Specific Aims

The CIMAC-CIDC Network will extend the currently limited benefits of immuno-therapy to majority of patients by providing insights into tumor-immune interactions at the cellular and molecular level within the context of the tumor microenvironment and at the level of systemic immune networks. The insights will accelerate the discovery of biomarkers for precision therapy, identification of new drug targets, and the discovery of drug combinations that will unleash the power of the immune system against cancer. Toward this goal, playing a key role within the Network, the CIDC will provide coordination and integrated informatics to enable correlative analyses of comprehensive molecular profiles and clinical variables for CIMACs-supported, NCI-supported and external clinical trials. Specifically, the CIDC will accomplish the following:

Aim 1. Establish a standards-based multi-cloud FAIR-compliant data commons for collection and management of clinical and tumor-immune profiling data. (A) Develop the first free open-source standards-based data commons software stack by implementing the Genboree Stack SaaS layer on top of the OpenStack IaaS layer and deploy it using Rackspace services across major commercial and private clouds. **(B)** Eliminate data "silo-ization" and enable data reuse and collaboration by establishing a FAIR-compliant OpenAPI- and dashboard-accessible database for tumor-immune profiling data. **(C)** Develop a virtual biorepository to capture clinical data and track biosamples across clinical trials.

Aim 2. Implement data collection standards, tools and processes for clinical tumor- and immuno-profiling assays. (A) In collaboration with the CIMACs, identify key assay platforms for clinical tumor- and immuno-profiling and establish (meta)data collection standards for the platforms. **(B)** Deploy tools and implement procedures for (meta)data modeling, validation and sharing.

Aim 3. Provide bioinformatics pipelines and interfaces to integrated tools and resources for multi-dimensional correlative analysis of immuno-therapy trials. (A) Deploy bioinformatics pipelines and continually adapt them to extract the increasing diversity of multi-layered molecular information relevant for immuno-oncology. **(B)** Provide interfaces to integrate local and external tools and resources required for correlative analyses; address the complexity of the tumor microenvironment; embrace a systemic approach by connecting tumor profiling and liquid biopsy.

Aim 4. Provide alignment, integration and synergy with external research resources and provide the wider research community with access to Network resources and services. (A) Align and integrate Network resources with external databases and computational resources. **(B)** Provide outside investigators and research community with access to CIDC resources and services.

Aim 5. Provide Network coordination and facilitate Network interactions. (A) Provide coordination, administrative and logistic support for the Network. **(B)** Facilitate Network interactions with the broader immune-oncology research, translational, and clinical communities.

Additionally, based on the experience gained during the first few years of this project and under guidance of an advisory board composed from leaders in industry and academia we will develop and implement a plan for transitioning CIDC into a self-sustainable public informatics resource for immuno-oncology.

Sub-section A: Overall vision.

Cancer immunotherapies have demonstrated remarkable clinical benefits for a fraction of cancer patients. It is anticipated that such benefits may be extended to many more patients through increased knowledge about clinically relevant tumor-immune interactions at the cellular and molecular level within the context of the tumor microenvironment, as well as through a deeper understanding of the systemic impact of immunotherapeutics. In pursuit of this opportunity, CIDC-CIMACs Network will perform correlative analyses of molecular profiles and clinical variables to identify biomarkers for precision therapy, new drug targets, and drug combinations that will unleash the power of the immune system against cancer and save many more patients' lives.

Most of the assays for high-throughput molecular profiling of both circulating immune cells, exosomes, cell-free protein, DNA and RNA, and the tumor-immune microenvironment require capabilities that are not readily available to clinical investigators. CIMACs will bridge this gap by providing NCI-supported Phase-1 and 2 trials and other clinical investigations with experimental platforms, informatics, and statistical support for deep molecular profiling of tumors. The primary goal of the CIDC will be to provide comprehensive support for all stages of correlative studies and create a Data Commons of reusable data (**Fig.1**). The first step will be to standardize and integrate clinical and biospecimen information across geographically dispersed CIMACs-supported trials. Toward that end, we will deploy a virtual biorepository system built on the model of the extracellular RNA virtual biorepository (EVB) system we deployed for the exRNA Communication consortium. To standardize the processing of diverse molecular profiling data, we will validate, customize, and deploy bioinformatics pipelines in a secure and scalable way in a multi-cloud fashion that will allow us to optimally combine resources offered by cloud providers. Because of the leading role of the Gerstein laboratory within ENCODE, we will have access to the most accurate and well characterized pipelines, and Gerstein's team will customize them for immunotherapy applications to extract the rapidly increasing diversity of multi-layered molecular information relevant for immuno-oncology.

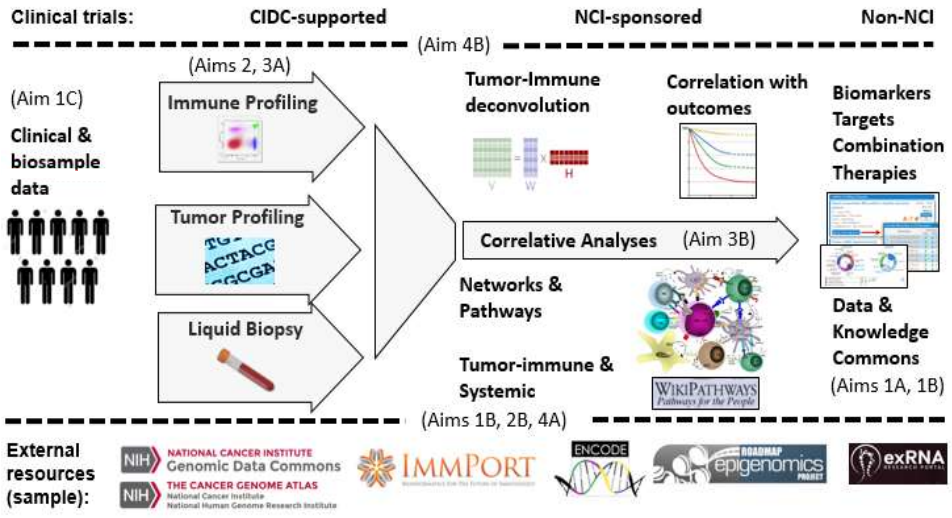

Figure 1. Correlative analyses of molecular profiles and clinical variables in cancer immuno-therapy trials. Note Specific Aim numbers.

To facilitate correlative analyses, we will utilize the established plug-in architecture and integration framework of the Genboree Commons. We will develop interfaces to integrate local and external tools and resources required for correlative analyses; by applying advanced tumor deconvolution methods that we recently developed[43], we will leverage the reference epigenomes from the NIH Roadmap Epigenomics and IHEC projects and those in the GEO archives and the TCGA profiles. Using our extensive experience in correlative network reconstruction, we will infer tissue-immune interaction networks that will inform clinical biomarker studies. By employing a Linked Data integration framework, our platform will establish links with pathway databases[52] allowing us to contribute tumor-immune interaction pathways to WikiPathways and utilize the pathways as a context for interpreting correlative results. To integrate tumor profiling data with data derived from potential liquid biopsy biomarkers such as exosomes and circulating DNA and RNA, we will deploy and extend the pipelines and integrated toolsets that we developed as part of the Extracellular RNA Atlas[3]. Through the synergy of these advanced analytical approaches, we will create a Data Commons of reusable data, empower clinical investigations, enable life-saving discoveries, as well as expand our knowledge and resources for immuno-oncology research.

Another CIDC core task is the storage, processing, and management of the large volume and extensive diversity of molecular and clinical information. To address the needs for economy, scalability, and security,

CIDC will be built on a scalable multi-cloud (AWS, Microsoft Azure, OpenStack, private/academic clouds) provided by Rackspace's low-latency, high-bandwidth, secure RackConnect Global multi-cloud integration services. The multi-cloud strategy will reduce reliance on a single vendor, mitigate against disasters, and facilitate integration with data and tools hosted across the clouds connected within a multi-cloud. The multi-cloud "Infrastructure-as-a-Service" (IaaS) layer will provide scalable storage, compute, and data transfer resources as well as security, compliance, and account management services. Sharing and reuse of tools will be enabled by virtualization and containerization. Our multi-cloud strategy will also enable high-bandwidth connections with the NCI GDC, ENCODE, and other critical data sets. This will facilitate integrative analyses and lower the barriers for the reuse of the CIDC-hosted tumor-immuno-profiling data by the immune-oncology community.

The CIDC will break new ground by deploying Genboree Stack, the first complete free open source system for providing a Data Commons "Software-as-a-Service" (SaaS) on top of the OpenStack Infrastructure-as-a-Service (IaaS) (**Fig.2**). Genboree Stack consists of the Genboree Commons, GenboreeKB, GenboreeAPIs, and the Genboree Workbench (not shown in **Fig.2**). Genboree Commons is an extended open source, Ruby-on-Rails-based system that provides authentication, group- and role-based authorization, and collaborative features, built on top of Redmine's[38] framework for integrating interactive tools such as dashboards and interactive analysis tools via plugins. GenboreeKB and the Genboree API enable sharing and reuse of data and tools based on FAIR (Findable, Accessible, Interoperable, Reusable) principles. Genboree Stack components have been extensively tested and deployed to serve multiple projects (ClinGen, NIH Roadmap Epigenomics, exRNA Communication Consortium) and will need to be only incrementally improved and extended to meet the needs of this project. Using this Data Commons infrastructure, CIDC will disseminate immune-oncology data and knowledge to the broader immune-oncology research communities and will effectively provide external investigators with access to CIDC services.

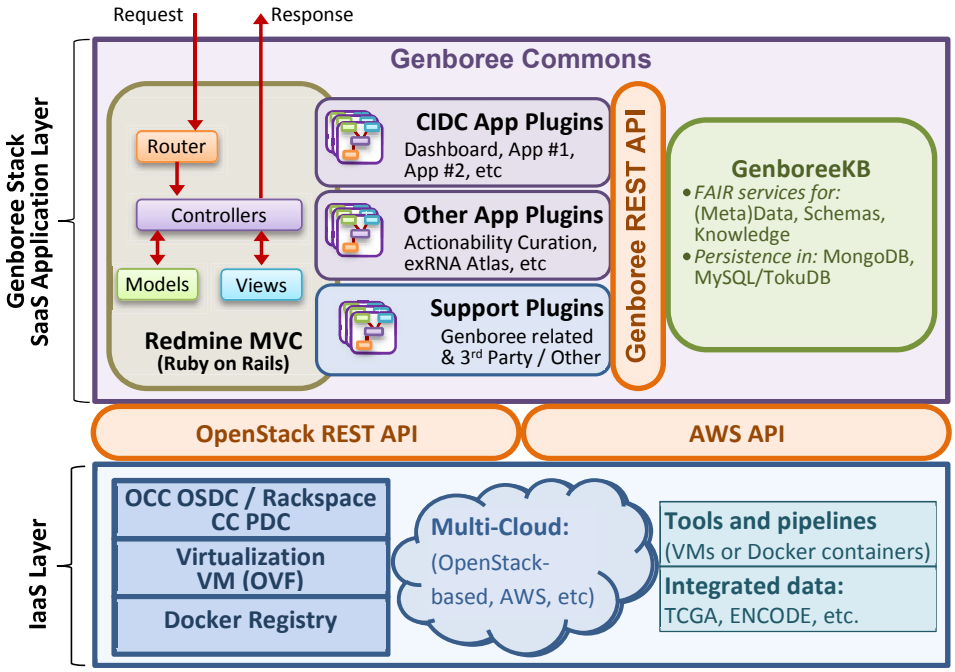

Figure 2. Overview of the Genboree Stack architecture and proposed integration with the IaaS layer.

Transition of the CIDC into a sustainable resource. Based on the experience gained during this project, with support from the Baylor College of Medicine and under guidance of an advisory board composed from leaders in the biotech/pharma and cloud-computing industries and academia, we will develop and implement a plan for transitioning CIDC into a self-sustaining public informatics resource for immuno-oncology.

During the first two years of the project we aim to identify key stakeholders and potential partners, assess unmet needs, chart the competitive landscape, and develop a business model that incorporates NCI/CTEP guidelines. During this initial period we will test the SaaS business model using Rackspace as the IaaS partner. To evaluate the potential revenue stream, a reseller business partnering program will be established with Rackspace Inc. To seamlessly pass the resource usage costs to end-users, billing and account management functions will be integrated between the SaaS and IaaS layers. This testing period will allow us to validate SaaS/IaaS integration, establish basic operations, quantify demand, make realistic estimates of revenue and cash flow, and address issues such as the need for consulting vs. self-service and gauge user perceptions that may affect demand such as information security.

By the end of year two, and based on input from the LCC and the NCI, we will recruit an Advisory Board including thought leaders from industry and academia. As evidenced from the letters (see attached LOS) of Dr. Kent Osborne, the Director of the Dan L. Duncan Comprehensive Cancer Center at BCM and Dr. Adam Kuspa, Senior VP for Research at BCM, the College is extremely interested in pursuing this venture and Dr. Kuspa has pledged $100K toward startup costs of informatics infrastructure. Both Drs. Osborne and Kuspa, who have extensive experience in spinning out both for-profit and non-profit startups and have extensive contacts in industry and academia have pledged to participate in an initial Advisory Board. The PIs also have access to a network of contacts from which they will recruit additional members of the Advisory Board.

In year three, under guidance of the Advisory Board, we anticipate making a decision on whether to partner with an established organization or form a new not-for-profit corporation with a mission to advance clinical immuno-oncology by providing a commons of immuno-oncology data, computational resources, and knowledge. Partnering opportunities may emerge from the pre-competitive data-sharing initiatives of forward-thinking pharmaceutical companies such as GlaxoSmithKline that created a clinical study register for sharing data with the scientific community[51]. Notably, such resources, while making valuable data available to the scientific community, do not provide readily available analytical tools and pipelines. We also see an opportunity for a private-public partnership modelled off of other successful efforts, such as the FNIH Biomarker Consortium[15]. In light of these developments, we anticipate multiple options and making an informed decision based on operational experience, input from stakeholders and advice from thought leaders and anticipate the completion of the partnering arrangement or launch of the non-profit by the end of the third year of the project.

Innovation. This project will develop the following innovative solutions:

1. Track biosamples and clinical data across trials by deploying a Virtual Biorepository.
2. Integrate and deploy Genboree Stack/OpenStack, the first free open source SaaS/IaaS combination to serve the needs of researchers wishing to launch new Data Commons.
3. Integrate data and resources across commercial and academic clouds via a multi-cloud IaaS strategy.
4. Eliminate data “silo-ization” and enable integration if CIDC into the NIH Data Commons and the emerging ecosystem of Linked Data by implementing FAIR principles.
5. Employ ENCODE and other advanced pipelines and adapt them to extract the rapidly diversity of multi-layered molecular information relevant for immuno-oncology.
6. Apply advanced tumor deconvolution methods to eliminate confounding in correlative analyses and uncover tumor-immune interactions leveraging NIH Epigenomics Roadmap, GEO and TCGA profiles.
7. Perform advanced correlative network reconstruction of deconvoluted profiles to uncover tissue-immune interaction networks from TCGA profiles that will inform clinical biomarker studies.
8. Create a tumor-immune knowledge commons by contributing interaction pathways to WikiPathways and by utilizing the pathways as a context for interpreting correlative results.
9. Correlate systemic liquid biopsy biomarkers such as exosomes and circulating DNA and RNA with tumor profiling data to reduce the need for tumor biopsy and enable clinical application of biomarkers.

Sub-section B: Scientific and technical capabilities.

Genboree Stack and SaaS/IaaS cloud deployment. Genboree Stack, a platform consisting of free open-source interoperable software components for implementing Data Commons emerged form a decade of data coordination experience by the Baylor team, including the TCGA pilot, NIH Roadmap Epigenomics, NIH Extracellular RNA Communication Consortium (exRNA.org), and the NIH-NHGRI Clinical Genome Resource (ClinGen). The team has a decade of experience in using Rackspace services for managed hosting and, more recently, cloud hosting of Genboree. ClinGen apps (reg.clinicalgenome.org, calculator.clinicalgenome.org, and actionability.clinicalgenome.org) currently run on a Rackspace-hosted (IaaS) Genboree Stack (SaaS).

(Meta)data modeling and alignment. Milosavljevic and Cheung have led multiple national and international (meta)data standardization efforts and collaborated on community-based ontology development[23] and data interoperability development[16] Milosavljevic led the effort to harmonize ENCODE and the NIH Roadmap Epigenomics data[45]. As the Steering Committee Member of the International Human Epigenome Consortium, Milosavljevic drafted metadata standards and coordinated their global adoption. Cheung chaired the BioRDF Task Force of the W3C Semantic Web for Health Care and Life Science Interest Group and is currently leading data modeling efforts for the Human Immunology Project Consortium.

Pipelines for advanced molecular profiling. Gerstein lab has extensive experience in developing advanced pipelines for extracting multi-layered molecular information of relevance for immuno-oncology such as DNAseq, RNAseq, and ChIPseq data. He is leading data analysis for ENCODE[31], and has previously led data analysis for mod/ENCODE[27, 32] and played a key role within the 1000 Genomes consortium[30]. Gerstein has also developed advanced analysis methods for proteomics[47, 49, 53] and metabolomics[39] data.

Molecular profiling of cancer. Gerstein is currently co-leading the International Cancer Genomics Consortium (ICGC) pan-cancer analysis-working group (PCAWG)-2 (analysis of mutations in regulatory regions) group. In addition, we have participated in two The Cancer Genome Atlas (TCGA) studies on comprehensive molecular characterizations of 333 primary prostate carcinomas[20] and 161 primary papillary renal-cell carcinomas[21].

Cancer genomics tools developed in Gerstein laboratory. Gerstein lab has developed Variant Annotation Tool (VAT)[33] that annotates the impact of protein sequence mutations; ALoFT tool that predicts the impact of potential loss of function (LOF) variants in protein-coding genes; STRESS[25] tool that employs models of conformational change to predict allosteric residues. The lab developed methods to predict variants that are disruptive to a TF-binding motif in a regulatory region[27] and has integrated these methods into a prioritization pipeline for variants from WGS profiling called FunSeq[29, 36] that identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast, and prostate cancer samples[36].

Transcriptome analysis tools developed in Gerstein laboratory. RNA-Seq provides a further layer of data which can provide valuable information, for instance the contribution of alternative splicing and non-coding RNAs to tumor and immune gene regulation. Gerstein lab has extensive experience in developing RNA-Seq processing pipelines as part of the mod/ENCODE consortia[27, 32]; has developed tools for identifying non-coding transcription and novel transcribed elements[18, 24, 32, 40, 46]; has developed the exceRpt[37] pipeline for extracellular small RNA-Seq profiling; and has developed tools and data formats for extremely large quantities of RNA-Seq data[34, 54].

Epigenome analysis and deconvolution in Milosavljevic laboratory. Milosavljevic led the Data Analysis and Coordination Center for the NIH Roadmap Epigenomics Consortium[45] and developed Epigenomic Deconvolution[43], a method that deconvolutes epigenomic, transcriptomic, and other “omic” profiles of complex tumors to infer cell type composition and “omic” profiles of constituent cell types.

Liquid biopsy. Gerstein and Milosavljevic lead data analysis and coordination for the Extracellular RNA Communication Consortium; have developed informatics methods, tools, and pipelines for the analysis of circulating RNA in human body fluids[52]; and have constructed the exRNA Atlas[2].

Network modeling in Gerstein laboratory. Gerstein lab has pioneered network frameworks for integrating a great variety of genomic data[22, 31, 55] and investigated the dynamics of networks[19, 48], thus laying a methodological groundwork for modeling tissue-level heterotypic tumor-immune interactions as well as modeling systemic effects of immuno-therapy.

Consortium administration and coordination. Dr. Matt Roth is the Administrative Core Director for the NIH Extracellular RNA Communication Consortium. The Core organized eight semi-annual two-day Consortium meetings (>100 attendees); interfaced directly with hotel management to secure meeting rooms and facilities; created a website to manage conference and workshop registration, and abstract submission; and conducted multiple on-line surveys to capture feedback and improve content and structure of subsequent meetings.

Development of data resource sharing policies. Serving on the steering committees of several consortia, Gerstein and Milosavljevic drafted data sharing policies, led the effort to collect feedback from the Consortium members, modify and adopt the policies. This experience will help articulate and build consensus on data resource sharing policies for the CIDC-CIMACs Network.

Qualifications of the team to lead transition to sustainability. Milosavljevic and Roth held executive-level positions in startup companies. Milosavljevic is a Registered Patent Agent with experience in patent

prosecution & licensing in biotechnology and is thus qualified to evaluate intellectual property issues. Gerstein is on SABs of several leading bioinformatics companies. As a VP of Research in the Otorhinolaryngology Department, member of the NCI Head and Neck Cancer Steering Committee, and PI & IND-holder for active cancer immuno-therapy trials, Sikora brings expertise and connections in the field of clinical immuno-oncology.

Sub-section C: Informatics infrastructure and bioinformatics support functions of the CIDC.

C.1. Aim 1. Establish a standards-based multi-cloud FAIR-compliant data commons for collection and management of clinical and tumor-immune profiling data.

C.1.a. Aim 1a. Develop the first free open-source standards-based data commons software stack by implementing the Genboree Stack SaaS layer on top of the OpenStack IaaS layer and deploy it using Rackspace services across major commercial and private clouds.

Deploy the Genboree Stack SaaS. Based on our extensive experience running data coordination and analysis centers for major projects (as described in **Sub-section B**) we have developed a set of interoperable software components that we refer to as "Genboree Stack" (**Fig.2**). They include Genboree Commons, GenboreeKB[52], Genboree APIs, and Genboree Workbench[26, 44] (not shown in **Fig.2**). Genboree Commons provides a platform for tool integration and extensive support for collaboration, including role- and group-based authorization, document sharing, forums, wikis, and an issue-tracking system for workflows such as software bug report tracking. Being an open source project, Redmine is extensible to meet specific needs as they emerge. Genboree APIs provide programmatic interoperability and GenboreeKB provides (meta)data modeling and storage. Genboree Workbench is a generic web-based UI for tools, pipelines, and data exposed via Genboree APIs. While many of our users find this generic interface useful, for CIDC purposes we anticipate building interfaces that are focused around CIDC-specific use cases such as dashboards (**Aims 1b** and **2b**) and correlative analysis tools (**Aim 3b**) using Bootstrap front-end framework and the Ruby-on-Rails back-end framework of the Genboree Commons.

Integrate the OpenStack IaaS layer. Genboree Stack (Software-as-a-Service layer) will be deployed on the IaaS layer using Rackspace's Managed Cloud services providing scalability. Compute-intensive pipelines will be deployed as templates of stacks of pre-configured virtual machines that are created on demand by Rackspace's Cloud Orchestration service.

For maximum portability, deployment scripts and IaaS-related functionalities will be implemented using OpenStack API[13], a REST-based standard API for interacting with cloud infrastructure, such as managing, creating and deleting virtual machines and network configurations. This approach enables easy migration of the systems between clouds in the future, including Open Science Data Cloud[12] that hosts the NCIs Genome Data Commons (GDC). Our preferred IaaS provider is Rackspace[14] because of the excellent service they provided to us over the years (see **Sub-section B**).

For full cross-cloud connectivity and multi-cloud computing we will employ Rackspace's RackConnect Global service. RackConnect provides private, fast and safe connections between servers in the Rackspace cloud and between clouds including Microsoft Azure or Amazon Web Services. The service supports cross-cloud activities such as data transfers, synchronization of data backups, and load balancing while also providing high levels of data security for transfers of sensitive information.

To address data security for clinical studies, Rackspace provides HITRUST CSF-certified hosting environment. HITRUST CSF[5] is the most widely-adopted security framework in the U.S. healthcare industry that includes and harmonizes nationally and internationally accepted ISO, NIST, PCI and HIPAA standards. Rackspace has already proven their expertise in managing sensitive systems by providing infrastructure for projects such as The National Kidney Registry[11].

While Rackspace's managed cloud services that we propose to utilize have been available for several years now, there is a risk that they may become unavailable during the course of this project. In this unlikely scenario, we will turn to a number of other commercial OpenStack cloud providers such as Internap[9] or DreamHost[1].

C.1.b. Aim 1b. Eliminate data "silo-ization" and enable data reuse and collaboration by establishing a FAIR-compliant and dashboard-accessible database for tumor-immune profiling data.

Implement a FAIR-compliant dashboard-accessible database. CIMACs-supported clinical studies will involve rich clinical information about patients as well as information about patient biosamples, molecular profiles, and downstream results. The utility of such information beyond a specific clinical study is too often limited because the information is not FAIR (Findable, Accessible, Interoperable, Reusable). To address this gap, we will align (meta)data with community standards, as described in **Aim 2a** and will deploy tools to validate, store, and distribute all clinical, immunological and "omic" (meta)data, as described in **Aim 2b**. Using the (meta)data and OpenAPIs (described below) we will develop immuno-therapy dashboards (**Fig.3**) that are

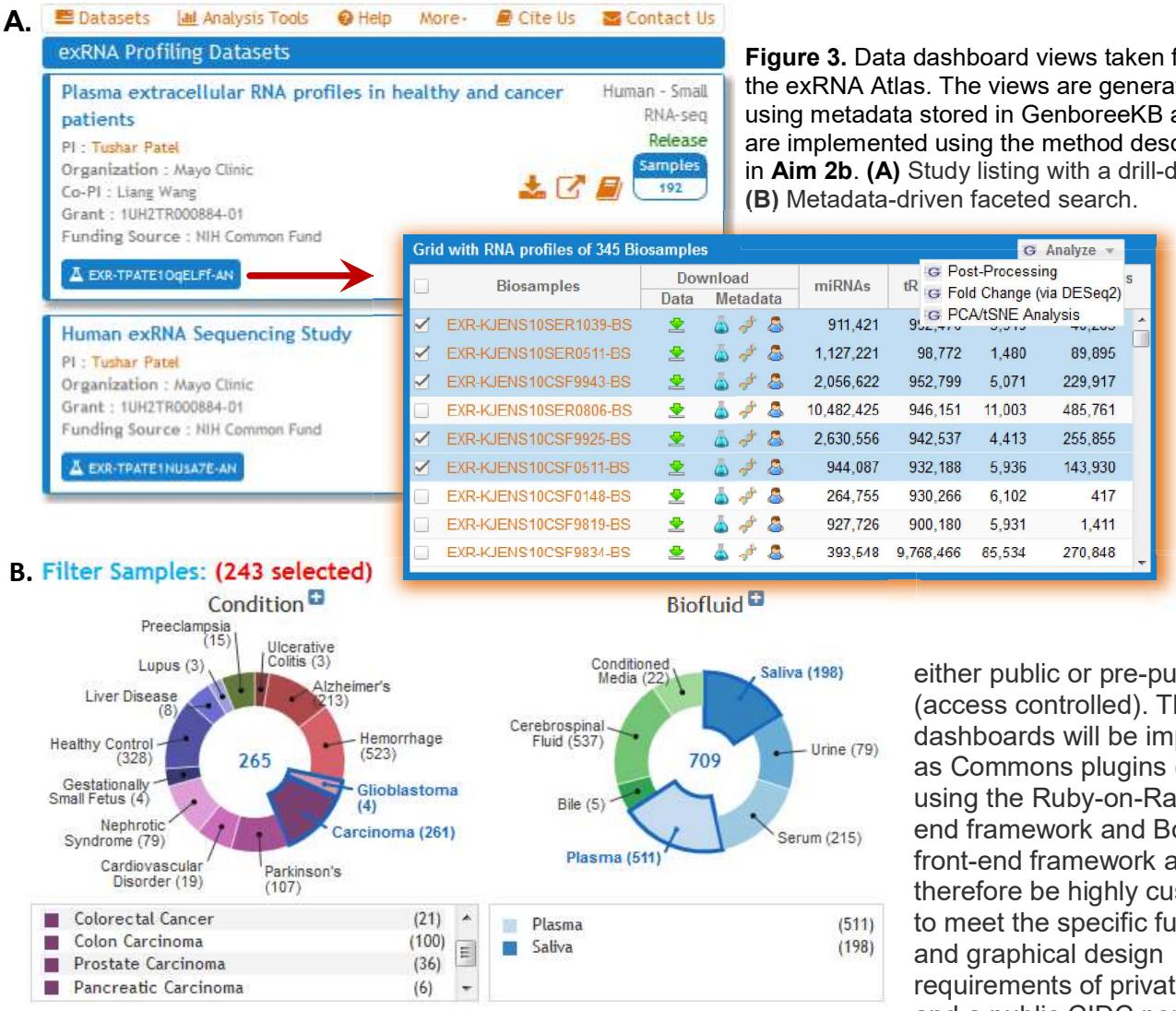

Figure 3. Data dashboard views taken from the exRNA Atlas. The views are generated using metadata stored in GenboreeKB and are implemented using the method described in **Aim 2b**. **(A)** Study listing with a drill-down. **(B)** Metadata-driven faceted search.

either public or pre-public (access controlled). The dashboards will be implemented as Commons plugins (**Fig.2**) using the Ruby-on-Rails back-end framework and Bootstrap front-end framework and will therefore be highly customizable to meet the specific functional and graphical design requirements of private projects and a public CIDC portal.

Address FAIR principles by developing OpenAPIs. Access to the CIDC data in standard formats will be provided via OpenAPI interfaces implemented as wrappers on top of existing Genboree REST APIs. The dashboards discussed above and public facing portals will utilize the same OpenAPIs. To facilitate integration, OpenAPIs include automatically generated documentation of methods, parameters and models. The APIs will support searching and downloading subsets of data files based on the metadata as well as BAM slicing and will support uploads and data submissions.

To facilitate data publication, we will utilize the GenboreeKB versioning system and will implement mapping of select dereferencable URIs to Document Object Identifiers (DOIs). For maximum interoperability, the new API developments will be coordinated with the GA4GH, and the FAIR-ness Working group of the NIH Common

Fund project where our laboratories are already represented. As indicated in the letter by Dr. Grossman, we will align the APIs with those of the GDC.

C.1.c. Aim 1c. Develop a virtual biorepository to capture clinical data and track biosamples across clinical trials.

One challenge for CIDC will be to track profiled biosamples and potentially heterogeneous clinical information across a multiplicity of clinical studies. We will address this by deploying the Virtual Biorepository (VB) application. Built for the Extracellular RNA Communication Consortium using the Genboree Stack (**Fig.2**), VB offers a distributed and extensible infrastructure for sharing information about biosamples and associated clinical information. Each group participating in the VB manages a cloud-hosted "node" that is implemented as a virtual machine and hosts group's biosample and some associated clinical data. The clinical data can be modeled locally using GenboreeKB with some help and instruction. Central VB hub communicates with the nodes via Genboree APIs and provides a dashboard, search and query tools for biosamples across the nodes. Because VB is designed to be generic, no programming will be required to develop a CIDC virtual biorepository and track biosamples and shared clinical data across participating clinical studies.

C.2. Aim 2. Implement data collection standards, tools and processes for key tumor- and immuno-profiling assays.

C.2.a. Aim 2a. In collaboration with CIMACs identify key assay platforms for tumor- and immune-profiling and establish (meta)data collection standards for the platforms.

We will develop a standards-based ontology-driven approach to data annotation and integration. Making use of established NCI data elements and ontologies in the immunology and cancer domains we will annotate and index metadata in CDIC and other related databases. Dr. Cheung is part of the data standards effort within the Human Immunology Project Consortium (HIPC)[4] as well as the CEDAR metadata effort[42] as part of the Big Data to Knowledge (BD2K) initiative. Both efforts involve interaction/collaboration with the Immunology Database and Analysis Portal (ImmPort)[8] that provides support in the archiving and exchange of scientific data for the diverse community of immunology researchers supported by NIAID/DAIT and serves as a long-term, sustainable archive of research and clinical data. Dr. Cheung has been working on mapping ImmPort metadata, data elements and value sets to ontologies available through BioPortal.

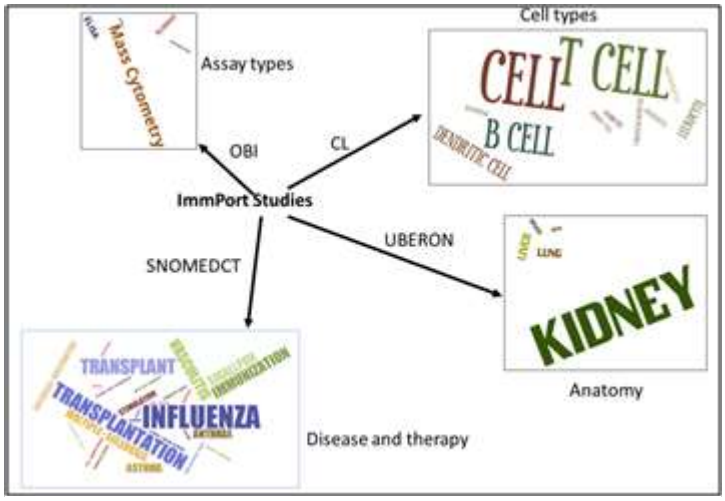

Figure 4. Ontology annotation of terms found in the description (metadata) of a subset of research/clinical studies with their data stored in ImmPort.

As an example, the ontology mapping/annotation shown in **Fig.4** was performed using NCBO Annotator[35] using free text description of each study and target ontologies (e.g., OBI[17], CL[28], UBERON[41], and SNOMEDCT[50]) as input to NCBO Annotator. The output in this example shows that the terms in the study description are mapped to concepts in different categories (e.g. cell types, body parts, assay types and diseases). Such ontology annotation can be used to index the data to support semantic data queries (e.g. faceted search). In this case, we can use the annotation to find ImmPort studies that feature a certain cell type (e.g., T cell) and assay type (e.g., ELISPOT). This annotation approach will enable integrated queries across other major immunology databases of relevance for cancer research, including IEDB[6], ImmGen[7], and ITN[10].

In collaboration with CIMACs and the NCI GDC (see letter from NCI GDC PI Bob Grossman), we will identify key assay platforms as well as biospecimen and clinical (meta)data and adopt standards for data interoperability. As described in **Sub-section B**, Drs. Cheung and Milosavljevic have extensive track record in leading (meta)data standard development at the national and international level and will use their experience toward this end.

C.2.b. Aim 2b. Deploy tools and implement procedures for (meta)data modeling, validation and sharing.

Making direct use of the work by Dr. Cheung within CEDAR, we will deploy CEDAR's template authoring tools to enable creation of ontology-linked templates for (meta)data submission and validation. The CEDAR-created templates will be adopted by CIDC for all metadata as well as low-volume data submissions and validation. The validated metadata and low-volume data will be permanently stored within GenboreeKB, a free open source NoSQL-based document-oriented system supported by MongoDB that offers a flexible service for hosting structured data with defined schemas such as those developed by Dr. Cheung and the CEDAR project. GenboreeKB embodies our extensive experience in metadata modeling, processing, storage and distribution, as outlined in **Sub-section B**. Through integration with other layers of the Genboree Stack, the metadata stored within the GenboreeKB can be used by Genboree Commons plug-ins and is exposed via Genboree REST APIs for interoperability and integration. For example, the dashboard plug-ins illustrated in **Fig.3** are built using the Genboree Commons Ruby-on-Rails back-end framework, Bootstrap front-end framework and access the metadata stored in GenboreeKB via Genboree REST APIs. This set of platforms will therefore be ideal for developing and deploying metadata-driven applications (whether developed by CIDC or others) and for bi-directional integration with NCI GDC and external resources.

C.3. Aim 3. Provide bioinformatics pipelines and interfaces to integrated tools and resources for multi-dimensional correlative analysis of immuno-therapy trials.

C.3.a. Aim 3a. Deploy bioinformatics pipelines and continually adapt them to extract the increasing diversity of multi-layered molecular information relevant for immuno-oncology.

Genomic profiling. At the level of DNA sequencing, we will help identify somatic variants, particularly in immune related genes such as MHC by developing a pipeline to process WGS data to call germline and somatic variants. From the germline variants, we will identify patient HLA type. From somatic variants, we will identify variants in immune-related genes, form a list of mutated protein coding sequences, and estimate tumor heterogeneity.

Transcriptomic profiling. RNAseq provides a further layer of information valuable to immuno-oncology. Because splicing variation is a significant determinant of tumor regulation, and differential splicing of immune-related genes can determine the interaction between the tumor and immune system, we will develop a pipeline to process RNAseq data, which will first quantify transcript expression levels and then identify splicing isoforms. From transcript expression levels, we will distinguish the relative expression of immune-related genes, identify which mutated protein coding sequences are transcribed, and deconvolute the immune cell make-up of the tumor microenvironment.

Integrative profiling. Information from other data types can support the DNA and RNAseq pipelines, as well as providing biomarkers. We will build tools to compare predicted neoantigens with TCR profiling data where available to benchmark and potentially improve methods for neoantigen prediction. Further, we will provide methods to compare predicted immune cell composition with that observed by immunohistochemistry staining where available to benchmark deconvolution methods and potentially enrich immune cell quantification data with immune cell spatial localization data.

Pipeline benchmarking and optimization. We plan to optimize methods of the kinds above by benchmarking tools and parameters against each other. In combination, the steps of this workflow become highly complex. We aim to distill this complexity by packaging appropriate tools in proper sequence through an OpenStack-like architecture for standardized, reproducible, and sound analysis.

CIMAC-CIDC Molecular Profiling Working Group (MPWG). We will help establish a MPWG including members across the Network to configure an environment of advanced, statistically rigorous, clinically relevant, readily-useable computational tools and pipelines, tailored to the computational needs of the CIMACs. The group will help prioritize profiling methods required for correlative analyses that are clinically impactful as well as bioinformatically and statistically robust. The group will continually adapt rapidly developing innovations from ENCODE and other consortia to gain maximal insights into tumor-immune interactions from profiling data. As described in **Aim 1a**, the pipelines will be deployed by the Baylor team as templates of stacks of pre-

configured virtual machines that are created on demand using Rackspace's Cloud Orchestration service and will be made accessible for use by the Network members.

C.3.b. Aim 3b. Provide interfaces to integrated local and external tools and resources required for correlative analyses; address the complexity of the tumor microenvironment; embrace a systemic approach by integrating tumor profiling and liquid biopsy.

Correlative analyses. We will further provide tools for correlative analyses of molecular profiles with clinical data. This will allow patient course on immunotherapy to be correlated with variants in immune-related genes, tumor heterogeneity, relative expression of immune-related genes, infiltrating immune cell phenotype and quantity, neoantigen load, recurrent neoantigens, and clinical and demographic information.

Addressing the confounding due to cellular heterogeneity. Immune response is a heterotypic interaction between immune cells and multiple cell types within tumor tissue. Because of cellular heterogeneity, the molecular profile of a tumor is inherently confounded by variation in cell type composition. To address this problem we have developed and recently validated advanced tumor tissue deconvolution method called Epigenomic Deconvolution (EDec)[43]. This and other related methods will deconvolute measurements, eliminate confounding, and provide information about cell type composition and state of constituent cells. We note that our deconvolution method leverages reference epigenomic profiles in the NIH Epigenomics Roadmap and GEO reference sets and utilizes the TCGA profiles, thus providing an example of highly synergistic data integration in action.

Tumor-immune network reconstruction to inform biomarker discovery. The deconvoluted measurements provide an opportunity to perform advanced correlative network reconstruction to uncover tumor-immune interaction networks from tumor profiles in the TCGA collection that may inform clinical biomarker discovery studies. Using our extensive experience in network modeling (described in **Sub-section B**), we will map such interactions from TCGA data and provide them to CIMACs for use in clinical trials.

Growing the pathway knowledge commons for immuno-oncology. Pathway models of tumor-immune interactions can significantly increase the power of correlative analyses by providing context for data interpretation. We therefore plan to capture and curate pathways of relevance for immuno-therapy in WikiPathways, following the practice we established in the Extracellular RNA Communication Consortium[52], thus expanding the "knowledge commons" that will empower clinical immuno-therapy studies.

Correlative analyses to discover clinically useful liquid biopsy markers. Clinical utility of biomarkers that require tumor biopsy is limited by the invasiveness of the procedure. "Liquid biopsy" that detects tumor-related nucleic acids in circulation has emerged as a promising alternative. As leaders of data analysis and coordination for the Extracellular RNA Communication Consortium, we have developed methods, tools, and pipelines for the analysis of circulating RNA in human body fluids and participate in projects to relate circulating exosomal RNA to tumor phenotypes. Building on these preliminary studies we plan to deploy these and other correlative analyses that involve "Liquid Biopsy" for use by CIMACs and thus enable discovery of clinically useful biomarkers.

UI development process. To support correlative analyses, we will develop prioritized applications and interfaces that expose relevant combinations of tools and data using the plugin framework of the Genboree Commons (**Aim 1b**; **Fig.2**). To accelerate development while retaining maximum flexibility, UIs will be developed using the Bootstrap approach that we have previously used for ClinGen and other applications (**Sub-section B**). In collaboration with target users, we will develop use scenarios and specific use cases that will drive the development of interfaces and integration of resources. One general scenario may include identification of response and resistance biomarkers (e.g. high lymphocyte / low neutrophil infiltration predicts better response to immunotherapy) or identification of targets for combination therapy to overcome resistance (potential target: neutrophil-stimulating and CD8+ T-cell inhibiting factor of stromal origin that correlates with low lymphocyte infiltration in TNBC). Such general scenarios will be translated into specific use cases that will help identify tools, UI designs, and resources to be integrated.

C.4. Aim 4. Provide alignment, integration and synergy with external research resources and provide the wider research community with access to Network resources and services.

C.4.a. Aim 4a. Align and integrate Network resources with external databases and computational resources.

Per letter of collaboration from Dr. Grossman, the NCI GDC PI, we will work with the GDC to align and integrate Network data and resources. In the following we define specific integration aspects that will be refined during the course of the project to address any unanticipated incompatibilities. Because of the high level of design concordance between our systems, we do not anticipate any major obstacles.

(Meta)data alignment. The alignment of Network resources will be accomplished by adherence to FAIR principles and OpenAPIs as described in **Aim 1b** and agreements on data elements and ontologies with CIMACs, NCI GDC and other resources as described in **Aim 2a**. We will adopt existing NCI GDC elements and implement them using the GenboreeKB (meta)data modeling framework described in **Aim 2b**. To facilitate data exchange, we will adhere to standard data formats for genomic data as defined by GA4GH and other bodies. Moreover, to align database schemas across the Network, for maximal compatibility, we will propose following the graphical schema defined by NCI GDC that is highly compatible and can be implemented within the document-oriented Linked Data modeling framework implemented in GenboreeKB.

(Meta)data integration. The alignment of (meta)data will form the basis for two forms of (meta)data integration. First, the data and metadata will be made query-able and exchangeable via OpenAPIs that are compliant with LinkedData standards such as JSON-LD. Second, high-volume data will be integrated by either arranging data hosting on the same cloud such as AWS or arrange high-bandwidth connection services between clouds such as RackConnect described in **Aim 1**. Connections with advanced data security levels (described in **Aim 1a**) will be established for highly sensitive data.

Tool alignment and integration. We will implement three types of tool integration. The easiest way to accomplish will be “serverless” integration where compute services will be exchanged via well documented HTTP REST API services. However, this method will not be adequate for high-volume data. For high data-volume pipelines and tools we anticipate applying the virtualization (OVF) and containerization (Docker) methods. The most resource-demanding tools will be exchanged as templates of stacks of pre-configured virtual machines that are created on demand, as described in **Aim 1a**.

C.4.b. Aim 4b. Provide outside investigators and research community with access to CIDC resources and services.

Develop data sharing policies and procedures. Working with the LCC, we will agree upon written policies for accessing Network data and develop procedures for their implementation. The policies will be disseminated to all Network stakeholders for discussion and approval, followed by annual reviews. The network data sharing policies will adhere to guidelines in the NCI Genomic Data Sharing Policies, and leverage extensive experience of the PIs in developing data sharing policies for several consortia, as described in **Sub-section B**.

Develop policies and procedures for access to CIDC services by outside investigators. We anticipate serving other NCI-supported trials as well as trials and investigators that are not supported by the NCI. Working with the LCC, we will develop policies and procedures for such services, including plans for funding such services. We will define a decision process for accepting new service requests, standardize the ways in which the results are served, say via private Genboree-provided study-specific dashboards, and will define closeout procedures. By implementing these processes we will gain valuable experience and “beta-test” operational procedures that will be important for eventual transition to sustainability.

Milestones	Y1	Y2	Y3	Y4	Y5
Aim 1. Establish a standards-based multi-cloud FAIR-compliant data commons for management of clinical profiling data.					
A(i). Implement the Genboree Stack SaaS layer on top of the OpenStack IaaS layer and deploy it using Rackspace services. **Task Description:** Integration, testing, and deployment complete; data transfers between CIDC and CIMAC by the end of Year 1.					
A(ii). Incremental improvements of Genboree Stack & SaaS/IaaS.					
B(i). Eliminate data "silo-ization". **Task Description:** Implement FAIR-compliant standards. CIDC-CIMAC designed interfaces demonstrated to CIMAC end-users with CIMAC data. Security review and systems activation for CIMAC network. CIMACs to submit and access profiling data at the beginning of Year 2.					
C. Deploy Virtual Biorepository.					
Aim 2. Implement data collection standards, tools and processes for key tumor- and immuno-profiling assays.					
A(i). Identify key assay platforms for tumor- and immune-profiling and establish (meta)data collection standards for the platforms. **Task Description:** Standardize CIMAC (meta)data formats used for seamless communication between CIDC-CIMAC and with NCI GDC.					
A(ii). Establish (meta)data collection standards for additional external resources based on availability of resources and use cases.					
B. Deploy tools and implement procedures for (meta)data modeling. **Task Description:** Initial (prioritized) tools are available via the Genboree Workbench for CIMAC use in Y1. Ongoing improvements.					
Aim 3. Provide bioinformatics pipelines and interfaces to integrated tools and resources for correlative analysis.					
A. Deploy bioinformatics pipelines and continually adapt them to extract the increasing diversity of multi-layered molecular information relevant for immuno-oncology. **Task Description:** Consultation with CIMACs will guide selection of pipelines to be implemented or upgraded. This will be ongoing iterative process with specific number of pipelines implemented or upgraded each year.					
B. Provide interfaces to integrated local and external tools and resources required for correlative analyses; address the complexity of the tumor microenvironment; embrace a systemic approach by integrating tumor profiling and liquid biopsy. **Task Description:** Consultation with CIMACs will guide selection of use cases and prioritize implementation. This will be ongoing iterative process with specific number of UIs and tools implemented each year.					
Aim 4. Provide alignment and integration with external research resources and provide access to Network resources & services.					
A. Align and integrate Network resources with NCI GDC and selected external databases and computational resources.					
B. Provide outside investigators and research community with access to CIDC resources and services.					
Aim 5. Provide Network coordination and facilitate interactions.					
A. Provide coordination, administrative and logistic support for the Network.					
B. Facilitate Network interactions with the immune-oncology research community.					
Transition of CIDC into sustainable research resource.					
A. Identify key stakeholders and partners, assess unmet needs, chart the competitive landscape, and develop a business model that incorporates NCI/CTEP guidelines.					
B. Recruit an Advisory Board including thought leaders from industry and academia.					
C. Partner with an organization or form an independent not-for-profit to advance clinical immuno-oncology by providing a commons of immuno-oncology data and computational resources.					

	= Planning		= In Progress		= Completing

Sub-Section D: Administrative Unit

D.1. Aim 5. Provide Network coordination and facilitate Network interactions.

D.1.a. Aim 5a. Provide coordination, administrative and logistic support for the Network.

Establish policies and monitor compliance. Working with the LCC and NCI, CIDC will evaluate relevant NCI/CTEP policies and guidelines for collaborating with industry and those covering data storage privacy and security. It will distribute relevant policies with all key Network stakeholders and provide guidance. Compliance status will be reported to NCI annually and via intermediate reports, as required. We will communicate policies to the CIMACs, and will monitor policy compliance.

Organize and host annual Network-wide grantees meeting, ensure communication within the Network. The CIDC Administrative Unit (Admin Unit) will contact all NIH/NCI, LCC, CIMAC, and Network key personnel prior to the Network kickoff meeting to identify needs and expectations. At the kickoff, we will describe our administrative and logistical plans, collect feedback and finalize the plans. We will give a live demo of the Genboree Commons to ensure immediate access and enable on-line communication within the Network. The Admin Unit will engage with NIH/NCI, LCC, and CIMACs to develop annual Network meeting format and content, and lead development efforts for workshop (2-3 hr) to be held on the evening before the main meeting to provide hands-on training on CIDC resources, including data resources and analytical tools/pipelines and computing services. Feedback will be collected via online surveys. CIDC will also coordinate LCC on-site visits to CIDC/CIMAC centers.

Coordinate Evaluation of the Network by External Scientific Panel. CIDC will report to the ESP all available performance metrics, including: projects completed/in-progress, number and type of samples processed and assays performed, biomarkers studied, turnaround time ("sample in" to "data out"), budgetary issues, and satisfaction of CIMAC-CIDC clients. Similarly, CIDC will solicit LCC/CIMAC feedback on CIDC performance: data management, tools, pipelines, infrastructure resources, and service request management, the results of which will be anonymous and collated by the CIMACs or LCC. CIDC will also work with LCC/ESP to organize teleconference and/or onsite review sessions at the annual Network meetings to assess Network performance.

D.1.b. Aim 5b. Facilitate Network interactions with the broader immune-oncology research, translational, and clinical communities.

Collaboratively develop a Network website and use it to post Network-generated protocols, publications, online materials, use cases, blogs, and data resource links. CIDC will create a Network-website (portal; login required for non-pubic information) for distributing information about Network projects, resources, publications, events, and blogs, in addition to establishing a newsletter and Twitter account to highlight Network accomplishments. Genboree Commons will serve as the online platform for Network information sharing via online topic-specific forums and folders for SOPs, protocols, and reagents.

Monitor and drive Network collaborations between CIMACs and interactions with the broader immuno-oncology community using social media. Network portal usage will be monitored using Google analytics and reported monthly. Metrics will include data uploads/downloads, top 10 portal pages visited, protocols viewed/downloaded, blog views, etc. Twitter engagement will be measured by number of followers and clicks, likes, replies, and retweets. Network activity in Genboree Commons (visits to project pages and topic-specific forums, number of posts) will be used to assess Network activity, helping identify project completion and inform resource allocation decisions.

Co-host workshops and symposia in coordination with scientific societies at major national meetings. To ensure wide impact of the Network, we will work with the LCC to identify opportunities to present CIDC/CIMAC-developed workshops and symposia at national meetings (American Society of Clinical Oncology (ASCO), American Society for Cancer Research (AACR), and similar conferences). Participation at national meetings may begin in the second half of year two, or early in year three, after the CIMACs have had a chance to establish operations and to allow adequate lead time to develop working relationships with national scientific societies.

Facilitate interactions with non-Network entities. CIDC infrastructure and processes (**Aim 4b**) will provide a foundation for contacting non-Network entities to establish collaborations and partnerships that extend beyond a service-provider relationship (see **Aim 4b**). CIDC infrastructure and analysis capabilities will position the Network for strategic partnering (Sub-section A, Transition to a Sustainable Resource). In pursuit of such opportunities, CIDC will work collaboratively with LCC/CIMACs to identify organizations likely to benefit from CIMAC services and will work with CIMACs to create documents (executive summary, position papers, and study dossiers) for use in contacting prospective collaborators. CIDC will ensure that all such collaborations conform to the NCI/CTEP guidelines, and amended as required, in consultation with the LCC and NIH/NCI program staff.

Multiple PD/PI Leadership Plan

This project will involve thee PIs and two institutions that bring highly complementary and synergistic capabilities necessary to accomplish the project. The corresponding applicant institution is Baylor College of Medicine (BCM). BCM performs the largest portion of the project, hosts the core informatics development team, implements the SaaS/IaaS integration with IaaS vendors and is the listed applicant, with a subcontract to Yale University. Drs. Milosavljevic, Sikora and Gerstein are the multi-PIs of this project. Dr. Milosavljevic will serve as the contact PI. The PIs will directly supervise their local administrative, technical, and scientific responsibilities on a day-to-day basis and control their own budgets to minimize disagreements and encourage accountability.

The three PIs and Dr. Roth will form the CIDC Management Committee. The decisions regarding the management of the CIDC will be made by a majority vote of the PIs present during their monthly meetings, usually via teleconference. The Management Committee meetings will require a quorum of two of the four participants and will be chaired by a PI. The meetings will be chaired by Milosavljevic and in his absence Gerstein.

Dr. Roth will serve as the administrative contact between the Management Committee, the Network, and the NCI. Dr. Roth will be preparing the agenda items for the meeting, including review of relevant communications from the Network and the NCI, and overseeing the execution of action items agreed during the meeting. Dr. Roth will overview yearly goals and milestones of CDIC agreed with the NIH and will coordinate yearly review and adjustment of goals and milestones. Dr. Roth will evaluate milestone metrics agreed with the NIH and will communicate them to the NCI.

If Milosavljevic steps down as the contact PI, Gerstein will step in. Any additional succession issues will be resolved by the remaining members of the Management Committee in consultation with the NIH and participating Universities. In case of disputes, every PI may request arbitration by an Arbitration Panel composed of three members, typically two senior executives from participating Universities and the third member being agreed by the two senior executives.

Bibliography

1. *DreamHost.* (URL). **https://www.dreamhost.com**.
2. *The exRNA Atlas.* (URL). **http://exrna-atlas.org**.
3. *exRNA Research Portal.* (URL). **http://www.exrna.org**.
4. *HIPC.* (URL). **https://www.immuneprofiling.org**.
5. *HITRUST CSF.* (URL). **https://hitrustalliance.net/hitrust-csf/**.
6. *IEDB.* (URL). **http://www.iedb.org/**.
7. *ImmGen.* (URL). **https://www.immgen.org/**.
8. *ImmPort.* (URL). **http://www.immport.org/immport-open/public/home/home**.
9. *Internap.* (URL). **http://www.internap.com/**.
10. *ITN.* (URL). **https://www.itntrialshare.org/**.
11. *National Kidney Registry.* (URL). **http://www.kidneyregistry.org**.
12. *Open Science Data Cloud.* (URL). **https://www.opensciencedatacloud.org**.
13. *OpenStack.* (URL). **https://www.openstack.org/),** .
14. *Rackspace.* (URL). **https://www.rackspace.com/**.
15. Altar, C.A., *The Biomarkers Consortium: on the critical path of drug discovery.* Clin Pharmacol Ther, 2008. **83**(2): p. 361-4.
16. Amin, V., et al., *Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs.* Nat Commun, 2015. **6**: p. 6370.
17. Bandrowski, A., et al., *The Ontology for Biomedical Investigations.* PLoS One, 2016. **11**(4): p. e0154556.
18. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.
19. Bhardwaj, N., P.M. Kim, and M.B. Gerstein, *Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators.* Sci Signal, 2010. **3**(146): p. ra79.
20. Cancer Genome Atlas Research, N., *The Molecular Taxonomy of Primary Prostate Cancer.* Cell, 2015. **163**(4): p. 1011-25.
21. Cancer Genome Atlas Research, N., et al., *Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma.* N Engl J Med, 2016. **374**(2): p. 135-45.
22. Cheng, C., R. Min, and M. Gerstein, *TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-7.
23. Cheung, K.H., et al., *Extending gene ontology in the context of extracellular RNA and vesicle communication.* J Biomed Semantics, 2016. **7**: p. 19.
24. Clark, M.B., et al., *The reality of pervasive transcription.* PLoS Biol, 2011. **9**(7): p. e1000625; discussion e1001102.
25. Clarke, D., et al., *Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation.* Structure, 2016. **24**(5): p. 826-37.
26. Coarfa, C., et al., *Analysis of interactions between the epigenome and structural mutability of the genome using Genboree Workbench tools.* BMC Bioinformatics, 2014. **15 Suppl 7**: p. S2.
27. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
28. Diehl, A.D., et al., *The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability.* J Biomed Semantics, 2016. **7**(1): p. 44.
29. Fu, Y., et al., *FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer.* Genome Biol, 2014. **15**(10): p. 480.
30. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
31. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
32. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.
33. Habegger, L., et al., *VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment.* Bioinformatics, 2012. **28**(17): p. 2267-9.
34. Habegger, L., et al., *RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.* Bioinformatics, 2011. **27**(2): p. 281-3.

35. Jonquet, C., N.H. Shah, and M.A. Musen, *The open biomedical annotator.* Summit on Translat Bioinforma, 2009: p. 56-60.
36. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.
37. Kitchen, R., M. Gerstein, and e. al., *The extra-cellular RNA processing toolkit (in preparation).* (URL), 2017. **http://github.gersteinlab.org/exceRpt/**.
38. Lang, J.-P., *Redmine.* (URL), 2017. **http://www.redmine.org**.
39. Li, X., et al., *Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses.* Cell, 2010. **143**(4): p. 639-50.
40. Lu, Z.J., et al., *Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data.* Genome Res, 2011. **21**(2): p. 276-85.
41. Mungall, C.J., et al., *Uberon, an integrative multi-species anatomy ontology.* Genome Biol, 2012. **13**(1): p. R5.
42. Musen, M.A., et al., *The center for expanded data annotation and retrieval.* J Am Med Inform Assoc, 2015. **22**(6): p. 1148-52.
43. Onuchic, V., et al., *Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types.* Cell Rep, 2016. **17**(8): p. 2075-2086.
44. Riehle, K., et al., *The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences.* BMC Bioinformatics, 2012. **13 Suppl 13**: p. S11.
45. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
46. Rozowsky, J.S., et al., *The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci.* Genome Res, 2007. **17**(6): p. 732-45.
47. Sboner, A., et al., *Robust-linear-model normalization to reduce technical variability in functional protein microarrays.* J Proteome Res, 2009. **8**(12): p. 5451-64.
48. Shou, C., et al., *Measuring the evolutionary rewiring of biological networks.* PLoS Comput Biol, 2011. **7**(1): p. e1001050.
49. Smith, A., et al., *Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics.* Bioinformatics, 2007. **23**(22): p. 3073-9.
50. Spackman, K., *SNOMED RT and SNOMEDCT. Promise of an international clinical terminology.* MD Comput, 2000. **17**(6): p. 29.
51. Strom, B.L., et al., *Data sharing, year 1--access to data from industry-sponsored clinical trials.* N Engl J Med, 2014. **371**(22): p. 2052-4.
52. Subramanian, S.L., et al., *Integration of extracellular RNA profiling data using metadata, biomedical ontologies and Linked Data technologies.* J Extracell Vesicles, 2015. **4**: p. 27497.
53. Vidal, M., et al., *The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare.* Clin Proteomics, 2012. **9**(1): p. 6.
54. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
55. Yan, K.K., et al., *OrthoClust: an orthology-based network framework for clustering data across multiple species.* Genome Biol, 2014. **15**(8): p. R100.