**PASS1A Analysis Plan (9/21/2021)**

This plan includes Data Inventory, Data Cleaning, and Temporal Modeling (i.e., Data Analysis). For considerations to incorporate PASS1C data when relevant, see 3D.

**1. Data Inventory**

**1A.** At the highest level, the inventory is a table-form manifest of all the datasets, arranged as a grid of Tissues by Data Types.

Some omics data types, such as the metabolomics data, are further divided by Targeted and Untargeted, with the latter divided into Named and Unnamed.

Datasets of a given data type and from different tissues of the same animals constitute a **Data Group**. Each Data Group is represented by its own grid of Tissue and Site/Platform.

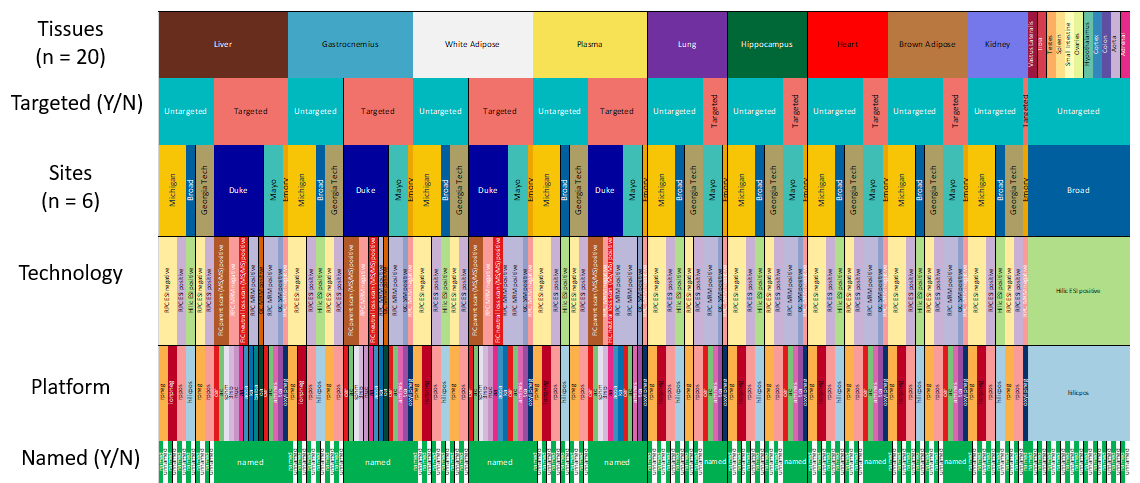
**1B**. Each dataset is a sample-feature data matrix. The samples are Bioanalytes, each linked to its Sample Metadata such as experimental batch and RNA quality measures.

**1C**. Animal metadata - not to be confused with the sample metadata - include sex, age, and exercise group of the animals (such as IPE, IPE-1h, etc.). Multiple tissues are collected from each animal, thus each animal maps to many tissue samples.

Currently, MoTrPAC adopts a metadata file format that merges animal metadata into sample metadata, essentially repeating the same animal attributes over all the tissues derived from each animal. In a different, more structured database schema, sample and animal metadata would be separated.

**1D**. The importance of having a reliable and complete data inventory is to ensure that discussions about data cleaning, data analysis (methods, results, and interpretations) can be made specifically for the tissue and the data type. A generic discussion without specifying tissue and datatype may lead to confusion.

An example of PASS1A metabolomic data, with 205 datasets, is below:



**2. Data Cleaning**

**2A**. Quality assessment starts with initial tasks to identify potential outliers and experimental batches and, if found, make decisions about how to address them. These procedures may be intertwined with normalization approaches.

**2B**. Data cleaning decisions must be made for each tissue and each data type separately. We will keep full documentation.

**2C**. The reason for separating Data Cleaning and Temporal Modeling (see below) is that they call for somewhat different skillsets. The former will deliver a data freeze without being locked in by a certain analysis task, such as a specific set of differential expression results. The latter involves declared analysis goals and plans, with the understanding that different stories can be built from the same data. Mixing data coordination with data analysis, or mixing analysis output with interpretation and story-telling, can lead to confusion.

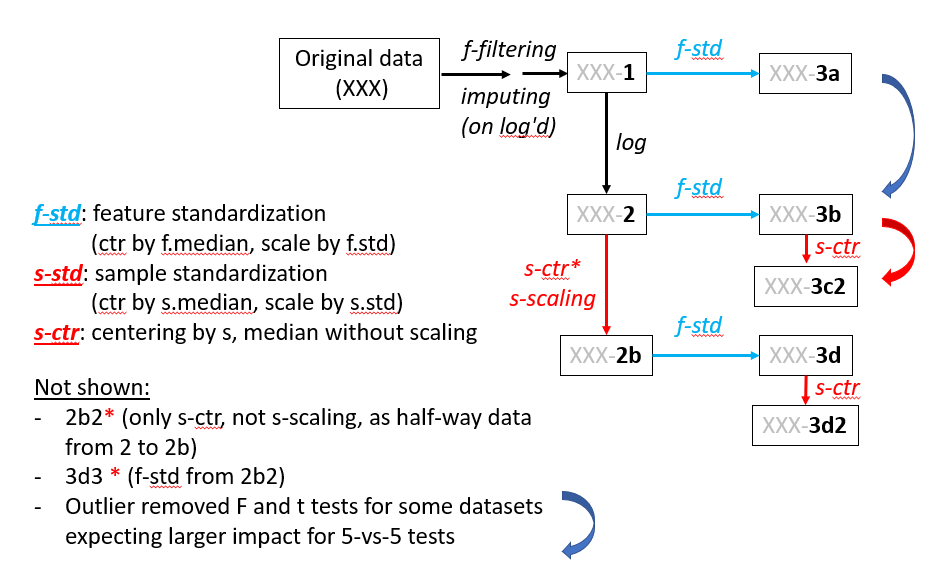
**2D**. Whether the outliers or batch effects are "big" or "small" depends on the empirically observed magnitude of these effects in comparison to the biological effects of interest, for that tissue and that data type. In this regard the data cleaning decisions can be indirectly related to the next task (3. Temporal modeling).

However, a measure of reproducibility can be established early, by (1) comparing repeated experiments on the same samples (when available) and (2) reporting within-group and between-group variance, without relying on the number of significant genes for a specific between-group contrast. We want to reduce the reliance on using DE genes in a prescribed comparison to inform data cleaning decisions.

**2E**. MoTrPAC has gained recent experience for cleaning metabolomic datasets by working through PASS1B (example in the figure below). For transcriptomic and proteomic data we may need to set up new procedures, or at least write out the plan.

During data cleaning for each dataset, there are opportunities to compare across tissues and across sites/platforms. The strategies vary by dataset and depend on the measurement technology and the use of different kinds of reference samples.

**2F**. Sample QC and feature QC are related processes and may need to be done iteratively until the results converge. For instance, removing outlier samples may alter the definition of outlier features, and vice versa. Batch effects in samples are usually driven by a set of batch-driving features, and vice versa.

**2G.** Features are often evaluated and filtered by (1) mean and variance - whether it has sufficient information to inform either technical or biological effects; (2) cohesion within biological groups - often measured by comparing between-group and within-group variance; (3) for temporal data, a measure of 'smoothness' - whether adjacent time points are more similar than distant timepoints. Sometimes a high-variance feature will show a lack of smoothness. This property is affected by the fact that we have a pseudo-time-series, not a true longitudinal series (see 3A1 below).

**2H**. Missing value distribution and imputation strategies need a separate set of analysis plans.

The order of running outlier/batch identification, normalization, and imputation is one of the difficulties we have wrestled with in PASS1B metabolomic data.

**3. Data analysis: temporal modeling** (i.e., finding features that respond to acute exercise)

The main goal of this step is to identify features that exhibit a response after acute exercise. The analysis plan will be based on the sampling pattern already implemented for PASS1A and PASS1C. Adoption of published tools and lessons learned in literature needs to be aware of the unique challenges in our sampling scheme.

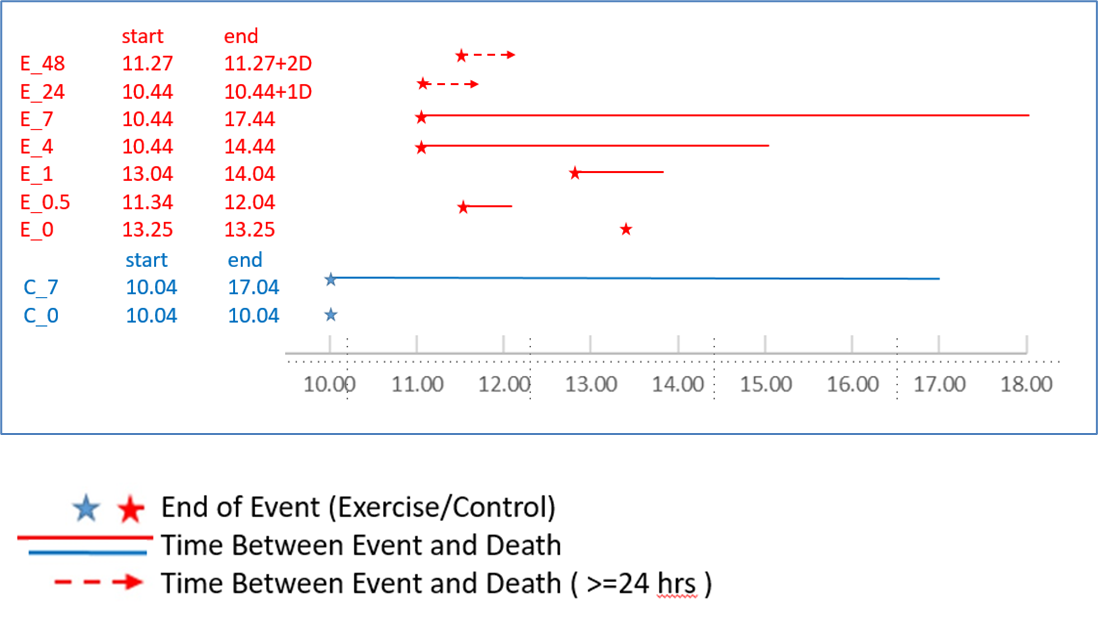
**3A. Challenges brought by the PASS1A sampling scheme.**

**3A1**. MoTrPAC's PASS1A/1C data are **not true longitudinal data**, which would entail measuring the same animals during an extended period. While some physiological data (such as body weight) were collected longitudinally, almost all the molecular profiles were from different animals, each presenting one time point. That is, across K time points, the K exercise groups consist of different animals.

As such, we have a **pseudo-time-series**, formed by independently sampled data. The notion of autoregressivity is much weakened. However, a notion of smoothness of data trajectories can still be developed, assuming that the animals are indeed similar enough to be biological replicates of each other: the notion of replaceability among animals.

**3A2.** Due to practical constraints, the sampling points are **not evenly spaced**. We covered IPE (aka E0, for 0h after exercise) and 0.5h, 1h, 4h, 7h, 24h and 48h after exercise. There are 7 time point. As they are concentrated in the first 4 hrs, PASS1A is designed to discover features that respond quickly: within the first 4-7 hours. To discover features that respond in the 7h-24h window would require a different sampling scheme.

**3A3**. There are **shifts in the real time** at which the exercise runs were conducted, as shown by the diagram below. For instance, the E\_0 animals were exercised at a later time of day than the E\_0.5 animals. Here the left ends of the horizontal lines are the stopping time of exercise (denoted by \*), and the right ends are the time of tissue collection.

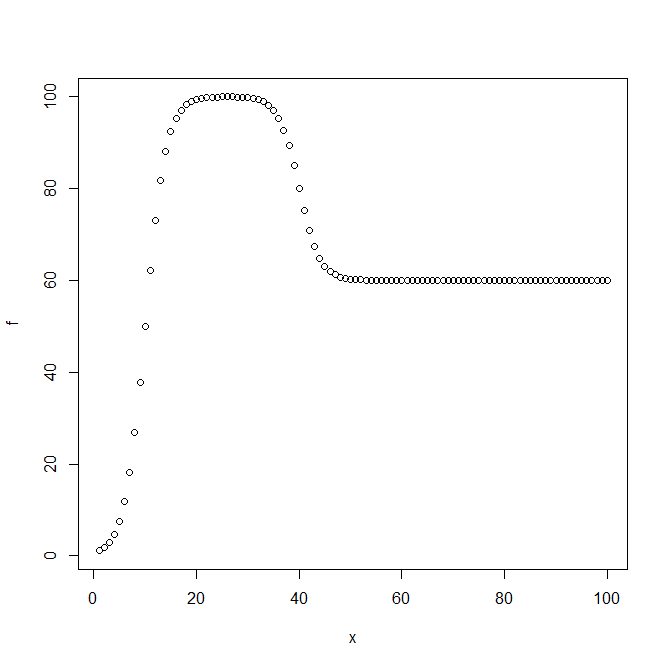


The **lack of synchrony** among some of the exercise groups becomes a complication when we consider potential contributions of baseline physiological changes due to feeding and circadian rhythms (more below, in **3C**).

Note that when we talk about potential contributions of circadian rhythms, we refer to the impact at the time of tissue collection, not the time of exercise. The alternative question, how the hour of exercise could affect the observed responses would require a different study design.

(Details of feeding time and related concerns can be added here)

**3A4**. Temporal modeling requires a **minimal number of effective sampling points** for a given functional pattern. For instance, a 3rd order polynomial function, ax3+bx2+cx+d, has 4 parameters. One of them is the baseline: coefficient "d", hence there are 3 degrees of freedom. A 4th order polynomial can model more complex "waveforms” but will cost one more degree of freedom. The advantage of polynomial functions is that they do not prescribe the likely shape of the trajectories. However, a drawback is that later timepoints tend to exert a greater weight if not properly supervised (because x3 and x2 are much larger than x for large x's. I can explain more).

An "Impulse" function (see figure to the right) with a sigmoidal up-phase and sigmoidal down-phase would need 3 parameters for the magnitude (the level before, peak, and after), 2 for the time of change (midpoints of up- and down-phase), and 2 for the rate (up-rate and down-rate), for a total of 7 parameters, even without modeling the curvature of the sigmoidal functions.

A sinusoidal curve with an assumed 24h period has 3 parameters: baseline, magnitude, and phase. An equivalent function is: a + b\*sin(x') + c\*cos(x'), where x' is 2\*pi\*T/24, T is the real time of tissue collection in the day, not the interval after exercise used in the polynomial or Impulse function. Note that the sine-cosine function above does not accommodate fast changes, either due to sudden changes in a narrow time window, or twice-daily rhythms (i.e., ultradian patterns).

Our data has 7 effective sampling points for modeling the response function, with the challenge that they are concentrated in the first 4 hours. For the Impulse function, the model is identifiable if we have at least 3 points describing the up-phase and at least 3 points describing the down-phase. Since we only have 7 points plus a pre-exercise baseline: C0, IPE (E0), E0.5, E1, E4, E7, E24, and E48, features that do not show an up-and-down pattern in the middle of series will not be adequately modeled (see below, 3B).

For modeling circadian patterns, we have 5 effective time points among the exercised groups, because the 24h and 48h are considered the same as the 0h control point.

To **jointly model** both the 3rd order polynomial (or Impulse) and circadian pattern would incur 6 (or 9) parameters, which is difficult to do with our 7-point time series.

A key component of the analysis plan is to deploy the right algorithms based on the understanding of the waveforms to be modeled and the **inherent limitations** of the sampling scheme.

**3B. Evaluating robustness of the modeling algorithms**

If we deploy two alternative approaches, (1) polynomial function and (2) "Impulse model", to estimate the functional parameters, we need to systematically test the algorithms for the robustness of estimation. More broadly speaking, we face two tasks: model selection and parameter estimation; identifiability for the former, uncertainties for the latter.

We need to use simulated, truth-known data to benchmark the computational algorithms.

Given a simulated temporal trajectory based on a known function of known parameters, we want to know if the computational modeling can (1) identify the true functional form amidst other, similar functions, and (2) accurately estimate the parameters, which are known only to the generator of the simulations. The plan is to create a large family of simulated data to cover the functions and their parameter space, and to run the model to obtain the best estimates.

This will start with the easiest case, with low levels of within-group noise, and denser or evenly spaced sampling points. Subsequently, we add difficulty in a controlled way, in at least two directions. The first is to increase the level of within-group noise, measured in relation to the between-group "signal". The second is to reduce the sampling points to mimic our real sampling points, which are sparse and distributed in the front-heavy way.

The performance from the simulated data will inform the expectations when we analyze real data. A given feature in each real dataset will have a measurable level of difficulty, based on its empirically observed noise level, smoothness, and putative signal levels. We will use the performance of the algorithms in simulated data to learn the expected robustness of the results.

**3C. Expected findings from PASS1A, and situations in which PASS1C is helpful.**

**3C1**. For a given dataset (a certain data type for a certain tissue), we may find N features to be responsive to acute exercise. They are characterized by a P value (where the null is No Response), and a series of effect size measures for individual time points. Heuristically, the trajectories of these N features may be loosely grouped by their principal patterns, such as

- fast up and down

- slow up and down

- fast up and stay up

- fast down and up

- fast down and stay down

- ...

Since the sampling points are concentrated in the first 7 hours, here "fast" refers to peaking at 1h or 4h, and returning at 7h; "slow" means peaking at 7h and returning at 24/48 hr. The study design is not equipped to find other types of dynamics, especially not the changes in the 7h-24h window.

**3C2**. The number of features deemed responsive and their principal curves will vary by tissue and by data type. We will document them separately. This is the **first level of results**. Note that the computational results are not yet “Findings,” which require us to state the biological interpretations (more below).

**3C3**. The feature-level results will be extended to pathway-level results to support the biological interpretation at both levels. A different set of expertise will be needed in the working groups to turn statistical results into themes and talking points. In this perspective, the collected statistical summaries are not the results per se; they are stepping-stones towards stating what we have found, and what findings are new.

**3C4**. The feature-level and pathway-level results will be compared (1) across tissues for a given 'ome, (2) across 'omes for a given tissue, (3) across both 'omes and tissue, with plans to uncover "slanted" interactions: correlations across tissues, with possibly a temporal lag, and across to a different 'ome. These higher-level analyses do not replace the analysis of individual datasets (described in 3C1-3C3), but will build on them. There are many "slices" along the different axes of data and result, and require consortium-wide coordination. Such coordination entails both bottom-up self-volunteered groups and top-down prescriptions that provide both organizational and statistical oversight.

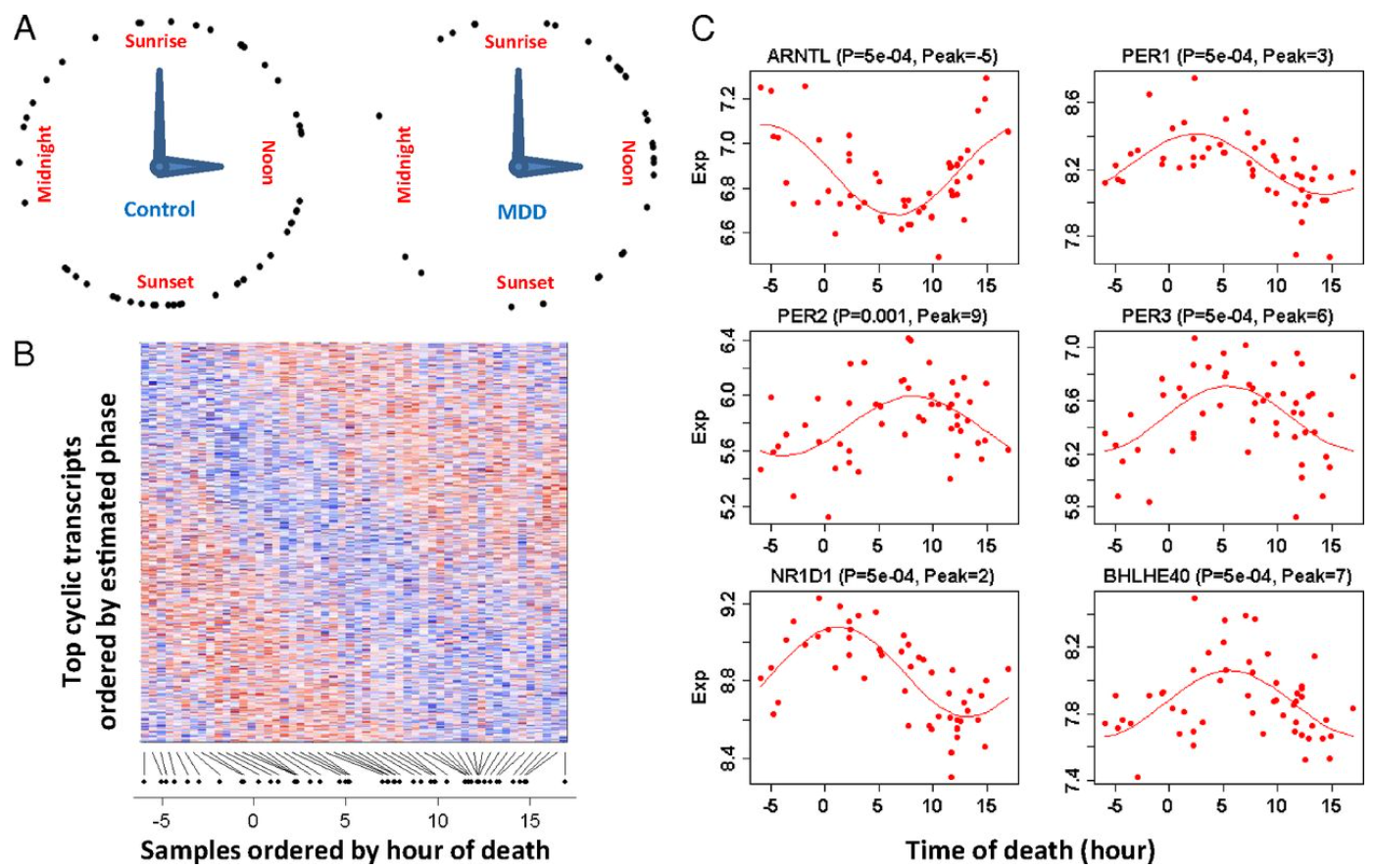
**3C5**. For concerns regarding **confounding factors**, such as the circadian and feeding effect, see 3D below.

**3D. Limitations of PASS1A and potential benefits from PASS1C data.**

A major concern was whether some of the findings in PASS1A could be driven by the dynamic patterns due to feeding, light cycle, and circadian rhythms. The most direct way to address the concern is to run a cohort of no-exercise animals in parallel, with group-wise matching to the 7 exercised groups. Such a dataset will establish the naturally occurring baseline pattern for the animals. However, in PASS1A we only have C0 and C7 for the no-exercise controls. The next question is: how much should we be concerned?

**3D1**. Hypothetically, for a certain dataset, if we find (say) 100 features to be "fast" responders, it's possible that 30 of them exhibit circadian patterns (this is a guess for the moment, for advancing a line of argument). While there are only dozens of genes to be known **drivers** of circadian patterns, some tissues have a far large number of **followers** that show 24h cyclicity. In literature, while some reference data exist for circadian genes, they are mostly for human and mouse, concentrated on some easily accessible tissues. We still lack directly relevant reference data for the rat, for most of the tissues we analyze, and for data types such as the metabolome. To establish such a reference data has not been a focus of MoTrPAC (PASS1A's control samples only covered 0h and 7h). Nonetheless, it is a reasonable guess that a minority of the fast-responding features, here put at 30 of 100, would show circadian patterns.

To continue this hypothetical case: for these 30 features, it is possible that only 5 (again a guess) show fast circadian changes. For instance, my past work (PMID: 23671070) on human brain gene expression patterns revealed that most circadian genes are not fast:



Out of the 5 fast circadian genes, it is possible that only 1 is also making the fast change in the same time window that we run the exercise experiment; the other 4 will be peaking in another time window around the clock.

Thus, in this hypothetical example, out of the 100 fast-acting features, only 1 is confounded by fast-acting circadian pattern in the same window, assuming that the magnitude of its circadian pattern is on par with the exercise-induced effect.

This one feature would benefit from the PASS1C data, assuming that we are not hampered by the potential batch effects.

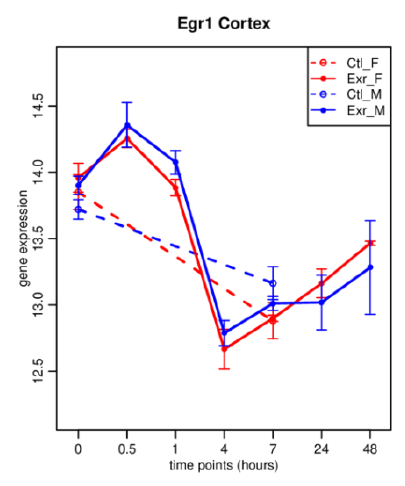
The other 4 fast-circadian features cannot by resolved with PASS1C data.

This hypothetical example invoked a 70-25-4-1 split of the 100 findings, with the 1 as a beneficiary of PASS1C data. In real data, the number of findings and the proportion of these categories will vary by tissue and by dataset. It remains an empirical question.

Stepping back: if we have another 100 slow responding features, some of them may be confounded by slow-circadian features. PASS1C data do not resolve them, as it requires more sampling points in the 7h-24h window. Importantly, the C7 and E7 groups in PASS1A are well positioned to resolve many of them.

An example of apparently slow-acting effect is gene expression of Artl in kidney. Artl’s continued to rise in 4h-7h (on the left), can be explained well by the sine-cosine function (on the right), and show similar patterns between C0 and IPE, and between C7 and E7. It can be ascribed to circadian pattern even without additional control samples.





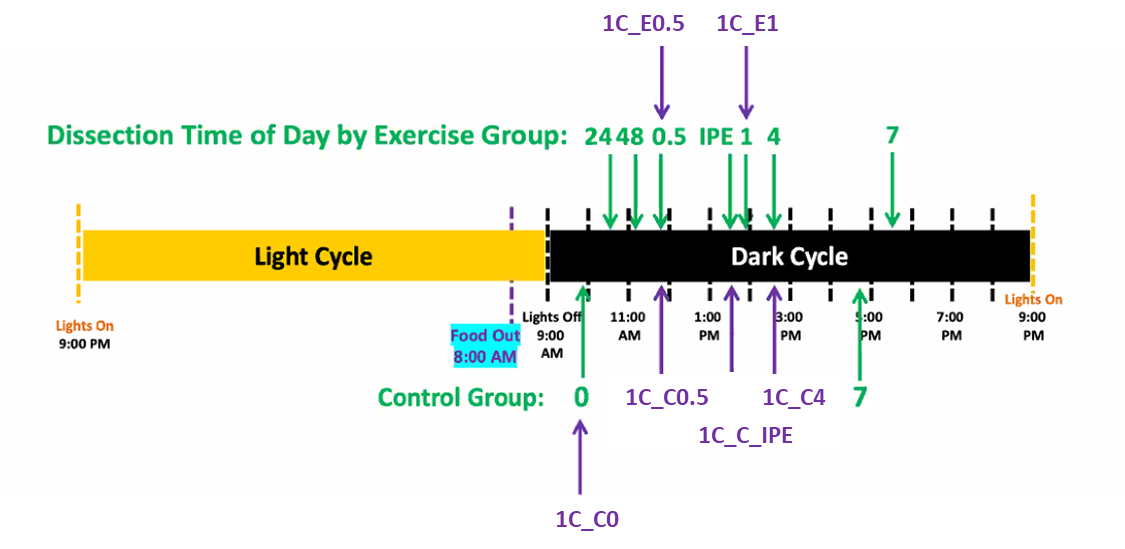
An example of apparently fast-acting effect is cited in the written proposal for PASS1C (Egr1 in Cortex). It peaks at E0.5 and returned at E4, and a concern was raised as to whether circadian pattern could have caused it. PASS1C added sampling points to address this concern.

In summary, PASS1C is designed to identify a special subset of confound due to the circadian effect: fast acting, in the same time window, and of sufficient magnitude. Slow-acting circadian patterns, and those fast-acting but in another window, would still be a potential confounder. However, what is possible may not be probably. A large fraction of exercise-induced features in PASS1A will likely NOT be affected by the concern. We have been slow to mine these findings for fear of widespread confounding by circadian patterns.

Note that PASS1C does not support a full time-course modeling, covering all the 7 points in PASS1A. It will be used as a **post hoc screening** for some of the PASS1A findings.

**3D2**. How can PASS1C data help address PASS1A's limitations? The question can be stated differently: for which dataset (for a certain tissue, 'ome, platform) do we have many fast responding features? And further: how likely are the fast-responding features also fast-circadian features?

For the PASS1A findings that are fast-responding, PASS1C offers **1C\_E0.5, 1C\_E1** and their matched controls: **1C\_C0.5, 1C\_C\_IPE** (assuming it is intended for the PASS1A's IPE, although it could also be intended to match E1, as **1C-C1**). The third control sample in PASS1C is intended as the pre-exercise control, here denoted as **1C\_C0**. The fourth control sample is **1C\_C4**. The prefix 1A is omitted for PASS1A samples.



The PASS1A fast-acting features, which can be explained by fast-circadian effect, can be assessed by the two "deltas" observed in PASS1C, 1C\_E0.5 versus 1C\_C0 and 1C\_E1 versus 1C\_C0. They will be compared to the two matched control deltas: 1C\_C0.5 versus 1C\_C0 and 1C\_C\_IPE versus 1C\_C0. The four deltas from 1C can be compared to the corresponding exercise-only deltas from PASS1A: E0.5 versus C0, and E\_IPE versus C0. This comparison can be done without merging the data between PASS1A and PASS1C. This is called a meta-analysis, albeit limited to comparing the fast-patterns between PASS1A and PASS1C, not for the full time course analyzed by PASS1A.

There are a few other options of meta-analysis that involve less well matched data at 4h. Here we have PASS1A's E4 versus PASS1C's 1C\_C4, without a C4 from PASS1A or E4 from PASS1C. This comparison can be useful if batch correction is successful.

**3D3**. How do we know if batch correction works?

The three matched pairs: E0.5 vs. 1C\_E0.5, E1 vs 1C\_E1, and C0 vs. 1C\_C0, provide three estimates of the PASS1A-PASS1C batch effect. If they agree well, relative to the effect of interest such as 1C\_E0.5 versus 1C\_C0 or E0.5 versus C0, we may merge them to obtain a batch-correction factor and apply them to both the 3 matched and the 3 unmatched sample groups in PASS1C. If they don't agree well, we have group-specific batch effect, and the utility of PASS1C data will be limited. This scenario would also manifest in meta-analysis, because they are mathematically equivalent: if A1-C1 does not match A2-C2 (divergent batch effect between time-1 and time-2), we know that A1-A2 would not replicate C1-C2 (divergent 1-2 difference between batch A and batch C).

The starting plan for evaluating whether batch correction is feasible or not is to analyze a few data groups, such as RNAseq data for a few tissues and compare between PASS1A and PASS1C, and metabolomic data for the same tissues. Some outcomes are:

1. PASS1A slow-acting features do not need to incorporate PASS1C data.

2. PASS1A fast-acting features may be divide into two groups: confounded, not-confounded.

3. PASS1A fast-acting features cannot be adequately divided, due to PASS1C data bringing group-specific batch effect.