NHLBI TOPMed WGS Program

Ethical, Legal, and Social Issues (ELSI) Committee Review: Imputation Server

Issue

This document summarizes ELSI Committee discussions regarding inclusion of data from TOPMed studies in the TOPMed imputation server. The discussions addressed whether, and to what degree, restrictions from individual consents should govern which data go into the imputation server.

The TOPMed Imputation server

In considering this issue, the Committee noted the following about an imputation server:

- Genotype imputation uses the linkage disequilibrium patterns in existing (reference) genomes as a means to generate new, imputed data in other genotyped sample sets.
- During imputation, the imputed results in previously genotyped samples will likely be obtained from components of many reference genomes (segments for each genomic region) rather than from a single reference sample.
- No phenotypes of reference genomes are maintained in a cloud-based imputation server, such as is being proposed for TOPMed, and no phenotypes are uploaded by an external user (only genotypes).

In practice, an imputation server "digests" a set of reference genomes into patterns (series of haplotypes). A user of an imputation server uploads a previously genotyped sample collection to have missing data (genotypes) filled in (imputed) using patterns in the "digested" reference genomes. The server returns the imputed data to the user, not keeping either the uploaded or imputed data. The user then conducts analyses of associations with phenotypic data using the increased density of genetic information afforded by the inclusion of imputed genotypes.

Why a TOPMed Imputation Server?

- TOPMed is a large collection of sequenced genomes, much larger than 1000 Genomes and, as such, would become the best imputation resource available to date.
- As with the TOPMed variant server (BRAVO), the imputation server is scientifically important and can be used not only for large GWAS cohorts but also for smaller/individual collections.
- It is feasible that the much reduced cost of GWAS compared with the cost of whole genome sequencing, followed by use of a well-powered imputation server, will permit many more investigators to obtain more comprehensive genomic characterization of their samples when funding is limited.
- Limitations of the TOPMed (or any) imputation server would occur in special situations, such as when imputing into samples from an isolated population, such that variation in that population, or in closely related populations, is not well-represented in the reference panel.

Potential concerns with an imputation server

The Committee noted two potential concerns related to studies contributing to TOPMed with respect to participant genomic data being used in the TOPMed Imputation Server.

- (1) Risk of re-identifiability
    - The risk is low that a user could determine whether a given individual was present in the imputation reference panel, though such an attack is theoretically possible. This risk is similar to the risk of re-identification from the original data set.
    - Assuming the scientific value of the imputation server and that data security protections are in place, this risk does not preclude submission of data to the imputation server.
- (2) Risk that imputed genomic data is used for an analysis that was not approved by the participant's original consent:
    - The imputation server is agnostic with regard to the phenotypes that the end-stage user might subject to analysis with the imputed genotypes. Consider the situation in which the initial set of reference genomes were collected under consents that restricted sample use to GRU and HMB categories. Once uploaded into the imputation server, there can be no guarantee about how the imputed genotypes from an external user will be used in downstream analyses and if the external users' downstream analyses will be consistent with the consents obtained from subjects contributing to the imputation panel reference set. The following examples were discussed regarding reference genome consents:
        - a participant consents to research on health and disease (HMB), which under NIH interpretations of this consent category would preclude studies of intelligence or ancestry; an investigator submits GWAS data to the imputation server that is later used in analysis of gene variants contributing to intelligence;
        - a participant consent indicated that data use should be limited to non-commercial purposes; a commercial entity utilizes the imputation server for analyses related to commercial objectives.

These two concerns reflect the distinction between a "risk-based" model (low for identifiability) and "respect-based" model (adherence to the restrictions agreed to in the consent at all times) for defining obligations of researchers to participants.

The ELSI Committee has previously endorsed the respect-based model as the appropriate general guide for determining obligations to participants. However, another factor, the publication analogy, must be considered when determining the relevance of the respect-based model to the issue of placing data in the imputation server.

Publication analogy

When data are published in a scientific article, they are in the public domain. Other researchers may utilize published data in their studies without reference to the original consent form governing the study from which data were published. Thus, the question is whether placement of data in the imputation server is equivalent to publication. If so, original consent restrictions would not apply.

The ELSI Committee has previously offered the opinion that the publication analogy applies to placement of TOPMed data into the TOPMed variant server (BRAVO). Factors that contributed to this opinion included: (1) the variant server provides data on frequency of variants but does not provide either phenotypic data or racial/ethnic identifiers; (2) similar summary data are often included as appendices to publications; and (3) the variant server provides descriptive/comparative data but does not enable analyses. The ELSI Committee further advised that although the publication analogy applies, the most prudent approach would be to include only those data that were consented for general research use and health and disease studies.

Does the publication analogy apply to the imputation server?

While the ELSI Committee considers the publication analogy to be applicable to the TOPMed variant server (BRAVO), the applicability of the publication analogy to the imputation server seems to be more nuanced. Potential concerns are as follows:

- An imputation server goes beyond provision of descriptive data. It constitutes an analytic resource that contributes (albeit indirectly) to the goals of any research study utilizing the imputation server.
- There is no provision (nor is it practical to implement one) for the TOPMed imputation server to include any review of study aims of external users of the imputation server. External studies could include aims falling outside the scope of the consents under which the reference genomes were obtained.
- Thus indirectly, the TOPMed imputation server could be construed as potentially contributing to the conduct of research studies outside the scope of the original consent(s).

One question raised was whether potential users of the imputation server could answer their research questions without imputation.

- In the fullness of time and with adequate funding, the users could conduct comparable research without the server – i.e., if they were able to perform sequencing in their own sample set. But imputation has the advantages of lower costs and quicker turnaround time.
- The key is that imputation enables analysis of less frequent and rare variants that would not be readily available in a typical GWAS dataset.

The ELSI Committee also heard arguments favoring application of the publication analogy to the imputation server.

- The imputation server can be likened to a statistical model deriving from TOPMed data. Like other statistical models, it is a product of a defined research activity and should be viewed as entering the public domain when complete. Its fundamental goal, in this sense, is to advance science.
- In this view, constraints from original consent(s) apply to the studies that enabled the creation of the model, but the model itself is a product of the research and no longer constrained by limitations imposed by the consent language.

Committee discussion further noted that the fundamental obligation of publicly funded research is to generate knowledge that is placed in the public domain. It can be argued that the imputation server provides such knowledge.

If the publication analogy is deemed applicable to the imputation server, then all TOPMed data could reasonably be submitted to it. Although again, in line with ELSI committee recommendations regarding the BRAVO (variant) server, the most prudent approach would be to include only those data that were consented for general research use and health and disease studies.

Advice of the TOPMed ELSI Committee

As an advisory body, the ELSI Committee wishes to communicate the following to the TOPMed Steering and Executive Committees:

- Given the applicability of the publication analogy, it would be reasonable for all TOPMed data to go into the variant server (BRAVO).
    - However, it is also reasonable to further restrict the deposition of allele frequency data in the BRAVO server only to those TOPMed samples consented for general research (GRU) use and studies of health and disease (HMB).
- For the TOPMed Imputation Server, the decision as to whether consent restrictions apply is more complicated, as noted above.
- Each TOPMed PI should, therefore, consider the language of study consent forms; their knowledge of study participants' wishes and expectations; and the arguments for and against the publication analogy when considering whether genotype data from samples in their study (and possibly from which consent group(s)) should be used in the imputation server.