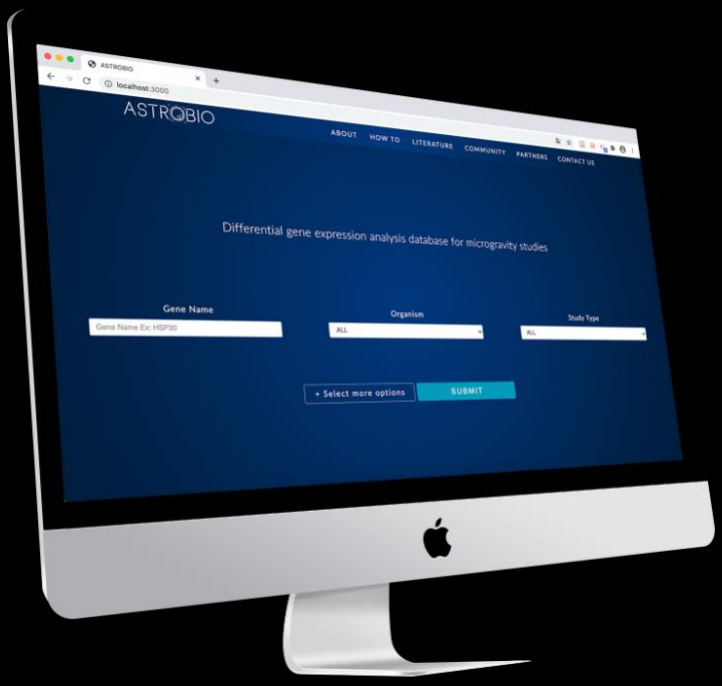
SOFTWARE

NOTEBOOK

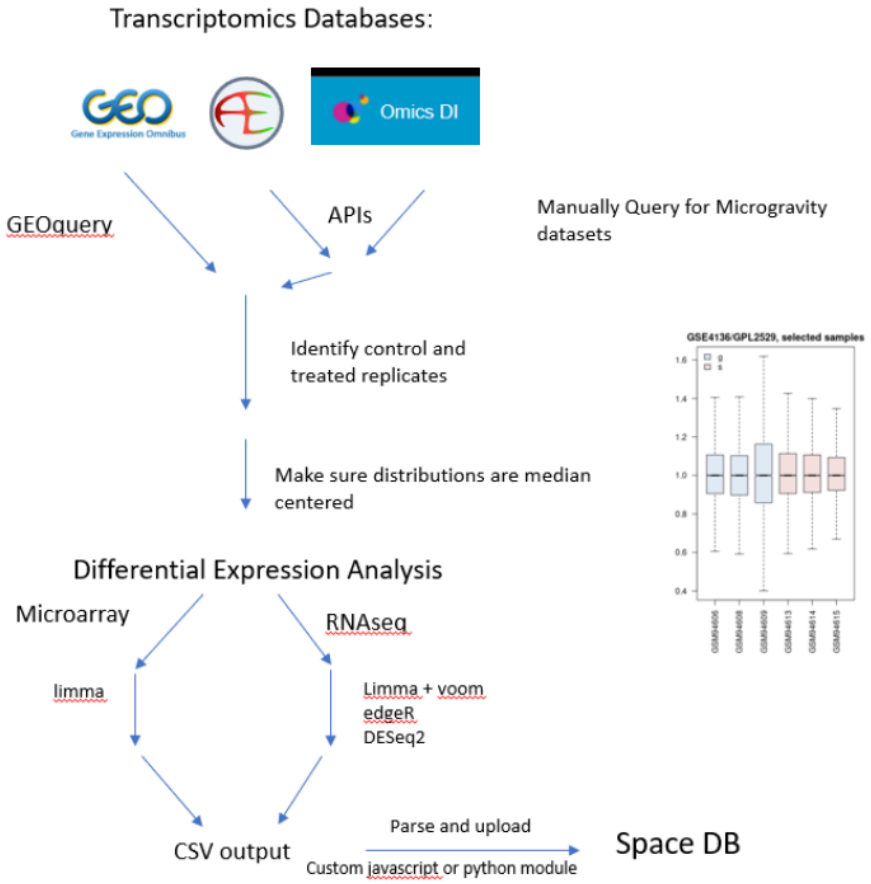


iGEM CONCORDIA

28-05-2020

Attendance: Maher/Ben/Hajar/Grecia

- Data is being collected from the NCBI GEO database, which has a huge collection of bio-information such as samples for micro-gravity related experiments.
- Strategy:
- Start with yeast studies in microgravity, then tackle other model organisms.
- Programming Language: R
- IDE: Rscripts
- Goal: Differential Expression Analysis on sample studies in control conditions vs in microgravity conditions and store in CSV files.



db schema:

```
{
"_id":"",
"gene_id":"",
"gene_name":"",
"description":"",
"method":"",
"regulation":"",
"amount":"",
"position":"",
"sequence":"",
"strain":"",
"link":""
}
```

29-05-2020

- Attendance: Maher - Hajar - Ben - Sam
    - Asif(late)
- Software to predict strain resistance is needed? Need to determine scope of the project
- Agile methodology -> MVP
    - User stories -> functional/non-functional
    - Rank (1-10) 10 is of highest importance
    - Arrange tasks
    - Set milestones
- API access -> GEO/Omics DI/AE
    - Research more into R package (automatic generation of CSV)
    - Omics DI API
    - How to access API? Omics DI
    - Do we allow the same gene, with different methods of study? What r the constraints of the project
- DB is up and running on the server!

Tasks:

Ben, Asif Maher : Setting up DE analysis -> CSV - > DB
Hajar and Maher Sam : User Story
EVERYONE: Ranking of story priority -> MVP

Next Friday: Clone / contribute to Github, discuss how we can move forward with respect to milestones

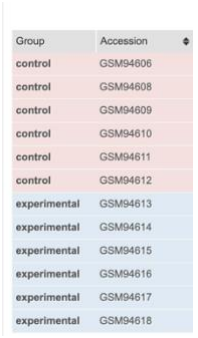
NCBI's data is organized into 4 components: platforms (GPL), samples (GSM), series (GSE), and DataSets (GDS).

GEO2R
Allows users to compare two or more groups of Samples in a GEO Series (GSE) in order to identify genes that are differentially expressed across experimental conditions. GEO2R performs comparisons on original submitter-supplied processed data tables using the GEOquery and limma R packages from the Bioconductor project. This tool provides a simple interface that allows users to perform R statistical analysis without command line expertise. Results are presented as a table of genes ordered by significance.

GSE: Lists of GSM files (samples) that together form a single experiment .

1. Enter NCBI ID (accession number) for the study of interest (accession number: GSE4136) [link]

   Note: each experimental/control sample has its own accession number



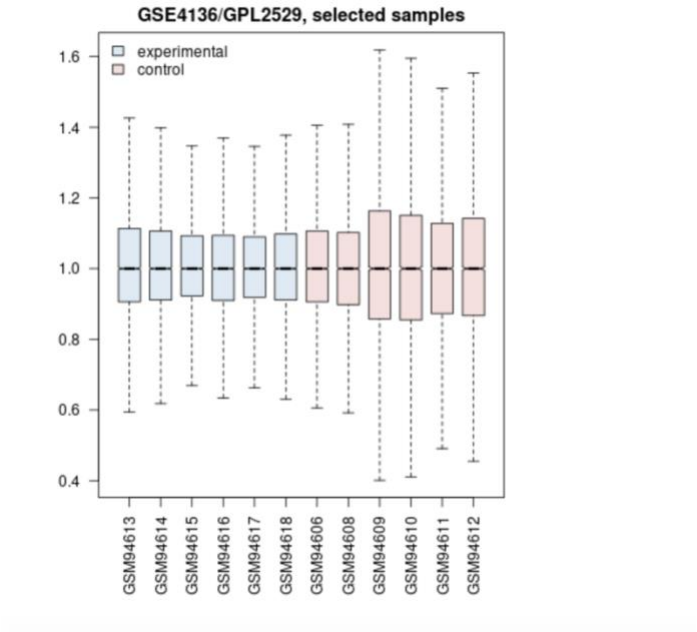
Group	Accession
control	GSM94606
control	GSM94608
control	GSM94609
control	GSM94610
control	GSM94611
control	GSM94612
experimental	GSM94613
experimental	GSM94614
experimental	GSM94615
experimental	GSM94616
experimental	GSM94617
experimental	GSM94618

2. Define groups tab:
   Enter group name "Control", select control samples, click on group name "control"
   Enter group name "Experimental", select low shear modeled microgravity samples, click on group name "experimental"

3. Check distribution of expression values for each sample to make sure they are comparable (good practice):

   Cross comparable and normalized data = median centered boxplots

To check this: under distribution tab, click view. Values normally distributed within a sample = midline for a box is centered within that box. Cross-group comparisons = midline is at about the same level for all boxes/ or at least overlapping between boxes.



Our data looks normalized and cross comparable. (Note: some groups have a wider distribution like GSM94609 but they also have wider error bars).

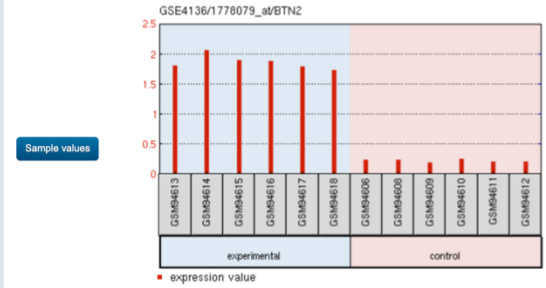
Question: if a sample isn't comparable to the rest, exclusion criteria for that sample? Midline/box overlap, whiskers, outliers? Need standards for submissions to our DB?

4. Perform analysis on GEO2R: Click Top 250 in GEO2R tab to perform DE analysis of top 250 genes [link]

   Results are listed by decreasing significance (first listed gene is the one that changed the most in the experimental groups)

   We will look at the first gene result (upregulated)

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
1778079_at	1.01e-10	9.25e-15	-38.6	23.23	-2.994	BTN2	Btn2p



Gene id: 1778079_at

From graph: this gene is upregulated in all experimental groups (blue) compared to control groups (pink).

t = t-statistic (comparison test between selected groups: experimental vs. control)

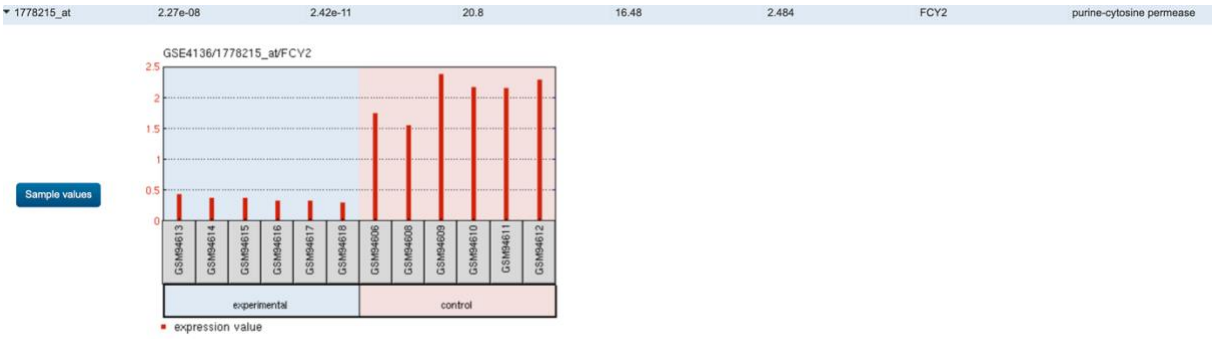
p value = raw p value / significance

adjusted p value = corrects for false positives (multiple comparisons at different significance levels) Takeaway: more stringent than just a p value. Recommended as the primary statistic by which to interpret results. Genes with the smallest P-values will be the most reliable. The default method is the Benjamini & Hochberg false discovery rate: most commonly used for microarray data. Other adjustments methods available under options.

B = B-statistic or log-odds that the gene is differentially expressed (doesn't seem useful! Log odds?)

Log FC = fold change transformed to log2. When raw data is not in logarithmic scale GEO2R transforms it automatically because Limma expects value in log space. It can be disabled in options (disable auto detect in options). Should we display fold change of log fold change? What are researchers interested in looking at?

Example 2 gene id: 1778215_at (downregulated):



R tab (info on the R code that GEO2R is running to provide the results we saw before). We want to extract info (differential expression, metadata (abstract, authors, study type, strain…) from GEO2R and import it to our DB. ? Need to extract this via R (using the code in R tab to create a csv file (table) to give to Maher so that he can upload it to our database). R code already written by Ben. We just need to use it to extract info from different studies. To do so, follow the next steps.

So far, the steps I have shown are not necessary for you to submit your results but it is important to check GEO2R's DE analysis results and know how to navigate it to compare them to the results of your analysis and make sure you did things right because you will be determining groups and some other variables… in R

5. Download R package : [link]
6. Go to [link]
   Copy the R setup code (it will download the R packages you need to run Ben's code).

7. Open R, paste the code there, hit enter to run it. Confirmation message look like this

```
The downloaded source packages are in
    '/private/var/folders/8h/30tbx3wn1kz2nxn6rqr4dkhm0000gn/T/RtmppDR18Y/downloaded_packages'
>
```

8. Open terminal
9. Type this: git clone [link]
   (this will copy all the folders that ben uploaded to our github), hit enter.
   The confirmation message look like this:

```
Last login: Thu Jun 25 21:26:16 on ttys000
ayaelmouddene@MBP-de-aya ~ % git clone https://github.com/Clarkvale/de-expression-igem.git
Cloning into 'de-expression-igem'...
remote: Enumerating objects: 61, done.
remote: Counting objects: 100% (61/61), done.
remote: Compressing objects: 100% (43/43), done.
remote: Total 61 (delta 22), reused 47 (delta 14), pack-reused 0
Unpacking objects: 100% (61/61), done.
```

10. The folder named “de-expression-igem” was downloaded somewhere on your computer. Search for it in finder and open it.
11. Follow the video instructions ben provided here: [link]

05-06-2020

Attendance: Maher, Ben, Asif
Brief summary:
Asif: Looking at HSP for genetics, Learning python and R
Ben: R code executed + generated CSV file + new repo in github + microarray dataset (need help)
Maher: User story + help ben with new repo + github contribution demo

R scripts:
Choose between short vs long term study (genetics team - lancia)
Repository for Rscripts
Yeast as a start - then go other organisms (GEO)
Need to focus on working on extracting data from datasets and generate in CSV files

Went over github for contribution - adding contribution name and commit changes to repo
Add everyone to new repo for R (two repos now, one for server and one for R)

This week:
Asif: Working on R + familiarize with GitHub
Benjamin: continue working on R
Hajar: Arranging contacts with Dr.Butler + user story finalization
Maher: User story finalization + R scripts + DB set up and running on server + Questions for Dr.Butler + writing description of the software work

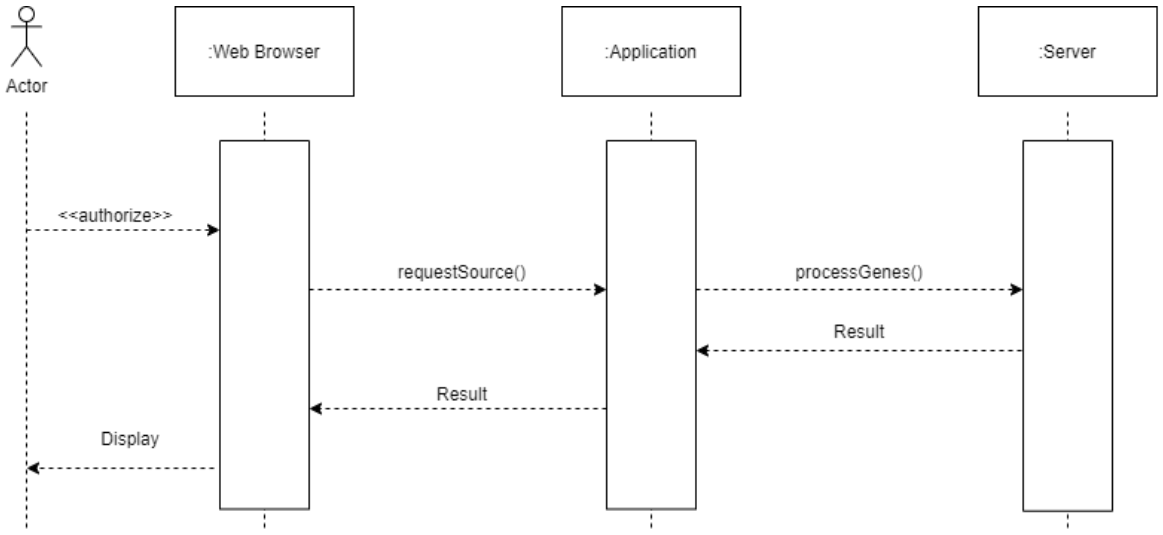
12-06-2020

- Attendance: Maher - Hajar - Ben - Asif - Sam
- Guest joining us:Brain
- Everyone to evaluate the user stories link posted on slack->software channel
- Go over what has been done
    - Maher : Software description, User stories, Software demo (JSON,DB data insertion, frontend EJS template, backend development) and presentation (writing and diagrams)
    - Ben : CSV generation of microarray study GSE4136 (new param, , working on RNAseq data gathering , reading about processing RNAseq data). Need for research into how to get genes accession number for gene

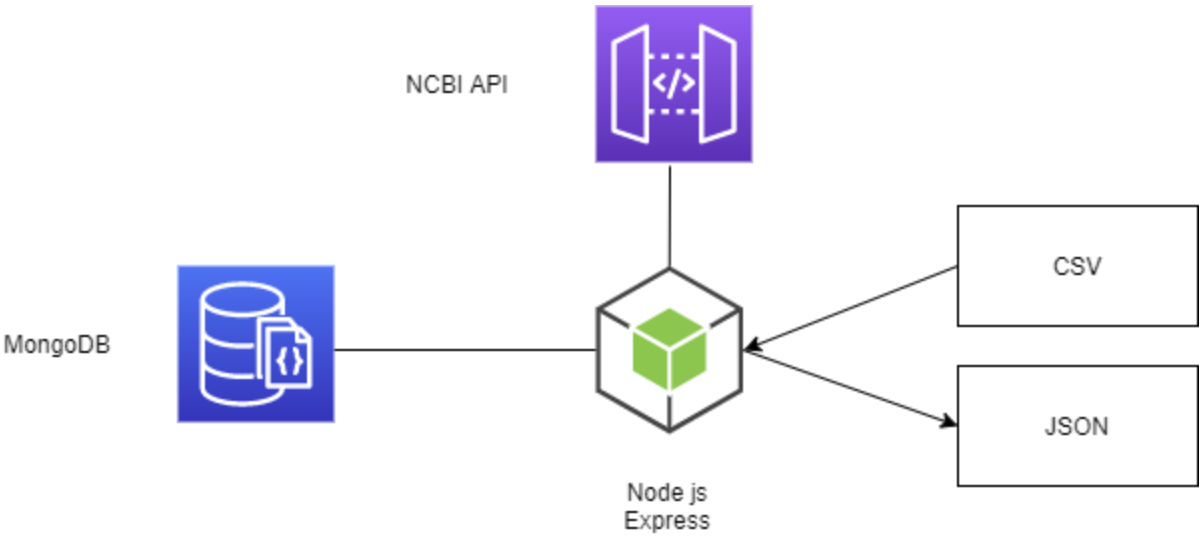
    - Hajar: User story, ranking, add some more stories looking into filters for research
    - Asif: Github contribution learning R
    - Sam: Learning R
- Go over difficulties
    - Maher: Learning R
    - Ben:RNAseq -> learning R more
    - Hajar: HP for software in process (need questions)
    - Asif:Learning R
    - Sam:Learning R
- Go over what's next
    - Maher: DB schema and filters options,
    - Ben:RNAseq
    - Hajar: Working to enhance user story,DB schema and filters options, R
    - Asif:R
    - Sam:R
    - Brian task: R + python

Github repo: [link]
Programming Language: Node.js
DB: NOSQL - Mongodb
Data gathering process: CSV -> JSON -> Mongodb

Sequence Diagram. Users will be able to access the service via a web browser, and make requests to get results from the server.



Architecture Diagram for data fetch / data authentication



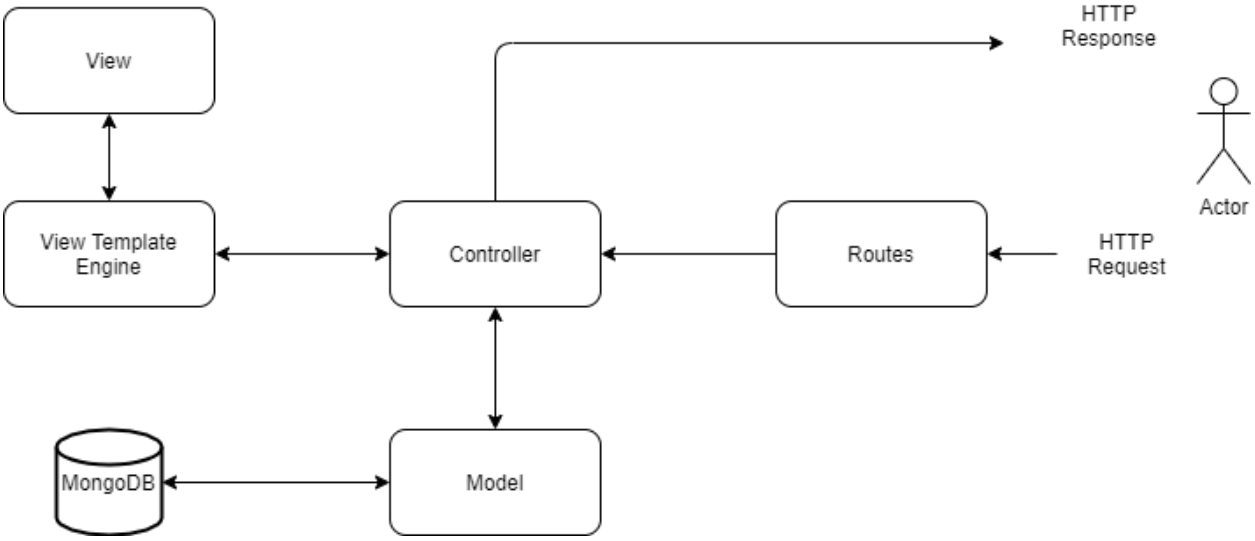
After collecting data using DE Analysis for sample studied in earth conditions vs micro-gravity conditions, data can be processed by converting it to JSON and inserting it into the mongoDB. NCBI API or bionode API will be used to process FASTA files and return list of genes present to compare expression for those genes with available genes in the DB.

18-06-2020

- Attendance: Maher,Sam, Ben
- Concerns and Feedback
    - RNAseq data using R (Ben)
    - DB schema notes (Maher)
    - User Input notes (Maher/Hajar)
    -

Concerns:
DB Schema
User input
RNAseq data using R

Software Design : Model-View-Controller (MVC) architecture that supports parallel and rapid development.



24-06-2020

Attendance: Maher/Ben/Hajar/Asif

Went over Rscripts and plan for yeast
Study papers to fetch data from found here [link]

- Filters recommendations
    - Type of study (parabolic flight vs space flight vs simulation)
    - Method (RNAseq / microarray)
    - Species
    - Strain
    - GEO Accession number
    - Source DB, filter by DB source
    - FASTA file read and interpretation
    - Reliability (count of sources for legitimacy)

- DB parameters addition, recommended to data and/or for filter (Filter requirement)
    - Type of study
    - Method
    - Species
    - Strain
    - GEO accession number (source) (Added)
    - Source of DB that data was extracted from.
    - DNA sequence (FASTA)

- Meeting with Kenza notes:

- Demo
- DB schema evaluation
- User interface notes and filter options

25-06-2020

- Attendance: Maher/Hajar
- Schedule
  - Run server locally
  - Go over frontend directories

Search by	GEO accession number	Genbank ID	Gene symbol	Gene keyword	Gene ontology	Chromosome	Chromosome location	Region
user inputs -->	E-GEOD-4136	850803	HOG1	Kinase	Heat shock protein binding	Y	20000 to 1000000	

Filters	Study type	Assay type	Organism	Species	Strain	Microgravity-induced regulation
user chooses -->	HARV	Microarray	Yeast	S. Cervisiae	BY4743	Up
	RPM	RNAseq	Yeast	Candida albicans		Down
	HYPERBOLIC		Bacterium	Escherichia coli		Unchanged
	SPACEFLOWN		...			

03-07-2020

- Attendance: Maher/Ben/Hajar/Samman
- Ben: second study is vague, check it further
- Stress responses (single vs 2 channel arrays)
  - Single -> 1 chip with probes -> 1 DNA sample
  - 2 channel -> mix with 2 different probes -> run on a single chip -> easier to normalize
- Co-expression libraries (packages) for R
  - Not clear how we can tackle it yet.
  - Needed for the genetics team
  - Last study has two-way channel, looking into how we can obtain the standard form for the DB schema
- Hajar: Issues with package installation in R
  - Ben supporting and helping debug issue
  - Install Mongodb and run locally
- Ben
  - Figure how to normalize the samples together and which ones to choose
  - Plots as pngs
- Maher
  - Continue working on software
  - Search engine completion with sort and filter
  - Run program in cloud
    - Nohup

        - Forever
        - Mongodb
        - Node
        - Git
        - Modules (body-parser, express, assert, bionode, etc)
    - Add existing S. cerevisiae data to software

15-07-2020

- Attendance:Maher/Hajar/Sam/Grecia/Asif
- Topics:
    - DE-Analysis progress
        - Ben: Data parsing - Fetch GO features
        - Hajar: GSE40648
        - Maher: GSM96146
        - Asif: Update, directory issue
    - Software
        - Search Criteria
            - Restricted search vs flex
            - If flex then need to confirm structure
        - Backend (Maher)
            - Added GSE64468 to DB
            - Planning search
        - Frontend (Maher)
            - Structure/Design
        - Discussion for software:
            - Maher: Added sample DB #2
            - Hajar: Colors selection, fonts selection, construct image
            - Grecia: Page design, JS JQUERY scripts

New Database schema for yeastData collection:

```
{
"_id" : "",
"adj" : { "P" : { "Val" : ""} },
"P" : { "Value" : "" } },
```

"t" : "",
"B" : "",
"logFC": "",
"Gene" :
{
"symbol" : "",
"title" : "",
}
"Platform_ORF" : "",
"GO" :
{
"Function" : "",
"Process" : "",
"Component" : ""
}
"Chromosome" : { "annotation" : "" } },
"EGEOD" : "",
"Organism" : "",
"Species" : "",
"Strain" : "",
"StudyType" : "",
"Gen" : "",
"meta_data" : ""
}

Database schema for metaData

{
"_id" : "",
"field" : "",
"accession" : "",
"treatment: "",
"description" : "",
"Link" : "",
"Experimenter" : "",
"Contact" : "",
"Title" : "",
"URL" : "",
"PMIDs" : "",
"Institute" : "",
"Design" : "",
"PlatformID" : "",

"Type" : "",
"Summary" : ""
}

20-07-2020

We are looking to hear from users:

1. Search criteria
    a. What would users need to search for from available data?
    b. What other criteria can be implemented for search?
        i. Adjust logFC?
            1. Not mandatory search criteria?
            2. Change to Up/Down/no diff in regulation/ no specifications?or other?
            3. Not important?
        ii. What other filters to be added? Platform_ORF? Or search based on Assay Type? Or experiment type? (HARV/Space-flows, etc), organism? species ?
        iii. Some existing criteria need to be removed?

2. Results table
    a. What should be shown as results in the table?
        i. Column should view which elements?
        ii. Sorting criteria for results should remain for logFC?
        iii. Modal vs pagination for more information
    b. What other information that is not relevant and can be removed?
3. What other information has to be gathered for search or results purposes?

12-08-2020

- Attendance: Maher/Hajar/Benja/Grecia
- Working on:
    - Hajar: Created pages for frontend. Curated some journals, and worked on finding a colleague later. Bring genetics onboard to help with content in AstroBio.
    - Grecia: Content for wiki, software layout from software processes. Read about igem criteria for documentations structure. Agile and NOT Waterfall. Picked what was relevant and asked maher to check for process workflow and information to display for process.
    - Ben: Working on yeast metaAnalysis for datasets so far. Making heatmap, working on PCA (Making R markdown file to present to Barker)

    - Maher: AstroBio fronend, gene page, added information. I am currently pagination.
- Issues:
    - Hajar: Issue with routing.
    - Grecia: Waterfall vs Agile. Most use waterfall, but using agile. DB modelling, what should be included. Requirements: Different ways to organize it, used functional and non-functional. Testing simple, no need for JEST.
    - Ben: RNAseq analysis issue, need server space.
    - Maher: Ajax issues.
- Next:
    - Hajar: Push changes, I can do the routing. Make GIF and screen shots. Content from genetics team.
    - Grecia work on software process
        - Separate data from application
        - Work with maher for user story
        - Work on logical workflow
        - Weekly sprints
        - Design thinking process.
        - Wiki Layout for consistency
        - Wire frames for wiki
    - Ben: Continue heatmaps and meta analysis, then shiny web app.
    - Maher: I will pagination, 255 db limit and skip - Get it done in a week.

20-08-2020

Maher/Hajar/Ben/Grecia

- Finalize db schema
- Fill user stories and update milestones
- Pagination implemented
- Create content for literature and practices
- Add GSE4136 analytical data to DB
- Add Experiment information as well as others such as organism, species, strain, study type, assay type and list of samples used for the obtained results.
- Dynamic display of results based on search query
- Show crucial information
    - Clarify what is more important to display -> HP
- Positive and negative expression values are good, check for unaffected values for logFC?
    - +/- LogFC values
    - Sort based on LogFC
- Search engine for Gene and Adjusted.P.Value can be done independently from LogFC
- Adjusted P Value not a crucial search field

28-08-2020

AstroBio Database Process Review
Grecia/Maher/Hajar/Ben

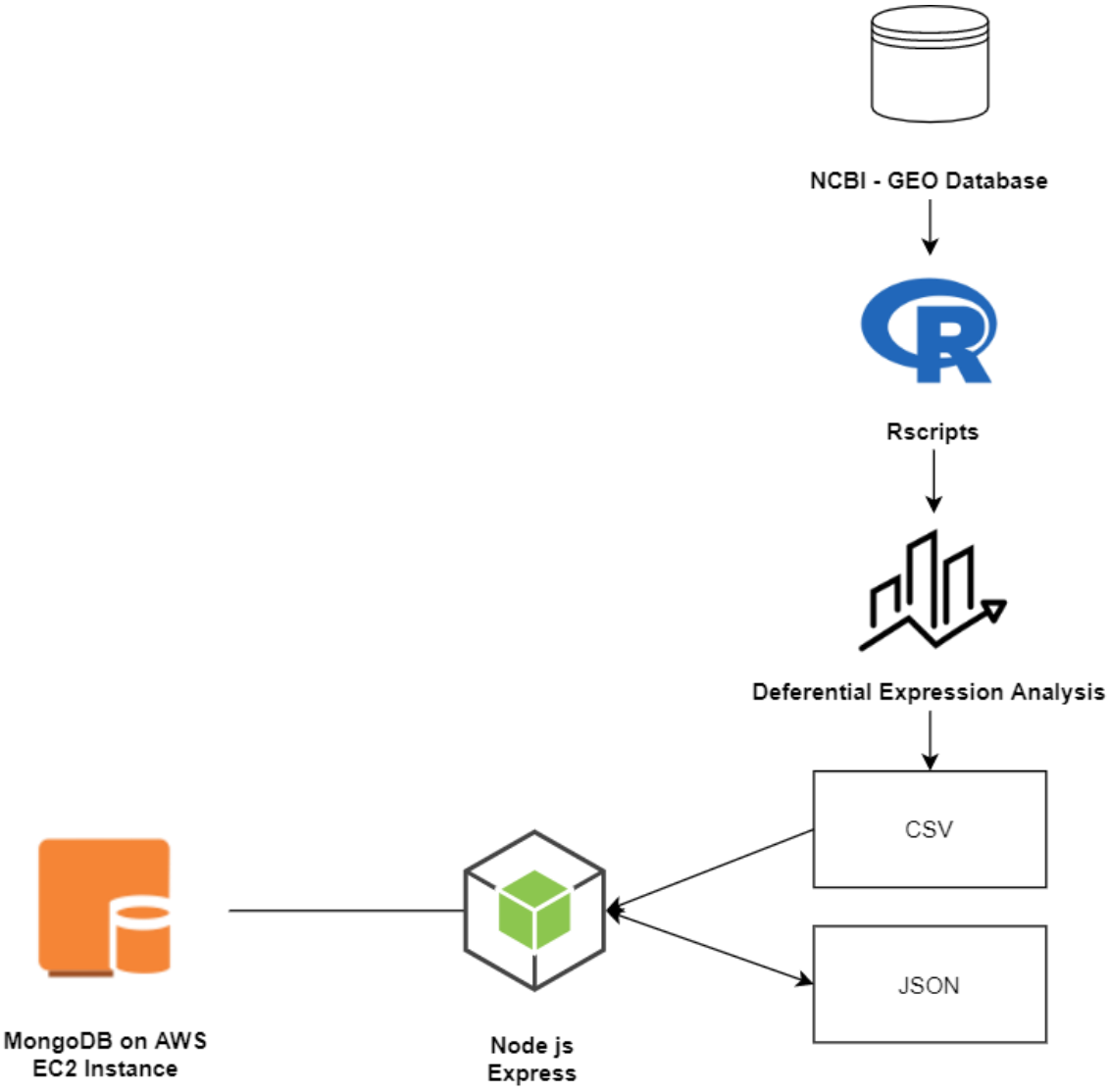
2 to 3 pages max
Written with figures and infographics as needed

1. Introduction
2. Project Description:
    a. Name: AstroBio
    b. Description:
3. Problem being addressed & proposed initial solution/ our first design
4. Goals and Constraints
    a. Functional requirements
    b. Domain model
    c. Constraints and Qualities
5. Resource Evaluation (we can add to this as we do HP)
    a. Human
    b. Technical
6. Architecture
    a. Insert diagram
7. Technologies in use
    a.
8. Risks
    a. During architecture and design phase
    b. During public use
    c. Data handling
9. Final choice and futures

11-09-2020

- Maher/Hajar/Grecia
- Need to create dedicated page for genes
- Prepare for mentors meeting and presentation

Updated software architecture



25-09-2020

Home page

- Gene symbol → change to 'Gene Name' with hover over "Gene symbol"
- Microgravity Type → change to 'Microgravity Method'
- Remove 'double text' ie in the image below, we don't need to write *Platform ORF Ex: YLR406C,* instead just have *Ex: YLR406C*

- Make the boxes smaller, less scrolling space

In the Search Results

Change 'Log2 Fold Change' to **Expression Change** (Log2 FC) or hover over for units

- Make Export & Print one button (same result)
- Eliminate dead scroll space at top, bring Search results higher
- When a search is submitted, have it so that it say Search Results in the same screen (to guide the user to the results and so they know it was successful)

Spelling
Saccharomyces cerevisiae (capitalize first letter)
Arabidopsis thaliana
Escherichia coli
Astro**B**io

Other
Add Concordia Student Union Logo (Google Concordia Student Union Logo)

AstroBio content: https://docs.google.com/document/d/1wy2X4-Hdk_ZlyPSM2hXwRbFKYC9tQBljkuSoRy5LFcs/edit?usp=sharing

02-10-2020

- Try to get as much feedback as possible
- Updated DB schemas:

Collections:

- geneResults
  - Contains basic gene information

        - Name
        - Title
        - Platform open reading frame
        - Strain
    - Gene ontology
        - Function
        - Process
        - Component
    - Gene location / annotation
    - Differential Expression Analysis results, such as
        - P Value
        - Adjusted P value
        - LogFC
        - B
        - t
        - F
        - AveExpr
- metaData
    - Experiment information
        - Accessions
        - Treatment
        - Description
        - Experimenter
        - PMID
        - Institution
        - Assay Type
        - Design
        - Summary
        - URL
- reducedGenes
    - Map-reduce operation to split and map genes with multiple names that point to the same statistical analysis results

In more details:

geneResults: different available schemas

1- GSE4136 schema

```
{
"_id" : "",
"ID":"",
"adj" : { "P" : { "Val" : ""} },
```

"P" : { "Value" : "" } },
"t" : "",
"B" : "",
"logFC": "",
"Gene" :
{
"symbol" : "",
"title" : "",
}
"Platform_ORF" : "",
"GO" :
{
"Function" : "",
"Process" : "",
"Component" : ""
}
"Chromosome" : { "annotation" : "" } },
"EGEOD" : "",
"Organism" : "",
"Species" : "",
"Strain" : "",
"StudyType" : "",
"AssayType":"",
"Gen" : "",
"EssentialforFlight":"",
"meta_data" : ""
}

2- GSE40648 and GSE95388 schemas
{
"_id" : "",
"adj" : { "P" : { "Val" : ""} },
"P" : { "Value" : "" } },
"t" : "",
"B" : "",
"logFC": "",
"Gene" :
{
"symbol" : "",
"title" : "",
}
"Platform_ORF" : "",

"GO" :
{
"Function" : "",
"Process" : "",
"Component" : ""
}
"Chromosome" : { "annotation" : "" } },
"EGEOD" : "",
"Organism" : "",
"Species" : "",
"Strain" : "",
"StudyType" : "",
"AssayType":"",
"Gen" : "",
"meta_data" : ""
}

3- GSE50881 schema
{
"_id" : "",
"adj" : { "P" : { "Val" : ""} },
"P" : { "Value" : "" } },
"AveExpr" : "",
"F" : "",
"logFC": "",
"Gene" :
{
"symbol" : "",
"title" : "",
}
"Platform_ORF" : "",
"GO" :
{
"Function" : "",
"Process" : "",
"Component" : ""
}
"Chromosome" : { "annotation" : "" } },
"EGEOD" : "",
"Organism" : "",
"Species" : "",
"Strain" : "",

"StudyType" : "",
"AssayType":"",
"Gen" : "",
"meta_data" : ""
}

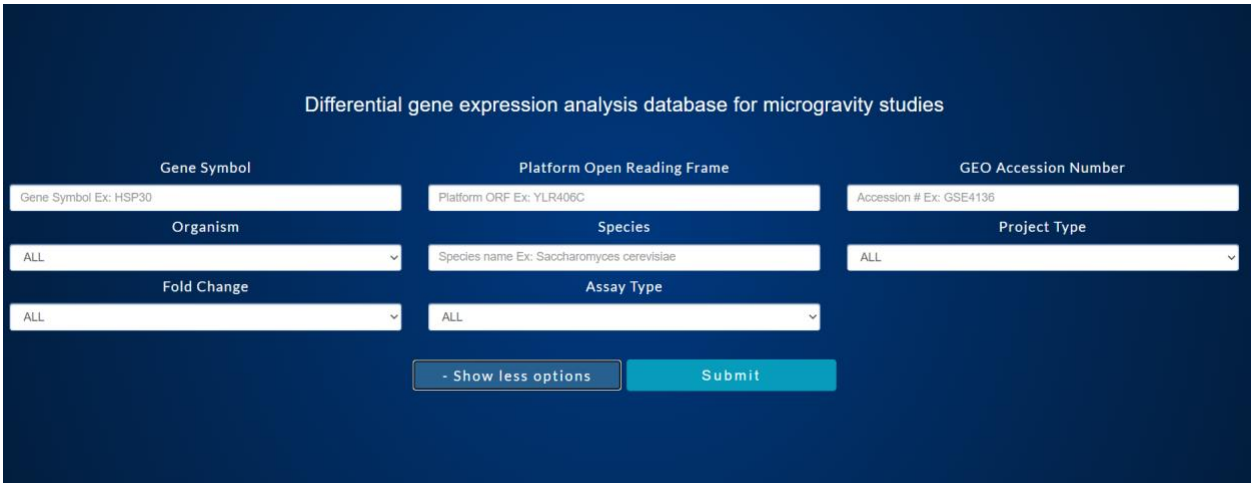
meta_data schema

{
"_id" : "",
"field" : "",
"accession" : "",
"treatment: "",
"description" : "",
"Link" : "",
"Experimenter" : "",
"Contact" : "",
"Title" : "",
"URL" : "",
"PMIDs" : "",
"Institute" : "",
"Design" : "",
"PlatformID" : "",
"Type" : "",
"Summary" : ""
}

Map reduce schema

{
"_id":"",
"value":""
}

- Map reduce is needed to map different gene names for bacteria
- Design for gene page and list is completed



SEARCH RESULTS

Email Export Print

3 studies found.

Note: results are sorted in ascending order by Adjusted P Value.

Gene Name	Author(s)	Study Type	Log_2FC	Organism	Species	Generation	Adjusted P Value	
PXR1	Kate,,McInnerney	HARV	-2.474477519	Yeast	S. Cervisiae	5	0.000432757	More Info

Gene Name	Author(s)	Study Type	Log_2FC	Organism	Species	Generation	Adjusted P Value	
PXR1	Kate,,McInnerney	HARV	-1.876765997	Yeast	S. Cervisiae	25	0.00045407	More Info

Gene Name	Author(s)	Study Type	Log_2FC	Organism	Species	Generation	Adjusted P Value	
PXR1	Katty,,Goossens	SPACEFLOWN	0.149998957	Yeast	S. Cervisiae	Not Specified	0.964812104	More Info

- Overview
- Gene Ontology
- Gene Location
- Study Overview
- Study Information

OVERVIEW

Gene Name	PXR1
Name Description:	telomerase inhibitor
ORF:	YGR280C
Log_2FC:	-2.474477519
t value:	-17.94689643
B value:	6.695724946
Organism:	Yeast
Species:	S. Cervisiae
Strain:	BY4743
Generation:	5
P Value:	8.38e-7
Adjusted P Value:	0.000432757
GEO Accession Number:	GSE4136