Effective study design for comparative functional genomics

Joanna L. Kelley[1] *and Yoav Gilad*[2]✉

Comparative studies struggle with the challenge of balancing technical properties with the requirement of obtaining samples from multiple species. Guidelines are needed to minimize the effect of confounders and increase the robustness of the inference. We argue for the paramount importance of extensive record keeping and reporting.

Comparative studies collect data and compare observations across multiple species. The data to be compared can be of any type, from physiological measurements, to behavioural observations, to molecular, genetic or functional genomic data. Such studies have traditionally addressed questions related to the environmental and evolutionary pressures that have shaped similarities and differences between species. When humans are included, comparative studies offer a way to empirically address the classic question: what makes us human? For example, by identifying genomic regions that have rapidly evolved exclusively along the human lineage, these studies can explain the basis for some human-specific traits and diseases[1]. Conversely, comparative genomic approaches have helped identify conserved and thus putatively functional genomic regions, which can have gene regulatory effects that modulate disease risk, and are thus useful for addressing outstanding questions relevant to human health.

In the era of cheap and rapid sequencing, comparative genomic approaches have gained momentum, and functional genomic data can now be collected from practically any species. Many inter-species differences are large and can be distinguished easily even with a small sample size. Perhaps because inter-species differences tend to often be conspicuous, comparative studies have not always adhered to common good practices with respect to study design, data collection and analysis. Large inter-species differences notwithstanding, it is highly unlikely that a sample of just one or two individuals can faithfully represent an entire population or species. Despite this caveat, comparative studies that report data from just a single sample from each species are still being published, albeit rarely.

As collecting suitable samples from non-laboratory animals is often challenging, comparative studies generally use modest sample sizes of four to a dozen individuals from each species. With such a small sample size, there is still a danger that inter-species differences are not faithfully represented, although one study that specifically addressed this question found that inter-primate differences can be reliably identified with a sample of half a dozen individuals from each species[2]. Nevertheless, special care must be taken to avoid confounding factors and biased study design, because even one or two unusual samples — that may have only minimal impact when the sample size is to the order of hundreds or thousands of samples — can have a profound effect in studies with small sample sizes.

Consider a comparative genomic study seeking to characterize gene regulation across multiple tissues and species. If different tissues were sampled across different individuals from the same species, the observed variance in gene regulation due to the tissue of origin and the individual donor would be completely confounded. This common confounder also affects studies of single species; in fact, even the Genotype-Tissue Expression (GTEx) Consortium study[3] was unable to sample all tissue types from the same individual donors. However, in large studies (with a few hundred samples from each tissue), the confounding of tissue and individual has a small effect on the observed regulatory differences between tissues, because ultimately most of the individual variance is random with respect to the sampled tissue. By contrast, for a small comparative study with only half a dozen samples from each tissue–species combination, it has been shown that the confounding variance due to the individual can have a significant impact on the observed regulatory difference between tissues[4].

Many other factors besides sample size must be considered when designing a comparative genomic study, including balancing potential confounders, such as age and sex, and avoiding batch effects[5]. One nearly inescapable difficulty of comparative studies is that samples from different species are obtained independently, almost always on different days, at different sites and by different people. This confounds species with sample collection batch, which cannot be effectively accounted for. Whenever possible, it is important to estimate the magnitude of such confounders by replicating the batch properties within each species; this is another reason that comparisons involving one or two individuals from

[1]*School of Biological Sciences, Washington State University, Pullman, WA, USA.*

[2]*Department of Medicine, University of Chicago, Chicago, IL, USA.*

✉*e-mail: gilad@uchicago.edu*

https://doi.org/10.1038/s41576-020-0242-z

each species are unlikely to provide faithful representation of inter-species differences. When it is impossible to replicate the batch conditions within a species, one must acknowledge this limitation explicitly when reporting the study.

Numerous properties can contribute to batch effects in comparative studies, and our experience suggests that most of these are not typically recorded and reported. Our general recommendation is to record as many details as possible on the origin of the sample, including age and sex of the donor, as well as every step of sample processing — from the moment the sample is obtained (noting how long postmortem the sample was sourced) until the final data are collected — and make these metadata available as part of the supplementary material for any relevant manuscript. Access to these data will allow investigators to identify confounders and, when possible, estimate their effects. Importantly, the sample record data often allow investigators to exclude potential confounders as possible explanations for observed inter-species differences and thus demonstrate that their conclusions are robust.

Another important consideration is that comparative studies sometimes collect samples from wild populations. Among the many inherent limitations to field sampling is the fact that populations with environmental exposures of interest (for example, populations in hydrogen sulfide-rich springs) are almost always sampled once, often on a single day. This approach makes it impossible to disentangle random events from the feature of interest. This limitation needs to be explicitly acknowledged, because reported conclusions might not be robust. To establish robustness, field studies need to consider sampling across multiple non-consecutive days, seasons or years. Of course, this approach is not always feasible when sampling requires special access to a field site (for example, use of a helicopter, extensive hiking or 1-day permits). In such cases, it is important to record environmental variables as thoroughly as possible so that they can be considered in subsequent analyses. Unfortunately, weather conditions and other similar types of event are only rarely recorded and reported[6]. Given that it is hard to know which environmental variables to measure, we recommend measuring and reporting as many variables as possible. Certain variables are often shared across sites, days or seasons, and keeping records helps to identify such instances and account for these factors.

Sample storage and transport are other potentially important sources of variation. Preservation methods (for example, liquid nitrogen versus other freezing methods) should not be mixed within a study, and manufacturer protocols should be consistently followed. The use of liquid nitrogen is not always practical in the field. Dry shippers, which enable the preservation and movement of samples in liquid nitrogen, are heavy and hold a limited number of samples. Moreover, the transport of dry shippers has become more restrictive in recent years. It is essential to consider sample storage and transportation over the course of a project to ensure that sampling will be robust to changes in infrastructure, including permits and airline restrictions. When the introduction of a confounder is unavoidable (for example, the day of collection or RNA quality), it is often possible to design a separate study to explicitly measure the variance associated with the confounder. Pilot studies not only allow us to estimate the magnitude of the confounder but also indicate the most effective strategy when forced to choose between imperfect designs.

A common error in study design is the non-random processing of samples once back in the laboratory. It is certainly tempting to start sample processing (for example, RNA extraction) before the completion of field sampling, especially when field sampling occurs over many months. However, it is imperative to think carefully about the timing of sample processing. In general, differences in proper storage times are associated with far less variation than independent processing of samples[7,8]. That said, some comparative studies are carried out over years, sometimes decades. We do not propose to routinely wait for years before samples are processed, but care should be taken to minimize processing-related batch effects as much as possible.

The properties and challenges of an effective study design are not unique to comparative functional genomics. However, because confounders can have a greater impact given the typically small sample sizes of such studies, it is both difficult and important to consider these properties. Moreover, the collection and processing of samples from different species creates multiple opportunities for sample properties to become confounded with species identity. From avoiding or accounting for batch effects, to randomizing sample processing, to establishing an unbiased analysis pipeline — although challenging, these are critical considerations for effective and robust comparative genomic studies.

1. Holloway, A. K., Bruneau, B. G., Sukonnik, T., Rubenstein, J. L. & Pollard, K. S. Accelerated evolution of enhancer hotspots in the mammal ancestor. *Mol. Biol. Evol.* **33**, 1008–1018 (2016).
2. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).
3. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
4. Blake, L. E. et al. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Res.* **30**, 250–262 (2020).
5. Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Res.* **4**, 121 (2015).
6. Deck, J. et al. The Genomic Observatories Metadatabase (GeOMe): a new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol.* **15**, e2002925 (2017).
7. Kim, S. J. et al. Effects of storage, RNA extraction, genechip type, and donor sex on gene expression profiling of human whole blood. *Clin. Chem.* **53**, 1038–1045 (2007).
8. Kelly, R. et al. RNA and DNA integrity remain stable in frozen tissue after long-term storage at cryogenic temperatures: a report from the ontario tumour bank. *Biopreserv. Biobank* **17**, 282–287 (2019).

Acknowledgements
The authors thank N. Gonzales, W. Dowd, M. Tobler, L. Barreiro, A. Pai, L. Blake and J. Tung for helpful comments on the manuscript. Y.G is. supported by NIGMS MIRA R35GM131726; J.L.K. is supported by NSF 1557795, 1906015 and 1931650.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.