Pipeline Development for DHArMA:

Generate csv containing:

1. dharma\_id
2. data\_type
   1. genomics\_types\_\_\_rnaseq = 1 or 0
   2. genomics\_types\_\_\_dnaseq = 1 or 0
   3. genomics\_types\_\_\_snparray = 1 or 0
   4. genomics\_types\_\_\_exparray = 1 or 0
   5. genomics\_types\_\_\_oth = 1 or 0
3. sample\_name
4. raw\_data\_dir

This csv needs to get passed to an array job script that calls the specific pipeline of interest for the data and hands each line of the csv to individual batch jobs in the array.

Script:

* checks first to make sure that data\_type is correct for the specific pipeline of interest, otherwise throws an error
* checks to make sure that there are no permissions issues for the raw data files, and if so throws an error
* checks to make sure minimal annotations in REDcap are complete otherwise throws an error or asks user to respond if it’s ok.
* Processes the files and saves relevant files and saves them in incremental file directories.
* Generates data provenance files:
  + RNA seq:
    - Pipeline script with parameters and associated database reference files for all steps clearly marked in the beginning of the file, \*\*\*including type of library made and type of sequencing performed if that’s not already specifically called out
    - RNAseQC report
    - Some sort of output file
  + DNA seq:
    - Pipeline script with parameters and associated database reference files for all steps clearly marked in the beginning of the file, including the type of library/capture/sequencing that was done
  + SNP arrays:
    - Explicit documentation of the formatting of the files exported from the array-specific software packages including any parameters chosen, the specific column names exported, so that someone could reproduce it
    - Some sort of version-specific information about the array (probe list, etc)
* Generates final output files:
  + RNA seq:
    - .hts file for gene expression pipeline
  + DNA sequencing:
    - .vcf file
    - Formatted and annotated file
    - (both will be put in Synapse and annotated accordingly with the same data provenance)
  + SNP microarrays:
    - Final\_report for illumina arrays
  + Expression microarrays:
    - ???
* Retrieves minimal and any additional annotations available for the dharma\_id from REDcap API
* Puts files up into Synapse into folders specific to the Data Access Group and type of data, annotates them with the annotations from REDcap, then deposits data provenance as well