Moreira-Rodrigues M, Faria B, Ferreira AP, Cerqueira C et al (2008) Attenuation of the cardiovascular and metabolic complications of obesity in CD14 knockout mice. Life Sci 83:502–510

Rosnow R, Rosenthal R (1996) Computing contrasts, effect sizes, and counternulls on other people's published data: procedures for research consumers. Pyschol Methods 1:331–340

Simpson A, Uitto J, Rodeck U, Mahoney MG (2001) Differential expression and subcellular distribution of the mouse metastasis-associated proteins Mta1 and Mta3. Gene 273:29–39

Storey J (2002) A direct approach to false discovery rates. J R Stat Soc B 64:479–498

Whitesall SE, Hoff JB, Vollmer AP, D Alecy LG (2004) Comparison of simultaneous measurement of mouse systolic arterial blood pressure by radiotelemetry and tail-cuff methods. Am J Physiol 286:H2408–H2415

Zambrowicz BP, Sands AT (2003) Knockouts model the 100 best-selling drugs—will they model the next 100? Nat Rev 2:38–51