**Key steps Documentation**

# Genome Structure Database (GSDB)

The first comprehensive database for 3D Genome Structure from Hi-C data.

GSDB was created to meet this need, to create a common Hi-C data structure repository using different prediction algorithms.

# Server Locations

The GSDB website is located on ***sysbio.rnet.missouri.edu***

path: /var/www/html/3dgenome/GSDB

The GSDB 3D structures and normalized data is located on ***calla.rnet.missouri.edu***

Find 3D structures here: /exports/store1/3dgenome/GSDB/Database

Find normalized Hi-C Data here: /exports/store1/3dgenome/GSDB/Datasets

# Database Update

Find the sql database in the website directory on sysbio named gsdb.sql

Check “**database\_readme.txt**” file for password

# Hi-C Datasets

Hi-C datasets were downloaded from majorly ENCODE database. The GEO database source of each of the data are included in the database website.

Data was grouped into 3 batches. The batches are described in the “***Database Data Info.xlsx***” file.

Batch 1 datasets are all from ENCODE and GEO datasets

Batch 2 datasets are 40kb mouse ES cell, mouse cortex, human ES cell (H1), and IMR90 fibroblasts) datasets from here:

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80. <http://chromosome.sdsc.edu/mouse/hi-c/download.html>

Batch 3 dataset is GM12878 data from Rao. et al. The GEO database link is included as well.

# Database Directory Structure

The database directory is organized as follow:

Top Directory -> Bottom:

GSDB ID Name -> File Name -> Normalization Technique Abbreviation (VC, KR) ->Algorithm Name.

# Structure Generation

Most of the structure generation was done of Lewis cluster. A sbatch script to generate each chromosome structure separately was created.

Please check the “GSDB\_Scripts” directory for more information about the structure generation scripts.

# GSDB Scripts

The “GSDB\_Scripts” directory can be considered as the powerhouse for the entire GSDB project. There you will find all the scripts used for data extraction, formatting, normalization, 3D structure generation, evaluation and other necessary task.

A “readme” file has been provided in the “GSDB\_Scripts” directory to descript the content of each script in it.

# GSDB Data

Find the original GSDB data: processed and extracted data; and scripts on Lewis server.

**Path to Data and scripts**: /storage/htc/bdm/tosin

You will find two directories at this Location:

* **GSDB** which contain Data and Structures
* **GSDB\_Scripts** which contains the scripts for processing and Structure generation

# Hi-C 3D structure prediction tools

As of May 13, 2019, 12 prediction tools were included in GSDB. They are LorDG, 3DMax, MOGEN, PASTIS, Chromososme3D, HAS, miniMDS, Shrec3D, GEM, ChromSDE, SIMBA3D, and InfMod3DGen.

Each of these tools were selected because they some of them are the most recent. In addition, the list included tools have been mostly used as benchmarks in the 3D structure prediction field.

The input format and input filename suffix used by each of the methods is provided in the readme file accompany each of the normalized Hi-C dataset and on the GSDB project GitHub page ( <https://github.com/BDM-Lab/GSDB#2-algorithms-input> ).