**Research Development Fund – Spring 2016 Cover Page Template**
**SUBMISSION DEADLINE: February 15, 2016 at 12 noon CDT to rdf@tamu.edu**
(All cover pages will be posted for the campus community to view at http://rdf.tamu.edu/abstracts)

**Application Title:** Petabyte data storage for the life sciences

**Lead contact for RDF Application:**
**Name** David Threadgill
**Department** Veterinary Pathobiology
**Email address** dwt@tamu.edu
**Phone number** 979-436-0850

**Key Participating Units:**
College of Veterinary Medicine & Biomedical Sciences
College of Medicine
School of Public Health
College of Agriculture and Life Sciences
College of Science
College of Liberal Arts
Texas A&M Institute for Genome Sciences and Society

**Anticipated Request Amount ($):** 600,000

**Executive summary of the intended application to utilize Research Development Funds.**

There is a growing need and requirement for mid-to-long term storage of big data in the life sciences that has been caused by the explosion in technologies that generate huge data sets. Genomic technologies in particular are being used in increasingly larger experiments across all life sciences. This has resulted in a flood of data the needs to be preserved. Federal awards to Texas A&M via principle investigators have a stipulation that all data be retained for at least three years after submission and acceptance of all final project closeout documents (OMB Circular A-110). At present, the university has no mechanism to provide investigators access to storage for large genomic and other –omic data despite a requirement by many granting agencies. Although numerous cloud-based options are available that are relatively inexpensive for storage, they are cost prohibitive for most investigators to move data back from storage. Consequently, onsite storage is an attractive option to facilitate investigator research activity, and to be in compliance with granting agency regulations. This application proposes to fund a 5 PB data storage system on campus for investigator use. This system is expandable as needed. The scale needed to support current storage needs is based on a 20X amplification factor of raw data to processed data files that need to be preserved. Additional high bandwidth lines will be required to connect the data storage system with key points across campus. The location of the data storage center will facilitate transfer of data to the two primary compute clusters on campus used for genomic analysis, Ada in the Super Computing Center and the TIGSS cluster.