Cancer Prevention Clinicogenomic Registry (CPCR) Core: Documentation of the genetic, environmental, and lifestyle, dimensions is crucial to validation and implementation of cancer prevention and treatment strategies. This is often a critical unmet need at academic health centers which limits the ability of cancer investigators to access biospecimens and complete in silico investigations. A discussion with TREC investigators and cancer researchers documented the many challenges they face accessing biospecimens with patient-matched clinical histories from other Texas Centers. Thus, a core facility is needed to provide prospective and retrospective collection and curation of specimens and clinical data from populations at-risk of cancer and those being screened for cancer at Texas A&M clinics. The service was modeled after the Interprofessional Pharmacogenomics Clinic (IPGx) Registry launched to study pharmacogenomic interventions in the Health Hub and Family Medicine clinics of Texas A&M Health.[1, 2]

Background and Rationale: The Framingham studies provided a robust platform from which to build risk models that leverage biomarkers to prevent cardiovascular disease. These seminal studies have led to marked reductions in mortality from cardiovascular disease over the last 50 years.[3] A major challenge to the implementation of chemoprevention strategies is the lack of reliable surrogate biomarkers of disease progression and resolution.[4] Using a definitive cancer diagnosis as a direct clinical endpoint renders most randomized controlled chemoprevention trials impractical due in part to a large study size and time requirements. Further, genetic, social, and environmental factors are difficult to study in the primary and community health settings. In most academic medical center and cancer center settings, where most robust clinicogenomic registries reside, the genetics of premalignant lesions and the natural progression of disease are often not available. In addition, the study of social and environmental factors intertwined with genetics is limited because biospecimens are collected *after* a cancer diagnosis is made and retrospective analysis of health records is limited to data obtained within the oncology ecosystem. As such, typical registries and biobanks do not currently serve the needs of **Cancer Interception and Precision Prevention** programs, thus limiting hypothesis generation. We posit that a more complete understanding of cancer development, and the validation of strategies that prevent cancer in at-risk populations would be aided by use of the Framingham approach that takes advantage of digital tools available today to contextualize social and economic disparities. Trust, patient agency, and data governance are currently major impediments to collection and curation of medical records and patient reported outcomes. We are implementing the Provenance Platform in support of our IPGx Registry to digitize governance and address these impediments to longitudinal registry development.

Impact: CPCR builds upon a robust collaboration between data scientists in the Texas A&M School of Public Health and TREC investigators to create a resource that provides context for implementation of translational strategies in support of TREC projects. Successful cancer prevention and treatment strategies require implementation of systems-based solutions that fully integrate the complexities of cancer.[4] As such, the core will make available medical informatics tools that include quality of life (QofL) assessments for at-risk populations and patient cohorts. We are already using the EQ-5D-5L QoL scale in our IPGx chronic disease assessment clinic and integrated this tool into the patient portal of our electronic medical records. The EQ-5D-5L has been widely evaluated for value-based care in oncology to impute quality adjusted life years (QALY).[5] CPCR will provide a systems-based patient-centric engagement framework to capture and address the cost and value data issues inherent to cancer prevention implementation. Importantly, CPCR will provide a means to interrogate gene-environment interactions at the root of cancer disparities. This approach is complementary to the Single Cell Data Science (SCDS) core which is also designed to examine cancer biology from a systems perspective.

CPCR Services: The core will be established as a Shared Resource Facility for easy access to data and populations serviced by the Texas A&M Family Medicine Center, including three CPRIT-funded cancer screening programs in women's health (**CPRIT PP200070**), colorectal cancer plus HCV (**CPRIT PP220013**), and lung cancer (**CPRIT PP210027**). We will leverage existing knowledge, capabilities, and technologies to service TREC investigators and the Texas A&M cancer research community.

Subject Engagement and Intake: Intakes will be completed in the Family Medicine Center in parallel with other engagement activities as part of the CPRIT Cancer Screening Programs or routine medical care of at-risk patients not participating in cancer screening. Specific criteria for inclusion and consenting will be specified in a new IRB protocol submitted by TREC. Patients will digitally consent to share data from their electronic medical record for inclusion in the registry, with subsequent sharing of information transforming into deidentified case-level data. Dynamic consent will give patients the opportunity to track use of their health data via LifeGraph® technology and modify their consent as appropriate. We are also building engagement programs with the Healthy Texas Medication Assistance Program and will work closely with community advocacy groups through health fairs and other community channels to extend the registry to underserved populations in other regions of the State. Under a base-case protocol for remote registry engagement, two micro blood samples and two buccal swabs will be obtained along with a digitally administered EQ-5D-5L, including a baseline at first sample collection and validated Quality of Life scales for any chronic diseases or risk factors such as COPD or liver disease. When in-clinic sample collection is possible, PAX tube blood samples will be collected and stored at -80^{o}C for up to 96 months.

Virtualized Clinical Research: The frequent use of telemedicine platforms (Mend® and Helios®) by the Texas A&M Family Health Center during the heights of the COVID pandemic enabled engagement of rural and semi-rural populations in clinical care. By extension, this patient and provider experience is being used to explore and develop virtual research capabilities. We are particularly interested in remote sample collection to enable periodic and longitudinal sample biobanking without the need to travel. Our collaborator at the clinical pharmacology laboratory at the University of Colorado has validated a sample procurement process for titration and monitoring of transplant medications and utilizing bloodspot[6] and microsampling[7] technology. We plan to adapt this sample collection protocol. Deploying such a capability will enable prospective recruitment of patients for participation and sample contributions for biobanking, 'omics, and immune cell profiling, if warranted. CPCR will use telemedicine capabilities to enable a virtual registry model whereby patient engagement and outcome assessments can be conducted remotely.

Table 1. *Current and Planned Clinicogenomic Registry Capabilities at Texas A&M University*

Capabilites of the Idealized CPC Clinicogenomic Registry

Operational (at IPGx Registry)	*Currently being implemented*	*Planning*
●Health economic outcomes	●Remote sample collection	●Machine learning capabilites
●AI-based EMR (NLP)	●e-Consent	●Rewards/tokens for data sharing
●Master agreement- HMRI	●Dynamic consent	
●Pilot Idealized Registry Protocol	●Revokable consent	
●Multimodal EQ-5D-5L in EMR and cloud	●Digital data governance	
●Pharmacogenomics and chronic disease management clinic	●Data indexing	
	●Master agreement- MDACC	
	●Case level longitudinal Lifegraphs®	

Biobanking: The capacity to build clinically annotated biorepositories and patient registries within the CPCR will provide incentives (data) for investigators to engage populations in clinical research in the future. A compelling capability of the Provenance Platform is that the BurstIQ technology allows for a LifeGraph® associated with a specimen. A LifeGraph® is a technology enabling connectivity of data that emanates from the patient journey (see **Fig.1** for a simplified schematic) and renders a visualization of that patient journey. Our views of best and ideal practices for biobanking are reviewed in detail elsewhere.[8] As part of this effort we will promote stronger partnerships to support outreach and education for human studies. The proposed registry model will provide non-tumor specimens, dysplastic specimens and tumor specimens working in concert with the appropriate referral teams. Our ability to track patients longitudinally before and after a cancer diagnosis represents a unique opportunity to build a distinctive resource in Texas. The CPCR model will be important for rural and disadvantaged communities in the Texas A&M catchment areas because they generally lack the resources and staff to participate in observational/interventional human studies. Creation of infrastructure that can provide community

oncology providers with access has the potential to make investigational studies available to their patients in collaboration with TREC investigators.

Data Governance: The use of blockchain tools and tokenization (turning a meaningful piece of data into a random set of characters without value if breached) to bring enhanced consent, trust, control, and transparency to the longitudinal linkage of health data and genomic data will inform emergent technologies used for research. This technology is being deployed as part of the IPGx initiative and will be the same approach used for collection of internal CPCR specimens and data. CRCR will also leverage the existing high level of engagement with Texas A&M compliance and privacy administrators to streamline specimen and data acquisition processes with collaborating institutions as well (external specimens and data). Texas A&M University has a master specimen and data transfer agreement with the Houston Methodist Cancer Center and is actively negotiating one with the MD Anderson Cancer

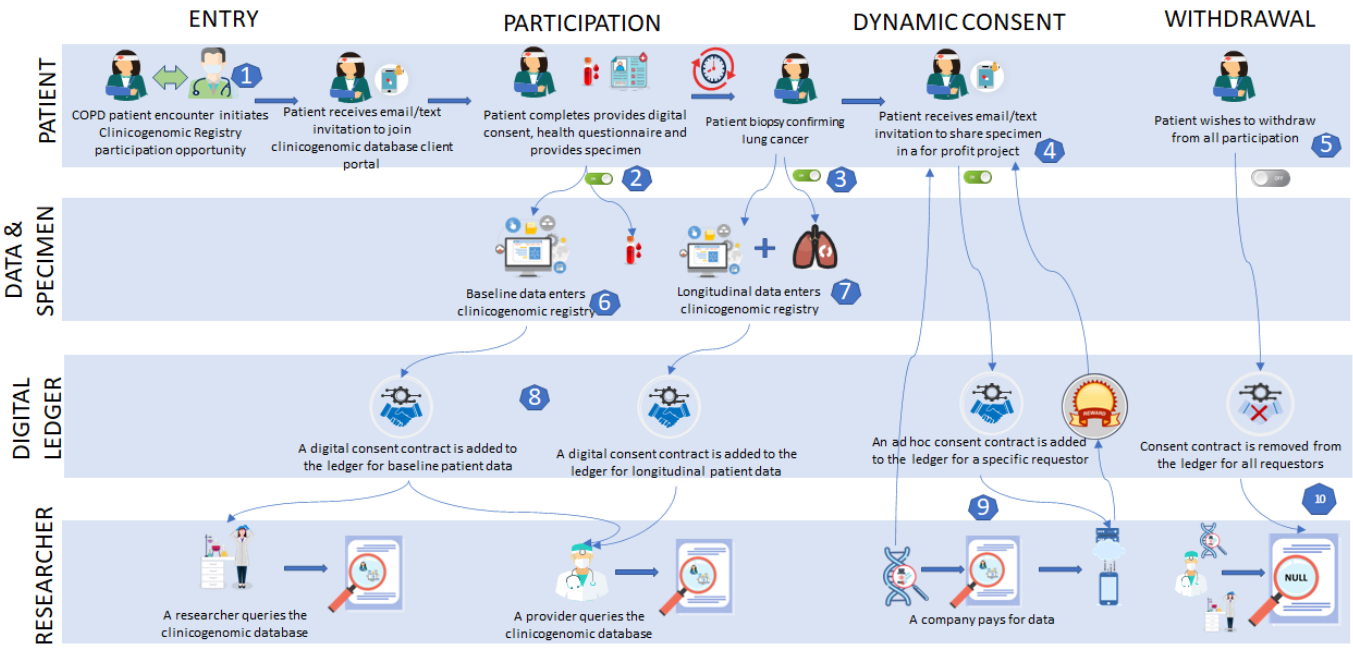

Figure 1. Schematic of the data journey in the Provenance Data Governance Platform.

Center. The legal obligations for data use in those agreements can readily be adopted and operationalized on the Provenance Platform which allows investigators, administrators, and research *subjects* to use digitally enforced smart contracts- in effect, digitizing part of the data governance process (**Fig. 1**). The Provenance Platform enables dynamic, revokable, and e-consent by research subjects perhaps ameliorating a major impediment to registry participation: agency and trust around indeterminant data and specimen use. While getting specimens and clinical annotation thereof from external collaborators is never trivial, the Provenance platform is designed to enhance transparency and provide third parties with additional risk management tools that can overcome the risk/benefit hurdles in specimen and clinical data collaborations. It is expected that greater patient control of data will translate into enhanced participation and engagement of patients in the continuum of care. Finally, these innovative data governance tools provide third parties with high visibility and control of how data is used and is expected to be highly catalytic to collaboration and generalizable to cancer research collaborations in Texas. Core performance will be monitored by the Evaluation Core.

Research Community Service Engagement: A CPCR data governance committee chaired by Dr. Silva and populated by key stakeholders will be established. Sharing multi-omic analysis of specimens from the CPCR biospecimens will be compulsory in alignment with NIH guidance and submitted to relevant repositories accordingly. Within 24 months of providing specimens, multi-omic data shall also be

submitted to the CPCR data warehouse and appended to the clinical cases from which they were derived to become a part of the LifeGraph® for that clinical case. TREC faculty will have priority access to specimens from the CPCR Biobank. CPRIT funded grants will have the next level of priority followed by external collaborators with CPRIT funded projects and other cancer investigators at Texas A&M University. Data acquisition services will be provided to TREC faculty at no cost, and to CPRIT investigators at a discounted hourly rate. We expect the early staging, deep social and lifestyle annotation, and other annotations in the registry to be unique, but also highly complementary to resources like CancerLinQ and other commercial repositories.

Application in TREC projects: CPCR will inform hypothesis generation for TREC investigators and advise on translational strategies at no cost. Several of the TREC projects have a component that mechanistically examines the early stages of carcinogenesis. The CPCR model will capture participant specimens and data that might provide a more complete clinical picture underlying the disease biology that individual research projects are capturing. As with the Single Cell Data Science core, CPCR will enable clinical data to drive hypothesis formation in a top-down fashion, moving from population phenomenology to biomolecular mechanistic studies. The CPCR infrastructure will also be leveraged prospectively to enable a bottom-up approach that can be used to test mechanistic hypotheses in specific populations engaged through the registry infrastructure and research networks. For example, breast cancer screening populations can easily be surveyed for circadian variables in support of Dr. Sato's project such that future immune cell or tumor samples would be accompanied by real-world circadian data. Aligning the woman's health, colorectal and lung cancer screening programs with the CPCR will allow us to target recruitment of well-annotated patients that transition from screening to cancer treatment, enabling select opportunities to study interactions of genetic, social, lifestyle, and economic dimensions of cancer with tumor microenvironment and immunity early in the disease process.

Organization and Core Personnel: CPCR will be an integral resource for the TREC program and the broader Texas A&M cancer research community. Core personnel: 1) Drs. Rick Silva and Gabe Neal will be Co-Directors at 1.8 and 1.2 months of effort, respectively. Their effort will be covered by the TREC program. Other members of the team will include: 2) a staff scientist (6 months effort, at the post-doctoral level), and 3) an experienced clinical research associate at 12 months effort. Drs. Ramos, Neal and Silva are founder members of the IPGx program, and co-investigators on the research protocol, and currently oversee its administration. CPCR will set up a consulting arrangement with other cancer researchers with clear boundaries and a specified budget to procure or provide data to individual labs and their trainees.

Education: The CPC Core will provide infrastructure to collect and curate relevant patient samples and clinical annotation from Texas A&M Health Family Medicine clinics and collaborating oncology clinics. Our clinic has a typical patient flow of 3800 patients per year thus, making initial sample procurement efforts substantial. The program will provide a platform for early career cancer investigators to access a resource that can help augment the clinical and translational impact of their research. CPCR will enable investigators to formulate a strong clinical and translational context for their research activities and empower highly competitive proposals and highly impactful publications arising from basic research.

Anticipated Publications/Grants & Data Management Plan: CPCR will spawn an increasing number of hypotheses as the resource grows in number of cases and length of longitudinal annotation. This will serve a unique niche in the cancer research ecosystem- early in the natural history of disease- to provide context and preliminary data that will enrich future proposals to federal agencies and industry with highly relevant clinical data. The registry will mirror the IPGx registry currently in place and adopt a data governance policy that is consistent with relevant state and federal privacy laws and NIH guidelines, as reviewed elsewhere.[8-10] It should be noted that the digital data governance capabilities of the Provenance Platform provide a facile mechanism for patients, administrators, and third parties to participate in data governance in real time.