Published on *NHLBI Trans-Omics for Precision Medicine* (https://topmed.nhlbi.nih.gov)

TOPMed Data Access for the Scientific Community

Updated 10/28/2021

Contents

Back to top [7]

Where are the data?

TOPMed genomic data and pre-existing Parent study phenotypic data are made available to the scientific community in study-specific accessions in the database of Genotypes and Phenotypes (dbGaP) [8] and in the NHLBI BioData Catalyst [9] cloud platform. Different types of data are organized within accessions as follows:

- *Phenotypes*: When the Parent study has a dbGaP accession that preceded the existence of the TOPMed program, phenotypic data are in the Parent accession. Otherwise, the phenotypic data are in the TOPMed accession. In addition, the TOPMed Data Coordinating Center (DCC) has harmonized select phenotypes across TOPMed; details and availability described under DCC-harmonized phenotypes for the scientific community [10].
- *Genotypes (WGS)*: Unphased genotype calls from TOPMed WGS are available in the TOPMed accession as Variant Call Format (VCF) files. Studies may have multiple sets of VCF files corresponding to the various TOPMed data freezes [11]. The VCF files contain variant-level quality metrics and a support vector machine (SVM) quality filter. The table below summarizes TOPMed WGS characteristics by freeze.
- *Read alignment data (WGS)*: Only a limited number of TOPMed Phase 1 CRAMs aligned to build 37 are available directly through the dbGaP Sequence Read Archive (SRA) [12]. These can be accessed through a dbGaP approval for their corresponding TOPMed accessions. All other CRAMs, including build 38 alignments for all TOPMed WGS samples, are hosted in NHLBI cloud buckets and accessed using the “fusera” software.

    - Instructions for controlled access to TOPMed sequence data on the cloud [13] (Provided by Tom Blackwell, TOPMed IRC)
    - Further documentation on dbGaP cloud access, including fusera [14] (Provided by NCBI)
- *Non-WGS omics*: TOPMed is generating a rich resource of multi-omics data that will include approximately 40K samples undergoing RNA-sequencing, 37K samples from metabolomics profiling, 57K samples from DNA methylation, and 4K samples from proteomics assaying. These projected totals include all stages of progress, from DNA/RNA that are currently being extracted, that are undergoing sequencing/profiling, or that have completed sequencing/profiling pipelines. Omics data will be released to the scientific community via NIH-designated repositories (dbGaP and BioData Catalyst).
    - Non-WGS omics pipelines and flowcharts [15]

Back to top [7]

TOPMed WGS characteristics by freeze

TOPMed freeze (methods documents linked)	Date	Genome Build	n_variants	n_samples	n_studies
freeze.5b [16]	Sep 2017	38	582M	56K	32
freeze.8 [17]	Feb 2019	38	1.02B	138K	72

Back to top [7]

How do I apply for access?

Users who want to apply for controlled-access TOPMed data should follow the dbGaP instructions for requesting controlled-access data [18]. In a dbGaP application, each TOPMed study-consent group will need to be requested individually. Note that participant consent and Data Use Limitations (DULs) differ within and across TOPMed studies. Therefore, dbGaP applicants will need to carefully review DULs and ensure that proposed Research Use Statements (RUS) are consistent with the study-consent group(s) being requested. Additionally, some TOPMed studies have consent modifiers that may require additional documentation, such as documentation of local IRB approval and/or letters of collaboration with the primary study PI(s).

Applicants should investigate whether phenotype data are deposited in the TOPMed or the Parent accession for the studies of interest. If the latter, then applicants will need to specifically apply for access to the Parent accession for phenotypes in addition to applying to the TOPMed accession for TOPMed WGS genotypes. Phs numbers for TOPMed and Parent accessions are available in the dbGaP methods documents [11].

Back to top [7]

How do I use the data?

Running mega analyses across TOPMed studies requires combining genotype and phenotype data across individual dbGaP accessions.

- *Combining genotypes*: The Informatics Research Center's (IRC) [19] joint calling process produces a multi-study VCF file for each chromosome, each of which is split into study-specific components. For studies with multiple consent groups, these components are further divided by consent groups and deposited in the study's TOPMed accession. The same variants occur in all VCF components of a given call set. To construct a multi-study VCF file for analysis, a user must apply for access to each study-consent group and reassemble the components. Note some TOPMed accessions will have VCF files for more than one data freeze [11]. Therefore, users must take care to select VCF files from the same freeze for their multi-study reassembly. Tools for combining VCF files include vcftools [20] and bcftools [21].
- *Combining phenotypes*: The Parent studies contributing to TOPMed have many phenotypic measures in common, thereby providing opportunities for cross-study analyses to gain power in detecting genetic effects. However, these studies' designs differ in how their phenotypic data were collected, and in how their data are annotated and structured. Creating harmonized phenotypic data sets for cross-study analyses is therefore a challenging and largely manual process. Users will need to carefully evaluate the source phenotypes and accompanying documentation before attempting to harmonize across studies. The TOPMed DCC's phenotype harmonization efforts are described under DCC-harmonized phenotypes for the scientific community [10], along with a phenotype tagging project that can assist members of the scientific community in finding related phenotype variables to perform their own harmonizations.
- *A note on Sample/subject identifiers*: The TOPMed ACC centrally assigns each molecular sample in the TOPMed program a unique sample identifier (e.g., for DNA, "NWD" followed by 6 digits), which is used in all files containing TOPMed sequence, genotype, or other molecular data. Subject (aka participant or individual) identifiers are assigned by study investigators and are not guaranteed to be unique across all studies. The subject identifiers are associated with individual-level phenotypic data and, in most cases, are consistent between the TOPMed and Parent accessions for a given study. Mappings between sample and subject identifiers, as well as subject ID aliases, are given in standard dbGaP files labeled as subject-sample mapping and subject consent files.

Back to top [7]

Where can I access variant summary data?

The following resources provide summary-level information on variants observed in TOPMed (e.g., allele frequencies, association results), or other non individual-level data (e.g., imputation server).

- BRAVO (BRowse All Variants Online) variant browser [22]
- gnomAD (Genome Aggregation Database [23]
- dbSNP (Database of Single Nucleotide Polymorphisms [24]
- Genomic Summary Results [25]
- TOPMed Imputation Server [26]

Back to top [7]

Where can I learn more?

- TOPMed Ancillary Session at 2020 ASHG annual meeting [27]
- TOPMed overview poster from the 2018 ASHG annual meeting [28]
- TOPMed design paper: Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program [29]
- TOPMed Publications [30]
- TOPMed Projects and Parent Studies [31]

Back to top [7]

Source URL (modified on 10/28/2021 - 9:50am):https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community

Links

[1] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#where-are-the-data- [2] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#topmed-wgs-characteristics-by-freeze [3] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#how-do-i-apply-for-access- [4] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#how-do-i-use-the-data- [5] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#where-can-i-access-variant-summary-data- [6] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#where-can-i-learn-more- [7] https://topmed.nhlbi.nih.gov/topmed-data-access-scientific-community#top [8] https://www.ncbi.nlm.nih.gov/gap/advanced_search/?TERM=topmed [9] https://biodatacatalyst.nhlbi.nih.gov/ [10] https://topmed.nhlbi.nih.gov/dcc-pheno [11] https://topmed.nhlbi.nih.gov/data-sets [12] https://www.ncbi.nlm.nih.gov/sra/?term=TOPMed [13] https://topmed.nhlbi.nih.gov/sites/default/files/TOPMed_cloud_access_instructions.pdf [14] https://www.ncbi.nlm.nih.gov/sra/docs/dbgap-cloud-access/ [15] https://topmed.nhlbi.nih.gov/standards [16] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2 [17] https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8 [18] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html#request-controlled [19] https://topmed.nhlbi.nih.gov/group/irc [20] https://vcftools.github.io/ [21] http://www.htslib.org/doc/bcftools.html [22] https://bravo.sph.umich.edu/ [23] http://gnomad.broadinstitute.org/ [24] https://www.ncbi.nlm.nih.gov/projects/SNP/ [25] https://topmed.nhlbi.nih.gov/topmed-genomic-summary-results-public [26] https://imputation.biodatacatalyst.nhlbi.nih.gov/#! [27] https://topmed.nhlbi.nih.gov/ashg-2020-ancillary-session [28] https://topmed.nhlbi.nih.gov/sites/default/files/ASHG2018_TOMed_overview_v10.pdf [29] https://pubmed.ncbi.nlm.nih.gov/33568819/ [30] https://topmed.nhlbi.nih.gov/publications [31] https://topmed.nhlbi.nih.gov/group/project-studies