Data Release Policy for the HIV+ Tumor Molecular Characterization Project (H+TMCP)

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 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 might identify novel non-human sequences that could suggest the presence of transcripts from hitherto undiscovered viral agents.

This is a "community resource project", with rapid data release to enable accelerated translation to enhance clinical impact. Therefore patenting on the PRIMARY data is discouraged to allow easy access and encourage its use. There is an expectation of a rapid initial "summary" publication by the group once the data are generated.

Two data types will be produced; 1) raw sequences from the tumor/normal genomes and tumor transcriptome, 2) analyzed data from those raw sequences. It is important to acknowledge that algorithms for sequence analysis to identify tumor-specific calls are still in the development stage and thus the results obtained require confirmation. Confirmation is defined in two ways:

- Verification: assessment of sequence quality before data release (e.g. identifying Illumina artifacts, performing sample swaps, etc.).
- Validation: confirmation of variants identified by the current analytical algorithms by using orthogonal experimental methodology such as Sanger sequencing. Validation will be performed; the scope will depend on the costs and the accuracy of the sequence-calling algorithms available at the specific time. It may be performed either for a subset or all variants found (the details will be developed on real time basis to take advantage of the best approaches). The criteria for selection of a subset of variants for validation will be developed by the cancer-specific working group based on all empirical data available at decision time.

Policy:

The data release policy should be consistent across all NCI-funded large-scale genomic characterization projects. The HIV+ cancers are hard to accrue and therefore the data generation

will span over a number of months or years. To best accomplish the goals of the project (generating and analyzing large enough data set to be able to draw statistically and biologically sound conclusions) and the Institute (to facilitate research and reduce redundancy by making primary data available to the scientific community in real time), the project members suggest the following policy:

- *Release of analyzed sequences (BAM files) will occur after a sample set (number to be determined) is complete, but not later than 4-6 month after they are generated.*
- *Table of the validated mutations (MAF) will be deposited to the Data Coordinating Center (DCC) after manuscript describing the findings of the dataset is submitted for publication.*

The DCC data portal (http://cgap.nci.nih.gov/cgci.html) will include a text about the philosophy of the rapid data release policy, "The Responsible use and publication of Data Generated by the Cancer Genome Characterization Initiative". The language will be aligned as much as possible to the one used for TCGA and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

An HIV+ tumor project manuscript(s) could include:

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

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 above, and journal editors who receive such manuscripts, are encouraged to coordinate their independent reports with the project's publication schedule described above. This may be done by contacting the Project Team (see below).

Once the first global analysis by the project members is in press, all other researchers are free, and indeed encouraged, to publish results based on integrating HIV+ tumor data with data from other sources. Researchers also are encouraged to use H+TMCP data to publish on the development of novel methods to analyze genomic data related to cancer and genotype-phenotype relationships in cancer.

NCI does not consider that deposition of data from the H+TMCP, 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.

Prior to the publication of the initial paper, the H+TMCP project requests that authors who use

data acknowledge the H+TMCP 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. Information about project and the investigators and institutions that constitute the HIV+ Tumor workgroups can be found at http://cgap.nci.nih.gov/cgci.html)"*. After initial publication, the paper and website should be referenced.

To ensure protection of genetic privacy for sample donors, data users will have to agree to certain conditions described in the H+TMCP Patient Protection Policy and Controlled Access Policy as to how the data will be used. For example, users will have to agree that they will share these data only with others who have also completed a data access agreement and that they will not patent discoveries in a way that prevents others from using the data. This means that reviewers of a manuscript who need to see any controlled-access H+TMCP data underlying a result must also agree to these user access conditions before they can see these data.

Meeting presentations of H+TMCP data and analyses by project team members are possible and encouraged. , We would request that the project team members inform the NCI of public meeting oral and poster presentations. The H+TMCP 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 H+TMCP and its many contributors; it is critical that the H+TMCP also be properly cited and identified in the meeting abstracts, and language will also be provided to accomplish this goal.

Project Team Representative:

Dr. Jean Claude Zenklusen
Scientific Programs Director
Office of Cancer Genomics
National Cancer Institute
31 Center Drive, Suite 10A07
Bethesda, MD 20892
Phone: 301-451-2144
Fax: 301-480-4368
e-mail: jz44m@nih.gov