1 **Slicing: a sustainable approach to the analysis of long-term biobanks**

2 Sil H.J. van Lieshout[1], Hannah Froy[2,3], Julia Schroeder[4], Terry Burke[5], Mirre J.P. Simons[5,6]* & Hannah L.

3 Dugdale[1]*

4 [1]School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK; [2]Institute of

5 Evolutionary Biology, University of Edinburgh, Edinburgh, UK; [3]Centre for Biodiversity Dynamics,

6 Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway;

7 [4]Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, UK; [5]Department

8 of Animal and Plant Sciences, University of Sheffield, Sheffield, UK; [6]The Bateson Centre, University of

9 Sheffield, Sheffield, UK.

10

11 *Joint last authors: these authors contributed equally to this study

12 Correspondence author: Sil H.J. van Lieshout

13 E-mail: sil.vanlieshout@gmail.com

14

15 ORCID SHJvL, 0000-0003-4136-265X; HF, 0000-0003-2965-3526; JS, 0000-0002-4136-843X; TB, 0000-

16 0003-3848-1244; MJPS, 0000-0001-7406-7708; HLD, 0000-0001-8769-0099

17

18 **Glossary**

19 **Batch –** a set of analysed samples that are inherently dependent on each other, e.g. all using the

20 sample standard curve, machine, time of day, technician or that are equally affected by any other

21 unknown variation.

22 **Cluster** – a set of samples that are distinct in the timing of their analysis; this typically includes multiple

23 batches.

24 **Within-individual effects** – Longitudinal changes within an individual.

Between-individual effects – Comparison of differences across individuals induced by demographic effects such as selective (dis)appearance.

Abstract (max 350 words)

1. The longitudinal study of populations is a core tool for understanding ecological and evolutionary processes. These studies typically collect samples over individual lifetimes and across multiple generations, building up a continuously growing biobank from which samples are then analysed in clusters over time in the laboratory. To ensure data are comparable among clusters we need to account for among-cluster variation and confounding variables, yet this is often ignored.
2. The commonly used approaches in structuring samples for analysis, sequential and randomisation, generate bias due to non-independence between their time of collection and cluster. We propose a new sample selection strategy, slicing, specifically designed to statistically account for this bias. Slicing would, however, be suboptimal if aggregating longitudinal samples of the same individual within a single batch reduces measurement error and thereby increases statistical power to detect within-individual effects, a notion we challenge using simulations.
3. Our slicing approach, whereby recently and previously collected samples are analysed in a cluster together, enables statistical separation of collection time and cluster effects through appropriate mixed models. Additionally, we recommend the use of internal controls (reference samples) to further assess among-cluster variation. Our simulations show similar precision and higher statistical power to detect cohort, within- and between-individual effects when samples are sliced across batches, compared with strategies that aggregate longitudinal samples or use randomised allocation.
4. While the best approach to analysing long-term datasets depends on the structure of the data and questions of interest, it is vital to account for among-cluster and batch variation. This can

be achieved through mixed models and appropriate sample selection strategies. Our slicing approach is simple to apply and creates the necessary statistical independence of batch and cluster from environmental or biological variables of interest. Crucially, it allows subsequent samples to be added in later analyses without completely confounding them with cluster. Our approach maximises the value of every sample, as each will optimally contribute to unbiased statistical inference from the data. Slicing therefore has the potential to maximise the power of growing biobanks to address important ecological, epidemiological and evolutionary questions.

Keywords: Ageing**,** biobank, internal controls, longitudinal, long-term studies, mixed models, slicing, telomeres

Introduction

Individuals and populations are shaped by ecological and evolutionary processes, for example, population structures regulated by demographic processes, and genetic variation and adaptation controlled by evolution. These processes commonly operate over multiple years, decades or even centuries (Clutton-Brock & Sheldon 2010). Consequently, long-term studies form a crucial basis for empirical studies into evolutionary processes and their interface with ecology.

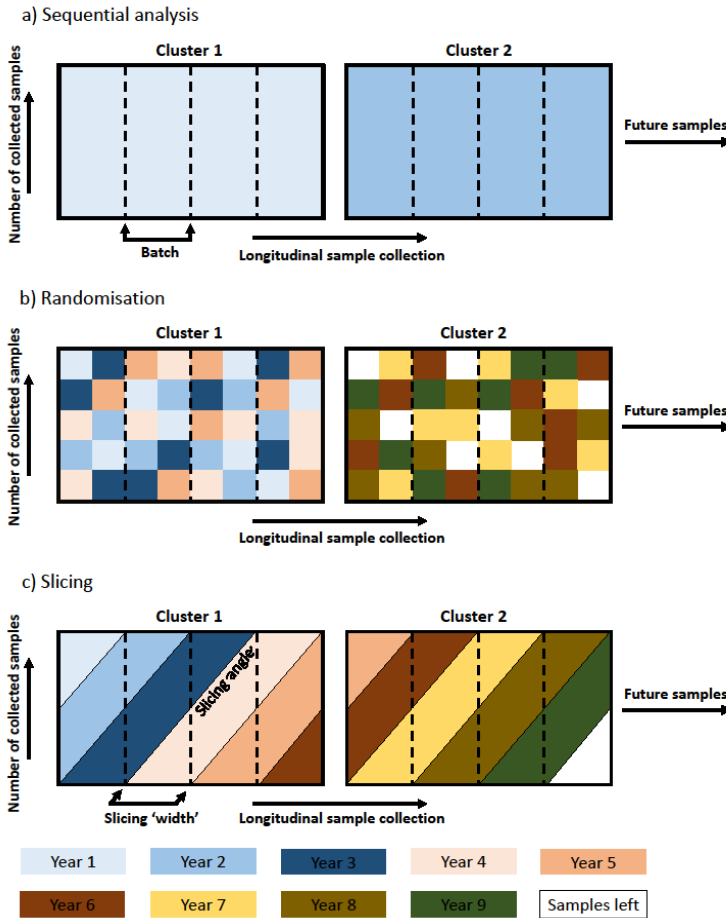
Long-term studies can be performed cross-sectionally, where individuals are sampled once. These data allow broad inferences on processes within and between populations. However, cross-sectional studies have the limitation that they cannot distinguish between, for example, selective disappearance due to intrinsic or environmental factors and within-individual changes such as ageing effects, or result in processes being masked by individual heterogeneity (Nussey *et al.* 2008). Another issue is that the assumptions required for the analysis of cross-sectional data, e.g. stationary age distribution and equal probability of sampling, are rarely met in natural populations (Gaillard *et al.* 1994). In contrast, longitudinal long-term studies, gathering data repeatedly over the lifetimes of

77 individuals, can eliminate biases associated with cross-sectional data. Moreover, these longitudinal
78 studies can distinguish within-individual patterns from between-individual variation, providing
79 opportunities to uncover the mechanisms and evolution behind many biological processes (Nussey *et*
80 *al.* 2008; Dugdale *et al.* 2011; Hammers *et al.* 2015; Fairlie *et al.* 2016). For example, individual-based
81 longitudinal studies can be used to identify proximate causes of changes in life-history traits (Clutton-
82 Brock & Sheldon 2010).

83 The individual-based collection of longitudinal data and biological samples from natural or
84 laboratory populations leads to large, continuously growing biobanks. Through laboratory analyses
85 these biobanks can provide vital information on, for example, individual telomere length (Boonekamp
86 *et al.* 2014; Fairlie *et al.* 2016), serological values (Telfer *et al.* 2008; Andraud *et al.* 2014) and genetic
87 variation (Berry *et al.* 2012; Tollenaere *et al.* 2012). The laboratory analysis of samples from growing
88 biobanks is often inevitably conducted on separate groups of samples over time (e.g. after each
89 fieldwork season, each year, coinciding with grant cycles). Such a group of samples – a cluster – will
90 be collectively analysed under similar conditions, but these conditions might differ between clusters
91 (e.g. different analyst, different machine, different month). Samples within a cluster are then
92 subdivided into batches (e.g. qPCR plates) where, again, samples are analysed under similar
93 conditions, but conditions may vary between batches (e.g. different reagents, different day). The
94 problem with this differential timing in analysis of batches and clusters is that it induces an unknown
95 level of variation that is often confounded with the independent variables of interest, which reduces
96 the ability to compare results across samples and draw reliable conclusions. For example, temporal
97 variation or where multiple populations are studied, spatial differences, in resource availability can be
98 confounded with cluster when samples are analysed after each period of collection, resulting in a
99 failure to separate the effects of resource availability and cluster on a response variable.

100 The two main approaches currently used to structure samples into clusters in long-term
101 studies suffer from differential timing in analysis. First, sequential analysis (i.e. running clusters in the
102 order in which they were collected) allows clustered analysis of samples (e.g. by year) and has the

103 advantage that samples can be analysed immediately, but this method confounds cluster with
104 organising variable (e.g. year) effects (Fig. 1a). Second, analysing multiple years within a cluster so that
105 year can be randomised among batches ensures that samples are sufficiently mixed to avoid
106 confounds, and should already be standard practice (Fig. 1b). However, this randomisation approach
107 requires a delay before analyses can be completed, so that samples collected at different time points
108 can be analysed together and organising variable and cluster effects can then be separated.
109 Furthermore, after applying this randomisation approach, any subsequently collected samples cannot
110 be compared to the already confounded samples as they will be subject to statistically uncontrollable
111 variation due to clustering of the samples already analysed. For example, separately randomising two
112 time periods of five years of sampling into two clusters results in uncontrollable variation between
113 these two clusters and confounds the first five years in cluster one with the subsequent years in cluster
114 two (Fig. 1b). Even though analysing the same samples multiple times in subsequent clusters can avoid
115 this issue, the additional costs and depletion of samples makes this an undesirable solution.
116 Additionally, the randomisation of samples among batches within a cluster is time-consuming and the
117 detailed reordering of samples from the biobank is prone to error due to sample labelling and placing.



118

119 **Fig. 1**. Schematic of three strategies to select samples from the biobank. The sequential analysis
120 strategy (a) can confound cluster and year, while randomisation of multiple years within a cluster (b)
121 prevents this confound but generates uncontrollable variation between clusters. The slicing approach
122 (c) combines the advantages of these approaches and can be used to sequentially analyse growing
123 biobanks, while maintaining independence between cluster and associated variables. The biobank is
124 sliced (e.g. by year), thereby analysing a set of continuously collected samples sequentially in each
125 subsequent cluster. Slicing width (frequency of new samples collected) and angle (degree of
126 independence between slices) determine the level of statistical independence between clusters.
127 'Samples left' are samples collected in subsequent years to highlight that, unlike with the sequential
128 strategy, in randomisation and slicing strategies samples cannot be analysed immediately.

129

130 The variation among clusters, created by the use of the sequential and randomisation
131 selection strategies, might be trivial, but there is no way to assess it. The randomisation strategy
132 applied to completed biobanks would enable separation of cluster and organising variable effects, but
133 this approach is not suited for biobanks that are still growing. In addition to confounding cluster and
134 organising variable effects, confounds can arise in many other ways (e.g. change in reagents or analyst
135 over time), which could coincide with changing population dynamics (e.g. age composition,
136 environmental effects). This uncontrollable variation among clusters and potential confounding
137 factors can lead to false positive or false negative results due to temporal correlations with cluster
138 effects. Hence, these methods cannot provide the comparable analyses of samples over time required
139 in long-term studies.

140 The differential timing in the analysis of a growing biobank that results in among-cluster
141 variation can be partially accounted for using appropriate normalisation of data, for example, by
142 including internal controls (or reference sample/calibrator) in all batches (e.g. qPCR plates), to which
143 sample data can be normalised (Cawthon 2002). However, normalisation through internal controls
144 can only be applied in specific forms of analyses (e.g. qPCR, antibody hormone analysis). The internal
145 controls provide measures of precision (e.g. repeatability) which should be reported as standard in
146 studies (Nussey *et al.* 2014) and allow inferences on among-batch variation. Still, the strong
147 dependence of multiple samples in one batch on a single reference sample can inflate noise and/or
148 be inadequate to fully correct for among-batch variation. When such effects are modelled using mixed
149 effects models, laboratory analysis effects (e.g. qPCR plate, Froy *et al.* 2017) still explain a considerable
150 proportion of the variance. This suggests that inclusion of a reference sample only partially accounts
151 for among-batch and among-cluster variation and thereby provides a potentially false sense of
152 confidence of the data acquired.

153 In addition to data normalisation, modelling statistical dependence through mixed models can
154 also increase the reliability of results from data collected in differentially analysed clusters (Bolker *et*
155 *al.* 2009). Mixed models allow flexible inclusion of random effects (Gelman & Hill 2006), such as

156 clusters and batches, which allow to control for among-cluster variation, and thus an accurate
157 estimate of repeatability. Accurate repeatability estimates, in turn, lead to more accurate estimates
158 of other fixed (e.g. age) or random effects (e.g. year) or parameters calculated from these effects (e.g.
159 heritability).

160 While internal controls and mixed models can partially account for among-cluster variation,
161 the differential timing in analysis is further complicated by the commonly applied sample allocation
162 strategy (i.e. allocating samples to batches), where longitudinal samples from a single individual need
163 to be analysed in the same batch. For example, a population monitored yearly provides multiple
164 samples per individual from which telomere lengths can be estimated. An individual's cohort and year
165 of sampling can have biological effects that impact telomere length (i.e. environmental variation that
166 we are interested in). However, the batch in which samples are analysed can affect the telomere
167 length estimate as well due to technical variation (i.e. experimental variation that we are not
168 interested in, but want to correct for). To account for this technical variation, samples from a single
169 individual are often analysed in the same batch (e.g. Beirne *et al.* 2014, Nettle *et al.* 2015), which is
170 thought to increase the statistical power to detect within-individual effects. The reasoning here is that
171 longitudinal samples are then exposed to the same technical noise, which allows higher statistical
172 power to dissect out the biology, without batch effects confounding longitudinal analysis of samples
173 (Nordfjall *et al.* 2005; Salomons *et al.* 2009; Rius-Ottenheim *et al.* 2012; Herborn *et al.* 2014; Nettle *et*
174 *al.* 2015).

175 However, the sample allocation strategy of aggregating longitudinal samples from individuals
176 within a single batch, has two disadvantages. Firstly, analyses need to be postponed until all samples
177 from a single individual have been collected. Second, it requires detailed picking and reordering of
178 samples, which increases the likelihood of human error, sample mix-ups and therefore false
179 interpretations of the data. Although it seems intuitive that aggregating samples from a single
180 individual in the same batch should improve precision, the reduction in statistical power to detect

181 within-individual effects in long-term studies might be negligible relative to randomly allocating
182 samples to batches, with appropriate statistical methodology, but this has not been quantified.
183 Here, we present an approach to the analysis of samples from growing biobanks that, while
184 maintaining statistical independence, accounts for among-cluster variation and controls for other
185 potentially confounding effects (Fig. 1c). Additionally, we test the assumption that aggregating
186 longitudinal samples within batches results in higher statistical power to detect within-individual
187 effects. We then discuss the analysis of long-term data, highlight the importance of appropriate
188 statistical mixed models in these studies, and elaborate on potential biases and the use of long-term
189 data in a meta-analytical context. While we will mainly consider the field of evolutionary biology, using
190 telomere dynamics as an illustrative example, these considerations and techniques can be applied to
191 a range of fields, including epidemiology, ecology and laboratory-based science.
192

193 **Materials and Methods**

194 *Slicing approach*

195 We have developed a slicing approach to select samples from growing biobanks, such that recently
196 added samples are analysed in clusters together with previously obtained samples, ensuring statistical
197 independence of collection time and cluster. The biobank is divided into slices (Fig. 1c), where a slice
198 resembles a group of collectively gathered samples (e.g. in the same year) analysed together. Slicing
199 uses a varying proportion of samples from each given sampling period (i.e. slices), sequentially
200 analysed in a single cluster, to statistically account for temporal and cluster bias. Slicing therefore
201 combines convenient sequential analysis with the maintenance of statistical independence.
202 Depending on the frequency at which new samples are obtained, the ‘width’ of the slices can
203 be changed (Fig. 1c). For example, low analysis frequency requires wider slices to account for among-
204 cluster variation. This decision is directly related to the slicing ‘angle’ (Fig. 1c), which determines the
205 degree of independence of sampling year from cluster. For example, if there are environmental effects
206 related to the collection time of samples, slicing samples by collection time (i.e. lower angle) removes

207 possible confounds with cluster effects. For slicing to be effective across clusters, it requires multiple
208 years/cohorts to be present within a single cluster and at least one of those years/cohorts to be
209 present in a different cluster. Depending on slicing width and angle, a slice covers approximately one-
210 third of each batch, when slicing across three batches, with three separate slices covering the same
211 batch (Fig. 1c and see simulations). Such a strategy also naturally allocates samples of certain slices to
212 batches in subsequent clusters, bridging clusters together (Fig. 1c) and allowing control of among-
213 cluster variation. Setting the slicing angle and width is a trade-off between statistical independence
214 (assessing statistical power in the case of confounding effects) and the number of samples that remain
215 unanalysed until addition of newly collected samples. This latter point is a constraint, as the number
216 of samples that can be analysed simultaneously will be reduced, if only slightly, by this approach unless
217 samples are analysed multiple times. We argue that the creation of statistical independence and
218 accounting for among-cluster variation are merits that outweigh this limitation.

219 Growing biobanks that store samples sequentially, can easily apply the slicing selection
220 strategy by using a slicing width that ensures the analysis of multiple years and cohorts in a single
221 cluster. If the number of samples exceeds the preferred slicing width, additional clusters can be
222 analysed using the same layout (e.g. slicing a year/cohort multiple times but analysing them in
223 separate clusters with slices from other years/cohorts). This approach minimises errors due to sample
224 selection when samples are already stored sequentially, as picking samples at random across a wide
225 biobank can be both impractical (multiple boxes, freezers), bad for sample integrity (sorting through
226 many samples can risk defrosting) and prone to error (due to transcription or pipetting errors).

227 Slicing of newly acquired samples is similar to slicing from an existing biobank, but to prevent
228 confounding effects, analysis needs to be postponed until samples from multiple time periods have
229 been collected (e.g. seasons, years and cohorts). In both existing and growing biobanks, slicing over
230 the potential confounding variable (e.g. year/cohort) is essential to statistically separate among-
231 cluster variation and confounding effects.

232 The slicing strategy, however, cannot be applied to clusters under the common assumption
233 that longitudinal samples need to be aggregated in a single batch, because longitudinal samples are
234 generally collected in different fieldwork sessions and therefore analysed in different batches.
235 Combining the slicing of clusters with aggregation of longitudinal samples in a batches or
236 randomisation of samples across batches maintains the disadvantage that samples still need to be
237 picked, increasing the likelihood for human error. We therefore simulated different strategies to test
238 the assumption that aggregating longitudinal samples results in higher statistical power to detect
239 within-individual effects. Besides aggregation of longitudinal samples and random allocation of
240 samples among batches, we also simulated slicing across batches which allows convenient sequential
241 analysis of clusters and batches, as long as, for example, multiple years are analysed in a single cluster
242 and a single batch (Fig. 1c), while maintaining statistical independence.
243
244 *Simulations*
245 We determined statistical power and precision to detect individual and cohort effects, using different
246 sample allocation strategies (i.e. samples from a single individual aggregated in the same batch,
247 randomly allocated to batches, or 'sliced' across batches), with simulations run in R 3.3.1 (R
248 Development Core Team 2018, see supplementary Data S2).
249 We simulated a population of 200 individuals in 10 cohorts that were sampled once a year for
250 a maximum of 5 years. 'Telomere length' was used as an example response variable; however, this is
251 applicable to any longitudinally measured continuous variable. Starting telomere length was drawn
252 from a Gaussian distribution to fix between-individual standard deviation (SD = 1.00) and all
253 individuals shared the same within-individual shortening rate of telomeres (0.06*1, scaled to SD = 1
254 parameter, = 0.06 per year).
255 We simulated cohort effects (20 individuals per cohort) by taking the fraction (0.9) of
256 generated values drawn from a uniform distribution (between 0 and 1) and added these to the
257 response variable. We chose to model 'cohort' as a possible biological confound with experimentally

258 induced variation. The choice to model such specific biology is rather arbitrary as we are simulating
259 the confounding effect of ‘batch of analysis’ and biology. Individual probability of death was then
260 modelled via telomere length associated with mortality (Eq. 1),

261 $$y_i = 0.25^{(-0.23 * x_i)} \quad (1)$$

262 where *x* is initial telomere length for *i*th individual, with a baseline probability of death of 0.25 and a
263 slope of -0.23, providing mortality (y_i) per year drawn from a uniform distribution. This resulted in the
264 probability of death varying with ±2 SD telomere length from 0.14 to 0.36 per year. Mortality was
265 partly determined by the response variable (to simulate selective disappearance from the population,
266 determined by the between-individual age component, see next paragraph), with variable telomere
267 lengths to start with (between-individual variation) and a set within-individual shortening (within-
268 individual age component, see next paragraph).

269 We simulated the relationship between telomere length and age (in years) both within and
270 between individuals. Between-individual effects were modelled using the mean age at which the
271 individual’s trait was measured, and within-individual effects as the age at which an individual’s trait
272 was measured minus the mean measurement age for that individual (van de Pol & Wright 2009).
273 Simulations were run 5,000 times, for a varying number of samples (12, 24, 36, 48) per batch and
274 simulated differences between batch means (batch attributable error, SD: 1, 2.5, 5, 10, 20, 40). This
275 error is relatively high to ensure we control for potential effects of batch attributable error when
276 determining the variation in statistical power among sample allocation strategies. Simulations were
277 repeated three times to obtain three separate results per sample allocation strategy. The slicing
278 strategy was simulated at an angle that resulted in at least three slices per batch. Note, to start the
279 sample allocation, the first batch was filled by 2/3 with the first slice and by 1/3 with the second slice,
280 where subsequent batches were filled by 1/3 with subsequent slices (Fig. 1c). Additional simulations
281 were run with a doubled sample size (*n* = 400) to assess the effects on statistical power and precision
282 estimates among sample allocation strategies.

283 The simulated data were analysed using linear mixed models in lme4 1.1-14 (Bates *et al.* 2015),
284 where the model included random effects (at the intercept level) for individual (to control for repeated
285 measurements on the same individual) and batch, and cohort was fitted as a fixed factor. Parameters
286 of the simulations were manually optimised so that a power of approximately 0.5 was achieved to
287 detect either within- or between-individual effects for the random allocation strategy, determined by
288 a t-value of less than -2 ($\alpha \approx 0.05$). This intermediate level of statistical power avoids thresholding
289 effects at either end of the power spectrum (0 or 1). Such a simulation strategy maximises the
290 sensitivity in detecting any modulation of statistical power by the sample allocation strategy used,
291 which is our focus. It is important to understand the effect of sample allocation strategy on precision
292 estimates, as well as statistical power. We therefore quantified precision as the width of the
293 distribution of parameter estimates from the models run on the repeated simulated datasets, as the
294 absolute difference between the 75% and 25% percentile divided by the median (note, a precision
295 value closer to zero means higher precision).

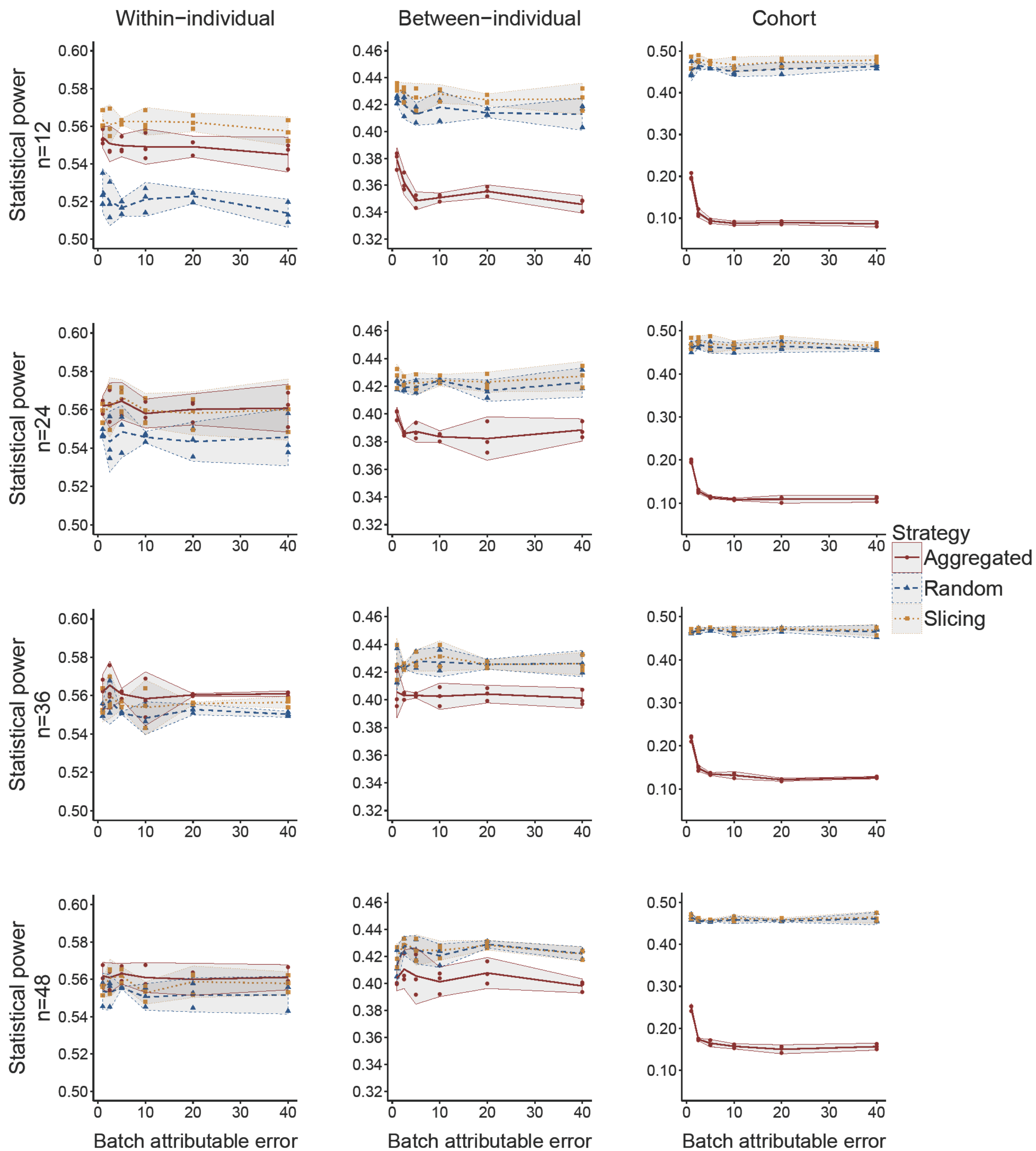
296

297 **Results**

298 Our simulations tested the widely held assumption that aggregating longitudinal samples of the same
299 individual in a single batch increases statistical power to detect within-individual effects (e.g. Herborn
300 *et al.* 2014; Nettle *et al.* 2015). The statistical power to detect within-individual effects was similar
301 when longitudinal samples were aggregated (mean statistical power ± SD across sample sizes and
302 three runs per simulation = 0.558 ± 0.007) and when samples were sliced (0.559 ± 0.006). It was
303 somewhat lower when samples were randomly allocated to batches, but only when batches were
304 small (i.e. n = 12; 0.543 ± 0.015; Fig. 2). For between-individual effects, however, the statistical power
305 was similar when samples were sliced across batches (0.426 ± 0.006), and randomly allocated to
306 batches (0.422 ± 0.007), but substantially lower (0.388 ± 0.021) when longitudinal samples were
307 aggregated in a single batch (Fig. 2). The statistical power to detect cohort effects was much higher
308 when samples were sliced across batches (0.469 ± 0.009) and randomly allocated to batches (0.461 ±

309 0.008) than when longitudinal samples were aggregated in a batch (0.140 ± 0.042; Fig. 2). A doubled
310 sample size (*n* = 400) increased statistical power but did not alter variation in statistical power among
311 sample allocation strategies (Fig. S1). The reduction in statistical power for aggregation of longitudinal
312 samples in a batch is explained by the confounding of the between-individual/cohort and batch effect.
313 Our slicing method outperforms both random allocation of samples and sample aggregation to
314 disentangle within- and between-individual effects, an objective shared by many longitudinal studies
315 (van de Pol & Wright 2009; Nussey *et al.* 2013).

316



317

318 **Fig. 2.** Statistical power analyses of simulated data for individual and cohort effects among a variety

319 of batch sizes (n=12 to 48) using three sample allocation strategies: (1) aggregating samples per

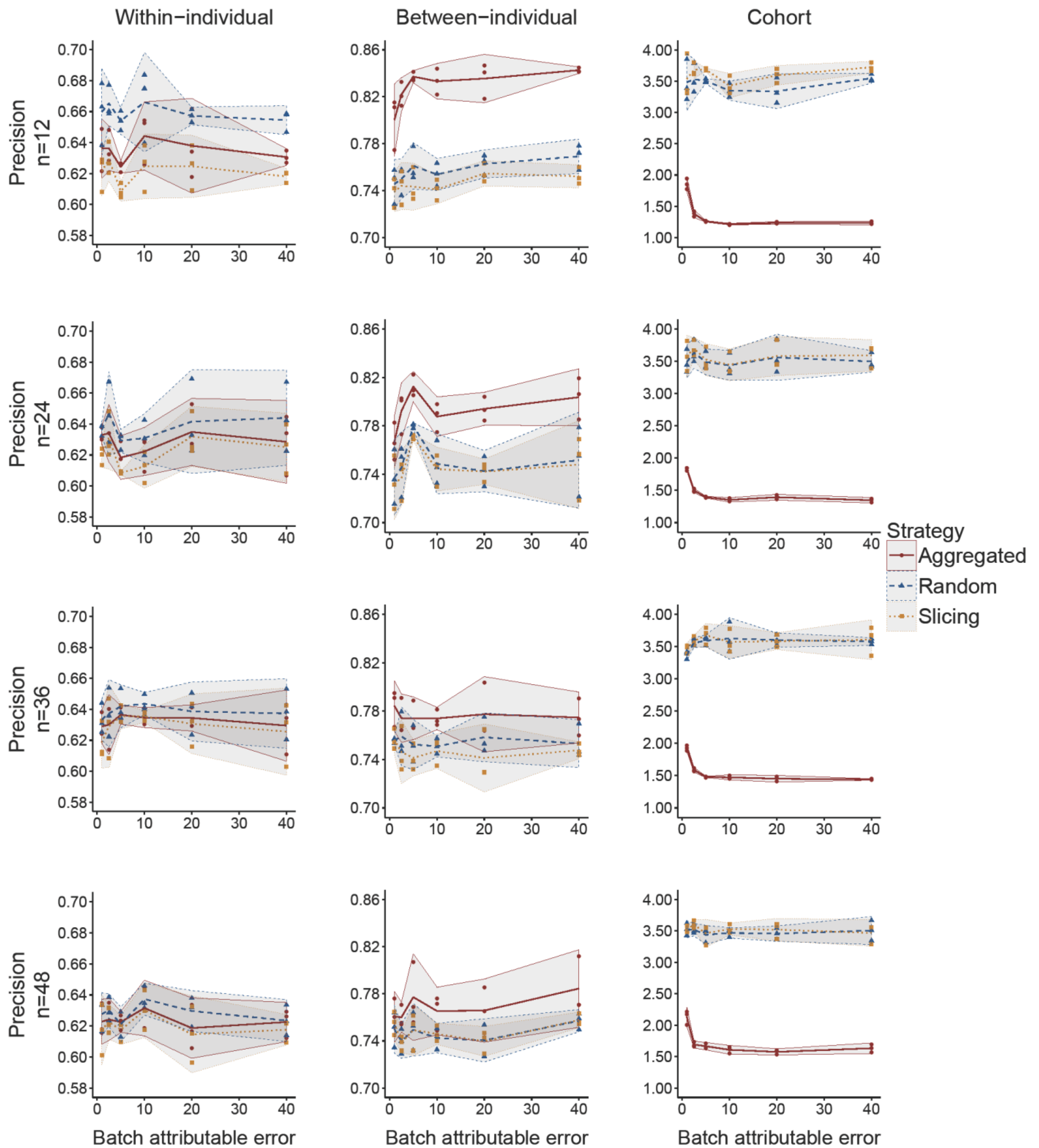
320 individual in the same batch (solid, red), (2) assigning samples randomly to batches (dashed, blue) or

321 (3) slicing samples across batches (dotted, yellow). Raw data points from three separate simulations
322 with mean statistical power per sample size are shown against among-batch variation. Statistical
323 power decreased in all simulations with increasing among-batch variation, as expected, and increased
324 when batch effects could be estimated more reliably with increasing numbers of samples per batch.
325 Scales differ between cohort, within- and between-individual effects.

326

327 The precision to estimate within-individual effects was similar when samples were sliced
328 across batches (mean precision ± SD across sample sizes and three runs per simulation = 0.623 ±
329 0.013), longitudinal samples were aggregated within batches (0.630 ± 0.012) and when samples were
330 randomly allocated to batches (0.641 ± 0.017; Fig. 3). For between-individual effects the precision was
331 also similar for slicing samples across batches (0.747 ± 0.013) and random allocation of samples to
332 batches (0.752 ± 0.015), but marginally more imprecise at a lower number of samples per batch for
333 aggregation of longitudinal samples in a single batch (0.792 ± 0.028; Fig. 3). Detection of cohort effects
334 was more precise when longitudinal samples were aggregated within batches (1.528 ± 0.234) than
335 when samples were randomly allocated to batches (3.505 ± 0.147) and sliced across batches (3.570 ±
336 0.151; Fig. 3). A doubled sample size (n = 400) increased precision but did not alter variation in
337 precision among sample allocation strategies (Fig. S2).

338 The slicing strategy we propose here thus performs similarly to randomising samples for
339 between-individual and cohort effects, but outperforms it on within-individual estimates. In general,
340 a strategy of aggregating samples does not outperform slicing or randomisation. Simulations were run
341 for a wide range of parameters and sample sizes (Figs 2, 3, S1, S2). When desirable, different
342 parameter sets specific to current or future datasets can be included in the script provided
343 (supplementary Data S2).



344

345 **Fig. 3.** Precision analyses of simulated data for individual and cohort effects among a variety of batch
346 sizes (n=12 to 48) using three sample allocation strategies: (1) aggregating samples per individual in
347 the same batch (solid, red), (2) assigning samples randomly to batches (dashed, blue) or (3) slicing
348 samples across batches (dotted, yellow). Raw data points from three separate simulations with mean

precision per sample size are shown against among-batch variation. Note, a precision value closer to zero means higher precision. Precision decreased in all simulations with increasing among-batch variation, as expected, and increased when batch effects could be estimated more reliably with increasing numbers of samples per batch. Scales differ between cohort, within- and between-individual effects.

Discussion

Our simulations clearly demonstrate that the slicing and randomisation strategies outperform aggregation of longitudinal samples in a batch, without any loss in statistical power or precision. The ability to reject the null hypothesis when false (i.e. statistical power) was marginally greater for within-individual effects when longitudinal samples were either aggregated within a batch or sliced across batches. Conversely, when samples were randomly allocated or sliced across batches, compared to aggregation of longitudinal samples within a batch, there was greater statistical power to detect between-individual effects.

The lack of variation in statistical power to detect within-individual effects between the slicing, random and aggregation allocation strategies was the consequence of appropriate statistical methodology, accounting for batch, individual and cohort through fixed and random effects. These results disprove the assumption that samples from a single individual need to be analysed in the same batch for greater statistical power to detect within-individual effects (Salomons *et al.* 2009; Beirne *et al.* 2014; Nettle *et al.* 2015). Such efforts will reduce the statistical power of the study and generate unnecessary effort in picking specific samples, which increases the likelihood of technical errors (e.g. sample mix-ups, freeze/thawing effects).

The notion that longitudinal samples should not be aggregated in the same batch becomes particularly pronounced when cohort effects apply. The effort of grouping samples from a single individual together collects cohorts together (an individual’s cohort is fixed) in a batch thus reducing the statistical power to distinguish between different cohorts. Random allocation of samples and

375 slicing have a substantially greater statistical power to detect cohort effects due to a higher mixture
376 of cohorts within the same batch. For telomere biology especially, estimating cohort effects reliably is
377 important as it affects telomere length quite strongly (Spurgin *et al.* 2017), but cohort effects are not
378 always estimated.

379 While it depends on specifications of the dataset, we conclude the patterns should hold in
380 general for the large majority of datasets. Thus, slicing batches provides optimal statistical power and
381 precision to detect individual and cohort effects, when combined with appropriate statistical
382 methodology (mixed model framework). Additionally, the benefits of applying slicing to clusters (i.e.
383 logistics, less error-prone, no uncontrollable variation among clusters) allows sequential analysis of
384 the biobank while ensuring statistical independence and accounting for among-cluster variation.
385 These benefits make slicing the preferred method for analysing longitudinal long-term datasets.

386

387 *Integral approach to growing biobank analysis*

388 The optimal sample selection strategy for analysing specific long-term datasets depends on the
389 structure of the data and questions of interest. However, in the majority of long-term datasets, slicing
390 provides a sustainable outcome. It overcomes the problems with differential timing in analysis of
391 clusters, which commonly occur in growing biobanks from long-term studies, allowing separate
392 analysis of current data and flexible inclusion of this into future analyses. Furthermore, slicing allows
393 sequential analysis of samples, preventing complicated sample labelling and placing among clusters
394 and therefore reducing the potential for human error. However, the assumption that longitudinal
395 samples should be aggregated in a single batch could hinder the slicing approach, but simulations in
396 this study have disproven this assumption. Thus, slicing is the selection strategy with optimal statistical
397 power to detect individual and cohort effects. Although slicing occasionally requires postponed
398 analysis (i.e. part of the biobank needs to stay in place to secure statistical independence), the higher
399 accuracy and comparability of analyses over time likely outweigh this limitation for the large majority
400 of datasets.

401 Even though slicing minimises the effects of among-cluster variation and confounding effects,
402 the continued use of appropriate internal controls is advised. The extent to which internal controls
403 account for among-batch variation remains questioned, as normalisation is based on a single
404 reference sample (Cawthon 2002). The single reference sample point estimate, to which every other
405 sample is corrected, has the same error as all the other samples and therefore inflates the error
406 estimates. Although comparing samples among batches relies on the usage of this same, repeatedly
407 analysed reference sample where large volumes are required, the combination with slicing increases
408 accuracy and accounts for among-batch variation and internal controls should therefore be included
409 in all batches. The inclusion of internal controls can at least alert the user of sudden changes in the
410 analysis setup and are needed in some forms of analysis, mainly those where a dilution curve is
411 required (e.g. qPCR, antibody-based hormone analysis).
412 In addition to internal controls, the use of statistical mixed models (Bolker *et al.* 2009) is
413 required to analyse samples collected in long-term studies because, for example, storage duration is
414 not accounted for by internal controls. Storage duration and batch effects are often recognised, but
415 rarely modelled using appropriate methodology. Long-term data should therefore be analysed in
416 mixed models, while including fixed and random effects (Bolker *et al.* 2009), to account for storage
417 duration, batch effects and other potentially confounding effects with cluster. The failure to include
418 these effects can inflate type I and type II errors when there is a temporal, spatial or other spurious
419 correlation with any independent variable.
420
421 *External variation and meta-analyses*
422 Long-term studies usually span multiple years or decades in which changes in equipment (e.g.
423 machines) or analyst among clusters can result in external variation that could compromise the
424 integrity of the data (e.g. Reichert *et al.* 2017). Consistency in the analysis of samples in long-term
425 studies (i.e. slicing) and testing for external variation allows for temporal intra-biobank comparisons

426 and flexible inclusion of data in future analyses, while accounting for among-cluster variation and
427 potential confounding effects.

428 Consistency in the analysis of growing biobanks also allows for inter-biobank comparisons
429 when comparing different studies in, for example, meta-analyses. These comparisons, however,
430 require caution as analyses are often conducted in different laboratories, resulting in inconsistencies
431 in the approaches to the analysis and high external variation. These effects do not necessarily affect
432 within-study effect sizes and their comparability, although differences in measurement error (i.e.
433 repeatability) will (see discussion in Simons, Cohen & Verhulst 2012). Caution is particularly warranted
434 for comparative studies as some methods are inherently not comparable between studies on a
435 continuous scale. For example, some methods are always compared to a standard, and are thus always
436 relative (e.g. qPCR). Such methods become non-comparable as the standard that samples are
437 compared to is not the same across studies (usually a pooled or reference sample). Conclusions
438 derived from the data can therefore be compared quantitatively across studies (e.g. by summarising
439 effect sizes across studies), but the raw data cannot be compared across species (i.e. comparing
440 telomere lengths as estimated by qPCR; Tricola *et al.* 2018; Wilbourn *et al.* 2018). Thus, because
441 telomere lengths estimated by qPCR are scaled to different internal controls and control genes,
442 samples between species and studies cannot be compared in absolute terms, but are comparable in
443 terms of effect size in a meta-analytic context (Tricola *et al.* 2018; Wilbourn *et al.* 2018). The several
444 levels of external variation (e.g. internal controls, reagents or analyst) can be severe and affect effect
445 sizes through differences in measurement error (Simons, Cohen & Verhulst 2012), as well as induce
446 bias. Meta-analyses should therefore be interpreted with caution and should explore moderators for
447 the methodology used where possible (Nakagawa & Santos 2012). Awareness of methodology,
448 internal controls and approach to analysis are important in determining the potential for comparison
449 among studies, populations and species.

450

451 **Conclusions**

452 A major current challenge in long-term studies is analysing data as it is collected while also being able
453 to include it in future analyses, without creating uncontrollable variation, allowing comparison of
454 results over multiple years or even decades. This requires the ability to compare differentially timed
455 analyses that are potentially biased by confounding cluster effects. This study shows the importance
456 of considering the structure of samples among clusters and batches in long-term studies. The slicing
457 approach proposed here retains statistical independence and accounts for among-cluster variation in
458 the sequential analysis of growing biobanks. This approach is further characterised by optimal
459 statistical power and precision to detect cohort, within- and between-individual effects, if analysed
460 using appropriate internal controls, statistical mixed models and consistent methodology to control
461 for confounding effects. A single sample's scientific value increases through this approach, as it can be
462 used separately in current studies, but can also be included in subsequent studies, providing
463 sustainable (re)use of collected data. The approach we propose here (slicing, internal controls, mixed
464 models) is easy to apply and improves the potential for these growing biobanks to address important
465 ecological and evolutionary questions.
466
467 **Acknowledgements**
468 The authors gratefully acknowledge feedback on an earlier version of the manuscript from Dan
469 Nussey. This work was supported by a Leeds Anniversary Research Scholarship to S.H.J.v.L., a Sir Henry
470 Wellcome (WT107400MA) Fellowship and a University of Sheffield Vice-Chancellor's Fellowship to
471 M.J.P.S., a NERC grant NE/J024597/1 to T.B. and J.S., and a NERC grant (NE/N013832/1) to T.B. and
472 M.J.P.S.
473
474 **Authors' contributions**
475 Conception/design: S.H.J.v.L., J.S., T.B., M.J.P.S., H.L.D. Methodology: S.H.J.v.L., M.J.P.S. Analysis and
476 interpretation: S.H.J.v.L, M.J.P.S., H.L.D. Drafting/Revising paper: S.H.J.v.L., H.F., J.S., T.B., M.J.P.S.,
477 H.L.D. All authors contributed critically to the manuscript and gave final approval for publication.

478

479 **Data Accessibility**

480 Supplementary figures and the R-script for simulations will be uploaded as online supporting
481 information upon acceptance, this paper does not include any empirical data.

482

483 **References**

484 Andraud, M., Casas, M., Pavio, N. & Rose, N. (2014). Early-life hepatitis E infection in pigs: The
485 importance of maternally-derived antibodies. *PLoS ONE, 9*, e105527.
486 https://doi.org/10.1371/journal.pone.0105527
487 Bates, D., Machler, M., Bolker, B.M. & Walker, S.C. (2015). Fitting linear mixed-effects models using
488 lme4. *Journal of Statistical Software, 67*, 1-48. https://doi.org/10.18637/jss.v067.i01
489 Beirne, C., Delahay, R., Hares, M. & Young, A. (2014). Age-related declines and disease-associated
490 variation in immune cell telomere length in a wild mammal. *PLoS ONE, 9*, e108964.
491 https://doi.org/10.1371/journal.pone.0108964
492 Berry, O., England, P., Marriott, R.J., Burridge, C.P. & Newman, S.J. (2012). Understanding age-specific
493 dispersal in fishes through hydrodynamic modelling, genetic simulations and microsatellite
494 DNA analysis. *Molecular Ecology, 21*, 2145-2159. https://doi.org/10.1111/j.1365-
495 294X.2012.05520.x
496 Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S.
497 (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends*
498 *in Ecology & Evolution, 24*, 127-135. https://doi.org/10.1016/j.tree.2008.10.008
499 Boonekamp, J.J., Mulder, G.A., Salomons, H.M., Dijkstra, C. & Verhulst, S. (2014). Nestling telomere
500 shortening, but not telomere length, reflects developmental stress and predicts survival in
501 wild birds. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20133287.
502 https://doi.org/10.1098/rspb.20133287
503 Cawthon, R.M. (2002). Telomere measurement by quantitative PCR. *Nucleic Acids Research, 30*, e47.
504 https://doi.org/10.1093/nar/30.10.e47
505 Clutton-Brock, T. & Sheldon, B.C. (2010). Individuals and populations: the role of long-term, individual-
506 based studies of animals in ecology and evolutionary biology. *Trends in Ecology & Evolution,*
507 *25*, 562-573. https://doi.org/10.1016/j.tree.2010.08.002
508 Dugdale, H.L., Pope, L.C., Newman, C., Macdonald, D.W. & Burke, T. (2011). Age-specific breeding
509 success in a wild mammalian population: selection, constraint, restraint and senescence.
510 *Molecular Ecology, 20*, 3261-3274. https://doi.org/10.1111/j.1365-294X.2011.05167.x
511 Fairlie, J., Holland, R., Pilkington, J.G., Pemberton, J.M., Harrington, L. & Nussey, D.H. (2016). Lifelong
512 leukocyte telomere dynamics and survival in a free-living mammal. *Aging Cell, 15*, 140-148.
513 https://doi.org/10.1111/acel.12417
514 Froy, H., Bird, E.J., Wilbourn, R.V., Fairlie, J., Underwood, S.L., Salvo-Chirnside, E., . . . Nussey, D.H.
515 (2017). No evidence for parental age effects on offspring leukocyte telomere length in free-
516 living Soay sheep. *Scientific Reports, 7*, 9991. https://doi.org/10.1038/s41598-017-09861-3
517 Gaillard, J.M., Allainé, D., Pontier, D., Yoccoz, N.G. & Promislow, D.E.L. (1994). Senescence in natural
518 populations of mammals - a re-analysis. *Evolution, 48*, 509-516.
519 https://doi.org/10.1111/j.1558-5646.1994.tb01329.x
520 Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New
521 York, Cambridge: Cambridge University Press.

522 Hammers, M., Kingma, S.A., Bebbington, K., van de Crommenacker, J., Spurgin, L.G., Richardson, D.S.,
523 . . . Komdeur, J. (2015). Senescence in the wild: Insights from a long-term study on Seychelles
524 warblers. *Experimental Gerontology, 71*, 69-79. https://doi.org/10.1016/j.exger.2015.08.019
525 Herborn, K.A., Heidinger, B.J., Boner, W., Noguera, J.C., Adam, A., Daunt, F. & Monaghan, P. (2014).
526 Stress exposure in early post-natal life reduces telomere length: an experimental
527 demonstration in a long-lived seabird. *Proceedings of the Royal Society B: Biological Sciences,*
528 *281*, 20133151. https://doi.org/10.1098/rspb.2013.3151
529 Nakagawa, S. & Santos, E.S.A. (2012). Methodological issues and advances in biological meta-analysis.
530 *Evolutionary Ecology, 26*, 1253-1274. https://doi.org/10.1007/s10682-012-9555-5
531 Nettle, D., Monaghan, P., Gillespie, R., Brilot, B., Bedford, T. & Bateson, M. (2015). An experimental
532 demonstration that early-life competitive disadvantage accelerates telomere loss.
533 *Proceedings of the Royal Society B: Biological Sciences, 282*, 20141610.
534 https://doi.org/10.1098/rspb.2014.1610
535 Nordfjall, K., Larefalk, A., Lindgren, P., Holmberg, D. & Roos, G. (2005). Telomere length and heredity:
536 Indications of paternal inheritance. *Proceedings of the National Academy of Sciences of the*
537 *United States of America, 102*, 16374-16378. https://doi.org/10.1073/pnas.0501724102
538 Nussey, D.H., Baird, D., Barrett, E., Boner, W., Fairlie, J., Gemmell, N., . . . Monaghan, P. (2014).
539 Measuring telomere length and telomere dynamics in evolutionary biology and ecology.
540 *Methods in Ecology and Evolution, 5*, 299-310. https://doi.org/10.1111/2041-210x.12161
541 Nussey, D.H., Coulson, T., Festa-Bianchet, M. & Gaillard, J.M. (2008). Measuring senescence in wild
542 animal populations: towards a longitudinal approach. *Functional Ecology, 22*, 393-406.
543 https://doi.org/10.1111/j.1365-2435.2008.01408.x
544 Nussey, D.H., Froy, H., Lemaitre, J.F., Gaillard, J.M. & Austad, S.N. (2013). Senescence in natural
545 populations of animals: Widespread evidence and its implications for bio-gerontology. *Ageing*
546 *Research Reviews, 12*, 214-225. https://doi.org/10.1016/j.arr.2012.07.004
547 R Development Core Team (2018). R: a language and environment for statistical computing. R
548 foundation for statistical computing, Vienna
549 Reichert, S., Froy, H., Boner, W., Burg, T.M., Daunt, F., Gillespie, R., . . . Monaghan, P. (2017). Telomere
550 length measurement by qPCR in birds is affected by storage method of blood samples.
551 *Oecologia, 184*, 341-350. https://doi.org/10.1007/s00442-017-3887-3
552 Rius-Ottenheim, N., Houben, J.M.J., Kromhout, D., Kafatos, A., van der Mast, R.C., Zitman, F.G., . . .
553 Giltay, E.J. (2012). Telomere length and mental well-being in elderly men from the
554 Netherlands and Greece. *Behavior Genetics, 42*, 278-286. https://doi.org/10.1007/s10519-
555 011-9498-6
556 Salomons, H.M., Mulder, G.A., van de Zande, L., Haussmann, M.F., Linskens, M.H.K. & Verhulst, S.
557 (2009). Telomere shortening and survival in free-living Corvids. *Proceedings of the Royal*
558 *Society B: Biological Sciences, 276*, 3157-3165. https://doi.org/10.1098/rspb.2009.0517
559 Simons, M.J.P., Cohen, A.A. & Verhulst, S. (2012). What does carotenoid-dependent coloration tell?
560 Plasma carotenoid level signals immunocompetence and oxidative stress state in birds - a
561 meta-analysis. *PLoS ONE, 7*, e43088. https://doi.org/10.1371/journal.pone.0043088
562 Spurgin, L.G., Bebbington, K., Fairfield, E.A., Hammers, M., Komdeur, J., Burke, T., . . . Richardson, D.S.
563 (2017). Spatio-temporal variation in lifelong telomere dynamics in a long-term ecological
564 study. *Journal of Animal Ecology, 87*, 187-198. https://doi.org/10.1111/1365-2656.12741
565 Telfer, S., Birtles, R., Bennett, M., Lambin, X., Paterson, S. & Begon, M. (2008). Parasite interactions in
566 natural populations: insights from longitudinal data. *Parasitology, 135*, 767-781.
567 https://doi.org/10.1017/S0031182008000395
568 Tollenaere, C., Ivanova, S., Duplantier, J.M., Loiseau, A., Rahalison, L., Rahelinirina, S. & Brouat, C.
569 (2012). Contrasted patterns of selection on MHC-linked microsatellites in natural populations
570 of the Malagasy plague reservoir. *PLoS ONE, 7*, e32814.
571 https://doi.org/10.1371/journal.pone.0032814

572 Tricola, G.M., Simons, M.J.P., Atema, E., Boughton, R.K., Brown, J.L., Dearborn, D.C., . . . Haussmann,
573 M.F. (2018). The rate of telomere loss is related to maximum lifespan in birds. *Philosophical*
574 *Transactions of the Royal Society B: Biological Sciences, 373*, 20160445.
575 https://doi.org/10.1098/rstb.2016.0445
576 van de Pol, M.V. & Wright, J. (2009). A simple method for distinguishing within- versus between-
577 subject effects using mixed models. *Animal Behaviour, 77*, 753-758.
578 https://doi.org/10.1016/j.anbehav.2008.11.006
579 Wilbourn, R.V., Moatt, J.P., Froy, H., Walling, C.A., Nussey, D.H. & Boonekamp, J.J. (2018). The
580 relationship between telomere length and mortality risk in non-model vertebrate systems: a
581 meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*,
582 20160447. https://doi.org/10.1098/rstb.2016.0447

583

584 **Supporting information**

585 **Supplementary figures S1:** Additional figures for power and precision estimates

586 **Data S2**: R-script for simulations

587

588 # Supporting information

589 **Have your slice and eat it too: a sustainable approach to the analysis of long-term biobanks**

590 *Sil H.J. van Lieshout, Hannah Froy, Julia Schroeder, Terry Burke, Mirre J.P. Simons, Hannah L. Dugdale*

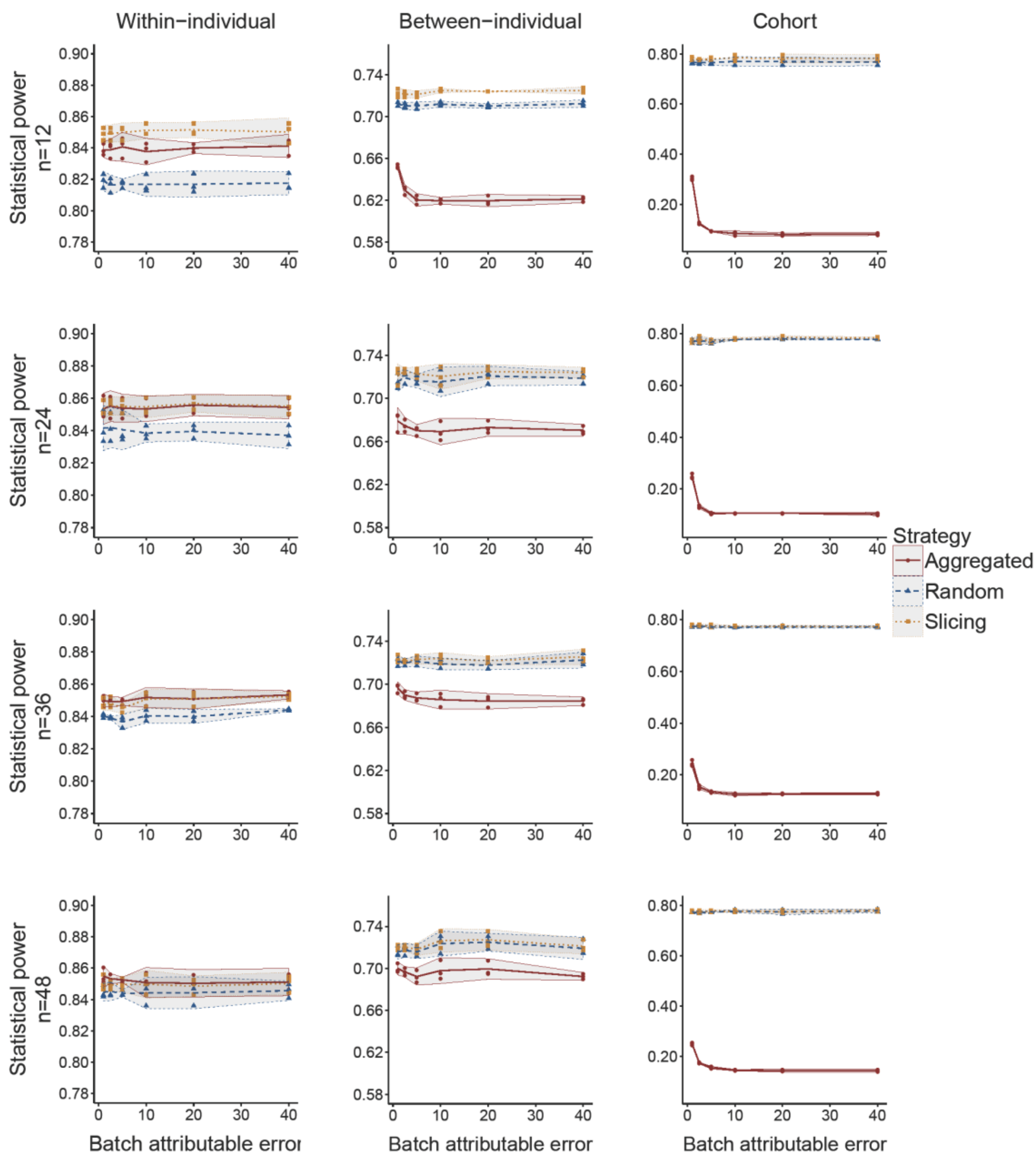
591

592 This document includes two supplementary figures, comprising power and precision estimates for

593 simulations with increased sample size (n = 400).

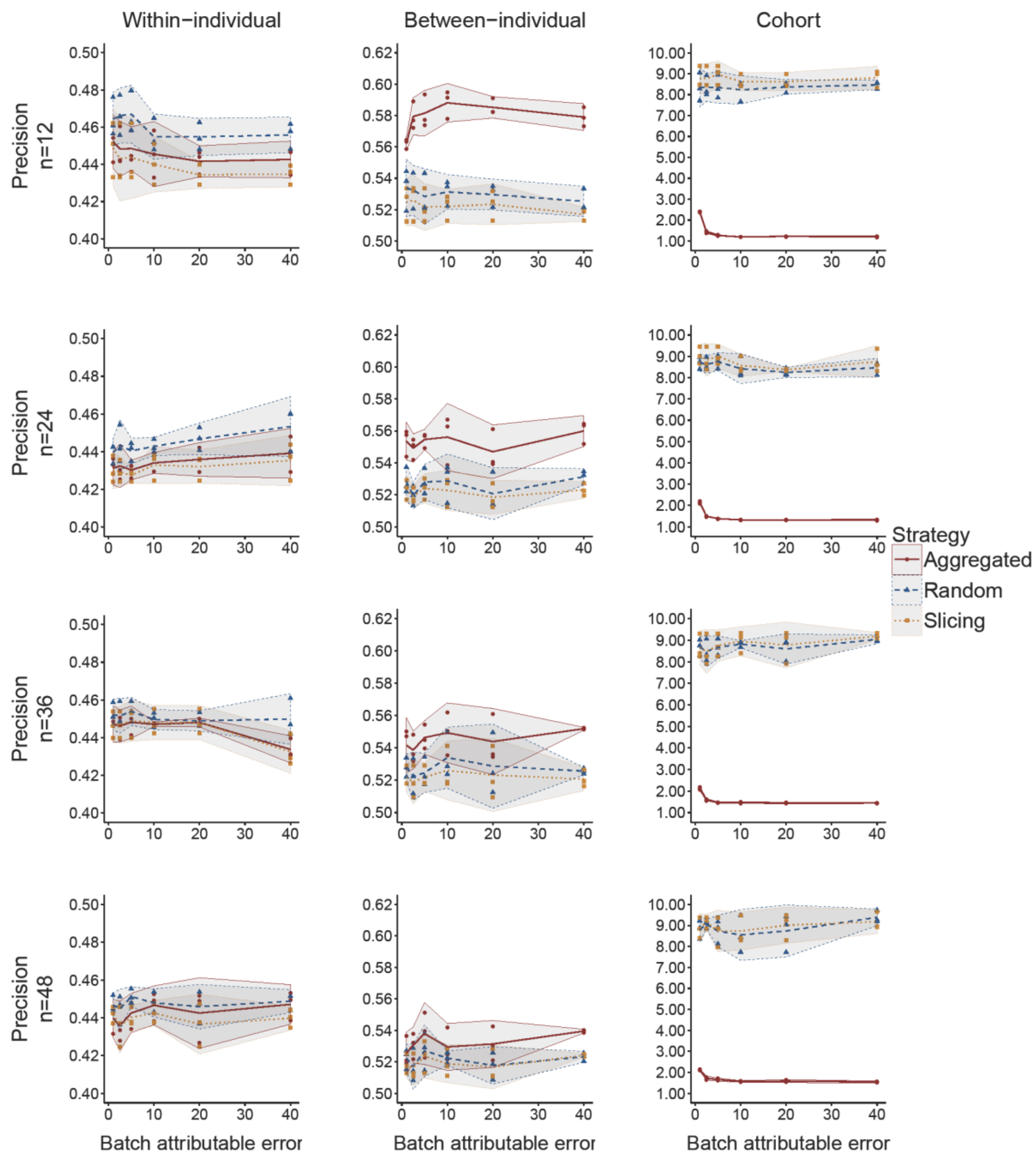
594

595 **Supplementary figures**



596

597 **Fig. S1.** Statistical power analyses (*n* = 400) for individual and cohort effects among a variety of batch
598 sizes (n=12 to 48) using three sample allocation strategies: (1) aggregating samples per individual in
599 the same batch (solid, red), (2) assigning samples randomly to batches (dashed, blue) or (3) slicing
600 samples across batches (dotted, yellow). Raw data points from three separate simulations with mean
601 statistical power per sample size are shown against among-batch variation. Statistical power
602 decreased in all simulations with increasing among-batch variation, as expected, and increased when
603 batch effects could be estimated more reliably with increasing numbers of samples per batch. Scales
604 differ between cohort, within- and between-individual effects.
605



606

607 **Fig. S2.** Precision analyses (*n* = 400) for individual and cohort effects among a variety of batch sizes
608 (n=12 to 48) using three sample allocation strategies: (1) aggregating samples per individual in the
609 same batch (solid, red), (2) assigning samples randomly to batches (dashed, blue) or (3) slicing samples
610 across batches (dotted, yellow). Raw data points from three separate simulations with mean precision
611 per sample size are shown against among-batch variation. Note, a precision value closer to zero means
612 higher precision. Precision decreased in all simulations with increasing among-batch variation, as

613 expected, and increased when batch effects could be estimated more reliably with increasing numbers

614 of samples per batch. Scales differ between cohort, within- and between-individual effects.

615