[[1]](#footnote-1)

Precision Medicine: Integrated and Relational Database Design Concept

Lei Jiang and Liang Huang

*Abstract*— In this project, we propose to build a proof-of-concept Integrated Relational Database for Precision Medicine. We integrate data from a variety of sources, such as patient phenotyping data, biopsy data, patient genome sequencing data or genotyping data and major genomic database data such as UCSC and RefSeq, for this database. The first step is designing ER diagram and use normalization techniques to improve the efficiency for storing and querying the data. Then we populate patient genotyping data by simulated data and patient biopsy data from online data source. We also demonstrate the database function by writing SQL queries to extract or subset for deeper analysis. It is also feasible to connect existing genomic databases by MySQL connections to Ensembl and UCSC. In addition, we propose to deploy machine learning model for prediction by Rest API, as well as GUI connect to database for patients to retrieve data on website.

*Index Terms*—Precision Medicine, Relational Database, SQL, Genotyping, Phenotyping, Genomics

# INTRODUCTION

T

HIS project is designing and populating a proof-of- concept integrated relational database for the application of precision medicine.

## Purpose of Precision Medicine

Precision medicine means selecting the most effective cancer treatments based on the presence of specific biomarkers in a patient’s tumor. Genomic testing is used to identify patient’s gene expression profiles to determine the corresponding sensitive targeted therapies. Precision medicine delivers individually tailored therapy based on the patient’s disease subtype [2]. This approach allows the patients to benefit from the treatments most, avoid unnecessary treatments, reduces toxicity, and significantly improves outcome.

Multiple studies have demonstrated the benefit of precision medicine [9, 13, 14]. Take some of the clinical use of Genomic tests on Biomarkers for example. Some of the standard genomic tests in clinical use including but not limited to Oncotype DX for Breast cancer, Colon cancer and Prostate cancer. Information from these genomic tests can help the patients and their doctors make decision on treatment method such as Chemotheragpy, Radiation, aggressive treatment, or surgery. There are over 1 million patients tested in more than 90 countries with Oncotype DX. EGFR mutations, ALK rearrangements, and ROS1 fusions, and PD-L1 expression testing are recommended for advanced non–small cell lung cancer. These tests help with decisions on Immunotherapies. In addition, a research just published this month suggested the potential genomic test for Pancreatic Cancer by identifying Biomarkers in Pancreatic Cancer patients [11].

Precision medicine is not only limited to cancer diagnosis and treatment. It can be applied to a variety of diseases and prevention of diseases from actionable insights gained from data analysis. Therefore, precision medicine depends largely on analysis of datasets from multiple sources such as clinical, genomic, transcriptomics or even environmental factor such as diet style and exercises habit, etc. Precision medicine also allow us the potential to find cure for currently lethal diseases.

## Integrated Relational Database Approach

The amount of genomics data increases rapidly due to the advancement of Next Generating Sequencing (NGS) technology making sequencing of human genomes cost effective. The cost falls from 10 million dollars a decade ago to 1 thousand dollars today. However, the rapid evolution of genomic testing platforms and emergence of NGS technologies make interpreting molecular testing reports more challenging [14]. There is enormous amount of data generated from genomic tests. For example, the size of existing genomics database Sequence Read Archive (SRA) grows 10,000 times in the last decade. We need to find solutions for efficiently storage and extract the large amount of data from clinical, genomics and transcriptomics, etc. The solution we proposed here is a relational database that integrated information from multiple sources.

In addition, deep learning algorithms emerge in the past few years and just started to spread into biology research to gain new understanding of the complex biological systems.

From previous researches, deep learning usually requires extremely large training sample size. Integration of biological data from various sources for analysis is becoming increasingly important due to the need for datasets for deep learning. Databases are the foundation of Artificial Intelligence (AI) in healthcare. In a few years, precision medicine may be widely applied in our daily life. The database we build can provide critical information to subtype cancers or predict drug responses.

The integration step is also critical in realization of precision medicine or personalized medicine. This idea is backed by multiple research papers as attached in the references. For example, a recent published research integrated the clinical, genomic, and transcriptomic data and performed integrative clustering to classify triple-negative breast cancers into more subtypes and suggested precision treatment strategies according [12].

Our goal for this project is to integrate multiple databases across omics, such as genomics, proteomics, and phenomics into one relational database. And then we show what novel information this newly built relational database can provide us by generating SQL queries to retrieve information. By doing so, we create great value for these existing data.

# Previous Work on Topic

## Existing Bioinformatics Databases

There are enormous amounts of genomic data available in

public genomic databases such as NCBI, ENCODE, UCSC

Genome Browser, TCGA, and Ensembl. The Sequence Read

Archive (SRA) database provides short reads of DNA

sequencing data generated by high throughput next-generation

sequencing (NGS) technology. It had around 2 Terabases of

data in 2009, and in 2019 it already contains 10,000 Terabases

of data. In addition to public databases, some companies own

genetic big data. The size of genome sequencing data

23andMe has was less than 100,000 in 2010, but in 2016 the

size increased to more than 1 million.

As we went through these databases, we found most of

them are search based and not connected to other level omics

data, even for the secondary and predictive databases. UCSC

Genome Browser combines information of UCSC genes,

RefSeq genes on chromosomes, Human mRNAs and

Transcription Factor CHIP-sq from ENCODE. Ensembl is

another genome browser for vertebrate genomes, which

pattern match DNA to protein. Online Mendelian Inheritance

in Man (OMIM) mainly consists of descriptive entries.

In the effort to combine information from multiple

bioinformatics databases, Roche Cancer Genome Database

(RCGDB) was created to integrate the information in multiple

databases such as Cancer Genome Atlas project (TCGA), the

IARC TP53 database, OMIM, KinMutBase and the L1CAM

mutation database. However, it is still not integrating

genomics data to proteomics and phenomics. In addition, this

integrated database is also search based. Montague et al. realized this problem and stated in a paper

in 2014 "Currently, data are scattered across single omics

repositories, stored in varying raw and processed formats, and

are often accompanied by limited or no metadata". The Multi-

Omics Profiling Expression Database (MOPED) was created.

It includes transcriptomics and proteomics information from

publicly available studies on model organisms and humans.

In conclusion, there are numerous omics databases nowadays, but most of them are in flat files structure and search-based. However, these omics information are much more useful if they are integrated. Relational databases allow integration of diverse types of information and efficient management of these large datasets. The relational database we proposed to build integrate multiple data sources and allow us to efficiently manipulate large datasets.

## Current Lack of Integrated Data Source

The lack of a complete relational database consisting of

“omics” and phenotyping databases creates bottle-neck in

the crafting of patient-specific medicines.

The design and creation of a Multi-omic database will

advance precision medicine research. Precision medicine requires predictive modeling based on a patient’s genomic data, protein expression data and phenotyping data. Kim et al. stated in their 2016 paper: “A significant obstacle in

training predictive cell models is the lack of integrated data

sources.” A relational biological database that integrates Multi-omics data is needed to advance precision medicine prediction research.

# Obstacles

There are number of obstacles for building this Integrated and Relational Biological Database. First of all, it is hard to access patient genotyping and phenotyping data due to privacy concern. But we overcome this by simulated data. Due to the nature of various non-trivial datatypes such as VCF and BAM, there would be a lot of data munging to get data from various Datatypes, File formats. Secondly, biological and clinical data size are large-scale, computing power was not enough. We may only create a smaller size database to prove the concept. In addition, biological data has extremely flexible schema. Normalization for biological data is impractical, but we will try our best.

# Research Methodology

We do not have enough resources to build out the entire

database as a class project, so we will instead focus on

building a simplified and demonstratable data model of the

conceptual relational database. Our project goal is building a Queryable Database so that we can easily integrate all the related data in one query and extract or subset the data for deeper analysis. Also, the efficiency for storing and querying the data can be improved.

First, we decide what entities to be included in our database.

We will determine attributes needed for each entity by biology domain knowledge. The tables could be expanded as

our knowledge grows and more scientific discoveries are

made. Data of interest are patient phenotyping data, patient genotyping data, and patient biopsy data. It is also feasible to connect existing genomic databases by MySQL connections to Ensembl and UCSC. So that we can extract data from current existing biological databases such as DNA-to-protein data from Ensembl. Afterwards, we assign Primary keys to each gene and this can be used as foreign key in the genotype or protein table. In a similar manner, we connect other entities in our database as time and resources permit. In summary, proposed entities for a simplified data model are GenoType data, Biopsy data, PhenoType and Gene\_Info.

After decided on the entities, we design the ER diagram for our database. Tables are connected with Primary and Foreign Keys and apply the database normalization techniques to reduce data redundancy. Then we populate patient genotyping data by simulated data and patient biopsy data from online data source. The R code and detailed method for data simulation is attached in Appendix. We use Data Definition Language (DDL) to build the database in mySQL. We also demonstrate use of Relational Databases such as xtract data using Data Manipulation Language (DML) such as combing with JOIN, filtering with WHERE. It also serves as the evaluation of our approach for storing and querying the data efficiently.

We would ideally add simulated phenotyping data for

systematic analysis of critical health factors. Phenotyping data

consists of physical traits, such as heart-rate, blood pressure,

BMI, body fat, muscle content, water content, etc.

Phenotyping also represents features on the molecular level,

such as blood glucose levels in diabetes patients or her2 gene

expression levels in breast cancer patients. This type of data

could be collected from hospitals (desensitized data) or on

wearable devices and smartphone apps. There are currently

several apps available for recoding phenotypes such as

MyHeart Counts, Hello Heart, and Yumai smart scale. We

have the technology to track many traits dynamically. Patient

genome sequencing data was collected by genealogy

companies (23andme and guardant health). Access to this data

is unlikely and will only be considered on a conceptual basis.

After the database was built, we could retrieve necessary

info to run deep learning algorithms on them and get insights.

# Results

A close up of text on a white background

Description automatically generated

Fig. 1. ER diagram for integrated relational biomedical database.

# Conclusions

We successfully designed the ER diagram and use normalization techniques to improve the efficiency for storing and querying the data. Then we populated the database with patient genotyping data by simulation and patient biopsy data from online data source. Demonstrate the database function by writing SQL queries to extract or subset for deeper analysis.

Significance of our database is gaining personalized biological information to diagnose and optimize treatment decisions for many types of cancers. In addition, we find trend from data for actionable health advices. For example, from predictive modeling we can provide information as for a given BloodPressure, what is the optimal BodyFat to have least risk for breast cancer. Our database can also potentially aid in biomarker or drug discovery. And it is easy to retrieve useful information for further such as deep learning algorithms.

# Next Step

More data from different sources can be integrated into this database. For example, create a table for Protein with expression levels, histone modification data, need to identify more detailed attributes. This entity can be used for clustering based on protein express level/histone modification. It is also feasible to populate genomics data by MySQL connections to Ensembl and UCSC.

What we also propose to do with our database is deploying machine learning model for prediction by Rest API, as well as build GUI connect to database for patients to retrieve data on website. For real world large datasets, we need to utilize cloud computing such as AWS or Google Cloud Platform.

Appendix

R code for getting the online breast cancer biopsy data is attached here in Appendix I.

R code for simulate patient genotyping data is attached here in appendix II.

## Appendix I

bc<-read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data",header=F,sep=",")

names(bc)<- c('id\_number', 'diagnosis', 'radius\_mean',

'texture\_mean', 'perimeter\_mean', 'area\_mean',

'smoothness\_mean', 'compactness\_mean',

'concavity\_mean','concave\_points\_mean',

'symmetry\_mean', 'fractal\_dimension\_mean',

'radius\_se', 'texture\_se', 'perimeter\_se',

'area\_se', 'smoothness\_se', 'compactness\_se',

'concavity\_se', 'concave\_points\_se',

'symmetry\_se', 'fractal\_dimension\_se',

'radius\_worst', 'texture\_worst',

'perimeter\_worst', 'area\_worst',

'smoothness\_worst', 'compactness\_worst',

'concavity\_worst', 'concave\_points\_worst',

'symmetry\_worst', 'fractal\_dimension\_worst')

write.csv(bc,file="/bcdataset.csv")

## Appendix II

#subset

bc2<-bc[,c(1,2)]

bc2

#simulation data 569 samples, 21 genes genomic test data. generate a 569x21 matrix for patients genomic test dataset with gene expression levels range from 1-3. (These are all cancer patients data, so we do not have 0 value since it means no cancer. we do not know the actual underlie distribution though. so just used a normal distribution)

# ref: Carlsson et al. paper in 2004. HER2 expression in breast cancer primary tumours and corresponding metastases. Original data and literature review

#HER2-scores: The HER2 expression was scored using the HercepTest criteria. The HER2-score was based on a 0 to 3+ scale. 0 corresponded to tumour cells that were completely negative, 1+ corresponded to faint perceptible staining of the tumour cell membranes, 2+ corresponded to moderate staining of the entire tumour cell membranes and 3+ indicated strong circumferential staining of the entire tumour cell membranes creating a fishnet pattern.

genomat=matrix(sample.int(3, 569\*21, TRUE),569,21)

genomat2=data.frame(genomat)

head(genomat2)

names(genomat2)<- c('Grb7', 'HER2', 'ER','PR', 'BCL2', 'SCUBE2', 'Ki67', 'STK15', 'Survivin', 'CyclinBI', 'MYBL2', 'MMPII', 'CTSL2', 'CD68', 'GSTMI', 'BAG1', 'Bactin', 'GAPDH', 'RPLPO', 'GUS', 'TFRC')

head(genomat2)

#bind patients ID and diagnosis data with genotyping data to form a complete table

GenoType=cbind(bc2, genomat2)

head(GenoType)

#write to csv

write.csv(GenoType,file="/GenoType.csv")

Acknowledgment

We would like to thank James Clay for the preliminary work. And we are very honored to get initial guidance of this project from Dr. George Bell. Last but not least, we would like to thank Dr. Engels and all the students in Database class for the fruitful semester. We help each other out and learn from each other a lot.

References

[1] Montague E, Stanberry L, Higdon R, et al. MOPED 2.5--an integrated

multi-omics resource: multi-omics profiling expression database

now includes transcriptomics data. OMICS. 2014;18(6):335-43.

[2] Servant N, Roméjon J, Gestraud P, et al. Bioinformatics for precision

medicine in oncology: principles and application to the SHIVA

clinical trial. Front Genet. 2014;5:152. Published 2014 May 30.

doi:10.3389/fgene.2014.00152

[3] Zhang, W., Li, F., & Nie, L. (2010). Integrating multiple 'omics'

analysis for microbial biology: application and methodologies.

Microbiology, 156 Pt 2, 287-301

[4] Kim, Minseung & Rai, Navneet & Zorraquino, Violeta & Tagkopoulos,

Ilias. (2016). Multi-omics integration accurately predicts cellular

state in unexplored conditions for Escherichia coli. Nature

Communications. 7. 13090. 10.1038/ncomms13090.

[5] Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects

somatic mutations to signalling in breast cancer. Nature.

2016;534(7605):55-62.

[6] Zhang H, Liu T, Zhang Z, et al. Integrated Proteogenomic

Characterization of Human High-Grade Serous Ovarian Cancer.

Cell. 2016;166(3):755-765.

[7] Mackey et al. Relational Databases for Biologists

[8] Janetzki et al. Genome Data Management using RDBMSs. Technical Report. 2015.

[9] McVeigh et al. Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. Breast Cancer - Targets and Therapy. 2017;9:393-400

[10] Carlsson et al. HER2 expression in breast cancer primary tumours and corresponding metastases. Original data and literature review. Br J Cancer. 2004; 90(12): 2344–2348.

[11] Dimitrakopoulos et al. Identification and Validation of a Biomarker Signature in Patients With Resectable Pancreatic Cancer via Genome-Wide Screening for Functional Genetic Variants. JAMA Surg. 2019;3: e190484. doi: 10.1001/jamasurg.2019.0484.

[12] Jiang et al. Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. Cancer Cell. 2019; 35: 428-440

[13]J Mascaux et al. Genomic Testing in Lung Cancer: Past, Present, and Future. Natl Compr Canc Netw. 2018 Mar;16(3):323-334. doi: 10.6004/jnccn.2017.7019.

[14] Katherine et al. Precision Oncology Decision Support: Current Approaches and Strategies for the Future. Clinical Cancer Research. 2018. DOI: 10.1158/1078-0432.CCR-17-2494

1. [↑](#footnote-ref-1)