CTD² GUEST EDITORIAL
Dissecting Cellular Heterogeneity using Single-cell RNA Sequencing

Single-cell RNA sequencing enables analysis of the transcriptome of individual cell, provides information on cell-states, and allows a high-resolution characterization of heterogeneous tumor microenvironment. It can be used to discover therapeutic targets and can enable mechanistic understanding of target inhibition.

HCMI PROGRAM HIGHLIGHTS
Scientific Applications of Next-generation Cancer Models

HCMI is providing the scientific community with next-gen cancer models that more closely resemble primary tumors, and that are annotated with genomic and clinical data. The article provides examples of how next-gen models have been applied in research.

CTD² GUEST EDITORIAL
GenePattern Notebook: Integration of Electronic Notebooks with Bioinformatics Tools for Genomic Data Analysis

The GenePattern Notebook is an electronic notebook that enables integrative genomic analyses. These analyses are displayed in a user-friendly form and allows scientists even without programming experience to share, collaborate, and publish the results.

HCMI PROGRAM HIGHLIGHTS
Collecting Uniform Clinical Data for a Community Resource

HCMI's clinical Case Report Forms (CRFs) standardize the clinical data that are collected from participating Tissue Source Sites (TSSs) collaborating with HCMI. The article discusses the process of developing cancer-type specific CRFs to ensure uniformity and compatibility for use at TSSs across the globe.

OCG PERSPECTIVE
Leveraging a Genomics Background to Facilitate Molecular Characterization of HCMI Models

This perspective article introduces Dr. Lauren Hurd, a new Scientific Program Manager for the Human Cancer Models Initiative. Dr. Hurd discusses her background in genomics and its applications in her current role.
CTD² GUEST EDITORIAL

Dissecting Cellular Heterogeneity using Single-cell RNA Sequencing

Anuja Sathe, M.B.B.S., Ph.D. and Hanlee P. Ji, M.D.
Stanford University

Gene expression analysis using RNA sequencing has contributed immensely to our knowledge of cancer. However, bulk sequencing methods average signals by pooling information from a mass of cells. These methods are thus unable to fully resolve the complexity of intra-tumor heterogeneity in cancer. Single-cell RNA sequencing (scRNA-seq) enables the analysis of the transcriptome of each individual cell and allows a high-resolution characterization of tumors. Moreover, tumors are not isolated masses of cancer cells but are surrounded by a unique microenvironment composed of different cell types. scRNA-seq enables the analysis of this heterogenous tumor microenvironment (TME) that is increasingly important in improving treatment strategies such as immunotherapy. Unlike other single-cell analysis methods such as mass cytometry or CyTOF, scRNA-seq allows an unbiased assessment of cellular phenotypes. This enables the identification of not just heterogenous cell types but provides information on individual cell states.

In recent years, scRNA-seq has been used to construct single cell atlases of several tumor types. These studies have revealed novel targets in sub-populations of cancer cells as well as in the TME. We have successfully used scRNA-seq in the characterization of lymphoma TME using patient biopsies and in an organoid mouse model of gastric cancer. We are applying it to understand the TME of gastrointestinal cancers from fresh surgical specimens as well as in cell line models to delineate clonal heterogeneity.

Several technologies have been developed for scRNA-seq. They differ in their method of cell isolation (e.g. plate or microfluidics-based), capture and length of transcript (full-length, 3’ or 5’ end) as well as the chemistry used for reverse transcription and amplification. Choosing a particular platform for an experiment depends on the question being investigated. For example, microfluidics-based techniques are more high-throughput than plate-based ones. Methods that capture the 3’ or 5’ transcript do not allow the detection of splicing events, isoforms, or quantifying allelic expression.

Following sequencing, a typical analytical workflow begins with data matrices containing entries of molecular counts corresponding to each gene and cell, which are represented in respective rows or columns. This high-dimensional data is generally analyzed using dimensionality reduction (e.g. with principal component analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE)) followed by clustering of cells with similar transcriptional profiles. Understanding differences across clusters is aided by differential expression testing.
Figure: Schematic representation of microfluidics based scRNA-seq workflow: Tissue specimens are dissociated into a single-cell suspension followed by a microfluidics based scRNA-seq protocol. cDNA from each individual cell is tagged with a unique barcode and made into a sequencing library. The resulting data is a matrix of molecular counts per transcript per cell. High dimensional data is processed using dimensionality reduction and clustering approaches. (Image credit: Created with BioRender)

A major challenge in scRNA-seq analysis results from the noise and variability of the assay owing to a high number of zero transcript counts. These dropouts can be biological, related to the stochasticity of mRNA expression, or technical, owing to the low amount of transcript material and limited efficiency of its capture. This can result in distorted false negative or false positive profiles. scRNA-seq requires careful attention to quality control and use of appropriate computational methods that are suited for such data distribution. Another limitation of scRNA-seq is that it does not retain any spatial information. Moreover, disaggregation methods are required to produce a single-cell suspension that could introduce artifacts in the gene expression program. Commercial assays and instrumentation can also be expensive.

A number of technical and computational approaches are enabling improvements in scRNA-seq that overcome many of these challenges. For example, conducting parallel experiments such as RNA in situ hybridization, imaging mass cytometry, or spatial transcriptomics can enable integration of spatial information. As sequencing costs reduce, the current cost for an scRNA-seq experiment with a microfluidics-based platform works out to be less than 1 USD/cell (https://satijalab.org/costpercell [2]). This is additionally aided by modifications to the assay using antibodies to tag each cell that enable sample multiplexing [5]. Several quality control and imputation methodologies are also being developed to improve data analytics [9, 10]. Novel assay developments such as single-cell DNA sequencing, single-cell Assay for Transposase Accessible Chromatin (ATAC) sequencing, single-cell T Cell Receptor (TCR) sequencing, and single-cell epitope sequencing can also be integrated with scRNA-seq to reveal a wealth of information at high granularity.

scRNA-seq is thus equipped to answer several important questions in cancer biology including target discovery and can enable a mechanistic understanding of target inhibition. It also has tremendous potential in translational applications such as longitudinal patient monitoring of treatment response [11].

References


...


HCMI PROGRAM HIGHLIGHTS

**Scientific Applications of Next-generation Cancer Models**

*Cindy Kyi, Ph.D.*

*Office of Cancer Genomics, NCI*

As a contribution to efforts in *precision oncology*[^13], the National Cancer Institute’s (NCI) [Human Cancer Models Initiative](https://www.nci.nih.gov/research-human-cancer-models) (HCMI) is developing next-generation (next-gen) cancer models, which include 3D organoids, neurospheres, 2D adherent, and conditionally reprogrammed cells. These next-gen models are derived from a variety of cancer types including poor outcome cancers, rare cancers, and cancers from ethnic and racial minorities, as well as pediatric populations. HCMI was founded by the NCI, Cancer Research UK, Hubrecht Organoid Technology, and Wellcome Sanger Institute to develop about 1,000 next-gen cancer models. The next-gen cancer models are annotated with molecular characterization data, as well as clinical data to address challenges of traditional cancer cell lines. The goal of HCMI is to provide the research community with a rich resource of diverse, fully annotated next-gen models (NGCMs) to better study disease biology.

Historically, traditional cancer cell lines have provided a platform to conduct large scale studies, such as investigating molecular regulations of cancer cell growth and progression, identifying genetic and biological markers, and predicting drug sensitivities. The [Cancer Cell Line Encyclopedia](https://www.cancerrxgene.org/)[^15], a repository of cancer cell lines with associated molecular data and analyses, is a valuable resource of over 1,100 cell lines generated from numerous cancer types. Another resource is the [Cancer Therapeutics Response Portal (CTRP)](https://cancer thermostatx.org/)[^16] which houses a large dataset of quantitative small-molecule sensitivity data of cancer cell lines. This resource could be used to mine for lineages or mutations[^17], enriched among cell lines, that are sensitive to small-molecules and identify new therapeutic vulnerabilities. The next-gen cancer models aim to address limitations of most available cell lines such as poor or unknown representation of the cellular architecture of the original tumor, heterogeneity of cell types and genetic drivers of cancer subtypes[^1].

Next-generation models

The successful culture and expansion of organoids from murine small intestinal tissue paved the way for early organoid models generated from mouse colon and human small intestine and colonic epithelium[^2]. Human intestinal organoids were observed to mimic *in vivo* cellular differentiation; however, adaptations of the culture media were needed to successfully grow organoids from different tissue types. Sato and colleagues reported
optimized cell culture methods utilizing growth factors, **Notch protein** [19] inhibitors, **nicotinamide** [20], and kinase enzyme inhibitors for culturing primary human epithelial [21] cells from small intestine, colon, adenoma [22], adenocarcinomas [23], and **Barrett’s esophagus** [24]. The models were hypothesized to be more representative of the tumor biology than colon cancer cell lines². The group recently published their protocols for generating next-gen cancer models (NGCMs) from breast normal and tumor tissues³.

Identifying an optimized culture medium for each cancer and tissue type is critical for successfully growing next-gen models which retain their originating tumor characteristics. According to Ince and colleagues, ovarian tumor cell lines grown in standard culture media: “(1) had very low success rate (less than one percent) [sic of being established in culture], (2) had long lag times for the first passage, (3) could only be propagated for up to 15 passages and (4) lacked the phenotype of original tumor⁴. The authors developed 25 diverse ovarian cell lines using optimized culture media compositions for each human ovarian cancer subtype. The resulting ovarian cancer cell lines retained the genomic landscape, histopathology, and molecular characteristics of the original tumors from which they were derived. The expression profiles and drug responses of these cell lines were also found to correlate with patient outcomes⁴.

**Applications of next-generation models**

Next-gen models have been shown to be excellent research tools to carry out **ex vivo** experiments as they recapitulate the biology and tissue architecture of primary tumors. A few examples of applications of next-gen models in research include studying disease progression, identifying genomic and molecular drivers of diseases, and screening compounds or small molecules for treatment sensitivity and/or resistance.

- Studying disease progression in pancreatic cancers can be challenging due to lack of patient-derived models that cover the full spectrum of disease progression, lack of clinical correlations, and accumulation of genetic aberrations. Boj and colleagues were able to generate human derived organoids from normal and neoplastic ductal cells using modified culture conditions⁵. Through targeted sequencing of cancer-associated genes on organoids derived from human normal and tumor tissues, the authors identified oncogenic KRAS mutations in majority of the tumor-derived models indicating the organoids represented the cancer driver mutations observed in the originating human tumors.

- Patient-derived organoid models of pancreatic ductal adenocarcinoma (PDAC) were used to test **chemosensitivity** [26] and chemoresistance of individual tumors. One of the limitations of traditional cell lines is that due to genetic drift after multiple passages, there are differences in genetic profiles between the original tumor and the derived cell lines; such as copy number alterations, DNA methylation [27], molecular subtypes, and resulting phenotypes. Tiriac and colleagues found that the patient-derived PDAC organoids harbored genetic alterations that are consistent with known pathogenic mutations in PDAC⁶. The authors concluded that the organoids recapitulated the mutational spectrum and molecular subtypes of primary pancreatic cancer and, therefore, are excellent models to accurately examine and predict responses to chemotherapeutic agents⁶.

- The lack of model systems that reflect the pathology of the primary disease and responses to therapy presents a challenge in studying esophageal adenocarcinoma [28] (EAC)¹. Using patient tumor-derived organoid models, Li and colleagues could identify tumor drivers of EAC through histological and genomic characterization¹. The molecular annotation of the EAC organoids showed that they retained patient-specific gene expression, disrupted cellular polarity, intra-tumor heterogeneity, and drug sensitivity¹. Based on these findings, the use of patient-derived organoids provides model systems to accurately study disease.

- The variability in drug response due to cellular heterogeneity presents another challenge in cancer research using traditional cell lines. Next-gen cancer models were shown to produce reliable responses that resemble those of the originating tumor when screening targeted therapy compounds². Similar to the responses found in mouse organoids, treatment of patient-derived organoids with Itraconazole, a cell cycle inhibitor compound⁸, led to inhibited organoid growth and cell death⁷. Buczacki and colleagues suggested that they found a therapeutic potential of Itraconazole in inducing cancer cell death and preventing late recurrence in colorectal cancer⁷. Hence, patient-derived organoids could be used to identify novel drug targets.

- The lack of cellular architecture to mirror the tumor microenvironment and stromal response presents difficulty in studying immune interactions in cancer. It is challenging to elicit tumor-immune specific
responses using traditional cell lines without a tumor microenvironment. Studying tumors using patient-derived xenografts (PDXs) in immunocompromised mice also does not reflect the immune interactions of human tumors due to lack of immune response. To overcome these challenges, Neal and colleagues used patient-derived 3D models from various tumor types using the air-liquid interface (ALI) method. The models included a mixture of cancer cells and several immune cell types, the latter expressing immune checkpoint surface receptor programmed cell death protein-1 (PD-1). These models enabled the study of immune interactions with cancer cells by providing in vitro tumor microenvironment. To test for anti-PD-1-dependent tumor cell killing, the ALI models were treated with PD-1 blocking antibody. The tumor-infiltrating lymphocytes in the patient-derived models were found to model the immune checkpoint blockade, resulting in cancer cell death. Hence, the use of NGCMs can aid in mirroring the tumor microenvironment in vitro and provide models for immune-oncology research.

The results summarized above, are just a few examples of important outcomes using the NGCMs. The models have a great potential in research to improve our understanding of cancer etiology and the improvement of treatment outcomes.

HCMI next-gen models

The advantages of using HCMI's NGCMs over traditional cell lines include the availability of clinical data such as patient and tumor information, histopathological biomarkers, and molecular characterization data. The comprehensive data increase accuracy in identifying driver mutations and targeted therapies in various cancer subtypes. If interested in using the next-gen models developed by HCMI, visit the HCMI Searchable Catalog to query and browse available cancer models based on a subset of clinical and molecular data. Currently, the catalog includes 35 NGCMs that are from the brain, bone, bronchus and lung, colon, pancreas, stomach, and rectum. Standardized protocols and optimized cell culture media formulations for model growth and expansion for each model are available through the third-party distributor, American Type Culture Collection. These HCMI resources should facilitate the repeatability and reproducibility in growing human cancer models specific to each cancer type. The model-associated genomic and clinical data are quality-controlled and harmonized at multiple checkpoints for reliability of associated data, and are available publicly at NCI's Genomic Data Commons. HCMI next-gen models with associated data provide the research community with a valuable resource to accelerate the translation of research findings to precision oncology.

References

GenePattern Notebook: Integration of Electronic Notebooks with Bioinformatics Tools for Genomic Data Analysis

Over the past several years, the electronic analysis notebook has emerged as an effective and versatile tool for the authoring, publishing, and sharing of scientific research. It allows scientists to combine the scientific exposition – text, images, and even multimedia – with the actual code that runs the analysis, creating a single “research narrative” document that is reproducible, containing all of the computational steps in an analysis; adaptable by other scientists to their own research; comprehensive, conveying research in a high level of detail, without the limitations of publications or paper media; and accessible, often requiring only a web browser to view and run.

The Jupyter Notebook system has become a de facto standard notebook environment in data science and genomic analysis. The community of Jupyter users extends well beyond these, reaching areas of science as diverse as physics, economics, and linguistics. However, the Jupyter notebook format assumes familiarity with a programming language in order to access analyses, and even text must be formatted using a programming-style language.

To extend the capabilities of notebooks to the needs of researchers at all levels of programming expertise, we developed the GenePattern Notebook environment with funding from the National Cancer Institute’s Informatics Technologies in Cancer Research (ITCR) program. GenePattern Notebook integrates Jupyter’s research narrative capabilities with the hundreds of genomic analysis and visualization tools available through the GenePattern platform. The GenePattern Notebook workspace is available for public use and requires registration to gain access. This tool allows scientists to develop, share, collaborate on, and publish their notebooks, requiring only a web browser. In this environment, investigators can design their in-silico experiments, perform and refine analyses, launch compute-intensive analyses on cloud-based and high-performance compute resources, and publish their results as electronic notebooks that other scientists can adopt to reproduce the original analyses and modify for their own work.

The GenePattern Notebook environment provides capabilities beyond standard notebook platforms:

**Access to a wide range of genomic analyses within a notebook.** GenePattern provides hundreds of analyses, from machine learning techniques such as clustering, classification, and dimension reduction, to omic-specific methods for gene expression analysis, proteomics, flow cytometry, sequence variation analysis, pathway analysis, and others. The analyses are launched from a user-friendly form in the notebook and run on a remote GenePattern server, which may be on a cloud provider or hosted at a high-performance compute site. This allows compute-intensive analyses to run in an environment where they are most suited. Results are available within the notebook and may easily be used in other analysis steps.

**Featured notebooks.** A library of featured genomic analysis notebooks is available on the GenePattern Notebook workspace (Figure 1). These notebooks include templates for common analysis tasks (e.g. hierarchical clustering of RNA-seq data, gene set enrichment analysis, non-negative matrix factorization (NMF)), as well as disease-specific research scenarios and compute-intensive methods. Featured notebooks also include those that were developed in collaboration with research labs as a means to disseminate their analysis methods, including the Coordinated Gene Activity in Pattern Sets (CoGAPS) Bayesian NMF method for inference of biological process activity and the AMARETTO multi-omics tool for inference of regulatory networks in cancer and other diseases. A cancer-focused notebook, “Genomic Discovery to Translation”, is aimed at providing insights into candidate drugs for patient therapy. This method combines RNA-Seq profiling data with expression and viability data from cell lines to identify compounds as candidate therapeutics. It uses publicly available data resources, including the Cancer Cell Line Encyclopedia (CCLE), Sanger Cell Line Project (SCLP), Cancer
Scientists can easily copy these notebooks, use them as is, or adapt them for their research purposes. Users with computational experience can modify their own versions of the notebook with variations, for example to try alternative analysis methods, additional data resources, or other ‘omics' data types. An example notebook for the analysis of copy number variation in methylation array data is shown in Figure 2. The GenePattern analysis shown there replaces a considerable amount of code and facilitates analysis for non-programming scientists. Researchers can upload and store up to 30 GB of data and GenePattern development team can increase the size if additional space is required for the analysis.
**Figure 2**: GenePattern Notebook for performing copy number variation analysis on Illumina 450k/EPIC methylation array data.

**Notebook enhancements.** GenePattern Notebooks have several features that enhance the original standard Jupyter notebook interface. First, a rich text editor allows scientists to enter and format text without knowing a text formatting language such as Markdown or LaTeX. Second, users can create a table of contents from the headings in a notebook, which updates automatically as headings are added or changed. It can either be embedded in the notebook or float alongside, allowing easy navigation to any point in a notebook. Third, a user interface-building tool (Figure 3) allows notebook developers to wrap their code so that it is displayed as a web form, with only the necessary inputs exposed. Users of the notebook are presented with a simplified display that allows them to run the analyses without needing to interact with the code behind them.

**Figure 3**: User Interface (UI) Builder: (a) A Python cell containing code to execute an analysis. (b) The UI Builder display hides the Python code, displaying only the required inputs.

**Publication and collaborative editing.** Notebook developers often wish to share their notebooks, either with the research community or among collaborators. To make a notebook publicly accessible, the author selects the “publish” feature and adds descriptive information and tags to make the notebook easy to find in a search query. The notebook is then made available on the “community” section of the workspace. An author can include a web link to a public notebook in a publication, and users who follow the link will see a read-only version of the notebook, with the option to log in to the workspace, where they can run, copy, and edit their own version. For collaborative editing, an author can send a sharing invitation to colleagues, who then can also view, run, and edit the notebook prior to its publication.

The GenePattern Notebook environment is freely available at [http://genepattern-notebook.org/](http://genepattern-notebook.org/). Researchers can make their tools available for public use through the GenePattern server or GenePattern Archive ([GParc](http://gparchive.org)), a community repository. A related GenePattern Notebook resource, the Human Cell Atlas Notebook Workspace, [https://hca.genepattern.org/](https://hca.genepattern.org/), is dedicated to the Human Cell Atlas and features a growing collection of notebooks providing single-cell analysis tools. For more details about the GenePattern Notebook, view the video tutorial at [https://genepattern-notebook.org/tutorials/](https://genepattern-notebook.org/tutorials/). If you have any further questions, please contact...
The use of cancer cell lines to model diverse cancer types have several challenges. Most of the commonly used cancer cell lines in research do not have clinical or therapeutic outcome data of the participant from which the cell line was derived. The genomic relatedness of the cell line to the parent tumor is unknown, and molecular characterization including assessment of genomes and transcriptomes of these cell lines, until recently, were mostly unavailable. Diverse racial and ethnic groups, as well as rare cancers, are seldom represented in currently available cell lines. Additionally, cancer cell lines often do not represent certain molecular subtypes. To address these challenges, the Human Cancer Models Initiative (HCMI) was formed. HCMI is an international consortium which focuses on generating novel, next-generation tumor-derived cancer models annotated with clinical, genomic, and molecular data as a community resource. The HCMI consortium founders include the National Cancer Institute (NCI), Wellcome Sanger Institute, Cancer Research UK, and Hubrecht Organoid Technology. The NCI arm of the consortium supports HCMI next-generation model development at four Cancer Model Development Centers (CMDCs).

To advance cancer research and further understand the relationship between in vitro findings and clinical biology, HCMI next-generation cancer models are associated with clinical data as well as molecular characterization data. To collect the associated clinical data, a cancer type-specific Case Report Form (CRF) is developed for each cancer type for which HCMI models are generated. As of June 2019, eighteen different enrollment and follow-up CRFs for glioblastoma, breast, colorectal, pancreas, pediatric cancers, and others have been developed.

The CRFs function to standardize the clinical data that are collected from participating Tissue Source Sites (TSSs) and are composed of standardized NCI common data elements (CDEs) with a controlled vocabulary or permissible values (PVs). Clinical Data Working Groups (CDWG) consist of cancer type-specific clinical experts including pathologists, oncologists, and surgeons from the United States, United Kingdom, and the Netherlands. They contribute to the content of the CRFs used to collect clinical data. A preliminary list of CDEs is generated according to guidelines from the World Health Organization, International Classification of Diseases for Oncology, and American Joint Committee on Cancer. The CDEs are used as a base on which the working groups build a
Clinical data include the participant’s demographics, prognostic factors, and specifics of the tumor including histological subtype, pathological staging, and grade. Each cancer type-specific CRF includes the current neoadjuvant and adjuvant therapy information and prognostic/predictive/lifestyle feature information based on the feedback from these experts.

A few patients donated tumor tissues from more than one anatomic site for model generation. Examples include a.) a primary tumor and a metastatic lesion, b.) a primary tumor and recurrence, c.) multiple metastases, or d.) a pre-malignant lesion and a primary tumor. With successful generation of multiple models per donor, the CRFs required updating to keep all the associated data in a single form. These CRFs utilize new CDEs to collect clinical data for distinct tissue types including “primary”, “metastatic/recurrent”, or “other” tissue as well as information linking a specific model to its corresponding originating tissue sample. Multiple model CRF design is available for several cancer types including melanoma, brain, hepatocellular, and rare cancers. It is not known if all cancer types will have donors with multiple models, however, the multiple model CRFs are available on the HCMI Resources webpage.

Because HCMI is an international consortium, consideration for differences in clinical data collection are also addressed during the CDWG meetings. Differences in terminology are discussed and incorporated into the CDEs. The Office of Cancer Genomics (OCG) works closely with the NCI's Cancer Data Standards Registry and Repository (caDSR) to ensure that the clinical data elements and metadata used in the HCMI CRFs adhere to a uniform vocabulary and meaning. This data standardization will enable HCMI clinical data to be compatible not only across the global HCMI TSS locations but also across different groups and programs. The use of PVs allows for mapping of the clinical data to NCI's Genomic Data Common's (GDC) data dictionary, enabling users to search and filter the data.

Once the CDEs have been finalized, and they are registered, updated, or modified with NCI's caDSR, the CDEs are submitted to the Clinical Data Center (CDC) at Information Management Systems. The CDEs are compiled into an interactive web-based electronic CRF (eCRF) where TSSs may submit the HCMI clinical data. The clinical data submitted by the TSSs are quality checked (QC’d) to ensure that no personal health information (PHI) is inadvertently submitted, and that the information submitted conforms to the PVs and type of information expected (e.g. numeric values for time interval questions). Once the clinical data are QC’d and any errors are addressed, the clinical data are approved and mapped to the GDC dictionary before submission. The finalized clinical data are submitted and stored at the GDC for access by the scientific community.

PDF versions of the CRFs are also generated and accessible on the HCMI Resources page: HCMI Case Report Forms. These CRFs can be utilized by anyone interested in collecting clinical data associated with cancer tissues for other projects. Check the HCMI Resources page for updates to the CRFs as new cancer types are added. The HCMI models and associated high-quality clinical data coupled with molecular characterization data provide the scientific community with a valuable resource for cancer research.

OCG PERSPECTIVE

Leveraging a Genomics Background to Facilitate Molecular Characterization of HCMI Models

Lauren Hurd, Ph.D.
Office of Cancer Genomics, NCI
My name is Lauren Hurd and I am a Scientific Program Manager for the Human Cancer Models Initiative (HCMI) within the Office of Cancer Genomics (OCG) at the National Cancer Institute (NCI), together with other consortium members, co-founded HCMI to provide a resource of ~1,000 clinically and molecularly characterized next-generation patient-derived cancer models to the research community. The cancer models are generated from parent tumors that span a range of subtypes originating from individuals of diverse ethnic and racial backgrounds as well as rare adult and pediatric tumors. The goal of this initiative is to provide a resource of diverse, fully annotated models which more accurately recapitulate the biology of their parent tumors. As a scientific program manager, I work with Daniela Gerhard, the Director of OCG to ensure that this goal can be met for the United States’ (U.S.) contribution to the HCMI.

For the last eight years, I have been working within the field of genomics in dynamic and challenging roles. My interest in genomics was cultivated in graduate school, where I focused on understanding the relationship between pathogenic variants in a nonselective cation ion channel and an autosomal dominant form of skeletal dysplasia, a disorder that affects the growth of bone and cartilage. I continued my career in genomics during my postdoctoral fellowship where I led the technical direction of molecular diagnostic testing for rare pediatric disorders in a clinical sequencing lab. It was in this role that I gained extensive knowledge of both genomic testing and genomic data. I developed Sanger and NGS panel testing from the ground up, interpreted variants identified during testing and reported on clinically actionable results. Meaningful interpretation or curation of the variants identified through the various sequencing technologies quickly became one of my favorite parts of the job. Prior to joining NCI, I managed a team of variant curation scientists who were responsible for interpreting variants identified during routine carrier screening of autosomal recessive and X-Linked pediatric disorders. I also collaborated with a multi-disciplinary team from bioinformatics, marketing and clinical operations to develop new versions of the carrier screening panel which required critical management of many deliverables on strict timelines.

HCMI models within the U.S. are molecularly characterized with genomic and transcriptomic data for the model as well as the associated normal, and parent-tumor. Annotating the models with this data, to the best of our ability, is a complex process. I leverage my strong background in molecular biology and genetics as well as my experience in project management to facilitate this process (Figure). I work closely with the Biospecimen Processing Center (BPC) to ensure that quality nucleic acid samples from model, normal and parent-tumor can be provided for downstream sequencing to the Genomic Characterization Centers (GCCs) and subsequently harmonized by the Genomic Data Commons (GDC). This involves monitoring hundreds of samples through the molecular characterization pipeline, tracking histopathology and nucleic acid metrics, and ensuring sequencing strategies are adjusted as needed. I also collaborate with these teams to identify potential challenges in the molecular characterization process and develop resources which address those challenges. I've developed standard operating procedures (SOPs) for the Data Coordinating Center (DCC) to increase the interoperability of model QC data and the BPC to streamline the nucleic acid isolation process. HCMI models are most valuable when they contain comprehensive and standardized datasets and I do the best I can to ensure that we provide these datasets to the research community.

**Programmatic Oversight of NCI HCMI Molecular Characterization**

<table>
<thead>
<tr>
<th>Sample Collection</th>
<th>Biospecimen Processing</th>
<th>Sequencing</th>
<th>Harmonization &amp; Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples of model (M), normal (N) &amp; tumor (T) provided by model developers</td>
<td>Histopathology &amp; DNA/RNA isolation completed by BPC</td>
<td>GCC performs 150X WXS &amp; 15X WGS of N, T, M and RNA-Seq on T &amp; M</td>
<td>Sequencing data harmonized by GDC analytical pipelines</td>
</tr>
<tr>
<td>Ensure all sample components are sent for biospecimen processing</td>
<td>Ensure quality metrics are met, adapt QC and SOPs as needed</td>
<td>Track &amp; adjust sequencing strategies as needed</td>
<td>Monitor data release</td>
</tr>
</tbody>
</table>

*Figure: Programmatic Oversight of NCI HCMI Molecular Characterization (A) Sample Collection*
While I am extremely familiar with pediatric and rare disease genomics, the field of cancer genomics is mostly uncharted territory for me. The HCMI program has provided a whole new perspective on learning about genomics. Providing models annotated with clinical, genomic and transcriptomic datasets provides researchers with a crucial resource to address large questions within the field. Study of the models will advance our basic understanding of tumor heterogeneity, genomic stability, tumor-immune microenvironment and mutational signatures on a more comprehensive scale. Large repositories of molecularly diverse models from the same tumor subtype will also allow for greater representation when screening for novel therapeutic targets as well as assessing drug sensitivities. All of this will certainly culminate in the advancement of both novel and improved precision therapies. It is exciting to be a part of this initiative where the possibilities for discovery seem endless.

The human nuclear genome consists of approximately 3.2 billion nucleotides. An extraordinary amount of biological information is brought together both in precise order and time to form the foundation of human life. It’s fascinating and logical, yet puzzling, all at the same time. It’s certainly what has driven me to pursue a career in genomics. There are so many ways in which one can work in the field of genomics and I have been fortunate to work in this field in many capacities. While the career path I have taken is varied in its relationship to the field of genomics, one commonality remains. I work with teams and initiatives who excel at providing answers to the greater scientific community and the HCMI is no exception.