Soft Windowing Application to Improve Analysis of High-throughput Phenotyping Data

Hamed Haselimashhadi[1,*], Jeremy C. Mason[1], Violeta Munoz-Fuentes[1], Federico López-Gómez[1], Kolawole Babalola[1], Elif F. Acar[2], Vivek Kumar[3], Jacqui White[3], Ann M. Flenniken[4], Ruairidh King[5], Ewan Straiton[5], John Richard Seavitt[6], Angelina Gaspero[6], Arturo Garza[6], Audrey E. Christianson[6], Chih-Wei Hsu[6], Corey L. Reynolds[6], Denise G. Lanza[6], Isabel Lorenzo[6], Jennie R. Green[6], Juan J. Gallegos[6], Ritu Bohat[6], Rodney C. Samaco[6], Surabi Veeraragavan[6], Jong Kyoung Kim[7], Gregor Miller[8], Helmult Fuchs[8], Lillian Garrett[8], Lore Becker[8], Yeon Kyung Kang[9], David Clary[10], Soo Young Cho[11], Masaru Tamura[12], Nobuhiko Tanaka[12], Kyung Dong Soo[13], Alexandr Bezginov[2], Ghina Bou About[14], Marie-France Champy[14], Laurent Vasseur[14], Sophie Leblanc[14], Hamid Meziane[14], Mohammed Selloum[14], Patrick T. Reilly[14], Nadine Spielmann[8], Holger Maier[8], Valerie Gailus-Durner[8], Tania Sorg[14], Masuya Hiroshi[12], Obata Yuichi[12], Jason D. Heaney[6], Mary E.Dickinson[6], Wurst Wolfgang[15], Glauco P. Tocchini-Valentini[16], Kevin C. Kent Lloyd[10], Colin McKerlie[2], Je Kyung Seong[13], Herault Yann[17], Martin Hrabé de Angelis [8], Steve D.M. Brown[5], Damian Smedley[18], Paul Flicek[1], Ann-Marie Mallon[5], Helen Parkinson[1], Terrence F. Meehan[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [2]The Centre for Phenogenomics, Toronto, Canada; The Hospital for Sick Children, Toronto, Canada, [3]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, [4]The Centre for Phenogenomics, Toronto, Canada; Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada, [5]MRC Harwell Institute, Harwell, OX11 0RD, UK, [6]Baylor College of Medicine, Houston, TX, USA, [7]Daegu Gyeongbuk Institute of Science & Technology (DGIST), Korea, [8]Helmholtz Center Munich, Neuherberg, Germany, [9]Korea Mouse Phenotyping Center (KMPC), Korea, [10]Mouse Biology Program, University of California Davis, [11]National Cancer Center (NCC) & Korea Mouse Phenotyping Center (KMPC), Korea, [12]RIKEN BioResource Research Center, Tsukuba, Japan, [13]Seoul National University & Korea Mouse Phenotyping Center(KMPC), Korea, [14]Université de Strasbourg, CNRS, INSERM, Institut Clinique de la Souris, PHENOMIN-ICS 1 rue Laurent Fries, 67404 ILLKIRCH, [15]Institute of Developmental Genetics, Helmholtz Centre Munich, Germany, [16]CNR EMMA Monterotondo, Italy, [17]Université de Strasbourg, CNRS, INSERM, Institut de Génétique, Biologie Moléculaire et Cellulaire, Institut Clinique de la Souris, IGBMC, PHENOMIN-ICS 1 rue Laurent Fries, 67404 ILLKIRCH, [18]William Harvey Research Institute, Charterhouse Square Barts and the London School of Medicine and Dentistry Queen Mary University of London, London EC1M 6BQ

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: High-throughput phenomic projects generate complex data from small treatment and large control groups that increase the power of the analyses but introduce variation over time. A method is needed to utlize a set of temporally local controls that maximises analytic power while minimising noise from unspecified environmental factors.

Results: Here we introduce “soft windowing”, a methodological approach that selects a window of time that includes the most appropriate controls for analysis. Using phenotype data from the International Mouse Phenotyping Consortium (IMPC), adaptive windows were applied such that control data collected proximally to mutants were assigned the maximal weight, while data collected earlier or later had less weight. We applied this method to IMPC data and compared the results with those obtained from a standard non-windowed approach. Validation was performed using a resampling approach in which we demonstrate a 10% reduction of false positives from 2.5 million analyses. We applied the method to our production analysis pipeline that establishes genotype-phenotype associations by comparing mutant versus control data. We report an increase of 30% in significant p-values, as well as linkage to 106 versus 99 disease models via phenotype overlap with the soft-windowed and non-windowed approaches, respectively, from a set of $2{,}082$ mutant mouse lines. Our method is generalisable and can benefit large-scale human phenomic projects such as the UK Biobank and the All of Us resources.

Availability and implementation: The method is freely available in the R package SmoothWin, available on CRAN http://CRAN.R-project.org/package=SmoothWin.

Contact: hamedhm@ebi.ac.uk

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

1 Introduction

High-throughput, large scale phenotyping studies evaluate variables of an organism's biological systems to examine the contribution of genetic and environmental factors to phenotypes. Standardised phenotyping screens that cover a wide range of biological systems have made useful insights for identifying new genetic contributors to robust phenotypes as compared to more focused studies that often target well-characterised genes with varying reproducibility (Prinz *et al.*, 2011; Edwards *et al.*, 2011; Stoeger *et al.*, 2018; Begley and Ellis, 2012; Freedman *et al.*, 2015). Leveraging economies of scale and using standardised procedures, high-throughput phenotyping screens addresses these challenges and have been applied in biological screening of chemical compound libraries, agricultural evaluation of crop plants, genome-wide CRISPR-based mutagenic cell line screens and multi-centre phenotypic screening of mutated model organisms (Viti *et al.*, 2015; Al-Tamimi *et al.*, 2016; Flood *et al.*, 2016; Friggens *et al.*, 2011; Vitak *et al.*, 2017; Malinowska *et al.*, 2017; Sun *et al.*, 2017; Dickinson *et al.*, 2016). The continuous generation of large volumes of data introduces new challenges affecting automated approaches to statistical analysis that have to scale with increasing data and address the underlying complexity inherent in large projects (Meyers *et al.*, 2017; Vaas *et al.*, 2013, 2012; Kurbatova *et al.*, 2015).

The International Mouse Phenotyping Consortium (IMPC) is a G7 recognised global research infrastructure dedicated to generating and characterising a knockout mouse line for every protein-coding gene (Brown and Moore, 2012; Bradley *et al.*, 2012; Hrabě de Angelis *et al.*, 2015). Currently, the IMPC has phenotyped over 148,000 knockouts and 43,000 control mice (data release 9.2, January 2019) across 12 research centres in 9 countries. These centres adhere to a set of standardised phenotype assays defined in the International Mouse Phenotyping Resource of Standardised Screens (IMPReSS), and designed to measure over 200 parameters on each mouse. As part of these standardised operating procedures, critical factors that can impact data collection, such as reagent type or equipment, are reported as required metadata. Phenotype data is then centrally collected and quality controlled by trained professionals before being released for analysis. All phenotype data is processed by the statistical analysis package PhenStat—a freely-available R package that provides a variety of statistical methods for the identification of genotype to phenotype associations by comparing mutant to control data that have the same critical attributes (Kurbatova *et al.*, 2015). For quantitative data, linear mixed models are typically employed with several factors modelled in including genotype, sex, sex-genotype interaction, body weight, and batch (i.e., phenotype measures collected on the same day). Mutant mouse lines found to have a significant deviation in phenotype measurements are assigned a phenotype term from the Mammalian Phenotype Ontology (Blake *et al.*, 2017). These associations, as well as the raw data, are disseminated via the web portal (https://www.mousephenotype.org) using application programming interfaces (APIs) and data downloads.

A challenge with high-throughput phenotyping efforts is the small sample size for the experimental group (i.e., the knockout mice) that is produced to maximise the use of finite resources, considering biological relevance and power analysis (Charan and Kantharia, 2013). All mice generated by the IMPC are on the inbred C57BL/6N strain. To reduce genetic drift, IMPC centers maintain wild-type C57BL/6N production colonies that are periodically rederived using commercial vendors (Kurbatova *et al.*, 2015; Dickinson *et al.*, 2016). Mutant F0 mice are bred with wild-type mice from the production colonies to reduce the confounding effects of any de novo, non-targeted mutations. In addition, the IMPC centres are encouraged to measure these knockout mice in two or more batches, as this improves the false discovery rate by modelling in the random effect of day-to-day variation (Karp *et al.*, 2014). In contrast, large control sample sizes accumulate as they provide a strong internal control of the pipeline and typically generated with every experimental batch. Such large control groups represent a unique dataset that increase the power of the subsequent analyses and allow the construction of a robust baseline (Bradley *et al.*, 2012). However, this can lead to the accumulation of heterogeneities including seasonal effects, changes in personnel, and unknown time-dependent environmental factors (Karp *et al.*, 2014).

A simple approach to cope with heterogeneity in the data is to set explicit time boundaries (e.g., one year) before and after experimental collection dates. This "hard windowing" approach will capture different time-frames depending on how much time elapses between the first and last batch of experimental data measured. This approach is unsatisfactory for IMPC data as some mutant lines had enough experimental mice to measure in one batch, while others needed multiple batches over 18 months due to breeding difficulties or other factors. This variation in time-frames can lead to a widely different number of controls being applied to an analysis, making it challenging to explore correlations between mutant lines. Thus, more tunable approaches were needed.

In this study, we address the complexity of the data collected over time by proposing a novel windowing strategy that we call "soft windowing". This approach utilises a weighting function to assign flexible weights, ranging from 0 to 1, to the control data points. Controls that are collected on or near the date of mutants are assigned the maximal weights, whereas controls at earlier or later dates are assigned less weight. In contrast to the hard windowing, the weighting function in the soft windowing allows for different shapes and bandwidths by alternating the tuning parameters. In addition, we demonstrate how to tune parameters and demonstrate the implementation of the soft windowing on the IMPC data.

2 System and methods

In high-throughput projects, such as the IMPC, the model parameters may not stay constant over time that can lead to misleading inferences. For example, Figure 1 illustrates changes to the control group trend and/or variation over time for the *Forelimb grip strength normalised against body weight* and *Mean cell volume*. One approach widely used in signal processing (Kervrann, 2011; Poularikas, 2019; Ford, 2003; Machado *et al.*, 2008) is to define a windowing function that includes the appropriate number of data points to capture the effect of interest while minimising the noise. This is defined by

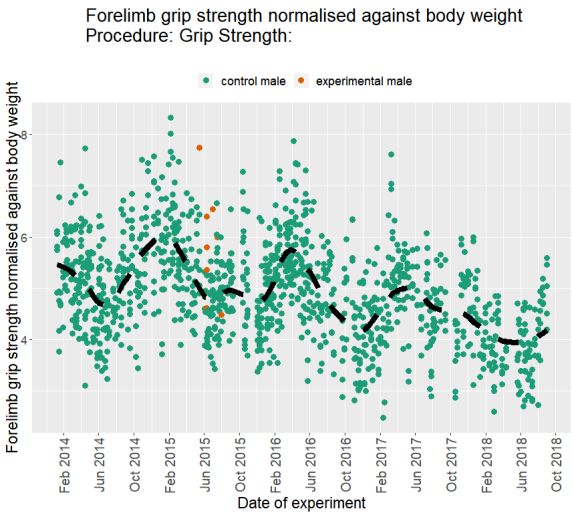

Fig 1: Examples of longitudinal data from the IMPC selected for high variance in control population. Scatter plot of the *Forelimb grip strength normalised against body weight* (top) and *Mean cell volume* (bottom) from the IMPC Grip Strength and Haematology procedures, respectively. The dashed black lines represent the overall trend of the controls (dark green). Mutant mice are in orange.

$$W(x,l_1,l_2) = \begin{cases} f(x) & l_1 \le x \le l_2 \\ 0 & o.w \end{cases}, \qquad (1)$$

where setting $f(x)$ to a constant, e.g., $f(x) = 1$, leads to hard windowing, while setting it to a smooth function results in the soft windowing. The same approach can be generalised to multiple signals (Tang *et al.*, 2009; Huang *et al.*, 2007; Li *et al.*, 2007) or applied as a rolling window (Jank and Shmueli, 2008) in the presence of exogenous variables to account for time dependency in the regression coefficients (Brown *et al.*, 2018). Alternatively, we propose a soft windowing approach for the regression methods by defining a weighting function that applies less weight to the residuals outside the window of interest. This leads to distinct advantages over the hard windowing. First, the entire dataset is included in the analysis in contrast to the limited data points in the hard windowing. Second, the windowing and the parameter estimation are coupled, which is a direct result of using the Weighted Least Squares (WLS). Critically, by bounding the controls in a window, we freeze the analysis and abrogate the need for further analysis assuming no new experimental data is generated within the time window.

3 Algorithm

Our novel windowing strategy explicitly defines the weighting function and proposes a simple but effective set of criteria to estimate the minimal window for the noise-power trade off.

3.1 Weight generating function

Let $t = (t_1,t_2,...,t_n)$ represent a set of n continuous time units, $m = (m_1,m_2,...,m_p)$ the time units when the treatments are measured (peaks in the windows), $l = \{(l_{1L},l_{1R}), (l_{2L},l_{2R}),..., (l_{pL},l_{pR})\}$ a set of p non-negative left and right *bandwidths* and $k = \{(k_{1L},k_{1R}),(k_{2L}, k_{2R}),...,(k_{pL},k_{pR})\}$ a set of p positive left and right *shape* parameters. We impose the continuity on the time to simplify the definition of a continuous function over the time units, e.g., by converting dates to UNIX timestamps. Furthermore, we introduce a peak generating function (PGF) of the form of $c_i = F(t;m_i - l_{iL},k_{iL})(1 - F(t;m_i + l_{iR},k_{iR}))$, $i = 1,2,...,p$ where $F(x;\mu,\sigma) = Pr_X(X \le x|\mu,\sigma)$ is selected from the family of cumulative distribution functions (cdf) with location μ and scale σ. In this study, we select F from the family of continuous and symmetric distributions (such as the Logistic, Gaussian, Cauchy and Laplace distributions). Then, we propose a weight generating function (WGF) of the form of

$$\begin{aligned} WGF(t,l,k,m) = &\sum_{i=1}^{p} c_i^* + \\ &\Bigg[\sum_{i \ne j \in \{1,2,...,p\}} \prod_{i,j} -c_i^* c_j^* + \\ &\sum_{i \ne j \ne h \in \{1,2,...,p\}} \prod_{i,j,h} c_i^* c_j^* c_h^* - \\ &\ldots + \\ &(-1)^{p+1} \sum c_1^* c_2^* \ldots c_p^* \Bigg], \\ t,l,m \in \mathbb{R}, k \in \mathbb{R}^+ & \end{aligned} \qquad (2)$$

where $c_i^* = \frac{c_i}{\max c_i}$ denotes the normalised peak generating function. The first term on the right hand side of Eq. 2 produces the individual windows and the second term accounts for merging the intersections amongst the windows. Figure 2 shows the symmetric weight generating function (SWGF), that is $l_{iR} = l_{iL}$ and $k_{iR} = k_{iL}$, $i = 1, 2,..., p$, for the different values of $k \in [0.2, 50]$ coloured from blue ($k = 50$) to red ($k = 0.2$) and for the different values of $l = 5, 10, 15$. The vertical black dashed lines show the hard window corresponding to the value of l. From this plot, the function is capable of generating a range of windows from hard (blue) to soft (red). Further, the weights lay in the $(0,1]$ interval for all values of time; however, they may not cover the entire $(0,1]$ spectrum in a bounded time domain. Then, the weights are normalised to be ranged in $(0,1]$ before inserting into the WGF as shown by c_i^* in Eq 2. Figure 3 shows the merge capability of the SWGF for the

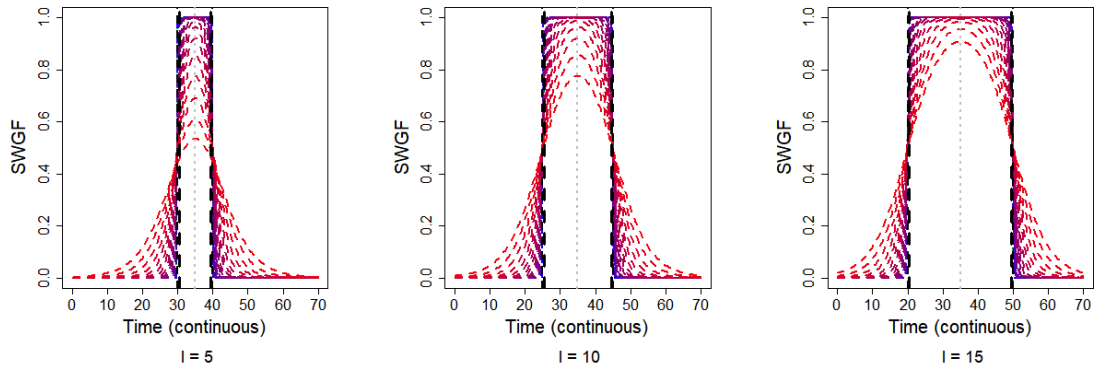

Fig 2: Behaviour of the Symmetric Weight Generating Function (SWGF) for a spectrum of values for the shape parameter, k, ranging from $k = 50$ (blue) to $k = 0.2$ (red), in intervals of $t = 1,2,...,70$, and for the different values of the bandwith $l = 5, 10, 15$ (left to right). The black dashed lines show the hard windows corresponding to l. The gray dotted vertical lines show the window peaks. These plots show the capability of the WGF to generate different forms of the window.

logistic F with $m = 15, 35$ and different values of $k = 0.5, 1.5, 3$ and $l = 6, 8, 10, 12$. From this figure, the function is capable of producing a range of flexible multimodal windows (top) as well as aggregated windows (bottom) if $|m_1 + l| > |m_2 - l|$ for all $m_1 < m_2, l \in \mathbb{R}$. In all cases, the weights lay in the $(0,1]$ interval.

3.2 Windowing regression

Let $y = x\beta + e$ denote a linear model, with y, x, β and e representing response, covariates, unknown parameters and independent random noise $e \sim N(0,\sigma^2 < \infty)$ respectively. Imposing the weights in Eq. 2 on the residuals leads to the following weighted least square (WLS)

$$Q(\beta) = \text{WGF}(t,l,k,m)||y - x\beta||_2^2 \quad (3)$$

where $||.||_2$ denotes the second norm of a vector. Minimising $Q(\beta)$ with respect to β leads to $\hat{\beta} = (x'wx)^{-1}x'wy$, where w is a diagonal matrix of weights from WGF and $(')$ denotes the transpose of a matrix. Weighted linear regression (WLR), in the context of this study, is equivalent to imposing less weight on the off modal time points with respect to m. We illustrate this in Figure 4, where 60 observations are simulated from the following model,

$$y_t = t\beta_1 I_{(t \le 20)} + t\beta_2 I_{(20 < t < 40)} + t\beta_3 I_{(t \ge 40)} + e,$$

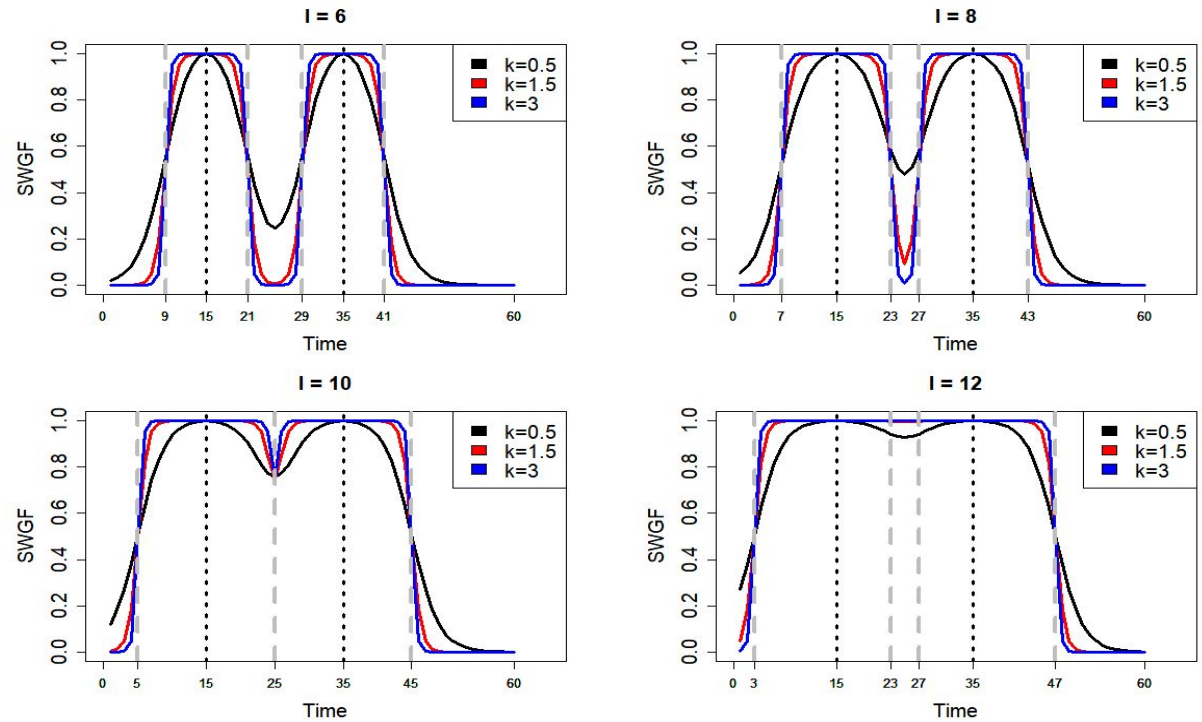

Fig 3: Merging behaviour of the SWGF for different values of the shape parameter $k = 0.5, 1.5, 3$ and the bandwidth $l = 6, 8, 10, 12$ on a sequence of time points $t = 1, 2, ..., 60$. The vertical dashed gray lines show the corresponding hard windows to l. This plot shows the capability of SWGF to generate multimodal windows as well as merging individual windows.

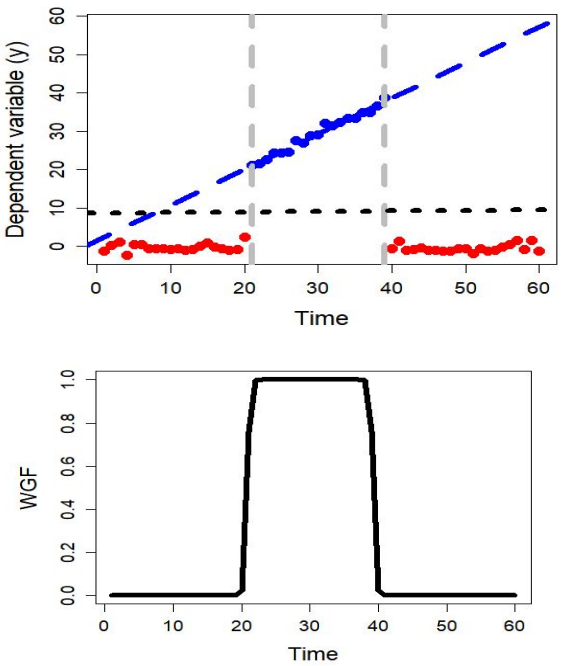

Fig 4: (Left) Comparison between the inferences from the windowed linear regression on the simulated data (blue dashed line) and without windowing (dotted black line). (Right) The corresponding weights from WGF centred on $m = 30$. With windowing, we attempt to model the effective section of the data (blue dots).

with $t = 1,2,...,60$, $\beta_1 = 0,\ \beta_2 = 1, \beta_3 = 0,\ e \overset{iid}{\sim} N(0,1)$ and I is the indicator function,

$$I(x \in [a,b]) = \begin{cases} 1 & x \in [a,b] \\ 0 & o.w \end{cases}.$$

In other words, the model is piecewise linear and only significant in the $t \in (20, 40)$ interval. Figure 4 (top) shows the global estimation of the linear regression from the entire data (dotted black line) and the WLR by $\mathrm{WGF}(t,9,5,30)$ (dashed blue line) as well as weights from the WGF on the bottom. This plot shows that the non-weighted linear regression leads to a horizontal line, where no significant gradient is detected, whereas the WLR tends to model the significant section of the data that leads to fitting the true line. Figure 4 compares the effect of windowing vs. considering the entire dataset, showing the different conclusions.

3.3 Selection of the tuning parameters

Selection of the tuning parameters k and l to define the soft window have a strong impact on the final estimations and consequently on the inferences that are made from the statistical results. Indeed, a wide or over-smooth window can lead to the inclusion of too much noise, whereas a small window can result in low power in the analysis. An additional challenge is the direct linear correlation between increasing the number of peaks, m, and to the total number of the parameters for the windows (l,k) that results in significant growth in the computational complexity of the final fitting. This is due to tuning the window in the general form of WLS in Eq. 3 requires $2p$ dimensions in space to search for the optimal l and k. To cope with this complexity, we propose to fix l and k so all windows are symmetric and have the same shape and bandwidth. We then select the tuning parameters by searching the space on the grid of (l,k) values and look for the most significant change in mean and/or variation of the residuals/predictions. The grid is searched by generating a series of scores from applying t-test (to detect changes in mean) and F-test (to detect change in variation) to the consecutive residuals/predictions at each step of expanding ($l \to l + \lambda,\ \lambda > 0$) and/or reshaping ($k \to k + \alpha,\ \alpha > 0$) the windows. This technique is based on the assumption that the mean and the variation of the residuals/predictions should remain unchanged in different time periods (Laurent and Berry, 2006).

To gain the necessary power in the analysis, we apply the statistical tests to the values of l that correspond to a minimum T observations in the windows. Then one can define the quantity of $T(l)$ that is the total number of observations that is included in the hard window corresponding to l. We should stress that the definition of $T(l)$ in the soft windowing can be challenging because the WGF assigns weights to the entire dataset in the final fitting. To address this complexity, we propose the Sum of Weights Score by $SWS(k,l) = \sum_{i=1}^{n} WGF(t_i,k,l,m)$, that is the summation of weights from WGF for specific l and k. Note that $SWS(l,k) \geq T(l)$ with the equality for sufficiently large k. Because l is generally unknown, a value of $T(l) = T$ independent of l needs to be decided before the analysis. Our experiments, inspired by the z-test minimal sample size $(n > 30)$, show that setting $SWS \geq T$ with

$$T \approx \begin{cases} \max(35, \sqrt{n\pi^2}) & \text{Single peak} \\ 35p & \text{Multiple peaks} \end{cases}$$

provides sufficient statistical power and precision for the analysis of each sex-parameter in IMPC.

Once the bandwidth, l, is selected, the shape parameter, k, can be optimised on a grid of values similar to l.

This algorithm is implemented for a broad range of models in the R package SmoothWin that is available from https://cran.r-project.org/package=SmoothWin. The main function of the package, $SmoothWin(...)$, allows an initial *model* for the input and, given a range of values for the bandwidth and shape, it performs soft windowing on the input model. Furthermore, it allows plotting of the results for diagnostics and further inspections. One also can generate the weights from SWGF using the $expWeigh(...)$ function.

4 Implementation

4.1 Sensitivity analysis

The sensitivity of the soft windowing to the tuning parameters in particular, the minimum observation required in the window (T), is tested on the two IMPC examples introduced in Figure 1 for *Mean cell volume* and *Forelimb grip strength normalised against body weight*. To this end, the tuning parameters l, k and T are set to

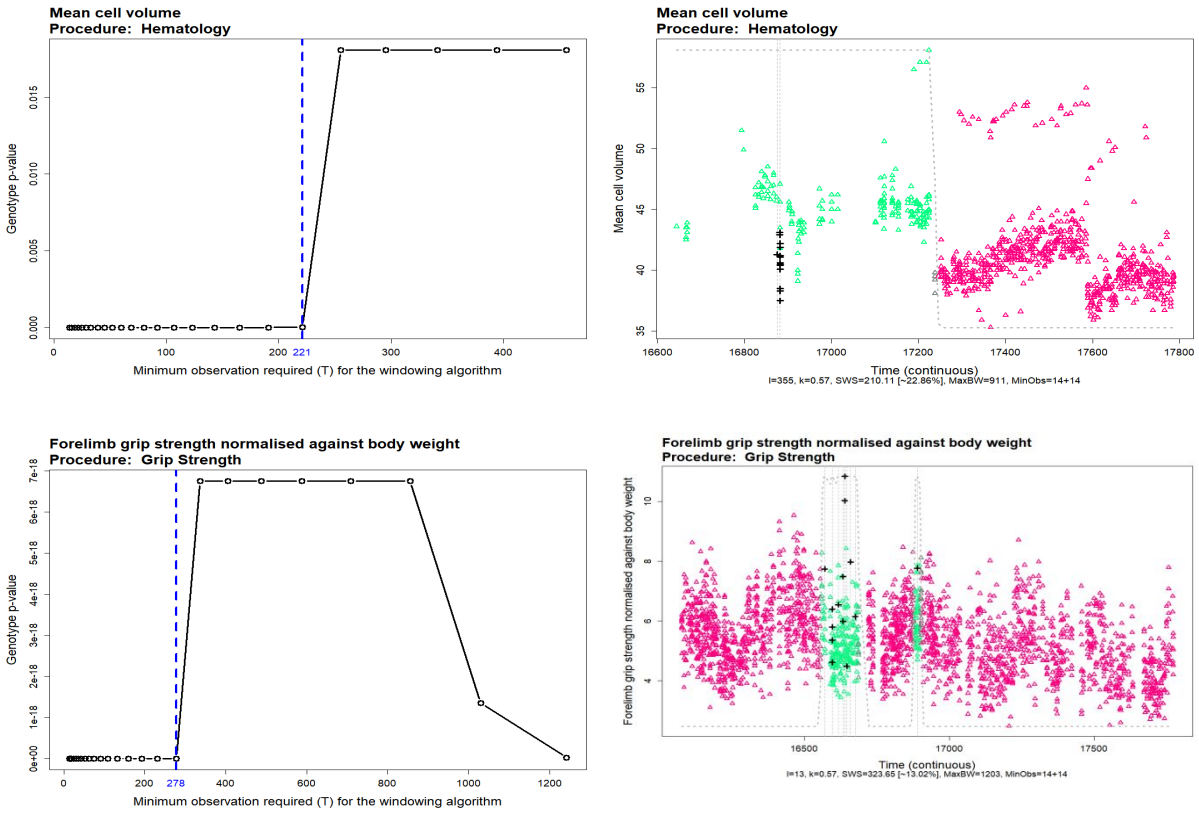

Fig 5. The sensitivity analysis of the soft windowing approach to the minimum observation required in the window. The left plots show the variation of the final Genotype p-values with different values of T. The vertical dashed blue lines show the maximum toleration of the algorithm before including too much noise in the final fittings. The right plots show the optimal soft windowed linear mixed model fitted to the data. The controls (triangles) weight are colour coded from green (inside the windows) to grey (on the window borders) and purple (outside the window). The mutants are shown with the black plus ($+$) on the plots.

- l The total range of the experiment time divided into 500 logarithmic distanced values
- k The values in $[0.5,10]$ interval divided into 50 logarithmic distanced values
- T The values from 14 to the n divided into 25 logarithmic distanced values

where n is the total observation in the dataset. We should stress that l and k are selected to cover the entire experiment range and avoid bias by selecting the incomplete ranges. Then we only study the effect of T on the final fittings.

Figure 5 shows the sensitivity of the p-values to the change in the minimum observation required for the soft windowing, T. The left plots show the change in the p-value corresponded to the genotype effect in the linear mixed model (with genotype, sex, genotype-sex interaction and body weight in the fixed effect term and the batch in the random effect) for different values of T. The dashed blue vertical lines show the maximum toleration of T before a step-change in the p-values being observed. The right-hand side plots show the final fitting of the windowed model. The controls (triangles) weight are colour coded on a spectrum of green-purple, inside the window (green), on the border (grey) and outside the window (purple). Figure 5 shows the sensitive of soft windowing to the T, for instance, selection of a high value for T could lead to including too much noise in the final fitting.

4.2 Simulation study

To assess the performance of the soft windowing method, we implemented a resampling approach to construct a sample of *artificial mutants* from the IMPC control data by relabelling some controls as mutant. We then examined the difference in the number of false positives that were detected by the standard (non-windowed) analysis versus the soft windowed approach. Since the resampling is only performed on the controls, we expect less false positives from the soft windowed results.

Mutant data in the IMPC has a special structure, resulting from mice being born in the same litters and being phenotyped closely together in time (batch effect), which must be replicated in the resampling approach. We address this by utilising *structured resampling* that replaces the mutants with the closest random controls in time. We create artificial mutant groups by randomly sliding the true mutant structure over the time domain of controls, collecting as many controls as there were mutants in the original set, and repeating this procedure five times per dataset (supplemental

Table 1: Top ten IMPC procedures with the highest change in the total number of false positives

Procedure name	#P-values[1]	NFP[2]	WFP[3]	Relative change[4]
Body Composition (IMPC_DXA)	167789	3809	2293	37.58
Clinical Blood Chemistry (IMPC_CBC)	320949	1472	2414	62.12
Open Field (IMPC_OFD)	182894	1507	830	35.52
Haematology (IMPC_HEM)	243640	3125	2746	46.77
Heart Weight (IMPC_HWT)	16236	553	409	42.52
Acoustic Startle and Pre-pulse Inhibition (IMPC_ACS)	73177	352	243	40.84
X-ray (IMPC_XRY)	7016	27	135	83.33
Insulin Blood Level (IMPC_INS)	9465	63	164	72.25
Electrocardiogram (IMPC_ECG)	122257	378	471	55.48
Eye Morphology (IMPC_EYE)	15739	86	153	64.02

[1]Total number of the analysis and p-values
[2]False positives from the non-windowed results
[3]False positives from the soft windowed results
[4]Relative percentage change of the false positives ($\frac{WFP}{NFP + WFP}$ %)

Figure 1 shows an illustration of three iterations of the structured resampling on the *Bone Mineral Content* parameter).

For non-windowed and soft windowed analyses, the same statistical model is fitted. That is the linear mixed model implemented in the R package PhenStat with genotype, sex, genotype-sex interactions and body weight for the fixed effect terms and the batch in the random effect. This setup implies that the difference in the results is a direct consequence of the control selection strategy by soft windowing. The outcome of the simulation study consists of 18 IMPC procedures across 11 centres and over $2.5\ million$ analyses and p-values. Comparing the results from the IMPC standard and soft windowed analyses on resampled data, we detect an overall of $14{,}201$ and $12{,}716$ false positives (FP), respectively, at the signficance level used by the IMPC, 0.0001. This constitutes more than a 10% relative improvement in FPs when the soft windowed method is applied. Table 1 shows the top ten IMPC procedures with the significant changes in the FPs. From this table, the procedures *Body Composition*, *Open Field*, *Urinalysis*, *Heart Weight*, *Acoustic Startle and Pre-pulse Inhibition* account for the highest relative reduction of 68% in FPs, whereas the *Clinical Blood Chemistry*, *X-Ray*, *Insulin Blood Levels*, *Electrocardiogram* and *Eye Morphology* account for the maximum increase of 32% in FPs. Supplemental Figure 2 shows parameters from the Body Composition and Clinical Blood Chemistry procudures that showed the biggest loss and gain in false positives for assocaited data parameters, respectively. This plot shows an improvement in decreasing FPs in all IMPC_DXA parameters, which contrasts with an increase in the FPs for IMPC_CBC parameters. We further examined the top two IMPC_CBC parameters, *Alanine aminotransferase* (IMPC_CBC_013) and *Aspartate aminotransferase* (IMPC_CBC_012) in Supplemental Figure 3, and noted a high level of randomly deviated points from the mean of controls that can bias the outcome of the structured resampling.

4.3 Soft windowing as part of the IMPC statistics pipeline

We next show the performance of the soft windowing approach on IMPC data by integrating it into the standard IMPC statistics pipeline in PhenStat (Kurbatova *et al.*, 2016). To this end, each dataset is processed by the PhenStat for the initial estimation of a fully saturated linear mixed model including genotype, sex, genotype-sex interaction and body weight in the fixed effect term and the batch in the random effect. The resulting fit is then passed into the soft windowing algorithm in the R package SmoothWin for the determination of the optimal windowing weights. After determining the optimal weights, the final model is fitted using a weighted linear mixed model and utilising a backward elimination approach to optimise the final model.

Using data release 9.2 (January 2019), we re-analysed $14\ million+$ data points from which $10\ million+$ are mutant animals across the range of IMPC phenotyping procedures. The original IMPC standard analysis that did not apply the soft windowing approach to select the control data encompassed $403{,}000+$ analyses and p-values. This analysis led to $12{,}728$ significant p-values (< 0.0001), compared to $16{,}415$ significant p-values when the soft windowing was applied, an increase of 30% in total significant p-values. The IMPC assigns mouse lines with phenotype terms from the Mouse Phenotype Ontology (MPO) when a significant deviation from the control data is detected for a given data parameter (Meehan *et al.*, 2017). Our windowing approach led to $17{,}391$ MPO associations gained and $15{,}996$ associations lost. To explore these differences further, we created an online tool that displays the entire control dataset for a given mouse line-parameter assay with the statistical summaries for both the non-windowed methodology and the soft windowed approach. Users may filter on a number of attributes, arrange filter order, zoom in on data visualisation, or navigate directly to the results (https://wwwdev.ebi.ac.uk/mi/impc/dev/phenotype-archive/media/images/windowing/).

Figure 6 shows the corresponding visualization on the IMPC website for the complete dataset (including males and females) previously shown for males only in Figure 1 (top) for the *Forelimb grip strength normalised against body weight* parameter from the IMPC Grip Strength procedure. The soft window is indicated, as well as changes in the total number of controls (here $1{,}572$ fewer after soft windowing – counting soft windowing weights greater than 10^{-7}). Further, the p-value corresponding to the genotype effect shows a significant change in magnitude, from 2.05×10^{-4} to 6.75×10^{-18} after applying the soft windowing. We then tested if our soft windowed analysis changed our human disease model discovery rate. We have previously described the IMPC Phenodigm translational pipeline that automatically detects phenotypic similarities between the IMPC strains and over $7{,}000$ rare diseases described in the Online Mendelian Inheritance in Man (OMIM), Orphanet and the Deciphering Developmental Disorders (DDD) databases (Meehan *et al.*, 2017). This pipeline generates qualitative

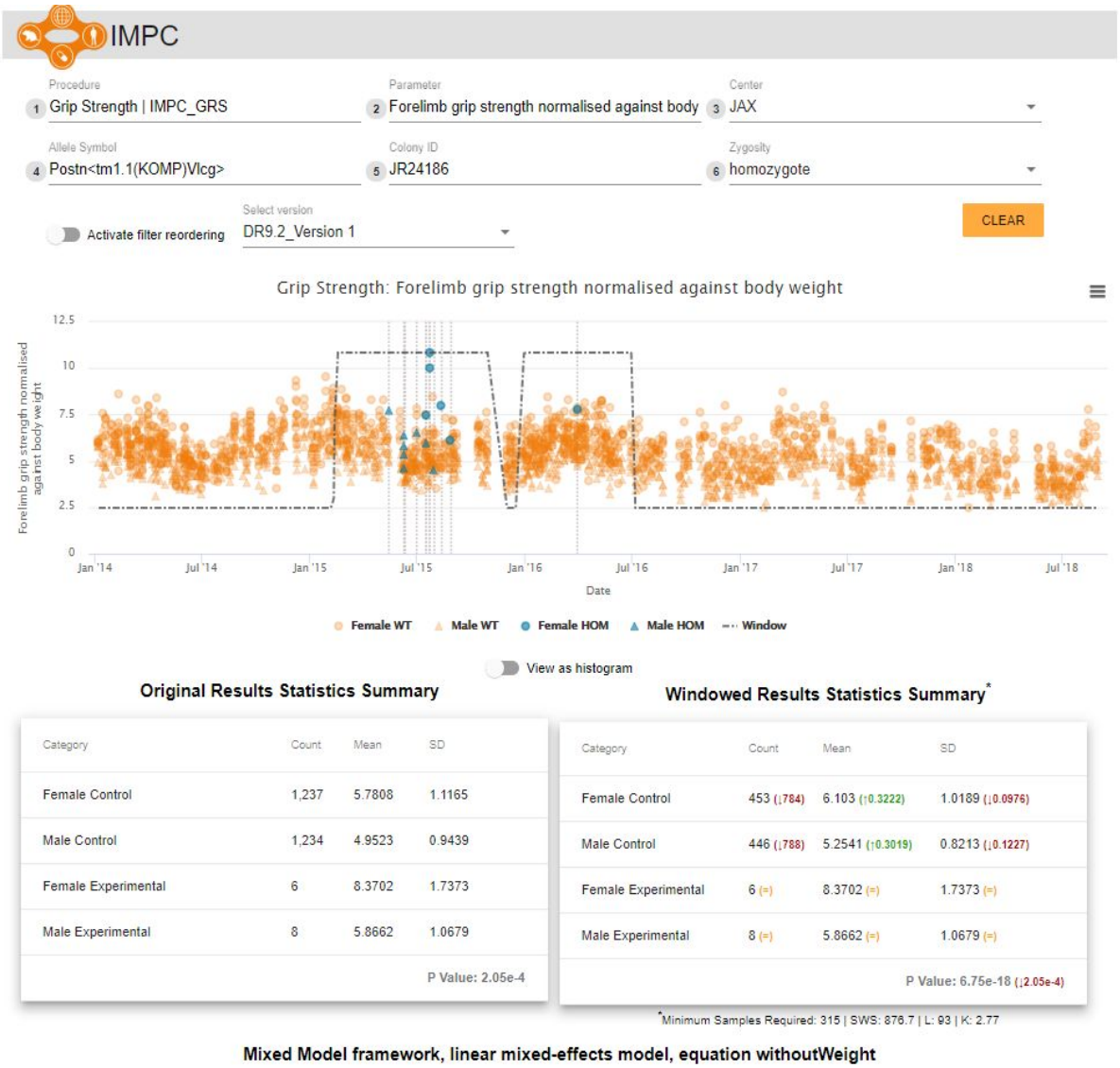

Fig 6. The soft windowing visualization in the IMPC website for the *Forelimb grip strength normalised against body weight* from the IMPC *Grip Strength* procedure. The plot shows the response over time as well as the fitted soft windows. The tables underneath show the comparison between the descriptive statistics obtained from the standard (non-windowed) analysis on the left and the soft windowed approach on the right. The p-values correspond to the genotype effect after applying the statistical analyses taking the corresponding controls based on the non-window and soft windowed approaches, respectively.

scores on how well a mouse line's associated phenotypes overlap with the phenotypes of the human rare disease populations (Meehan *et al.*, 2017; OMIM Browser, 2017; Rath *et al.*, 2012; Firth *et al.*, 2009; Mungall *et al.*, 2015; Akawi *et al.*, 2015). By comparing the disease model resulting from our soft windowed analysis vs non-windowed analysis for IMPC data release 9.2, we find a slight increase in the number of disease models (106 vs 99 models using a threshold of 50% phenotype overlap from a set of 2,082 mouse lines that contain mutations- Supplemental File I).

5 Discussion

High-throughput phenomics is a powerful tool for the discovery of new genotype-phenotype associations and there is an increasing need for innovative analyses that make effective use of the voluminous data being generated. Batch effects are inevitable when a large amount of data is collected at different times and/or sites and, therefore, need to be accounted for in the statistical analysis. In this study, we developed a novel "soft windowing" method which selects a window of time to include controls that are locally selected with respect to experimental animals, thus reducing the noise level in the data collected over long periods of time (years). Soft windowing has notable advantages over a more traditional hard windowing approach. In contrast to the limited data points included in the hard windowing method, the entire dataset is considered for the analysis. To this end, we engineered a weighting function to produce weights in the form of a window of time. Control data collected proximally to mutants were assigned the maximal weight, while data collected earlier or later had less weight. This method has the capability of producing indivdual windows as well as merging

intersected ones. Moreover, the method was implemented to automatically select window size and shape.

The performance of the method was shown on a simulated scenario that uses real control data collected by the IMPC high-throughput pipelines to assess detection of false positives. We also showed the enhancements to the IMPC statistical pipeline that establishes genotype-phenotype associations by comparing mutants vs control data using our soft windowed approach.

There are two known conditions that affect the method: (1) The weight generating function can be slow when there are too many (> 20) distinct windows, however, we have optimised the algorithm to be fast enough for the typical IMPC number of peaks (≈ 3 seconds for 1500 samples and 16 peaks under $k = 1$ and $l = 30$); and (2) Our resampling scenario indiciated that our soft windowing approach is sensitive to the data that has a high level of outliers or random deviation from the mean. This may result from a bias in the design of the resampling but may also indicate that using all available controls may be appropriate for the cases with extreme variability.

Our soft windowing approach addresses the scaling issues associated with analysing an ever-increasing set of control data in long-term projects by eliminating controls with weights sufficiently close to zero from future analysis. In the case of the IMPC, once a window of control data is determined for a dataset, there would be no further requirement to re-analyse the dataset with each subsequent data release. This will reduce the computational resources needed with the resulting gene-phenotype associations remaining stable, greatly facilitating data exchange with research groups trying to functionally validate genes and their disease variants. Our findings also have important implications for such efforts as the UK BioBank and the All of Us initiatives where large cohort sizes coupled with mobile medical sensors are generating phenotype data at an unprecedented rate (Sankar and Parker, 2017; Sudlow *et al.*, 2015). Researchers performing restrospective analysis to analyse exposures for a defined outcome group (e.g. metabolic disease) are challenged by the variability and longitudinal characteristics associated with these datasets. The methods described here can be used with these human health resources to maximise analytical power and help researchers find the genetic and environmental contributers to human diseases.

Funding Information

This work was supported by: [HH, MCJ, VMF, FLG, KB, RK, EW, SDB, DS, PF, AMMHP, TFM-NIH:UM1 HG006370], [EFA, AMF, AB, CM - NIH; UM1 OD023221; Genome Canada and Ontario Genomics (OGI-051 & 137)], [VK, JW -NIH:UM1OD023222], [DC, KCL- NIH: UM1 OD023221], [JS, AG, AG, AEC, CH, CLR, DGL, IL, JRG, JJG, RB, RCS, SV, JDH, MED- NIH:UM1 HG006348; U42 OD011174; U54 HG005348], [MT, NT, MH, OY- Management Expenses Grant for RIKEN BioResource Research Center, MEXT], [JK, SC, YK, JS- Korea Mouse Phenotyping Project (2017M3A9D5A01052447) of the Ministry of Science, ICT and Future Planning through the National Research Foundation], [GB, MC, LV, SL, HM, MS, PTR, TS, HY- We are grateful to members of the Mouse Clinical institute (MCI-ICS) for their help and helpful discussion during the project. The project was supported by the French National Centre for Scientific Research (CNRS), the French National Institute of Health and Medical Research (INSERM), the University of Strasbourg and the "Centre Europeen de Recherche en Biomedecine", and the French state funds through the "Agence Nationale de la Recherche" under the frame programme Investissements d'Avenir labelled (ANR-10-IDEX-0002-02, ANR-10-LABX-0030-INRT, ANR-10-INBS-07 PHENOMIN)], [GM, HF, LG, LB, NS, HM, VG, HM- German Federal Ministry of Education and Research: Infrafrontier [no. 01KX1012] (M.HdA.), the German Center for Diabetes Research (DZD), EU Horizon2020: IPAD-MD [no 653961] (M.HdA.)], [WW EUCOMM: Tools for Functional Annotation of the Mouse Genome' (EUCOMMTOOLS) project - grant agreement no [FP7-HEALTH-F4-2010-261492]]

References

Akawi,N. *et al.* (2015) Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet.*, **47**, 1363–1369.

Al-Tamimi,N. *et al.* (2016) Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. *Nat. Commun.*, **7**.

Begley,C.G. and Ellis,L.M. (2012) Drug development: Raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Blake,J.A. *et al.* (2017) Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Res.*, **45**, D723–D729.

Bradley,A. *et al.* (2012) The mammalian gene function resource: The International Knockout Mouse Consortium. *Mamm. Genome*, **23**, 580–586.

Brown,R.L. *et al.* (2018) Techniques for Testing the Constancy of Regression Relationships Over Time. *J. R. Stat. Soc. Ser. B*, **37**, 149–163.

Brown,S.D.M. and Moore,M.W. (2012) The International Mouse Phenotyping Consortium: Past and future perspectives on mouse phenotyping. *Mamm. Genome*, **23**, 632–640.

Charan,J. and Kantharia,N. (2013) How to calculate sample size in animal studies? *J. Pharmacol. Pharmacother.*, **4**, 303.

Dickinson,M.E. *et al.* (2016) High-throughput discovery of novel developmental phenotypes. *Nature*, **537**, 508–514.

Edwards,A.M. *et al.* (2011) Too many roads not taken. *Nature*, **470**, 163–165.

Firth,H. V. *et al.* (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.*, **84**, 524–533.

Flood,P.J. *et al.* (2016) Phenomics for photosynthesis, growth and reflectance in Arabidopsis thaliana reveals circadian and long-term fluctuations in heritability. *Plant Methods*, **12**, 14.

Ford,M.S. (2003) The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science.

Freedman,L.P. *et al.* (2015) The economics of reproducibility in preclinical research. *PLoS Biol.*, **13**, 1–9.

Friggens,N.C. *et al.* (2011) Extracting biologically meaningful features from time-series measurements of individual animals: towards quantitative description of animal status. In, *Modelling nutrient digestion and utilisation in farm animals*. Wageningen Academic Publishers, Wageningen, pp. 40–48.

Hrabě de Angelis,M. *et al.* (2015) Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. *Nat. Genet.*, **47**, 969–978.

Huang,B.E. *et al.* (2007) Detecting haplotype effects in genomewide association studies. *Genet. Epidemiol.*, **31**, 803–812.

Jank,W. and Shmueli,G. (2008) Statistical Methods in e-Commerce Research Jank,W. and Shmueli,G. (eds) John Wiley & Sons, Inc., Hoboken, NJ, USA.

Karp,N.A. *et al.* (2014) Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. *PLoS One*, **9**, e111239.

Kervrann,C. (2011) An Adaptive Window Approach for Image Smoothing and Structures Preserving. Springer, Berlin, Heidelberg, pp. 132–144.

Kurbatova,N. *et al.* (2016) PhenStat : statistical analysis of phenotypic data. *Bioc.Ism.Ac.Jp*, 1–9.

Kurbatova,N. *et al.* (2015) PhenStat a tool kit for standardized analysis of high throughput phenotypic data. *PLoS One*, **10**, e0131274.

Laurent,R.T. St. and Berry,W.D. (2006) Understanding Regression Assumptions.

Li,Y. *et al.* (2007) Association Mapping via Regularized Regression Analysis of Single-Nucleotide–Polymorphism Haplotypes in Variable-Sized Sliding Windows. *Am. J. Hum. Genet.*, **80**, 705–715.

Machado,J. *et al.* (2008) Intelligent Engineering Systems and Computational Cybernetics.

Malinowska,M. *et al.* (2017) Phenomics analysis of drought responses in Miscanthus collected from different geographical locations. *GCB Bioenergy*, **9**, 78–91.

Meehan,T.F. *et al.* (2017) Disease model discovery from 3,328 gene knockouts by the International Mouse Phenotyping Consortium. *Nat. Genet.*, **49**, 1231–1238.

Meyers,R.M. *et al.* (2017) Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.*, **49**, 1779–1784.

Mungall,C.J. *et al.* (2015) Use of Model Organism and Disease Databases to Support Matchmaking for Human Disease Gene Discovery. *Hum. Mutat.*, **36**, 979–984.

OMIM Browser (2017) Online Mendelian Inheritance in Man - An Online Catalog of Human Genes and Genetic Disorders. *academic.oup.com*.

Poularikas,A. (2019) Transforms and Applications Handbook.

Prinz,F. *et al.* (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.*, **10**, 712–712.

Rath,A. *et al.* (2012) Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Hum. Mutat.*, **33**, 803–808.

Sankar,P.L. and Parker,L.S. (2017) The Precision Medicine Initiative's All of Us Research Program: An agenda for research on its ethical, legal, and social issues. *Genet. Med.*, **19**, 743–750.

Stoeger,T. *et al.* (2018) Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.*, **16**.

Sudlow,C. *et al.* (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.*, **12**, e1001779.

Sun,J. *et al.* (2017) Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *Plant Genome*, **10**, 0.

Tang,R. *et al.* (2009) A variable-sized sliding-window approach for genetic association studies via principal component analysis. *Ann. Hum. Genet.*, **73**, 631–637.

Vaas,L.A.I. *et al.* (2013) Opm: An R package for analysing OmniLog® phenotype microarray data. *Bioinformatics*, **29**, 1823–1824.

Vaas,L.A.I. *et al.* (2012) Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS One*, **7**, e34846.

Vitak,S.A. *et al.* (2017) Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*, **14**, 302–308.

Viti,C. *et al.* (2015) High-throughput phenomics. In, *Methods in Molecular Biology*., pp. 99–123.