BUILDING PLANT 3D GENOME COMPUTING RESOURCES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Chanaka Sampath Cooray Bulathsinghalage

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

April 2021

Fargo, North Dakota

North Dakota State University
Graduate School

Title

BUILDING PLANT 3D GENOME COMPUTING RESOURCES

By

Chanaka Sampath Cooray Bulathsinghalage

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Lu Liu
Chair

Anne Denton

Changhui Yan

Xiwen Cai

Approved:

04/09/2021
Date

Simone Ludwig
Department Chair

ABSTRACT

Chromatin interactions play increasingly important roles in three-dimensional genome organization and long-range gene regulation. Analyzing the three-dimensional structure of the plants is currently a growing field and we noticed that there is lack of computing resources on chromatin interactions for the plants. So, we are introducing a database of statistically significant chromatin interactions processed using Hi-C experimental approach. The users can search in the database using a set of genes or regions for a selected plant organism through a web browser and it lists down all the statistically significant chromatin interactions involved those genes or regions with the confidence scores, Gene Ontology information and pathway information. It serves as a computing resource for biologists and scientists who want to study plant genomes under the context of three-dimensional structure without any programming experience.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Lu Liu for his continued support, help and direction. This would not have been possible without his constant support. I wish to covey my gratitude to Dr. Xiwen Cai, Dr. Anne Denton and Dr. Changhui Yan for being on my graduate committee and for the valuable support. I would also like to thank my wife, Wathsala Jayawardana, for all the support and all my family members, friends who encouraged to complete this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

DNA stores the genetic information of an organism and consists of two complementary strands which form a double-helix structure. Mainly DNA is made of a sugar phosphate backbone and four types of nucleotides: Adenine, Thymine, Cytosine and Guanine as the ATCG bases. Adenine pairs with Thymine and Cytosine pairs with Guanine inside DNA.

In the Eukaryotic cells, DNA is wrapped around the histone proteins, which are again re-wrapped as a very tight coil structure called a chromatin. Chromatin is a very long strand that tightly fits inside the tiny space of a nucleus and it was discovered that the different parts of the chromatin that are in proximity interact with each other. These are called Locus-Locus interactions or chromatin interactions. Chromatin interactions are further broken down into the interactions that occur within the same chromosome (intra chromosomal interactions) and interaction that occur between different chromosomes (inter chromosomal interactions).

With the development of the technologies to analyze the three-dimensional structure of the genome and the increasing need to analyze the proximal regions led the number of chromatin interaction experiments to grow rapidly. Considering the genome wide analyzing techniques such as Hi-C or 4C, the raw data size is enormous, and the processing takes ample amount of computational power, memory and time. In addition, the people need to have programming skills and knowledge about the available tools in order to process raw data and get the result. So, the availability of computing resources for three-dimensional data is crucial for the people who do not have enough computational power or time. We analyzed the available computing resources on chromatin interactions for the plants, and we noticed that there is a lack of computing resources available online. Since analyzing plant genomes plays an important role in agriculture,

we figured that the availability of three-dimensional genome computing resources is essential for the plants and that is our main motivation to perform this study.

In this study, we first performed a thorough analysis of all the available chromatin conformation capture experiments for the plants. Then we selected three plants (Arabidopsis thaliana, rice, corn) that have majority of Hi-C experiments for the initial version of this application. As the results, we gathered publicly available 10 experiment publications for A. thaliana, 6 experiment publications for rice and 5 experiment publications for corn. Then we used separate tools to process the raw sequence data of Hi-C and extracted the significant interactions using a statistical measurement. In addition, we processed the gene distribution among the whole genome for each plant organism along with the gene ontology and pathway information. We built separate collections of Hi-C data with other necessary information related to the experiments and interactions. Finally, we developed a web interface for users to query statistically chromatin interactions.

The users have two main options to query the interactions which are either using a collection of genes or using a single or pair of regions for a selected organism. The results consist of multiple tabs. The web application first lists down all the interactions that are common to all the selected genes or regions and it will separately show interactions by individual gene or region as well. It will show additional information for each interaction such as gene ontology, pathway, significant level and normalized counts. Each experiment linked to its publication and raw experiment data as well.

2. BACKGROUND AND RELATED WORK

2.1. The importance of studying chromatin interactions

The approximately two-meter-long DNA is packed in a tiny nucleus of about six micrometers in diameter. In order to fit into this micro space, chromatin regions that are far apart from each other in the genome, are inevitably packed in very close proximity inside the nucleus. Deciphering the packaging and organization of the DNA in this tiny space is important to understand the functions of the different loci of the genome and their roles in gene regulation. More importantly this organization plays a critical role in determining the cell functions such as which genes are turned on or off in the cell and at what times those genes are active or inactive.

A recent study introduced a new method called Micro-c, to provide a better imaging technology on how smaller regions of DNA are organized in the three-dimensional space [1]. The study demonstrated that the smaller loci of DNA can control gene activities and when the regions are close enough, they can affect the process of recruiting proteins, which affects how genes are turned on or off. Thus, analyzing the close regions will help us to discover how the genes are controlled in the three-dimensional space and may provide opportunities to develop treatments according to the understanding of the gene activity regulation in the healthy and diseased cells.

Phanstiel et al. compared genome-wide high-resolution looping maps using Hi-C [2]. They discovered that the genes at loops that are newly formed or newly activated have increased gene expression levels, which leads us to believe that the complex network of chromatin loops is involved in coordinating changes in transcription during cell development. To identify a chromatin loop, we need to identify the proximal starting and ending regions of the corresponding loops and those can be identified by studying chromatin interactions.

Another team of researchers focused on the positions of the genes of the nucleus during different cell cycles [3]. In this study, the researchers analyzed thousands of different cells in different cell cycles. They discovered that the genes do not have a fixed location of the nucleus and the genes change positions in different stages of the cell cycle. This shows us the importance of analyzing the chromatin interactions in different cell cycles which illustrates how changes of the locations of the gene affect normal development and diseases.

Finn et al. combined Hi-C with an imaging method – high-throughput fluorescence in situ hybridization to physically map and visualize regions of DNA [4]. Their findings demonstrated that the organization and packaging of the DNA inside the nucleus varies among different cells and because of that the interaction counts between different regions are highly diversified across single cells. This is important because interactions between number of different single cells need to be analyzed in order to get an understanding of how the organization of genes affects in healthy and disease conditions in the cells.

2.2. Arabidopsis Thaliana, Rice(Oryza sativa) and Corn(Zea mays)

Arabidopsis Thaliana is a widely used organism in biology as a model plant to identify genes and their functions and it is the first plant to have its genome sequenced. It is a small flowering plant and a member of the mustard (Brassicaceae) family. Arabidopsis Thaliana has a relatively small genome with around 135 mega base pairs, and only five chromosomes. The latest genome release of Arabidopsis Thaliana is called TAIR10 and it contains 27,416 protein coding genes, 4827 pseudogenes or transposable element genes and 1359 ncRNAs (33,602 genes in all, 41,671 gene models). Apart from having a short genome, Arabidopsis Thaliana has a very short generation time and a large number of offspring. These combined facts constitute advantages for genome analysis.

Oryza sativa commonly known as rice, is widely used as a food for human. It also acts as a model organism for cereal biology and popular for being easy to genetically modify. It has a genome with around 373 mega base pairs and 12 chromosomes. The genome release of Oryza sativa is called IRGSP-1.0. The genome has 49,066 gene models and 39,045 total genes.

Zea mays commonly known as corn, is one of the most importance crops for human throughout the world. It is used as a food for humans as well as a source of biofuel. It has relatively large genome with around 2.1 giga bases and 10 chromosomes. We are using the genome version B73-AGPv4. The genome has 55801 total genes.

2.3. Experimental methods for chromatin interactions

2.3.1. Chromosome conformation capture (3C)

3C is one of the pioneering methods to identify the locations of the chromosomal interactions. It has later become the foundation for many of the other techniques [6]. 3C is used to identify the interaction frequency counts between two specific regions of the genome (one-to-one mapping). 3C process can be used to prove the existence of the chromatin loops between chromatin regions that are in proximity, and therefore show that they are involved in gene regulation. 3C method also provides high resolution visualization of the interactions.

The steps of the 3C process involves crosslinking the regions that are spatially proximal in the nucleus with formaldehyde which helps to freeze the contacts, cutting DNA with the restriction enzyme to separate out the contacts, ligating DNA fragments, purifying and detecting the site with Polymerase chain reaction (PCR). PCR is a method to make huge amounts of copies of a DNA sample which helps to study a small amount of DNA fragments.

2.3.2. Circular chromosome conformation capture (4C)

4C method is a derivative of 3C method and is used to identify the genomic sites in the whole genome interacting with a specific genomic site of interest (one-to-many mapping) [5]. 4C method can be used to provide high resolution contact maps around the genomic site of interest and involves a smaller number of reads compare to the other methods like Hi-C. The main difference between 3C and 4C method is cutting the fragment using two restriction enzymes instead of one, which helps to ligate the fragment in circular. In 4C, the genomic site of interest is called the viewpoint and the interacting sites are called captures.

Similar to the 3C method, 4C method also has the steps of crosslinking the ligation sites and then cutting DNA into fragments using the first restriction enzyme which is called primary restriction enzyme. After the fragments are ligated in situ, the crosslinks are removed and purified. Then the resulting fragments are trimmed using a second restriction enzyme and ligate again to form circularized ligations. The circularized ligations are processed with inverse PCR method to break the ligations and bind the primers to the viewpoint. Finally, the captures are sequenced using the next generation sequencing method and the contact frequencies are calculated using the proportion of the population mapped to the specific genomic sites.

2.3.3. Chromosome conformation capture with high throughput sequencing (Hi-C)

Hi-C is also an extension of the 3C method and it has proven to be more successful than the other methods when finding genome-wide pairwise chromatin interactions. Unlike the 3C and 4C methods, Hi-C method is used to identify interaction frequencies between all possible genomic sites in the whole genome (many-to-many mapping) [7]. So, it is very useful when the purpose of the study is to analyze the chromatin interaction map across the whole genome. The drawback of this method is the low resolution of the contact maps because sequencing cost

increases tremendously to cover the whole genome. Therefore, frequencies are grouped together into fixed-size bins.

Hi-C process has the same initial steps as 3C with crosslinking and cutting them with the restriction enzyme. It extends the 3C process by filling and marking the four ends with biotins which help to identify the ligation sites. Then it includes the ligation and shearing the DNA from the crosslinks. The process ends with reading the chimeric reads using the high throughput paired-end sequencing.

2.4. Computational methods

2.4.1. HiC-Pro

HiC-Pro [5] is one of the many tools that are capable of processing Hi-C data. Unlike the other available tools to process Hi-C data such as HOMER [6], HICUP [7], HiC-inspector [8], HiCdat [9] and HiC-box [10], HiC-Pro offers a full pipeline to process Hi-C data from raw reads to normalized contact maps including recovery of chimeric reads and correction of systematic biases. The hiclib [11] package is also capable of offering full pipeline as HiC-Pro but not being a standalone tool rather a python package and its limitations of parallelization and normalization of high-resolution data made HiC-Pro a better solution over hiclib for our study. Besides hiclib is not an actively maintained library.

HiC-Pro's workflow consists of four steps (Figure 1); read alignment, detection and filtering of valid interactions, binning and normalization. HiC-Pro first uses bowtie2 to map read pairs and then it rescues chimeric reads using an exact matching procedure. It detects valid interactions by choosing pairs near restriction sites and discard invalid ligation products such as dangling end, self-circle ligation and duplicate fragments. Then user can specify a resolution size in order to generate contact map divided into bins of equal sizes. Finally, the normalization is

done considering different biases such as GC content, mappability and effective fragment length [12] [13]. It is mentioned that HiC-Pro proposes a fast sparse-based implementation of the iterative correction method [14] in order to perform the normalization in a short time with reasonable memory requirements. In addition, HiC-Pro carryout a variety of quality controls in each step considering the metrics such as alignment statistics, fragment size distribution, fraction of intra and inter chromosomal interactions and long-range versus short range intra chromosomal interactions. The output of HiC-Pro workflow consists of non-null contact frequencies from half of the contact matrix. HiC-Pro is implemented in Python and C++ programming languages and freely available under BSD license.

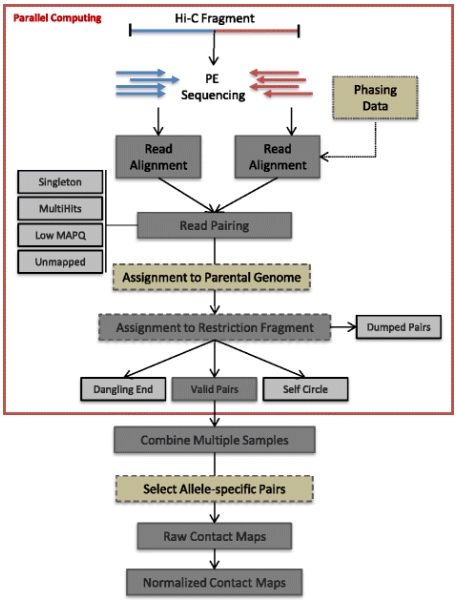

Figure 1. HiC-Pro Workflow; read alignment, detection and filtering, binning and contact map normalization. [5]

2.4.2. Fit-Hi-C and FitHiC2

Fit-Hi-C [15] is a computational tool which describes the statistical significance of mid-range chromosomal contacts and it is proven to be more successful than the previous methodologies such as the statistical model used by Duan et al. [16] in which every interaction is assumed to be equally likely. Fit-Hi-C only focuses on intra-chromosomal interactions.

First it fits an initial spline from the Hi-C contact maps using observed contact counts and genomic distances. Shape of the initial spline is the basis for the initial null model and using that, a threshold is determined to identify outliers. After filtering the outliers, a second spline is calculated to estimate the prior contact probabilities. Then it calculates P-values for all contacts, including null and outlier pairs using a binomial distribution and finally a Q-value is computed for each P-value by applying multiple hypothesis testing correction (Figure 2).

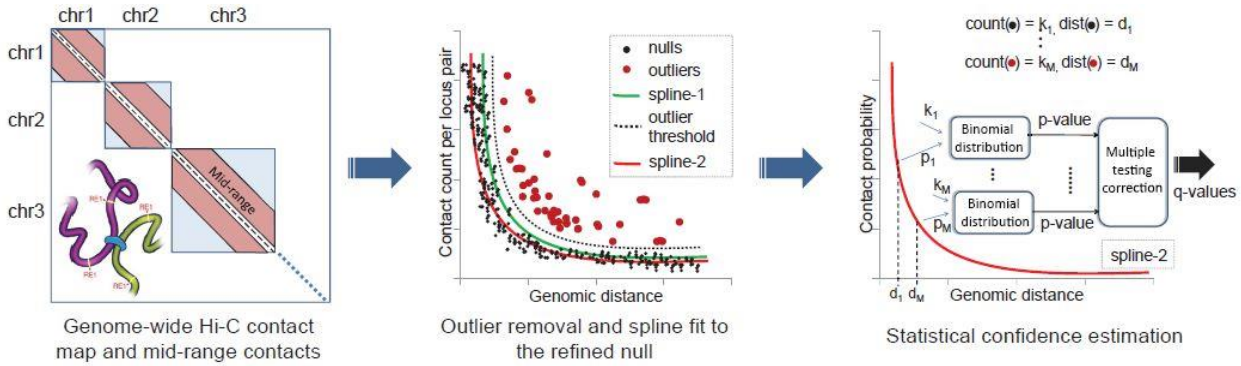

Figure 2. [15] Fit-Hi-C workflow. Compute the first spline from the Hi-C contact map. Compute p-value from the refined model and generate q-values using multiple testing correction.

One of the drawbacks of Fit-Hi-C is that it is only calculating statistical significance of intra-chromosomal interactions. So, they introduced FitHiC2 [17], the latest release of Fit-Hi-C, which is capable of describing the statistical significance of both inter and intra chromosomal interactions. In addition, FitHiC2 is capable of processing the Hi-C data on higher resolution such as 1 and 5kb. Due to these reasons, in our study, we used FitHiC2 to compute the statistical confidence estimates of both intra and inter chromosomal interactions.

2.5. Existing computing resources on chromatin interactions

2.5.1. 4DGenome

4DGenome [18] is the work most closely related to our study. 4DGenome is a comprehensive database of chromatin interactions and they claim that it is the first database that comprehensively documents and curates chromatin interactions generated by both experimental

and computational approaches. 4DGenome covers the experiment data processed via the methods including 3C, 4C, 5C, ChIA-PET, Hi-C, Capture-C, and IM-PET only in selected five organisms; Plasmodium falciparum (3D7), Drosophila melanogaster (dm3), Homo sapiens (hg19), Mus musculus (mm9) and Saccharomyces cerevisiae (sacCer3). It does not include any experimental data related to plants which is our main focus on this study.

4DGenome provides two main methods for users to query the data such as using multiple regions or using multiple genes. In the output, it provides confidence scores along with the contact frequencies, gene and tissue data. But in our study, we provide gene ontologies and pathway information along with the confidence scores and these details are explained in the latter sections. Also, in 4DGenome, it is showing all the interactions regardless of the significance level of the data. But in our study, we are filtering out only the statistically significant interactions based on the confidence scores.

2.5.2. LOGIQA

To the best of our knowledge, LOGIQA [19] is the only online database that references three-dimensional genome experiment data related to some plants. The database is designed to only show the quality score of the experiments. Experiments can be searched basically using organism, protocol, sample and restriction enzyme and the browser will list down the experiments related to the query with their relative quality score. Also, it provides the references to the experiment data and the publication. But LOGIQA is not capable of representing any individual interactions and cannot be searched by any genes or region information.

3. METHODOLOGY

3.1. Data collection

Table 1. Hi-C experiments of Arabidopsis Thaliana

Publication	**Experiment Data**	**Number of experiments/replicates**
Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. [20]	https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP043612	10
Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution [21]	https://www.ncbi.nlm.nih.gov/sra/?term=SRP064711	2
The effects of Arabidopsis genome duplication on the chromatin organization and transcriptional regulation [22]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114950	4
Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila [23]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55960	3
Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. [24]	https://www.ncbi.nlm.nih.gov/sra/?term=SRP032990	6
Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid [25]	https://www.ncbi.nlm.nih.gov/sra/?term=SRP095993	4
De Novo Plant Genome Assembly Based on Chromatin Interactions: A Case Study of Arabidopsis thaliana [26]	http://ibi.hzau.edu.cn/3dmodel/download/mp2014_raw_data.tar.gz	2
MORC Family ATPases Required for Heterochromatin Condensation and Gene Silencing [27]	https://www.ncbi.nlm.nih.gov/sra?term=SRP012587	2
Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis [28]	https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=545383	4
Long-range control of gene expression via RNA-directed DNA methylation [29]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64389	9

We gathered existing publicly available Hi-C experiments for each plant to perform this study. Table 1,Table 2 and Table 3 list the experiments for each plant with the link to corresponding experiment data and the number of experiments or replicates in each study. We used TAIR10 reference genome for A. Thaliana, IRGSP-1.0 reference genome for Oryza sativa

and B73-AGPv4 reference genome for Zea mays to process the data. In addition, we collected gene information, gene ontology information and pathway information as metadata to represent with the experiment data for each plant.

Table 2. Hi-C experiments of Oryza sativa

Publication	**Experiment Data**	**Number of experiments/replicates**
3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments [30]	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA391551/	4
Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis [31]	https://www.ncbi.nlm.nih.gov/sra/?term=SRP093806	4
Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice [32]	https://www.ncbi.nlm.nih.gov/sra/?term=SRP129302	5
Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains [33]	https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA486213	5
Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data [34]	https://bigd.big.ac.cn/gsa/browse/CRA001597	2
Population Genomic Analysis and De Novo Assembly Reveal the Origin of Weedy Rice as an Evolutionary Game [35]	https://www.ncbi.nlm.nih.gov/sra/SRS3098796	1

Table 3. Hi-C experiments of Zea mays

Publication	**Experiment Data**	**Number of experiments/replicates**
3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments [30]	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA391551/	27
Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains [33]	https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA486213	12
Widespread long-range cis-regulatory elements in the maize genome [36]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120304	2
3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize [37]	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA599454/	12

3.2. Data processing

3.2.1. HiC

First, Hi-C data was downloaded using fastq-dump tool [38]. Each publication has multiple experiments and those has to be downloaded and processed separately. Each experiment data has two fastq files to represent the two ends of each read. Then we processed the experiment data separately by each publication using HiC-Pro tool [5]. We generated multiple BED files with the list of restriction fragments digested using restriction enzymes dpnII, hindIII, MboI, MseI since those are the only restriction enzymes used in all of the experiments. Then we used these BED files, 20,000 bin resolution as parameters to run HiC-Pro tool.

We analyzed the contact maps generated using HiC-Pro and applied fithic2 [17] evaluation on top of that for both inter and intra chromosomal interactions. We used ICED normalized contact maps as inputs for fithic2. Finally, we filtered the records having qvalue<0.05 to filter out statistically significant interactions. We processed total records of 246,758,782 experiment data for A. thaliana, total records of 561,115,995 experiment data for Oryza sativa and total records of 367,241,504 for Zea mays.

3.2.2. Gene information

A. Thaliana TAIR 10 genome contains 27,416 protein coding genes, 4827 pseudogenes or transposable element genes and 1359 ncRNAs (33,602 genes in all, 41,671 gene models). Oryza sativa IRGSP-1.0 genome contains 49,066 gene models and 39,045 total genes. Zea mays B73-AGPv4 genome contains 55801 total number of genes. We do not consider the gene separation as gene models for all the plants and divide the genes into equal size bins same as experiment data. It helps to map the experiment data with the corresponding genes.

3.2.3. Gene ontology

Gene ontology (GO) is a way to capture biological knowledge in a written and computable format [39]. So, it is arranged as a hierarchical relationship between a set of concepts. It defines the relationship from less specific concepts to more specific concepts and captures information such as biological processes, molecular functions and cellular components. Gene ontology helps to identify the functional information of a gene, validate the experimental techniques, explore functional information for novel genes and etc.

In this study, gene ontology results are processed separately. Each gene associated with multiple gene ontologies. Each bin is associated with multiple genes. So, when we consider an interactive bin pair, there can be many gene ontologies associated with the interaction. To filter out the statistically significant gene ontologies we use hypergeometric test on each gene ontology of a bin pair.

We calculated Hypergeometric cumulative distribution function for each gene ontology in each bin pair as in equation 1. We declared the total number of genes in the bin pair mapped with the corresponding gene ontology subtracted by 1 as the value of x. M is the total number of genes in the whole genome, which is a constant for the corresponding plant. K is the total number of genes mapped with the corresponding gene ontology. N is the total number of genes included in the corresponding bin pair. All the values are counted without duplication. After getting the results we subtract that by 1 to get the pvalues (Equation 1) and gene ontologies whose pvalues are less than or equal 0.05 are selected for the corresponding bin pair.

$$p = F(x|M,K,N) = \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}} \quad (1)$$

$$pvalue = 1 - p \quad (2)$$

3.2.4. Pathway

A Biological pathway is series of actions occurring between molecules in a cell that results in a product or change in a cell such as assembly of a new molecules, turning genes on and off, moving a cell and etc. Studying pathways lead to learn more about diseases such as cancers by identifying which genes and other molecules involved in the corresponding biological pathway [40].

Pathways are processed in the same way as gene ontologies. We used the same statistical threshold as in gene ontologies to filter out significant pathways for a bin pair. So, the same equation 1 is used for the calculation. Only x and K variables are changed as x is now the total number of genes in the bin pair mapped with the corresponding pathway subtracted by 1 and K is the total number of genes mapped with the corresponding pathway. Other variable definitions stay same. In here, we used 0.05 threshold to filter out statistically significant pathways per bin pair using pvalue as well.

3.3. Database architecture

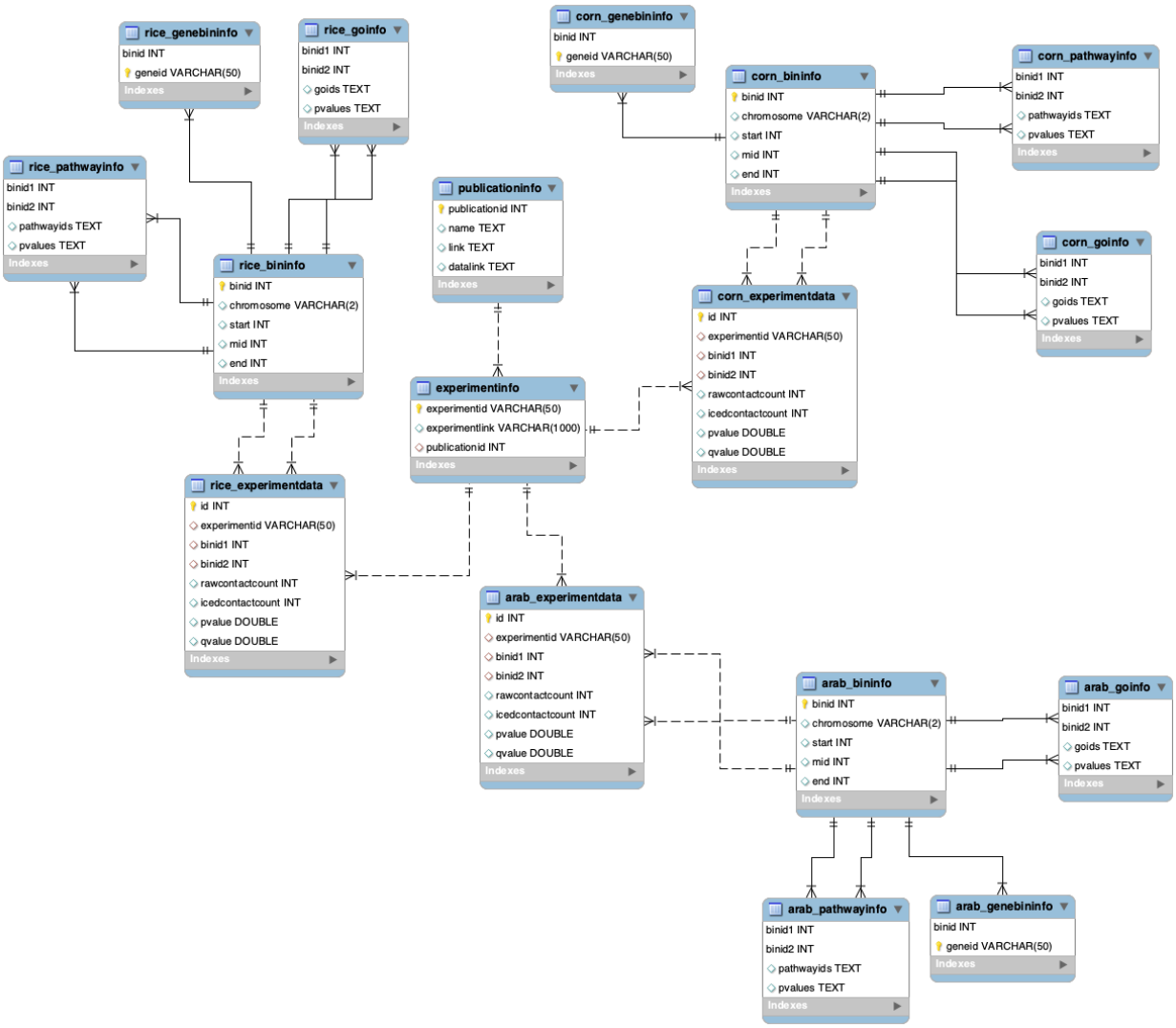

Figure 3. Database Schema Diagram generated using MySQL WorkBench v8.0 [41]

We used the MySQL version of 8.0.22 as the database management system. Figure 3 shows the full schema diagram of the database which is generated using MySQL WorkBench v8.0 [41]. The database only stores the statistically significant interactions filtered using qvalue. Gene ontology and pathway information are stored for each bin pair and the application aggregates the results at the time of querying from the database with the experiment data.

3.4. Web application architecture and data querying

We used the LAMP(Linux,Apache,MySQL, PHP) architecture to build our web application. We used the Ubuntu 20.04.1 LTS as the operating system, Apache v2.4.41 as the server, PHP v7.4.3 and MySQL v8.0.22.

To query the experiment data, we provide two methods for querying such as using a set of genes or using a pair of ranges for the corresponding selected plant organism. We have implemented the functionality of query using a set of genes in a way that it shows the interactions involving all the genes in the set and at least one bin of the interaction pair should include one gene from the set. So, we will not show the experiment data if all the genes in the set belongs to only one bin. Figure 4 explains possible valid pairs in more detail. The theory behind this is that we are interested in representing the significant interactions between genes in the genome. Most common use case of this is to find the interactions between two genes and in that case gene set only includes two genes. In that case, the results should include the interaction pairs in which each bin contains one of the two genes. In addition, we show interactions categorized by individual genes as well.

Functionality of querying using regions is same as the functionality of querying using genes as well. In this method, we only allow users to enter up to two regions and the interactions are filtered as both of the regions are included in the interactive pair. We consider overlapping of

the regions with the bins as well. Also, we show interactions by individual region as well. Web application can be acceded via the web site http://3dgenome.cs.ndsu.edu/.

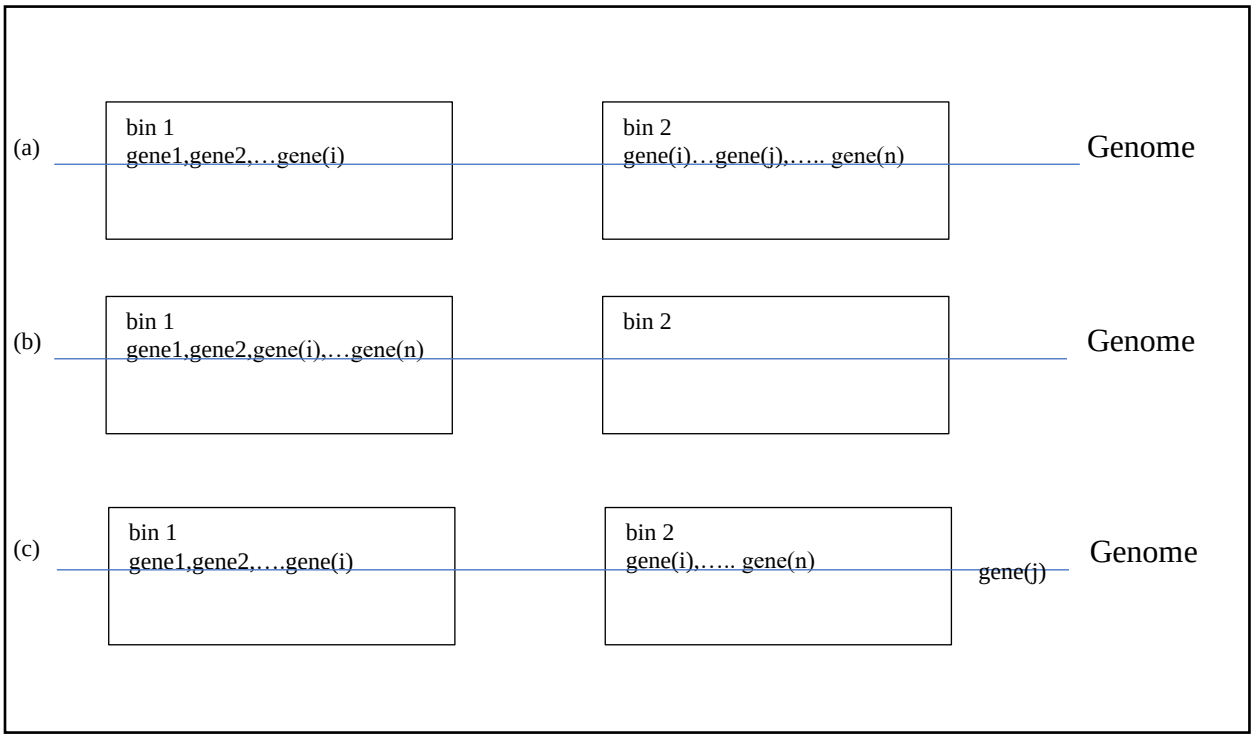

Figure 4. Bin pair analysis for a given set of genes (gene1,gene2,... gene(i),gene(j).....gene(n)). (a) one or more genes of the gene set included in bin1 and rest of the genes included in bin2, overlapping of genes possible. So bin1,bin2 is a valid pair. (b) all of the genes in the gene set included in bin1 so none of the genes are in bin2. So bin1,bin2 is not a valid pair. (c) bin1 has some of the genes of the gene set and bin2 has some of the genes of the gene set. But there are some genes outside of bin1 and bin2. So bin1,bin2 is not a valid pair.

4. RESULTS AND DISCUSSION

Table 4, Table 5 and Table 6 list all the publications we have used for this study with their total records and filtered significant interactions. According to the number of filtered records 0.56% for A. thaliana, 0.10% for Oryza sativa and 0.20% for Zea mays of total records has been identified as significant interactions.

Table 4. Total number of records processed per publication and total number of records identified as significant interactions for each publication for A. Thaliana

Publication	**Total Records**	**Significant Interactions Count**
Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. [20]	97,263,781	1,083,238
Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution [21]	11,917,993	55,362
The effects of Arabidopsis genome duplication on the chromatin organization and transcriptional regulation [22]	22,982,597	27,428
Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila [23]	18,726,981	46,379
Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. [24]	17,630,390	104,347
Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid [25]	9,718,100	16,034
De Novo Plant Genome Assembly Based on Chromatin Interactions: A Case Study of Arabidopsis thaliana [26]	6,355,952	9,918
MORC Family ATPases Required for Heterochromatin Condensation and Gene Silencing [27]	8,742,893	25,927
Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis [28]	11,948,592	14,108
Long-range control of gene expression via RNA-directed DNA methylation [29]	41,471,503	8,850
Sum	**246,758,782**	**1,391,591**

Table 5. Total number of records processed per publication and total number of records identified as significant interactions for each publication for Oryza sativa

Publication	**Total Records**	**Significant Interactions Count**
3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments [30]	63,880,802	24,910
Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis [31]	173,406,840	169,239
Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice [32]	202,035,105	303,956
Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains [33]	91,882,258	64,363
Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data [34]	11,953,341	5,937
Population Genomic Analysis and De Novo Assembly Reveal the Origin of Weedy Rice as an Evolutionary Game [35]	17,957,649	8,634
Sum	**561,115,995**	**577,039**

Table 6. Total number of records processed per publication and total number of records identified as significant interactions for each publication for Zea mays

Publication	**Total Records**	**Significant Interactions Count**
3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments [30]	207,256,614	9,380
Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains [33]	250,411,872	64,317
Widespread long-range cis-regulatory elements in the maize genome [36]	46,856,292	12,408
3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize [37]	74,304,499	35,136
Sum	**578,829,277**	**121,241**

From these results we can observe that the percentage of significant interaction count decreases for larger genomes. The reason for this could be that it needs significantly larger

amount of reads in order to cover a large genome. Also using Figure 5, we can see that even though A. thaliana has very less total interaction count compared to the other two plants, A. thaliana has the most number of significant interactions after we filter the interactions using confidence score.

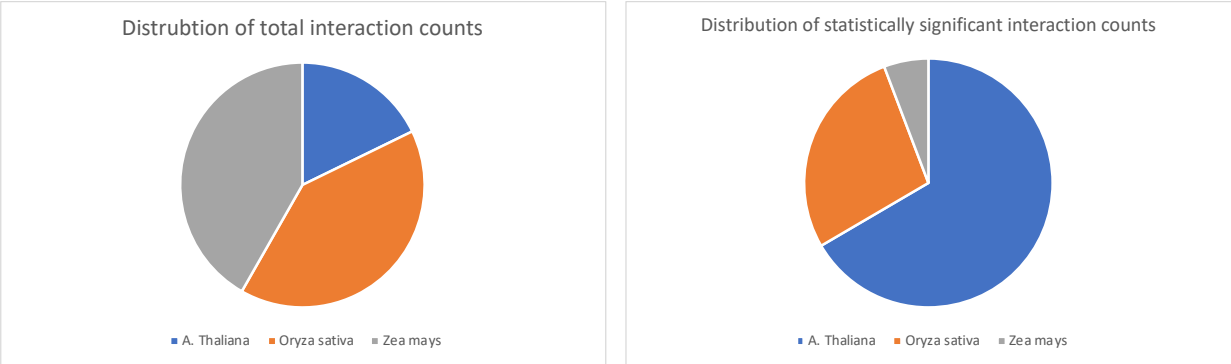

Figure 5. Distribution of total interaction counts among the three plants vs distribution of significant interaction counts

4.1. Understanding the results

First user has to select the plant organism to query the results in the home webpage. There are two methods to query the results; search using genes and search using regions. Considering the search using genes method, the user has to select the search using genes as the method and enter the comma separated gene list in the next text field(Ex: AT4G03290,AT4G05410) as shows in Figure 6. In the result webpage(Figure 7), it first shows all the interactions occurring between the given gene sets in a tab named common interactions. And then there are multiple tabs separated for each gene in the gene list. These tabs show all the interactions involving the corresponding gene. The result data are grouped by experiment name and under that experiment it shows the link to the raw experiment data and the corresponding publication article. Interaction data is organized in a table in a way that one row represent one

interacting pair of the genome. In the table, it shows the two loci that are interacting, normalized contact count, pvalue and qvalue of the corresponding interaction. As the metadata, for each interaction it shows the relevant gene ontology ids and pathway ids and their corresponding pvalues separated by commas.

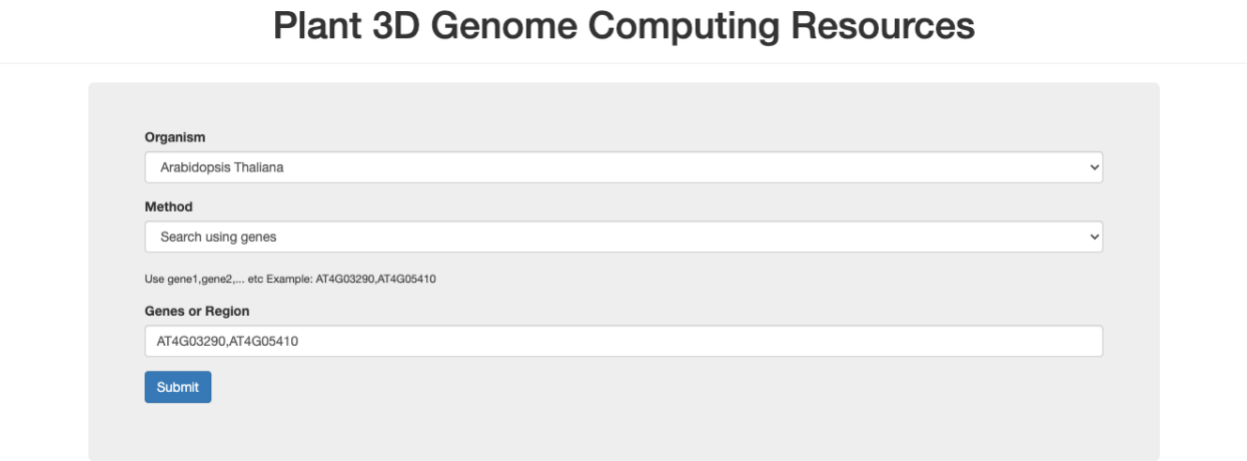

Figure 6. Example of search using genes; AT4G03290,AT4G05410 for plant A. thaliana

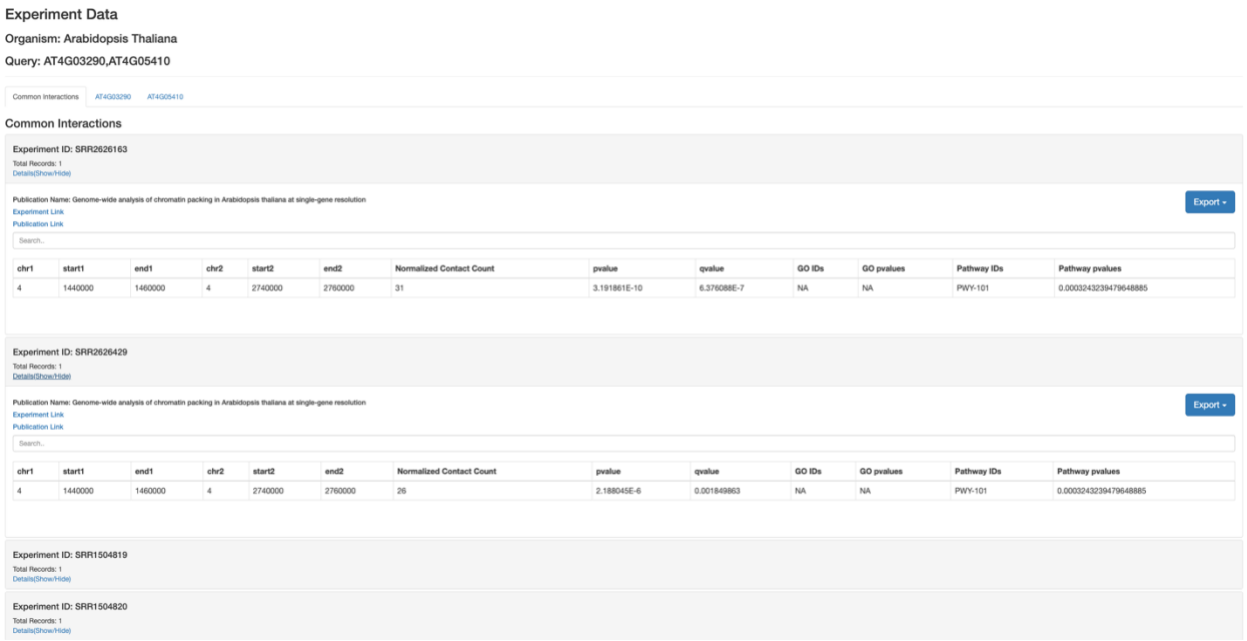

Figure 7. Result page for the query of search using genes; AT4G03290,AT4G05410 for A. thaliana

Functionality of search using regions works same as search using genes. In the text field the user has to enter the comma separated regions using the format chromosome number:start position of the region:end position of the region (Ex: 6:12820000:12920000,10:3480000:3580000) as shown in Figure 8 for organism Zea mays. User can enter up to two regions only. Result page is similar to the search using genes result page and it shows common interactions and the individual results separated by tabs(Figure 9).

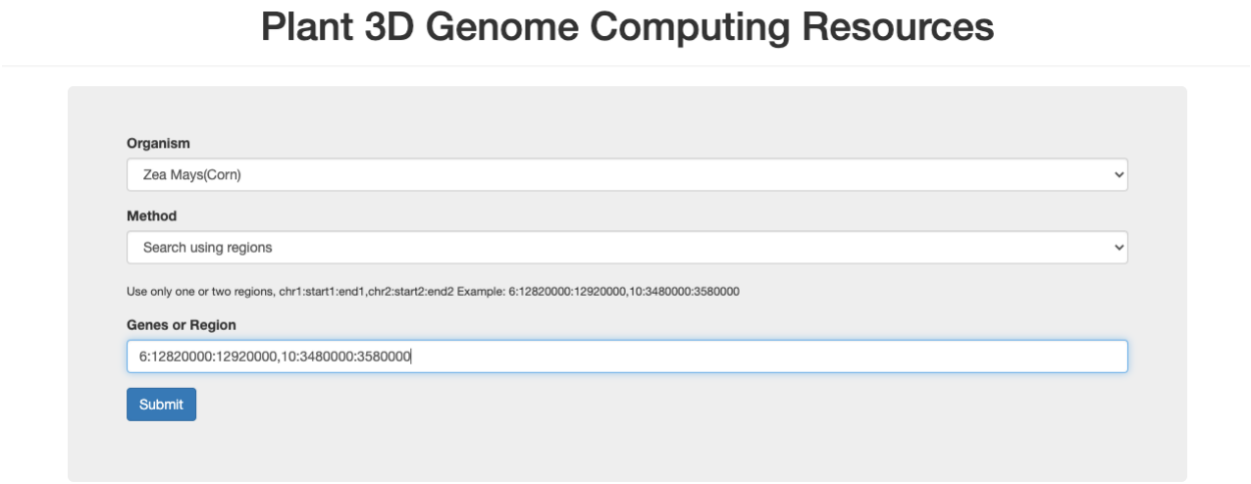

Figure 8. Example of search using regions for Zea mays; Using two regions; 1) chromosome 6, start index 12820000, end index 12920000, 2) chromosome 10, start index 3480000, end index 3580000

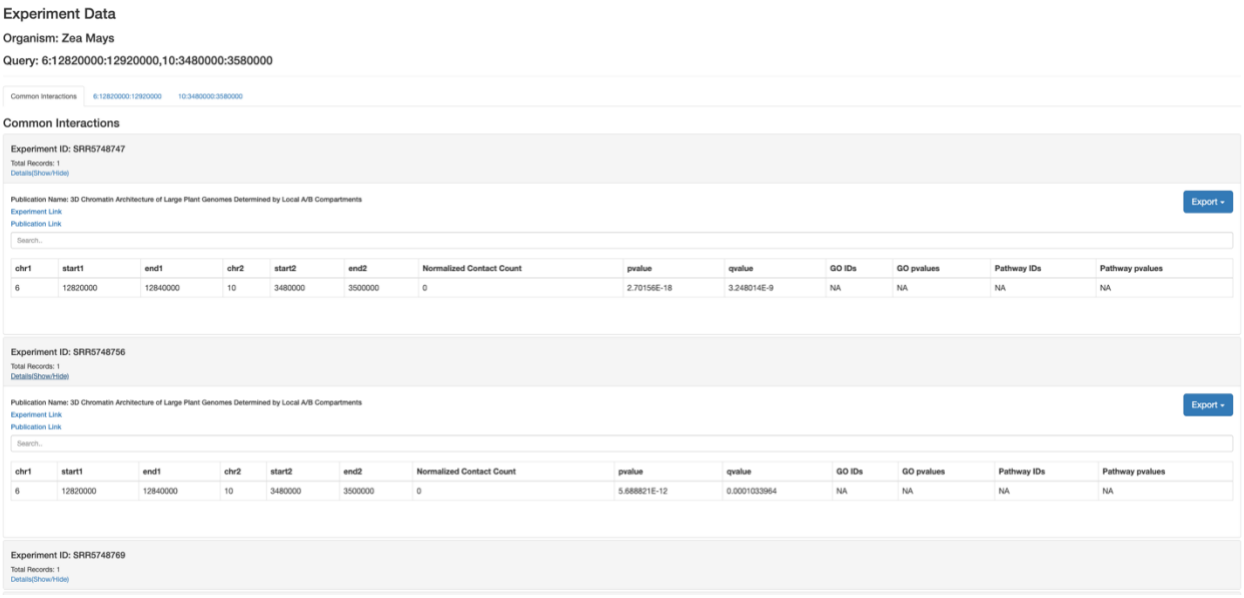

Figure 9. Experiment results for the query search using regions; 6:12820000:12920000,10:3480000:3580000

4.2. Exporting the records

For the moment, the users can export individual experiment results as a comma separated file. In the future, we are considering of implementing an FTP server so that the users can download the unfiltered records as well.

4.3. Example use case

Sample use cases include all the reasons to analyze the chromatin interactions as mentioned above. Assume that a user wants to understand if there is a relationship between two genes and wants to analyze whether one gene can affect the other genes behavior(gene regulation). As the initial step, the user can search if there are significant interactions between the two genes. We can use the gene AT4G03290 and gene AT4G05410 in A. thaliana organism as an example. From the results we can observe that multiple experiments have shown that the region 1440000 – 1460000 of chromosome 4 is highly interacting with region 2740000 – 2760000 of chromosome 4(Figure 9). This observation shows an interesting fact that even though the two regions are more than 1 million base pairs separated in the genome, the two

regions are actually in proximity inside the nucleus and are significantly interacting with each other. In addition, since the same interaction is detected by multiple experiments, it increases the confidence in the corresponding interaction. Then as the next step user can analyze the gene ontology and their annotation trees and pathway information to identify the behaviors and functionalities of the two interacting regions.

5. FUTURE WORK

We want to expand this study in order to cover more plant organisms such as Foxtail millet, Tomato and etc and make this database as the standard 3D genome browser for plants. Currently, users can not download unfiltered experiment data through the website. So, an FTP server connecting to this web application is useful to transfer the files to the end users who want to analyze unfiltered data. As an additional information, we are planning to add Single Nucleotide Polymorphisms (SNPs) information associated with the interactions. Connecting to the other genome browsers such as UCSC genome browser is also essential for the users to further analyze the interacting regions, their functionalities and genes. Currently, we have included almost every Hi-C experiment results available to this date for the corresponding plants in our data collection and we will continue to update the database in the future as well. In order to continuously update the database, automated system or semi-automated system will be essential and we will look for the ways to implement an automated system to update the data.

6. REFERENCES

[1] T. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. Hansen, O. Rando, R. Tjian and X. Darzacq, "Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding," *Mol Cell,* vol. 78, no. 3, pp. 539-553, 2020.

[2] D. H. Phanstiel, K. V. Bortle, D. Spacek, G. T. Hess, M. S. Shamim, I. Machol, M. I. Love, E. L. Aiden, M. C. Bassik and M. P. Snyder, "Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development.," *Molecular cell ,* vol. 67, pp. 1037-1048, 2017.

[3] T. Nagano, Y. Lubling, C. Várnai, C. Dudley, W. Leung, Y. Baran, N. M. Cohen, S. Wingett, P. Fraser and A. Tanay, "Cell-cycle dynamics of chromosomal organization at single-cell resolution," *Nature,* vol. 547, no. 7661, pp. 61-67, 2017.

[4] E. H. Finn, G. Pegoraro, H. B. Brandão, A.-L. Valton, M. E. Oomen, J. Dekker, L. Mirny and T. Misteli, "Extensive heterogeneity and intrinsic variation in spatial genome organization," *Cell,* vol. 176, no. 6, pp. 1502-1515, 2019.

[5] N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C.-J. Chen, J.-P. Vert, E. Heard, J. Dekker and E. Barillot , "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing," *Genome Biology,* vol. 16, 2015.

[6] S. Heinz, C. Benner , N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. . K. Glass, "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities," *Mol Cell,* vol. 38, no. 4, pp. 576-589, 2010.

[7] B. Bioinformatics, "HiCUP (Hi-C User Pipeline)," Babraham Institute, [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/hicup/. [Accessed 15 11 2020].

[8] G. Castellano, F. Le Dily, A. H. Pulido, M. Beato and G. Roma, "HiC-inspector: a toolkit for high-throughput chromosome capture data," *bioRxiv*, 2015.

[9] M. . W. Schmid, S. Grob and U. Grossniklaus, "HiCdat: a fast and easy-to-use Hi-C data analysis tool," *BMC Bioinformatics*, vol. 16, 2015.

[10] "HiC-Box," [Online]. Available: https://github.com/koszullab/HiC-Box. [Accessed 15 11 2020].

[11] "hiclib-legacy," [Online]. Available: https://github.com/mirnylab/hiclib-legacy. [Accessed 15 11 2020].

[12] E. Yaffe and A. Tanay, "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture," *Nature Genetics*, vol. 43, pp. 1059-1065, 2011.

[13] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren and J. . S. Liu, "HiCNorm: removing biases in Hi-C data via Poisson regression," *Bioinformatics*, vol. 28, no. 23, pp. 3131-3133, 2012.

[14] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker and L. A. Mirny, "Iterative correction of Hi-C data reveals hallmarks of chromosome organization," *Nature Methods*, vol. 9, pp. 999-1003, 2012.

[15] F. Ay, . T. L. Bailey and W. S. Noble, "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts," *Genome research*, vol. 24, no. 6, p. 999–1011, 2014.

[16] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau and W. S. Noble, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, p. 363–367, 2010.

[17] A. Kaul, S. Bhattacharyya and F. Ay, "Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2.," *Nature Protocols*, vol. 15, pp. 991-1012, 2020.

[18] L. Teng, B. He, J. Wang and K. Tan, "4DGenome: a comprehensive database of chromatin interactions," *Bioinformatics*, vol. 31, no. 15, pp. 2560-2564, 2015.

[19] M.-A. Mendoza-Parra, M. Blum, V. Malysheva, P.-E. Cholley and H. Gronemeyer, "LOGIQA: a database dedicated to long-range genome interactions quality assessment," *BMC Genomics*, vol. 17, 2016.

[20] S. Feng, S. J. Cokus, V. Schubert, J. Zhai, M. Pellegrini and S. E. Jacobsen, "Genome-wide Hi-C analyses in wild type and mutants reveal high-resolution chromatin interactions in Arabidopsis," *Mol Cell.*, vol. 55, no. 5, p. 694–707, 2015.

[21] C. Liu, . C. Wang, G. Wang, C. Becker, M. Zaidem and D. Weigel, "Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution," *Genome Res*, vol. 26, no. 8, p. 1057–1068, 2016.

[22] H. Zhang, R. Zheng, Y. Wang, Y. Zhang, P. Hong, Y. Fang, G. Li and Y. Fang, "The effects of Arabidopsis genome duplication on the chromatin organization and transcriptional regulation," *Nucleic Acids Research,* vol. 47, no. 15, p. 7857–7869, 2019.

[23] S. Grob, M. W. Schmid and . U. Grossniklaus, "Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila," *Molecular Cell*, vol. 55, no. 5, pp. 678-693, 2014.

[24] C. Wang, C. Liu, D. Roqueiro, D. Grimm, R. Schwab, C. Becker, C. Lanz and D. Weigel, "Genome-wide analysis of local chromatin packing in Arabidopsis thaliana," *Genome Res,* vol. 25, no. 2, p. 246–256, 2015.

[25] W. Zhu, B. Hu, C. Becker, E. S. Doğan, K. W. Berendzen, D. Weigel and C. Liu, "Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid," *Genome Biology,* vol. 18, 2017.

[26] T. Xie, J.-F. Zheng, . S. Liu, . C. Peng, Y.-M. Zhou, . Q.-Y. Yang and H.-Y. Zhang, "De Novo Plant Genome Assembly Based on Chromatin Interactions: A Case Study of Arabidopsis thaliana," *Molecular Plant,* vol. 8, no. 3, pp. 489-492, 2015.

[27] G. Moissiard, S. J. Cokus, J. Cary, S. Feng, A. C. Billi, H. Stroud, D. Husmann, Y. Zhan, B. R. Lajoie, R. P. McCord, C. J. Hale, W. Feng, S. D. Michaels, A. R. Frand and M. Pellegrini, "MORC Family ATPases Required for Heterochromatin Condensation and Gene Silencing," *Science,* vol. 336, no. 6087, p. 1448–1451, 2012.

[28] L. Sun, Y. Jing, X. Liu, . Q. Li, Z. Xue, Z. Cheng, D. Wang, . H. He and W. Qian, "Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis," *Nature Communications,* vol. 11, 2020.

[29] M. J. Rowley, . M. H. Rothi, G. Böhmdorfer, J. Kuciński and A. T. Wierzbicki, "Long-range control of gene expression via RNA-directed DNA methylation," *PLoS Genet,* vol. 13, no. 5, 2017.

[30] P. Dong, X. Tu, P.-Y. Chu, P. Lü, N. Zhu, D. Grierson, B. Du, P. Li and S. Zhong, "3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments," *Molecular Plant,* vol. 10, no. 12, pp. 1497-1509, 2017.

[31] C. Liu, Y.-J. Cheng, J.-W. Wang and D. Weigel, "Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis," *Nature Plants,* vol. 3, p. 742–748, 2017.

[32] Q. Dong, N. Li, X. Li, Z. Yuan, D. Xie, X. Wang, J. Li, Y. Yu, J. Wang, B. Ding, Z. Zhang, C. Li, . Y. Bian, . A. Zhang, Y. Wu, B. Liu and L. Gong, "Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice," *the plant journal,* vol. 94, no. 6, pp. 1141-1156, 2018.

[33] P. Dong, X. Tu, H. Li, J. Zhang, D. Grierson, P. Li and S. Zhong, "Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains," *Integrative Plant Biology,* 2019.

[34] X. Zhang, . S. Zhang, . Q. Zhao, R. Ming and H. Tang, "Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data," *Nature Plants,* vol. 5, p. 833–845, 2019.

[35] J. Sun, D. Ma, L. Tang, M. Zhao, G. Zhang, W. Wang, J. Song, X. Li, Z. Liu, W. Zhang, Q. Xu, Y. Zhou, J. Wu, T. Yamamoto, F. Dai, Y. Lei, S. Li, G. Zhou and H. Zheng, "Population Genomic Analysis and De Novo Assembly Reveal the Origin of Weedy Rice as an Evolutionary Game," *Molecular Plant,* vol. 12, no. 5, pp. 632-647, 2019.

[36] W. L. Z. J. L. e. a. Ricci, "Widespread long-range cis-regulatory elements in the maize genome," *Nature Plants,* vol. 5, p. 1237–1249, 2019.

[37] Y. D. L. Z. Y. e. a. Sun, "3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize," *Genome Biology,* vol. 21, 2020.

[38] "SRA Toolkit Documentation," National Center for Biotechnology Information, 2020. [Online]. Available: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=fastq-dump. [Accessed 30 11 2020].

[39] G. O. Consortium, "The gene ontology project in 2008," *Nucleic Acids Res,* vol. 36, pp. D440-D444, 2008.

[40] F. Zheng, L. Wei, L. Zhao and F. Ni, "Pathway Network Analysis of Complex Diseases Based on Multiple Biological Networks," *BioMed Research International,* 2018.

[41] "MySQL Workbench," Oracle Corporation and/or its affiliates, 2020. [Online]. Available: https://www.mysql.com/products/workbench/.