Draft Data Release Policy for the HIV+ Tumor Molecular Characterization Project (HTMCP)

Background

Rapidly evolving sequencing and informatics tools are substantially diminishing costs of comprehensive characterization of tumor transcriptomes and tumor genomes. These advances have resulted in detailed information on the repertoire of alterations in tumors. NCI already supports tumor genome characterization projects for several common cancers, as part of the Cancer Genome Characterization Initiative (CGCI) and the Cancer Genome Atlas (TCGA). Comprehensive sequencing of genomes and transcriptomes in cancers that arise in HIV infected individuals may provide a starting point for a systems biology approach towards understanding differences in etiologies among identical histological subtypes of cancers in HIV+ and HIV- patients. The results obtained could provide important clues to the pathways that either allow tumors to counteract immune surveillance mechanisms or are redundant in the presence of an extrinsic oncogenic influence such as viruses. It is also possible that the comparison of transcriptomes and genomes between tumors from HIV+ and HIV- individuals may identify novel non-human sequences that could point to the presence of transcripts from hitherto undiscovered viral agents.

HTMCP

HTMCP is a "community resource program" made up of three individual projects, each of which studies a particular cancer type. Through rapid data release, HTMCP enables accelerated translation to enhance clinical impact. Therefore, patenting on the PRIMARY data is discouraged to allow easy access and encourage its use. The Project team for each cancer type will publish the "summary" manuscript once the data generation and analyses are complete.

The project will be in two phases, discovery and validation (for more information, visit the CGCI Overview webpage, https://ocg.cancer.gov/programs/cgci/overview)

Two data types will be produced: 1) raw sequences from the tumor/normal genomes and tumor transcriptome; 2) results of analyses from those raw sequences. We acknowledge that algorithms for sequence analysis to identify tumor-specific alterations (e.g. chromosome rearrangements, single base mutations etc.) are still in the development stage and thus the results will be available for "mining" with new algorithms.

Policy

The HTMCP data release policy is aligned with other NCI-funded large-scale genomic characterization projects. To best accomplish the goals of the project (generating and analyzing a large enough dataset to be able to draw statistically and biologically sound conclusions) and the mission of the Institute (to facilitate research and reduce redundancy by making primary data available to the scientific community in real time), the Project team members agreed to the following data release policy:

- Sequence files (BAMs for DNA and FASTQs for RNA) will be deposited into a public repository after quality control is complete, but no later than 4-6 month after they are generated.

- Table of the verified mutations (MAF) will be deposited to the OCG Data Coordinating Center (DCC, http://cgci.nci.nih.gov/dataMatrix/CGCI_DataMatrix.html) after manuscript describing the findings of the dataset is accepted for publication.

The "Using CGCI Data" webpage (https://ocg.cancer.gov/programs/cgci/using-cgci-data) includes information about the philosophy of the rapid data release policy. To support the continued prompt public release of large-scale genomic data prior to publication, researchers who plan to prepare manuscripts that would be comparable to the analyses described below, and journal editors who receive such manuscripts, are requested to coordinate their independent reports with the project's first publication by contacting the HTMCP Program representative or OCG (see below).

HTMCP's first manuscript for each individual project (i.e., each cancer type studied) could include:

- Commentary detailing the scientific aims and organization of HIV+ tumor molecular characterization project
- Analysis of paired DNA sequencing data for the sample set
- Analysis of the RNA sequencing data for the sample set
- Validation of a subset of variant calls found by either DNA or RNA sequencing of the sample set

NCI does not consider that deposition of data from the HTMCP, like those from other large-scale genomic projects, into its own or public databases to be the equivalent of publication in a peer-reviewed journal. Therefore, although the data are available to others, the producers still consider them to be formally unpublished and expect that the data will be used in accord with standard scientific etiquette and practices concerning unpublished data.

For each individual HTMCP project, once the first manuscript describing global analysis by the Project team members is in press, all other researchers can publish results based on integrating HIV+ tumor data with data from other sources. The use HTMCP data to develop novel methods to analyze genomic data related to cancer and genotype-phenotype relationships in cancer, and publication of manuscripts that describe these methods, is permitted.

Prior to the publication of the initial papers, the HTMCP requests that authors who use the data acknowledge the HTMCP as follows: "The results published here are in whole or part based upon data generated by The HIV+ Tumor Molecular Characterization Project established by the Office of Cancer Genomics and Office of HIV and AIDS malignancies of the NCI; https://ocg.cancer.gov/programs/cgci." After initial publication, the first paper per cancer type and website should be referenced.

To ensure protection of privacy and confidentiality for sample donors, access to the primary genetic data is controlled by policy developed by the NIH (http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf). For example, users have to agree that they will share these data only with others who have received data use certification (DUC) and that they will not try to identify or contact the patients. Reviewers of a manuscript who need to see any controlled-access HTMCP data underlying a result must also agree to these access conditions before they can see these data. To apply for DUC, visit the database of Genotypes and Phenotypes (dbGaP) website (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login).

Meeting presentations of HTMCP data and analyses by project team members are possible and encouraged. We request that the project team members inform the NCI of public meeting oral and poster presentations. The HTMCP Project team will develop two-three slides that should be used for oral presentations, posters, etc. They will provide a standard method of citing the HTMCP and its many contributors; it is critical that the HTMCP also be properly cited and identified in the meeting abstracts, and language will be provided to accomplish this goal.

HTMCP Program Representative

Dr. Nicholas Griner
Office of Cancer Genomics
National Cancer Institute
31 Center Drive, Suite 10A07
Bethesda, MD 20892
Phone: 339-502-8280
Alternative phone: 301-451-8027
Fax: 301-480-4368
Email: nicholas.griner@nih.gov