Integrating genomics data with cancer registry records
Midterm Report

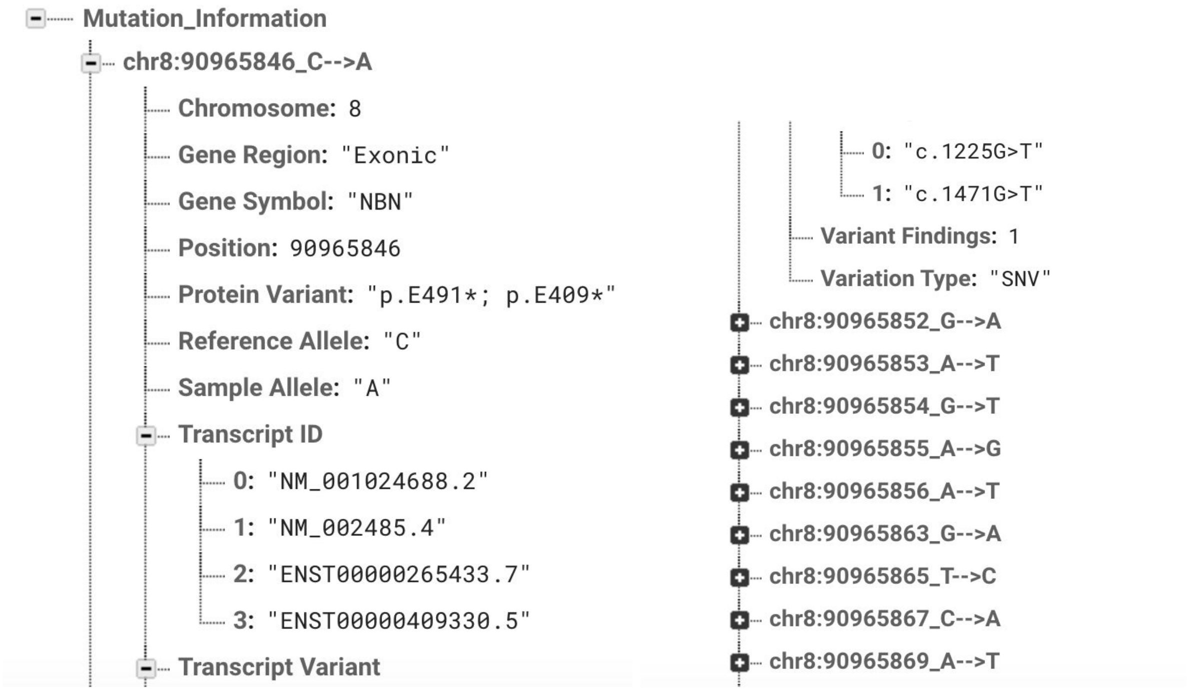
Jae Young Kim, Kevin Tsang

Modification to Original Plan

After working with the data, we have decided to split the two data sets (genomics data and cancer registry data) into 1 table and 2 json files: mutation information json, patient mutation table and patient registry json. Below are the columns of each table:

I. Mutation information data (json file):
This json file shows the biological information of each mutation.
Columns: ‘Mutation_ID’*, ‘Chromosome’, ‘Position’, 'Reference Allele', 'Sample Allele', 'Variation Type', 'Gene Region', ‘Gene Symbol’ (Gene name), ‘Transcript ID’ (transcript IDs from various biological databases, cannot use as key because it has multiple IDs for each mutation), ‘Transcript Variant’ (mutation notation in DNA format), ‘Protein Variant’ (mutation notation in Protein format)

* We have created Mutation_ID from Chromosome, Position, Reference Allele, and Sample Allele; this will truly create a unique ID for each mutation as we learned that more than one mutation can happen in a chromosome position. We will use Mutation_ID to connect with the Patient Mutation table.



```
Mutation_Information
  chr8:90965846_C-->A
    Chromosome: 8
    Gene Region: "Exonic"
    Gene Symbol: "NBN"
    Position: 90965846
    Protein Variant: "p.E491*; p.E409*"
    Reference Allele: "C"
    Sample Allele: "A"
    Transcript ID
      0: "NM_001024688.2"
      1: "NM_002485.4"
      2: "ENST00000265433.7"
      3: "ENST00000409330.5"
    Transcript Variant
      0: "c.1225G>T"
      1: "c.1471G>T"
    Variant Findings: 1
    Variation Type: "SNV"
  chr8:90965852_G-->A
  chr8:90965853_A-->T
  chr8:90965854_G-->T
  chr8:90965855_A-->G
  chr8:90965856_A-->T
  chr8:90965863_G-->A
  chr8:90965865_T-->C
  chr8:90965867_C-->A
  chr8:90965869_A-->T
```

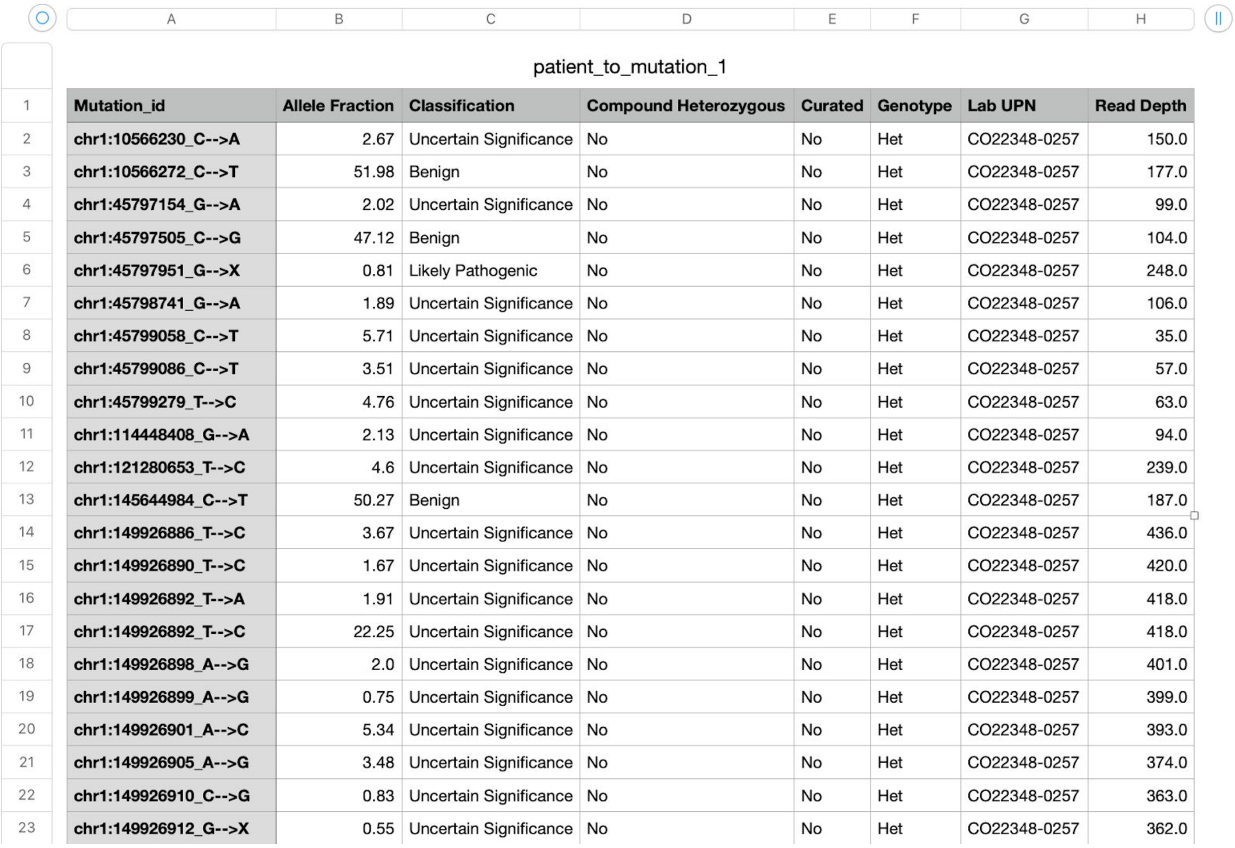
II. Patient - Mutation table (csv / pandas dataframe):
This table shows which mutation the patient has. Note that Genotype, Compound Heterozygous, Read Depth, Allele Fraction depends on both Lab UPN (patient id) and Mutation id. Therefore, there is no redundancy for the mutation information described in the

data. For each spreadsheet, there are “final call” and “all out” sheets; primarily, we extracted the majority of the data from the “all out” sheet, which comprises all variants found in a run. The “final call” includes final results manually curated by genetics experts. We have replaced the respective rows in the “all out” sheet with the “final call” data because they are more relevant to clinical management and patient care.

Columns : LabUPN (Sample unique ID), ‘Mutation_ID*’, ‘Genotype’, ‘Compound Heterozygous’, ‘Read Depth’, ‘Allele Fraction’, ‘Classification’, ‘Curated’ (Marked yes if it was manually curated by a genetics expert, marked no if the Classification was created by software algorithm.)

*Mutational_ID is used to connect with the Mutation information json and LabUPN connects the patient registry data. We have uploaded the data into Firebase.

Data example and screenshot:



patient_to_mutation_1

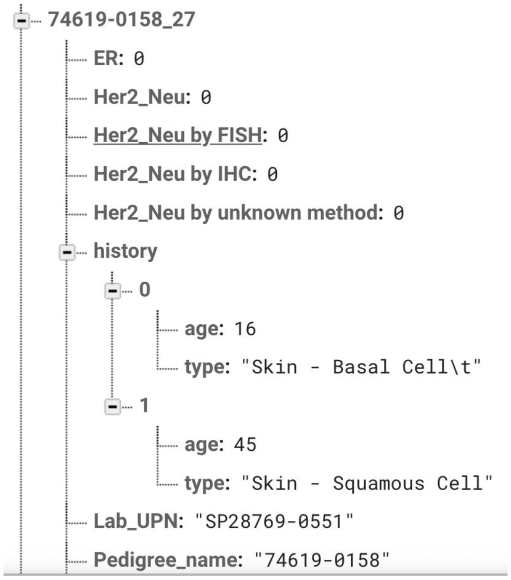
Mutation_id	Allele Fraction	Classification	Compound Heterozygous	Curated	Genotype	Lab UPN	Read Depth
chr1:10566230_C-->A	2.67	Uncertain Significance	No	No	Het	CO22348-0257	150.0
chr1:10566272_C-->T	51.98	Benign	No	No	Het	CO22348-0257	177.0
chr1:45797154_G-->A	2.02	Uncertain Significance	No	No	Het	CO22348-0257	99.0
chr1:45797505_C-->G	47.12	Benign	No	No	Het	CO22348-0257	104.0
chr1:45797951_G-->X	0.81	Likely Pathogenic	No	No	Het	CO22348-0257	248.0
chr1:45798741_G-->A	1.89	Uncertain Significance	No	No	Het	CO22348-0257	106.0
chr1:45799058_C-->T	5.71	Uncertain Significance	No	No	Het	CO22348-0257	35.0
chr1:45799086_C-->T	3.51	Uncertain Significance	No	No	Het	CO22348-0257	57.0
chr1:45799279_T-->C	4.76	Uncertain Significance	No	No	Het	CO22348-0257	63.0
chr1:114448408_G-->A	2.13	Uncertain Significance	No	No	Het	CO22348-0257	94.0
chr1:121280653_T-->C	4.6	Uncertain Significance	No	No	Het	CO22348-0257	239.0
chr1:145644984_C-->T	50.27	Benign	No	No	Het	CO22348-0257	187.0
chr1:149926886_T-->C	3.67	Uncertain Significance	No	No	Het	CO22348-0257	436.0
chr1:149926890_T-->C	1.67	Uncertain Significance	No	No	Het	CO22348-0257	420.0
chr1:149926892_T-->A	1.91	Uncertain Significance	No	No	Het	CO22348-0257	418.0
chr1:149926892_T-->C	22.25	Uncertain Significance	No	No	Het	CO22348-0257	418.0
chr1:149926898_A-->G	2.0	Uncertain Significance	No	No	Het	CO22348-0257	401.0
chr1:149926899_A-->G	0.75	Uncertain Significance	No	No	Het	CO22348-0257	399.0
chr1:149926901_A-->C	5.34	Uncertain Significance	No	No	Het	CO22348-0257	393.0
chr1:149926905_A-->G	3.48	Uncertain Significance	No	No	Het	CO22348-0257	374.0
chr1:149926910_C-->G	0.83	Uncertain Significance	No	No	Het	CO22348-0257	363.0
chr1:149926912_G-->X	0.55	Uncertain Significance	No	No	Het	CO22348-0257	362.0

III. Patient Registry data (json file)
This data shows the cancer history and various cancer biomarker statuses of patients in the City of Hope Registry.

‘Patient_ID’*, ‘Pedigree name’, ‘UPN’, ‘Lab UPN’, ‘ER’, ‘PR’, ‘Her2/Neu', 'Her2/Neu by FISH', 'Her2/Neu by IHC', “history” (cancer history: cancer type and age of diagnosis)

*Patient_ID is created from Pedigree name and UPN (Unique Patient Number); A pedigree is a family history diagram and since we have enrolled family members into the registry, we use UPN to identify each individual. Lab UPN is used to connect with Patient Mutation data.

json object example and firebase screenshot:



```
74619-0158_27
    ER: 0
    Her2_Neu: 0
    Her2_Neu by FISH: 0
    Her2_Neu by IHC: 0
    Her2_Neu by unknown method: 0
    history
        0
            age: 16
            type: "Skin - Basal Cell\t"
        1
            age: 45
            type: "Skin - Squamous Cell"
    Lab_UPN: "SP28769-0551"
    Pedigree_name: "74619-0158"
```

Motivation behind the modifications

Though the original plan of uploading Cancer Registry data and Genomics mutation data to two different databases is viable, we learned that we need to avoid redundancy in storing data in the ER and Relational lecture. Therefore, we separated the mutation information (which is universal to the specific mutation) from the genomics mutation data into the mutation information table. Since many patients can share the same mutations, we avoided many data redundancy by just having a record of each mutation once. Patient mutation table then indicates how many mutations the patient has and its corresponding relevant statistics and information.

Milestones and timelines (checklist)

Clean Genomics Table and upload to Firebase - Completed Cleansing, not uploaded yet
Clean Cancer Registry data and upload to database - Complete
*Clean Mutation information Date and upload to database - Completed
Integration of the two data - Complete (We have thought of the linkage between json objects and table)
--------------Midterm -----------------
Build database and upload data - not started
Build webapp to show query (requires time to learn) - not started

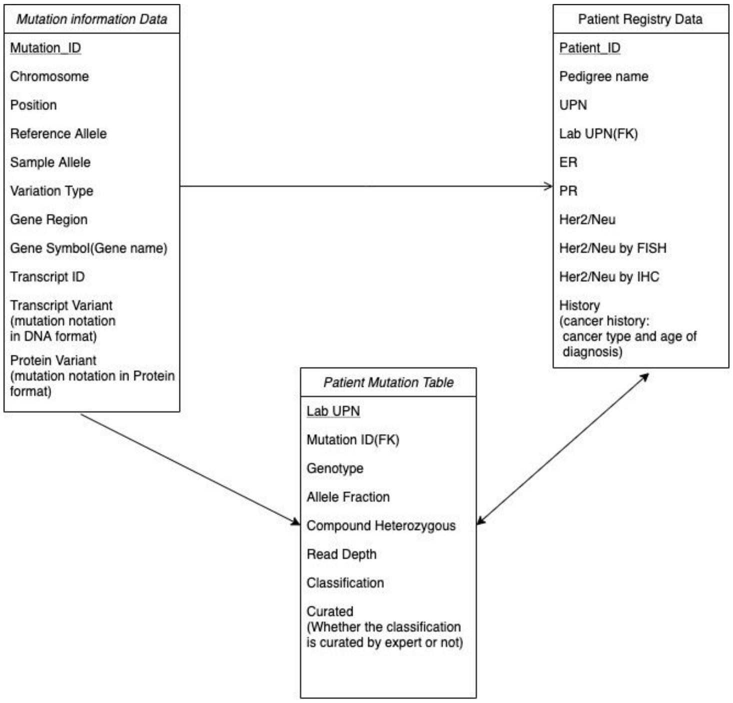
Progress Evaluation

We have spent significant time perfecting the Python code that helps us perform data cleaning tasks; therefore, all of the data will be uploaded to the databases in bulk after the midterm. In this report, we have demonstrated the proof of concept and made sure it works. We are on track for the milestones as we believe the hardest part is cleaning and structure the data in a manner that is efficient and understandable. Our goal is to create a dashboard for search and query so the remaining tasks will not be very intense. We will need to learn how to build the web application which will be the most challenging part for the rest. We also need to learn how to incorporate Spark into the project as well.

Challenges

For building out the web application for the final dashboard, since we have little to none experience in that, we would appreciate some help locating tutorials to get us up to speed so we can successfully build the web application.

Data Architecture Diagram



Data Link :
https://drive.google.com/drive/folders/1e3KhRyGq5Ae5KPtccpnVDraB1rB1ooZo?usp=sharing