**CCREQ ID**

CCREQ-2017-03-00012

**Project Title**

Comprehensive detection of transposable element mobilization in healthy and tumor samples leveraging cloud enviroments

**Submitter**

Mark Gerstein

**Submission Date**

1/5/2017

**Grant ID***Check that the listed Grant ID can be validated in NIH RePORTER. The Grant ID in NIH RePORTER (*[*https://projectreporter.nih.gov/reporter.cfm*](https://projectreporter.nih.gov/reporter.cfm)*) must populate active identified grant.*

1R01HG008126-01A1

**NIH Program Officer Name***(First Middle Last) If there is no program officer for this grant, please put N/A and fill out the Alternative Grant Verification Contact fields (Limit 200 characters)*

PAZIN, MICHAEL J

**NIH Program Officer Email***Provide active NIH-official email account. This is required for validation of your NIH-awarded grant (Limit 200 characters)*

pazinm@mail.nih.gov

**Proposed Research and Impact***Describe and provide rationale for proposed research. Identify research communities impacted and likely to reuse the results of your research. Attach external document (s), if this field space is insufficient. Attach any Data Use Agreements (DUA) here if applicable. (Limit 3000 characters). Attachments should be uploaded as one zipped archive labeled "supporting documents"*

We propose to develop and implement an integrative computational platform to discover the link between genomic retrotranspositions and neoplasm. Retrotransposition is a genetic variation in which transcripts are integrated in the genome, process promoted by retrotransposable elements (i.e., LINEs and ERVs). This is one of the major process remodelling the non-coding region of the genome and is responsible for much non-coding variation. The majority of retrotranspositions comprehend in the mobilization of protein-coding genes (processed pseudogenes), and repetitive elements (Alus, LINEs and HERVs) all of which can have major impacts on genome architecture and significantly contribute to human phenotypic variation and disease development and susceptibility. Despite the importance of retrotransposition for human health, the development of computational methods for their discovery has proven challenging. Most currently available algorithms lack the comprehensiveness and scalability necessary for large scale analysis.
We propose to develop and integrate methods to comprehensive analyze retrotransposition with unprecedented depth and resolution. We have a proven record of collaboration and innovation in the field of retrotransposition discovery and large-scale functional genomics analysis. We will establish a novel platform for discovering and genotyping retrotransposition events from thousands of neoplasm genomes. By applying this platform to data from the PCAWG, we will provide insight into the functional relevance of retrotransposition. We will achieve these goals through the following aims:
- Build an integrative pipeline for large-scale discovery of retrotransposition. We will build an integrated, smart and scalable pipeline of retrotransposition detection algorithms, developed by our group and others, to discover all classes of retrotransposition in a select cohort of samples sequenced as part of the PCAWG consortium. These studies will deliver the largest reference library of retrotranspositions discovered in humans and will allow us to make novel biological inferences in the cancer cohorts.
- Use tools to analyze the functional impact of retrotransposition. We will use a framework similar the one previously developed in our lab to evaluate retrotranspositions that impact protein-coding genes, non-coding RNAs and non-coding regulatory regions. This framework integrates information about evolutionary conservation and existing genomic annotations. In particular, we will up-weight the impact score of retrotranspositions that overlap elements with ubiquitous activity, high network connectivity and strong allelic activity.
This systematic and comprehensive investigation of retrotransposition will yield valuable new resources, including a reference catalogue of somatic and germinative retrotransposition events from thousands of individuals, and a plataform for performing functional SV analysis and association of retrotransposition with cancer.

**Productivity gain through the use of Commons***If Yes: Provide your assessment of how much more research you can accomplish for the same credit award using Commons credits versus other approaches you would normally adopt (Limit 300 characters)*

Yes

Common credits productivity gain are associated with direct access to datasets available in the cloud - allowing to save 20% of processing time associated to downloading, decompressing and cleaning datasets. Also we would be able scale the number of processing nodes without constraints of local HPC.

**Unique Resource Access***If Yes: Explain how the credit obtained will provide computational or data resources not currently available at your institution datacenter or laboratory (Limit 300 characters)*

Yes

Downloading the PCAWG dataset locally would take at least 730Tb, therefore it is not available in our institution. Furthermore, Yale HPC environment is not docker compatible and is under heavy usage. The obtained credits will provide a docker friendly environment and scale up processing units.

**Past experience with Commons conformant objects***If Yes: Describe research projects in which you have already used Commons objects (Limit 500 characters)*

No

Digital Objects generated from biomedical research such as Datasets, Applications/Tools, or Workflows that are pledged to the Commons, should the credit be approved and research completed. Investigators are not expected to share all Data Objects generated by this research, but rather the Data Objects that may be deemed valuable to the Commons community, which is at the discretion of the researcher. For each of the following, please leave blank if not applicable.

**Applications/Tools***List all Application/Tool functionalities and potential value to the Commons. If possible, include information on compatible operating systems, input/output data vocabularies and formats, size footprint, existence of user and installation manuals, scalability, licensing issues, available source code, and whether any of these Apps./Tools is currently indexed (Limit 1500 characters)*

We will port our method to detect mRNAs presence/absence polymorphism (retroCNV) to a cloud compatible solution. The retroCNV pipeline is based in short read WGS and protein coding gene annotation. Source code is available on private Github repository.
Furthermore, the detection of Long Interspaced Repetitive Elements (LINE1) retrotransposition has long been addressed by the scientific literature. These methods use short read WGS datasets to detect germinative or somatic mobilizations. To leverage the qualities of many methods, we will integrate MELT (http://melt.igs.umaryland.edu), Traffic (https://github.com/cancerit/TraFiC) and TranspoSeq (http://archive.broadinstitute.org/cancer/cga/transposeq). All three methods are well established and widely used by large consortium projects.
We will adapt both methods to be docker compatible, allowing them to be portable across different platforms. As a byproduct we will also create docker images for software dependencies that are not available or not maintained in docker repositories such as dockerHub or BioContainers. We will also port their specific output formats to a VCF compatible format integrating information across different methods into a single comprehensive retrotransposition element VCF. Size footprint from both tools should be minimal, however, intermediate files generated during the processing of WGS datasets is significantly large.

**Has access/usage restrictions**

Yes

**Workflows***List all Workflows descriptions and potential value to the Commons. If possible, include information on compatible operating systems, input/output data vocabularies and formats, size footprint, existence of user and installation manuals, scalability, licensing issues, available source code, and whether any of these Workflows has been currently indexed (Limit 1500 characters)*

We will develop workflows for the detection of all classes of retrotranspostion. We intend to use Common Workflow Language (CWL) to integrate intermediate processing steps. When possible, we will use small atomic docker images to partially process input files. We will also use Whole Genome Sequencing reads already aligned to the human reference genome by the PCAWG consortium to avoid aligning raw sequence files. The workflow should be compatible to any platform implementing CWL (AWS, GCP, Azure or OpenStack). Due to the huge amount of data provided by PCAWG, we intend to make these workflow scalable and compatible to large scale processing, therefore, we are aiming to keep the end size footprint as small as possible as well as scalable as possible. We also will make the pipelines publicly available when most of the processing is done. We will to make the source code, docker images and CWL implementation freely available. Software manuals and installation instructions will also be available.

**Has access/usage restrictions**

Yes

Cloud Service Request *Indicate the quantity and price for each service indicated below. You must append vendor calculator snapshots, documentation, or a filled*[*Service Worksheet*](https://www.commons-credit-portal.org/nihcp_commons_credit_request/service-request-worksheet)*to substantiate credit requests. However, you are not limited to select vendor(s) used in making these resource estimates. Attachments for service estimation should be uploaded as one zipped archive labelled "service estimates"*