Randomizing samples in TOPMed

Proposal

Nov 4, 2015

Based on the points below, the DCC proposes the following plan for randomizing sequencing. While we have aimed to make this plan practical, we recognize that all TOPMed investigators will bring their specific expertise to the challenges of avoiding artefacts.

Essentially, the proposal is to randomize each study's samples (both cases and controls; also cohort samples) within each of a small number of large batches. Larger batches are preferred, ideally just one batch per study where this is possible. In more detail:

1. Studies will send their samples to their sequencing center as soon as possible, either all together or in sets that are representative of the entire sample (e.g. with respect to case-control balance and ancestry groups). Members of families should be shipped together.
2. When enough samples are at the center to justify a batch, the composition of samples (e.g. case-control and sub-study balance) will be assessed and a study investigator will approve proceeding with that batch. Generally, a batch will include at least a few hundred samples from a study.
3. Plate maps will be made for the batch by randomizing assignments of samples to plates and wells. This step can be done by either the DCC, sequencing center, or study; the assignments are recorded, for later validation. Note that this formal randomization step breaks any relationship between phenotype (and/or relevant covariates, e.g. ancestry) and any genotyping artefact within this batch.
4. In family studies, the unit of randomization will be families, rather than individuals, in order to keep family members together as they move through the process. So, for example, all members of a family will remain in the same batch. Where large families are too big to keep together, the family will be broken into smaller sub-pedigrees.
5. Samples will be robotically re-arrayed according to the plate maps, prior to library construction and sequencing. This will typically happen at the sequencing center, but in a few cases will be practical at the sample repository.
6. Within sequencing center, informal "blending" of multiple studies' samples, i.e. typing more than one study's samples at a time, is encouraged where it is practical. This will help minimize artefacts in cross-study analyses.
7. After sequencing, regression analysis within a single study that has multiple batches will typically adjust for batch; studies sequenced in one batch need not do this. Assuming that the batches are representative of the overall study population at large, how well this adjustment works is governed, primarily, by the size of the smallest batch. Therefore, every effort should be made to maximize the size of each batch.

Background

Genome sequencing involves complex and evolving technologies. Call rates and call accuracy can therefore be expected to differ *somewhat*, by sequencing center, by time within sequencing center, and according to sample handling within center and across time, due to variation in reagent batches, instrumentation, etc. When these genotyping artefacts correlate with traits, or adjustment factors important in the analysis of traits, then Type I error rates (i.e. false positives) and Type II error rates (i.e. power) may be affected.

Removing or reducing the impact of these artefacts can be done in two approaches: through *design*, i.e. careful choices of which samples are sequenced when, and through *statistical analysis*, such as regression adjustment or weighting, that account for artefacts. These two approaches can be combined – TOPMed need not rely entirely on one or the other. But in practice neither approach can solve all problems TOPMed will face, and the appropriateness of either approach, or a combination of them, will vary with the analysis being implemented. For example, what is acceptable for a single-study analysis may not be adequate for a cross-TOPMed analysis. This discussion below describes important theoretical and practical issues underlying the rationale for this proposal above.

Design

Almost all design solutions implement randomization, i.e. “breaking” the relationship between genotyping artefacts and trait values. For example, typing subjects in a random order, or allocating them randomly to sequencing machines. *Formal* randomization occurs when a random number generator is used to make these allocations, and a record is kept of how allocations occurred. *Informal* randomization or *blending* instead follows no formal rule; sequencing centers make an approximately random sample as they go, while also ensuring that a minimum number of ancestry or other groups are represented in a batch, and also taking into account current capacity and sequencing demand.

- Randomizing across sequencing centers is not feasible, formally or informally: except for a small number of duplicates, each TOPMed study sends all their samples to just one single sequencing center
- Randomizing across phases of TOPMed is not feasible, formally or informally: studies in different phases of TOPMed will be sequenced in different years
- Manually pulling samples at random for sequencing is physically laborious and error-prone. Robotic control of the process makes such errors less common. Robotics are available within sequencing centers but available less frequently within studies’ sample repositories.
- Randomizing within a sequencing center, either formally or informally, takes time and resources. While it is true that sufficiently small artefacts will not invalidate results, compared to common variant work, rare variant research is more susceptible to artefacts as results typically rely on a small number of variant-carriers with extreme phenotypes. Hence, the costs of randomizing should be viewed as “insurance”, bought in advance to hedge against later problems with high cost.
- When sequencing more than one study in a single center, it is typically not practical to wait until all studies have submitted samples before beginning sequencing, making randomization across studies within-center challenging.
- For large studies, having all samples ready simultaneously is not practical, making it difficult to randomize across an entire study. However, “batches” of samples can be randomized, and these batches can be taken into account in subsequent analysis (see below).
- For analysis of family data, it is helpful to keep families together, i.e. sequenced either together or close in time. This reduces the impact of differential genotype misclassification on analyses that rely on Mendelian inheritance.

- Just as with randomization processes in clinical trials, blending is inherently less robust than formal randomization. Informal processes are particularly difficult for case-control designs; sequencing more cases early and more controls late may – with no clear specification of how much more early vs late – is extremely challenging to fix, post hoc (see below). The same concern applies to any phenotype that might end up correlating with aspects of the sequencing process.

Statistical analysis

We focus on using regression adjustments, which attempt to assess how trait values vary with genotype, among subjects in whom other factors such as study, sequencing center, sample storage etc. are held constant. (Alternative methods based on weighting, e.g. weighting individual genotypes by plausible validity of the SNP's typing when each individual was measured, requires considerable extra information, and are unlikely to be compatible with the analysis tools typical for sequence data.)

- To be adjusted for, all sources of artefact should be recorded and should be available to the analyst.
- Successful regression adjustment requires knowing or estimating the *extent* to which artefacts reflect sequence center, time, reagent batch, instrument, etc; if an artefact is not correctly represented in the statistical analysis, its impact cannot be successfully removed. Adjusting for multiple possible artefacts with as much flexibility as possible is therefore prudent, given limited information on the impact of these artefacts. A consequence of this prudent approach is that some over-adjustment and conservatism (i.e. loss of power) should be expected.
- Particularly for rare variant analysis, the performance of adjustment methods is usually dependent on the ability to measure the impact of artefacts. When sample size is limited – as in many TOPMed settings – statistical adjustment may result in extensive "filtering" of results, i.e. discarding possible novel findings, as these cannot be reliably distinguished from Type I errors.
- Adjustment methods vary in the strength of their assumptions (e.g. identical effects of study regardless of genotype, versus study effects that vary by genotype) and power-robustness tradeoffs are typical. Deciding how to adjust, and defending one's choice adequately, can be a challenge to the publication of results.
- Regression adjustment makes analyses more complex and therefore somewhat slower. For some analyses (e.g. some permutation tests) the requirement to adjust is a significant hindrance.

Combining design and statistical analysis

Given a particular design that avoids some artefacts, careful choices of statistical adjustment can help remove others.

- If studies submit *batches* of randomly-chosen subjects, which are then randomized within-center, the analysis need only adjust for batch to remove batch artefacts, in single-study analysis. In cross-study analyses additional adjustment for study would remove artefacts due to e.g. sample storage prior to sequencing.
- Within batches, it is important to keep a balance of samples who will be analyzed together (e.g. cases and controls) or the statistical adjustment for batch, while necessary to avoid Type I

errors, will result in low power to find novel association. This form of confounding is known as *aliasing*.