Justification for requesting 50 TB

(1) Project space needs and file sizes

The space needed will depend on the number our users and the number of projects each users has. As an example, Saranga Wijeratne (System Developer and Bioinformatics Analyst at MCIC) is currently working on a RNASeq project and a Genome Assembly project, and the files counts and sizes for each step for these two pipelines are shown in the table below.

Pipeline steps	Steps and Tools	File Type	Total File Size
RNA-Seq			
RAW Data		fastq.gz	80GB
Short read pre-processing	Fastqc report	.html, .gz	
	Adapter trimming and quality filtering using Trimmomatics	fastq.gz	75GB
	Fastqc on the Trimmomatics output file	.html, .gz	
Transcriptome assembly	Transcriptome assembly using Trinity	.fasta, other	350GB
	Transcriptome assembly using rnaSpades	.fasta, other	200GB
Transcriptome assembly quality assessment	Assess assembly by reads alignment using Bowtie	.bam	~ 100GB
	Counting full length transcripts using NCBI Blast	.txt,.xml	~1GB
	Calculate Transcriptome Contig Nx and ExNy Statisitics	.csv,txt	
	Estimate isoform abundance using RSEM	.tsv,csv,txt	
Downstream Analysis	Coding region identification in Trinity using TransDetector	.txt	
	Differential gene expression using R package DESeq2	.R,.Rdata	
	Transcritome annotations using local Blast	.b2g,txt, nr database	100-200GB
	Gene networks analysis (WGCNA)	txt,.R	
Total			**1006GB**
Genome Assembly			
RAW Data		fastq.gz	20GB
Short read pre-processing	Fastqc report	.html, .gz	
	Adapter trimming and quality filtering using Trimmomatics	fastq.gz	18GB
	Fastqc on the Trimmomatics output file	.html, .gz	
	Removing contamination by aligning reads to genome,db,etc	.fastq, fasta, databases	~400GB
Read assembly	Genome Assembly using Spades	.fasta, other	66GB
	Genome Assembly using IDBA	.fasta, other	80GB
Annotation	Assess assembly by reads alignment using Bowtie	.bam	~ 100GB
	Annotating Assembled Contigs against nr	nr database	100-200GB
	Annotating Assembled Contigs against Silva	silva database	100GB
Gap filling and Completeness	Filling gaps in aseemble genome using RNASeq data	fastq,bam,	200GB
Total			**984GB**

A total of ~1.9 TB of HDD space is required for these two projects. As the MCIC usually has about 20 users in a year, we estimate that we need ~50 TB of space for projects.

(2) Measures we are and will continue to take to optimize the space usage

- ✓ To reduce file size we will keep large files (fastq, etc) in compressed format (.gz)
- ✓ We will keep Alignment files (.sam) in a binary file format (.bam)*
- ✓ Project folders will be moved to Buckeye Box (https://box.osu.edu) for archiving once the project is completed.

(3) Why we need this space

- ✓ We need a working space to do data analysis similar to the ones listed in the table above.
- ✓ We need storage space to store project data (Raw fastq, genomic assembly data, etc) till the project is completed.
- ✓ We need space for hosting custom Blast databases for MCIC users (nr, swissprot, etc).
- ✓ We need space to host the MCIC Galaxy

(4) Length of time needed

- ✓ A minimum of one year with possibility to renew it.