Clustering analysis for the data in the manuscript:

Neurothreads: development of supportive carriers for mature dopaminergic neuron differentiation and implantation

Aleksandra Filippova, Fabien Bonini, Liudmila Efremova, Manon Locatelli, Olivier Preynat-Seauve, Amélie Béduer, Karl-Heinz Krause, Thomas Braschler

Table of Content

Summary

This document provides clustering analysis[1] of the quantitative data associated with the manuscript "Neurothreads: development of supportive carriers for mature dopaminergic neuron differentiation and implantation". The overall conclusion is that clustering to biological replicates[2] tends to be statistically relevant in experiments involving cell differentiation, morphology or *in-vivo* injections, but not in experiments concerning primarily physical or chemical processes such as viability testing after injection. This document provides the analysis underpinning this conclusion along with further technical details and results.

As a result of this analysis, we generally report error bars and statistical hypothesis testing with values per independent biological experiment ("biological replicate"[2]) for cell differentiation, morphology and *in-vivo* experiments, whereas for experiments with anticipated primarily physical or chemical variability, we use directly the "technical replicates"[2] for both the calculation of error bars and statistical hypothesis testing. This is succinctly outlined in the manuscript (statistics section). Quantitatively justified in this document, the approach is also in line with literature practice: For dopaminergic cell differentiation experiments, the results are typically reported per independent experiment.[3,4] On the contrary, for studies with a focus on cell handling, or the physical or chemical environment of cells, the technical replicates are often used directly[5–7].

Introduction

Hierarchical data structure

The manuscript "Neurothreads: development of supportive carriers for mature dopaminergic neuron differentiation and implantation" presents both *in-vitro* and *in-vivo* data about minimally invasively injectable cryogel carriers for neuronal differentiation, culture and implantation. Some of this data presents hierarchical structure[1], also referred to as clustered[1] data, which is known to pose particular challenges in its statistical analysis[1,8]. The clustering arises here in the form of "biological" and "technical" replicates[2]. The biological replicates here reflect serial execution of distinct experiments, typically at different dates and therefore comprise at least some variation due to pipetting, changes in environmental conditions, and aliquots. The technical replicates on the other hand represent different wells, gels, image acquisitions or assay repetition within the same experiment. Since the technical replicates share are larger part of the experimental process, one can generally anticipate the variability between the biological replicates to be larger than the variability between technical replicates, although as shall become evident in our analysis, this tendency does apply uniformly.

Statistical inference with clustered data

Statistical inference in the presence of hierarchical data poses well-known problems[1,8]. Diminished variance within the clusters (here, the technical replicates belonging to a given biological replicate for a given condition) leads to decreased variability in the overall analysis. If no countermeasures are taken, this in turn leads to inflated statistics and incorrect inference[8]. The problem is particularly pronounced in imbalanced designs where the clusters are limited to individual conditions[1].

Given the known hierarchical schemes with biological and technical replicates used to obtain some of the data presented here, it is necessary to check on a case-by-case basis whether or not this indeed leads to clustered data. Importantly, the mere fact of having used hierarchical schemes does not necessarily imply effectively clustered data[1]. Whether or not this results in diminished intra-cluster variability as compared to inter-cluster variability needs to be assessed explicitly on a case-per-case basis[1]. If substantial clustering is found on the basis of the analysis of the actual data, the statistical approach needs to be adapted. Averaging over the technical replicates prior to statistical testing can be used to resolve the problem in these case[1]; another possibility is the application of a corrective Moulton factor[1].

Methods

Intraclass correlation

The most commonly used measure to assess presence of statistically relevant clustering (as opposed to mere hierarchical experimental design) is the so-called "intraclass correlation"[1], originally introduced in sibling calculations by R. Fisher[9]. Conceptually, the intraclass describes the balance between variance within and between clusters.[1] Strong intraclass correlation signifies that the clusters (i.e. biological replicates) are nearly homogeneous and should thus be taken as the statistical unit of calculation, while weak intraclass correlation signifies that the variability within the clusters is similar to the general variability and that the clusters are in fact irrelevant for statistical analysis. In this second case, technical replicates can be used instead.

Various approaches have been proposed to calculate intraclass correlations, and their suitability depends on the exact requirements[1,10–12]. We use here the generalized formula[1] based on Fisher's pioneering work in sibling problems.[9] For a collection of sample values z_{ig}, where i indicates the i-th member of the group g (i.e. cluster), the expression is (see reference [1], eq. 8.2.5):

$$\rho(z) = \frac{\sum_g \sum_{i \neq j} (z_{ig} - \overline{z})(z_{jg} - \overline{z})}{V(z_{ig}) \cdot \sum_g n_g (n_g - 1)} \qquad \text{eq. 1}$$

where $V(z_{ig})$ denotes the variance of the sample values regardless of the group structure (i.e. $V(z_{ig}) = \frac{\sum_g \sum_i (z_{ig} - \overline{z})^2}{n}$), with $n = \sum_g n_g$ the total number of observations. n_g designates the individual group sizes and $\overline{z}$ is the arithmetic mean of all the observations across all the groups.

Eq. 1 materialises the notion of intraclass correlation. In the presence of strong clustering all the z_{ig} values in a given cluster g are similar, and so all the crossed product terms $(z_{ig} - \overline{z})(z_{jg} - \overline{z})$ will be positive and sum up to a substantial contribution to the intraclass correlation coefficient $\rho(z)$. On the contrary, in the absence of intra-cluster correlation, the signs of the $(z_{ig} - \overline{z})$ and $(z_{jg} - \overline{z})$ are random, and so summing up the crossed terms $(z_{ig} - \overline{z})(z_{jg} - \overline{z})$ will mostly lead to cancellation of terms of opposing signs and ultimately to a $\rho(z)$ close to 0. The denominator in eq. 1 serves to normalize $\rho(z)$ to an anticipated range of approximately 0 (no intraclass correlation) to 1 (homogeneous clusters).

Moulton factor

Intraclass correlation (clustering) by itself does not directly lead to inflation of the test statistics, it generates problems particularly in unbalanced designs where the cluster groups do not regularly span the values of the regressors[1]. The Moulton factor[1,8] takes into account both intra-class correlation of the observed values y_{ig} and regressors x_{ig}. Its general expression is[1]:

$$M = 1 + \left[\frac{V(n_g)}{\overline{n}} + n - 1\right] \cdot \rho(x) \cdot \rho(y) \qquad \text{eq. 2}$$

with an explicitly symmetric contribution of intra-class correlation of both the observed values y and the regressor values x. In our case, the experimentally observed values y are numerical observations, for example the fraction of viable cells or the number of neurites per cell. The regressors x can be binary in the case of t-testing (i.e. 0 for condition A and 1 for condition B), unordered factors (i.e. laminin, Matrigel, fibronectin, …, coating) or numerical values (i.e. EDC concentrations). The clusters g are the experimental runs and thus represent the biological replicates, while the values within the clusters are the technical replicates. A Moulton factor of 1 indicates no inflation of test statistics.

The worst-case scenario arises where groups are uniquely associated with given conditions[1], resulting in $\rho(x) = 1$. In this case, the Moulton factor M assumes its maximum value:

$$M_{\max} = 1 + \left[\frac{V(n_g)}{\overline{n}} + n - 1\right] \cdot \rho(y) \qquad \text{eq. 3}$$

Since we generally perform a series of statistical comparisons on any given dataset, with only partially balanced designs, we use the worst-case Moulton factor values given by eq. 3 to assess whether we there is overall low intra-class correlation ($M_{\max} \leq 1.1$) or whether it is advisable to aggregate the observed values within the clusters and perform statistics solely on averaged per-cluster values. In this latter case, the fundamental unit of the statistical analysis becomes the biological replicate rather than the technical replicate.

Analysis of residual error

The clustering analysis explicitly refers to the residual error structure[1,13], not the data structure by itself. When fitting hierarchical models,[2,13] this is automatically taken into account. Here, we separate the clustering analysis from actual modelling and testing, and so to be accurate, it is necessary to remove the main effect of the conditions. We do so by subtracting the mean for the appropriate conditions before performing the clustering analysis.

Results

Main figures

Table 1 below shows the results of the clustering analysis for the main figures. For some figures (Fig. 2b, 2c, 6c), the dataset corresponds to a single biological experiment and thus

clustering analysis is not applicable by definition. For the remainder of the data, the general observation is that where cell differentiation and morphology is important, there is relevant clustering as evidenced by at least some M_{max} values substantially above one (i.e. Fig. 3, 4 and 6b). On the contrary, for viability testing after injection (Fig. 5), there is no sign of clustering along biological replication. This should not be surprising: while cell passage number and exact media composition are an important source of variation regarding the aspects of cell morphology and differentiation, the main variability in injection testing is expected to arise from exact gel handling, dominating the one arising from the biological gel preparation prior to injectability testing.

Figure	M_{max} (eq. 3)	Approach
Fig. 2b, 2c	N.A. (unique experiment)	This is a single experiment, no clustering => use values directly
Fig. 3c	1.7	Use aggregated values (1 value per biological replicate obtained by averaging of corresponding technical replicates). 3e: We included cell density as a mechanistically interfering covariate into the regression
Fig. 3d	6.3	
Fig. 3e	1.6	
Fig. 4b	SOX2: 3.8 Ki67: 1.0 TH: 1.34	Use aggregated values (1 value per biological replicate obtained by averaging of corresponding technical replicates)
Fig. 4c	SOX2: 1.0 Ki67: 2.7 Map2: 1.2	
Fig. 5b	1.0	Negligible clustering, use technical replicate values directly
Fig. 6b	Transcripts with positive intra-class correlation (M_{max}>1): EN1: 1.9 MAP2: 1.1 TH: 1.03	By precaution, since there are individual genes with higher M_{max}, average corresponding technical replicates to single values per biological replicate and condition
Fig. 6c	N.A. (unique experiment)	This is a single experiment, no clustering => use values directly

Table 1. Results of the clustering analysis for the main figures. Values of M_{max}<1.0 truncated to M_{max}=1.0 (these result from small negative intraclass correlation values as a result of random effects with the unbiased expression for intraclass correlation given by eq. 1).

Supplementary figures

Table 2 below shows the results of the clustering analysis for the supplementary figures portraying quantitative data. Fig. Fig. S1, S2 and S3 analyse additional readout and data similar to the corresponding main Fig. 3 and 5, and their clustering analysis confirms the tendencies found for the main figures. Fig. S6-S9 concern experiments for which at least part

of the results stem from single experiments. These figures are therefore not amenable to aggregation at the level of biological replicates, and additional considerations for clustering apply. The quantitative *in-vivo* data provided in Fig. S10 and S11 reflects low numbers of mice and remains preliminary regardless of aggregation.

Supplementary Figure	M_{max} (eq. 3)	Approach
Fig. S1a	3.2	Use aggregated values (1 value per biological replicate obtained by averaging of corresponding technical replicates), in agreement with main Fig. 3.
Fig. S1b	3.0	
Fig. S1c	1.14	
Fig. S1d	1.6	
Fig. S2	1.01	Negligible clustering, use technical replicate values directly. This is in agreement with main Fig. 5.
Fig. S3	1.08	Negligible clustering, use technical replicate values directly. This is in agreement with main Fig. 5.
Fig. S6	N.A. (unique experiment)	This is a single experiment, no clustering => use values directly
Fig. S7	N.A. (unique experiment)	This is a single experiment, no clustering => use values directly
Fig. S8	Unique experiment for pH 6 => cannot use aggregation to biological replicates for testing. M_{max}=1.5	The pH 6 condition was added from a separate, unique experiment during the review process. Aggregating to biological replicates (experiments) would therefore lead to n=1 for the pH 6 condition, removing the possibility to perform t-tests. However, the actual statistical test result is non-significant (see Supplementary S8), so potential inflation is irrelevant, implementing the Moulton correction would indeed only increase the P-value[1].
Fig. S9	Unique experiment for hESC => cannot use aggregation to biological replicates for testing. Neurites per DAPI: M_{max}=1.0 Neurite length : M_{max}=2.1	Similar case to S8: The hESC data corresponds to a separate, unique experiment. Aggregating to biological replicates (experiments) would therefore lead to n=1 for the hESC, removing the possibility to perform t-tests. With M_{max}=2.1, there is substantial clustering, and therefore using technical replicates only would unduly inflate the test statistics. To address the problem, we average the data per biological replicates for the Luhmes data. This reduces M_{max} to 1. The statistical question regarding the hESC then becomes whether or not the various hESC values from the one experiment performed are significantly different from what one would expect for a typical Luhmes experiment.

Fig. S10-1c	1.5	Use aggregated values: 1 value per mouse (=biological replicate) obtained by averaging of corresponding histological slices (=technical replicates).
Fig. S10-3	1.85	Use aggregated values: 1 value per mouse (=biological replicate) obtained by averaging of corresponding histological slices (=technical replicates). An additional analysis based on individual images is also carried out, but necessitates confirmatory studies due to the low number of biological replicates with possible inflation of inference due to clustering (i.e. in Fig. S10-3a, non-significant P=0.13 with aggregation per mouse, significant P=0.022 when using the technical replicates, and marginal P=0.077 with Moulton correction)
Fig. S11	(1.0)	Use cumulated cell counts for assessment of proportions of Ki67+ cells. Limited sample.
Fig. S14-2	Beta-III tubulin positive fraction: M_{max}=1.0 Syn1-positive among BIII-positive: M_{max}=1.0	Use technical replicates directly; there is relatively high inherent variability in the limited regions covered by the confocal images overshadowing possible biological variation.
Fig. S14-3	N.A. (unique experiment)	This is a single experiment, no clustering => use values directly

Table 2. Results of the clustering analysis for the supplementary figures. Values of $M_{max}<1.0$ limited to $M_{max}=1.0$ (these result from small negative intraclass correlation values as a result of random effects with the unbiased expression for intraclass correlation given by eq. 1).

Discussion and Conclusions

Globally, the clustering analysis described in this document confirms that clustering in biological units or biological experimental runs is more important in fundamentally biological experiments such as cell differentiation and morphology, and also *in-vivo* experiments. On the contrary, experiments regarding mainly physical processes such as viability assessment after injectability testing show little clustering along the biological repetitions. We therefore empirically confirm the anticipated main sources of variability.

The statistical problem of data clustering is particularly important with large, unbalanced experimental designs, and therefore has been a longstanding subject in social sciences, from where the techniques and concepts used here stem[1]. In experimental biological sciences, it is more typically addressed by the choice of "technical" vs. "biological" replicates as the primary unit of statistical evaluation[2]. Here, we use the Moulton factor formalism[1] to decide

between the two scenarios[2] based on the actual data. Reassuringly, the overall conclusion drawn corresponds to commonly encountered literature practice[3–7]: focus on biological replicates for experiments involving primarily biological processes such as dopaminergic cell differentiation[3,4], and direct use of technical replicates for experiments addressing physical processes such as cell handling[5–7].

Overall, our approach allows to rationally navigate between statistical power and conservativeness. Where clustering to biological units is statistically relevant, we use aggregation to the biological units in both statistical evaluation and figure display. This approach is conservative[1], preventing potential inflation of statistical test results due to undue data clustering. Where clustering effects are quantitatively negligible, the more numerous technical replicates can be used to provide statistical power. Finally, by using clustering analysis upstream of actual data analysis, statistical hypothesis testing remains simple and transparent.

Replicating the clustering analysis

Software environment

We performed the clustering analysis in R, so if not already installed, the first step would be to wownload and install the R statistics program (i.e. from https://cran.r-project.org). In addition, two R packages are needed. The first is the CRAN package "readxl", the second is the "moultonTools" library from our public Github software repository. "readxl" can be installed via the usual R package installation (Package installer from the R menu or at R the command line with an "install.packages" command). The difficulty with readxl is its reliance on Perl for reading Excel files, it may be necessary to install Perl as well (various packages depending on operating system, see https://www.perl.org).

To install the "moultonTools" package from Github, there are several options. The simplest, out-of-the-box approach is to use the "devtools" library to automatically download and install from Github. The devtools library itself may already be installed on a given R installation, otherwise, it can be downloaded in the usual way (Package manager or R command line). With devtools installed, the moultonTools library can be installed in an automated fashion:

```
library(devtools)
install_github("tbgitoo/moultonTools")
```

This should install the public R library moultonTools. As given above, the install_github command will install the most recent issue of moultonTools. For reference, to download and install the version used to estimate the Moulton factor values in this document, the specific Github commit used here can unambiguously be installed via:

```
install_github("tbgitoo/moultonTools",
ref="2b6346d7866790743b6d0218be7192df25e2c5a2")
```

We have had issues with the "install_github" function itself due to encoding problems, but could resolve this by explicitly setting the default encoding in R. On MacOSX, we used the system command line ("Terminal") for this, the command is something of the type "defaults

write org.R-project.R force.LANG en_US.UTF-8" depending on the desired default encoding.

An alternative is to download the entire source code of the moultonTool package from https://github.com/tbgitoo/moultonTools (green "Code" button, download as zip, unzip locally) and install it with the R CMD utility from a system command prompt. As for the automatic installation with install_github, the specific Github version used here can be retrieved by accessing the suitable specific Github commit (at https://github.com/tbgitoo/moultonTools/tree/2b6346d7866790743b6d0218be7192df25e2c5a2).

Scripts and data files

With the moultonTools and readxl library in place, the software environment for the scripts in "clustering_analysis_R_scripts.zip" is set. In addition to the general software environment, the scripts contained in the "clustering_analysis_R_scripts.zip" archive need access to specific data files. The easiest approach is to set the working directory in R (command "setwd" or from the menus) to the directory where you download the raw data files from this repository (https://doi.org/10.5281/zenodo.4441090). Table 3 lists the data files required for the various scripts.

Script	**Data files**
Clustering analysis for main figures	
Rscript_3c.R	Fig_3c.xlsx
Rscript_3d.R	Fig_3d.xlsx
Rscript_3e.R	Fig_3e.xlsx
Rscript_4b.R	Fig_4b.xlsx
Rscript_4c.R	Fig_4c.xlsx
Rscript_5b.R	Fig_5b.xlsx
Rscript_6b.R	Fig_6b.xlsx
Clustering analysis for supplementary figures	
Rscript_S1a.R	Supplementary_Fig_S1.xlsx
Rscript_S1b.R	
Rscript_S1c.R	
Rscript_S1d.R	
Rscript_S2.R	Supplementary_Fig_S2.xlsx
Rscript_S3.R	Supplementary_Fig_S3.xlsx
Rscript_S8.R	Supplementary_Fig_S8.xlsx
Rscript_S9.R	Fig_3c.xlsx, Supplementary_Fig_S9.xlsx
Rscript_S10_1c.R	Supplementary_Fig_S10.xlsx
Rscript_S10_3a.R	Supplementary_Fig_S10.xlsx
Rscript_S11.R	Supplementary_Fig_S11.xlsx
Rscript_S14_2.R	Supplementary_Fig_S14_2_and_3.xlsx

Table 3. Results of the clustering analysis for the supplementary figures. Values of $M_{max}<1.0$ limited to $M_{max}=1.0$ (these result from small negative intraclass correlation values as a result of random effects with the unbiased expression for intraclass correlation given by eq. 1).

References

1. Angrist, J. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. (Princeton University Press, 2009).
2. Tsvetkov, D., Kolpakov, E., Kassmann, M., Schubert, R. & Gollasch, M. Distinguishing Between Biological and Technical Replicates in Hypertension Research on Isolated Arteries. *Front. Med.* **6**, (2019).
3. Kriks, S. *et al.* Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature* **480**, 547–551 (2011).
4. Cui, J. *et al.* Quantification of dopaminergic neuron differentiation and neurotoxicity via a genetic reporter. *Sci. Rep.* **6**, 25181 (2016).
5. Ning, L., Betancourt, N., Schreyer, D. J. & Chen, X. Characterization of Cell Damage and Proliferative Ability during and after Bioprinting. *ACS Biomater. Sci. Eng.* **4**, 3906–3918 (2018).
6. Han, W. T. *et al.* Improved cell viability for large-scale biofabrication with photo-crosslinkable hydrogel systems through a dual-photoinitiator approach. *Biomater. Sci.* **8**, 450–461 (2019).
7. Weiden, J. *et al.* Injectable Biomimetic Hydrogels as Tools for Efficient T Cell Expansion and Delivery. *Front. Immunol.* **9**, (2018).
8. Moulton, B. R. Random group effects and the precision of regression estimates. *J. Econom.* **32**, 385–397 (1986).
9. Fisher, R. A. On the 'probable error' or a coefficient of correlation deduced from a small sample. *Metron* **1**, 3–32 (1921).
10. Fleiss, J. L. & Shrout, P. E. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika* **43**, 259–262 (1978).
11. HARRIS, J. A. ON THE CALCULATION OF INTRA-CLASS AND INTER-CLASS COEFFICIENTS OF CORRELATION FROM CLASS MOMENTS WHEN THE NUMBER OF POSSIBLE COMBINATIONS IS LARGE. *Biometrika* **9**, 446–472 (1913).
12. Kish, L. *Survey Sampling*. (John Wiley & Sons, 1965).
13. Stanish, W. M. & Taylor, N. Estimation of the Intraclass Correlation Coefficient for the Analysis of Covariance Model. *Am. Stat.* **37**, 221–224 (1983).