# कॉपी संख्या विविधता के लिए डेटा मार्ट का विकास

# DEVELOPMENT OF DATA MART FOR COPY NUMBER VARIATION

By

**RAMESH PRAJAPAT**

**Master of Science**

In

**Computer Application**


**ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE**
**ICAR-INDIAN AGRICULTURAL RESEARCH INSTITUTE**
**NEW DELHI – 110012**
**2017**

# कॉपी संख्या विविधता के लिए डेटा मार्ट का विकास

# DEVELOPMENT OF DATA MART FOR COPY NUMBER VARIATION

**By**

***RAMESH PRAJAPAT***

*Thesis submitted to the Faculty of Post-Graduate School,*
*ICAR-Indian Agricultural Research Institute, New Delhi,*
*In partial fulfilment of the requirements for the degree of*

**MASTER OF SCIENCE**
**IN**
**COMPUTER APPLICATION**

**Approved by:**

**Chairman:** _____

**(Dr. Krishna Kumar Chaturvedi)**

**Co-Chairman:** _____

**(Mr. Mohammad Samir Farooqi)**

**Member:** _____

**(Dr. Shashi Bhushan Lal)**

**Member:** _____

**(Mr. Sanjeev Kumar)**

**Member:** _____

**(Dr. Mohammad Wasi Alam)**

फैक्स@FAX : 011-25841564
दूरभाष@Phones: संस्थान@Off.: 011-25847122-24 (4306)
घर @ Res.: 011-25273935
09868085105 (M)
ईमेल@Email :*kk.chaturvedi@icar.gov.in*

**भा. कृ. अनु. प. – भारतीय कृषि सांख्यिकी अनुसन्धान संस्थान**
**लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली – ११००१२ (भारत)**
**ICAR-Indian Agricultural Statistics Research Institute**
**Library Avenue, Pusa, New Delhi-110012 (India)**

**डा. कृष्ण कुमार चतुर्वेदी**
वरिष्ठ वैज्ञानिक
**Dr. K. K. Chaturvedi**
*Senior Scientist*

# CERTIFICATE

This is to certify that the work incorporated in the thesis entitled **"Development of Data Mart for Copy Number Variation"** submitted in partial fulfillment of the requirement for the degree of **Master of Science** in **Computer Application** of the **Post-Graduate School, Indian Agricultural Research Institute, New Delhi,** is a record of bonafide research carried out by ***Mr. Ramesh prajapat*** under my guidance and supervision and no part of this dissertation has been submitted for any other degree or diploma.

All assistance and help received during the course of this investigation has been duly acknowledged.

New Delhi
Date:

(**Dr. K. K. Chaturvedi**)
Chairperson
Advisory Committee

# ***Acknowledgements***

*This thesis is the fruitful outcome of the knowledge gained over the entire period of my M.Sc. study at ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi during which I have been in touch with a great number of people whose contribution in varied yet myriad ways has led to the research and making of the thesis which deserves special mention. It is a pleasure to convey my gratitude to all of them by way of my humble acknowledgements.*

*First and foremost with reverence, I want to express deepest sense of gratitude to* ***Dr. Krishna Kumar Chaturvedi*** *Senior Scientist, Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi and Chairman of my Advisory Committee for his initiative, benevolence, endurance, constructive criticism and constant monitoring during the period of my investigation and also during preparation of this thesis. Above all and most needed, he provided me constant support and encouragement in various ways. I consider myself blessed having the privilege of being guided by him. I am really indebted to him.*

*I am also equally indebted to* ***Md. Samir Farooqi****, Scientist, Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI) and Co-Chairman of my advisory committee for his moral support during the course of my research work. He gave me constant encouragement from time to time for completion of my research work. I am really thankful to him for providing his valuable time and helping me in understanding the basic concepts regarding my research work. Without his support and encouragement it would have been quite difficult for me to reach the goal in time.*

*It is great privilege for me to express my esteem and profound sense of gratitude to* ***Dr. Shashi Bhushan Lal****, Senior Scientist, Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi and Member of my Advisory Committee for his valuable suggestions and help.*

*It is great privilege for me to express my esteem and profound sense of gratitude to* ***Mr. Sanjeev Kumar****, Scientist, Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi and Member of my Advisory Committee for his valuable suggestions and help.*

*I would also like to express my sincere thanks to* ***Dr. Mohammad Wasi Alam****, Scientist, Forecasting Techniques and Agricultural Systems Modeling, ICAR-IASRI, New Delhi and Member of my Advisory Committee. I am really grateful to him for the support and encouragement he provided during the course of my studies.*

*I would like to convey my deep sense of gratitude and appreciation to* ***Dr. Sudeep Marwaha****, Professor, Computer Application, ICAR-IASRI, New Delhi, for his help and support during my entire course work. I am indebted to him for his valuable advice and constructive criticism during preparation of this thesis.*

*I am also thankful to* ***Dr. A.K. Choubey****, Head, Computer Application for his kind hearted support and encouragement in completing my research work.*

*I am deeply indebted to* ***Dr. U. C. Sud****, Director, ICAR-IASRI for the help and infrastructures provided by him.*

*I want to express my respect and gratitude to my family, especially* ***Papa, Mummy, Bhaiya, Bhabhi, Didi, Jija ji, Suman, Rohit, Ruchika, Pradhuman, kunal*** *for their sacrifices they did for me. These were the inspirations and moral boosting which gave me sufficient energy to complete this thesis in time.*

*I am thankful to my all batch mates specially* ***Arpan, Dilip****,* ***Himanshu, Asit, Sapna, Ronit, Aamir, Yeasin, Dipankar, Samir, Omprakash, Akhilesh, Aashis, Shweta, Sneha, Ritwika, James*** *and* ***Fabrice*** *for their friendly approach and moral support throughout my entire M.Sc. tenure.*

*I do not have any words to express the help and support given by my seniors specially* ***Chandan sir, Sreekumar sir, Tanuj sir, Pramod Sir, Rahul sir, Asif Sir, Somanna Sir, Nitin Sir, Sanober Sir, Murari Sir, Rajiv Sir, Animesh Sir, Sanchita ma'am , Supriya ma'am, Parvez sir, Shabana ma'am, Sonica ma'am, Soumya ma'am, Gopal sir, Amit sir, Bubul sir, Anubhav sir, Nalini sir, Sandeep sir, Nabin sir, Priyanka ma'am, Subhrajit sir, Kuldeep sir, Prakash sir*** *and my juniors* ***Lovkush, Kapil, Subin, Mahalinga, Vaijunath, Vivek, Jitendra, Rohit, Sandeepan, Amit, Ankita, Debdali, Garima, Lashmi, Sayanti*** *for all their affectionate support and help.*

*I am especially thankful to* ***Arpan***, ***Nitin sir, Rahul Sir, Murari Sir, Animesh Sir, Shabana ma'am, Ankita*** *for their constant support and encouragement everytime.*

*I take this opportunity to appreciate the help rendered by the staff of TAC, and CABin, ICAR-IASRI and special thanks to* ***Sanjeev ji, Thakur ji and Gagan ji***.

*I am indebted Dean and Joint-Director (Education), ICAR-IARI, New Delhi for providing facilities to carry out this research work.*

*I like to thank all the staff of PG School for their helpful attitude and cooperation, throughout the period of study.*

*Last but not the least I am thankful to ICAR-IASRI for the financial assistance provided to me in the form of Fellowship during the tenure of my study.*

*Finally, I would like to thank everybody who was imperative to the successful realization of this thesis, as well as articulate my apology that I could not mention personally one by one.*

*Date :*

*Place: New Delhi-110012* (***Ramesh Prajapat***)

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER I

# INTRODUCTION

## 1.1 Background

India is a country with a predominantly agrarian economy. Nearly 70% of the population of the country is being directly or indirectly engaged in agricultural practices for meeting their daily livelihood. In India, agriculture contributes about seventeen percent (17%) of total GDP and nearly ten percent (10%) of total export from the country. India is a second largest country in terms of total arable land with 60% of the total land area. The growth rate of population is continuously increasing but the area under agriculture is decreasing due to urbanisation and also because of increase in barren due to excessive use of fertilizers, salinity and other stress related factors. So, to feed this increasing population of the country, the traditional method of farming is not sufficient. Modern agriculture technique is the need of the hour to support the growing population. Newer high yielding varieties are needed to be developed by using modern breeding techniques. The advances in genome sequencing technologies are helpful in identification of different types of markers which can help in development of high yielding varieties. Single reference genome is not able to provide the representation of genetic diversity in a given species. The diversity can be identified and discovered using the study of structural variation in the form of copy number variants (CNVs). This will account for complete value of genetic information that is present in individual species. Copy number variation (CNV) plays an important role in identifying the genetic and phenotypic variation in the breeding population.

Copy number variation (CNV) is a one of the important genetic variant that makes a large genomic region with multiple number of copies. The number of repeats in the genome varies between various individuals exists in the population (McCarroll and Altshuler, 2007). Copy number variation is a type of intra-specific genetic and source of phenotypic variation (Yu *et al.,* 2013). Specifically, it is a type of duplication or deletion in the sequence of genetic material event that affects a considerable number of base pairs. Copy number variation occurs in human, plant, and variety of organisms including

*E. coli*. Approximately two-thirds of the entire human genome is composed of repeats and 4.8-9.5% of the human genome can be classified as copy number variations. Copy number variations play an important role in generating necessary variation in the population.

Based on the size of repeats, copy number variation is generally categorized into two groups namely short repeats and long repeats. This classification depends on the nature of the loci of interest. Short repeats include mainly bi-nucleotide repeats (two repeating nucleotides e.g. A-B-A-B-A-B...) and tri-nucleotide repeats (three repeating nucleotides e.g. A-B-C-A-B-C…). Long repeats include repeats of even entire gene sequences. CNV can be detected with the help of Next Generation Sequencing-based (NGS) method.

For detection of the copy number variation, the most common technique is cytoplasmic technique. These techniques allow one to observe the physical structure of the chromosome. One of the cytoplasmic technique is fluorescent in situ hybridization (FISH) which involves inserting fluorescent probes that bind with a high degree of complementary sequence in the genome. Comparative genomic hybridization is also another method in which fluorescent probes used to detect copy number variations by fluorophore visualization and then comparing the length of the chromosomes. This method has one important drawback *i.e.* the genomic resolution is relatively very low with compare to the new detection method and only large repeat sequences such as whole gene repeats can be detected.

Recently, so many detection methods are developed in the Biotechnology and these methods have advantages over the traditional methods. The CNV can be detected very short copy number variation or short repeats. One of the advance methods is bacterial artificial chromosome (BAC) array. NGS-based method is also being used to detect CNVs.

The NGS-based CNV detection methods are categorized into five different strategies namely

i. Paired-end mapping (PEM),
ii. Split read (SR),

iii. Read depth (RD),
iv. *De novo* assembly of a genome (AS) and
v. Combination of the above approaches (CB).

Researcher and scientists want to produce disease resistance and high yielding varieties. This is possible by making the genetic improvement by studying copy number variation. Thus, it is very important to study the variation produced by genomic sequence and genetic improvement. CNVs are associated with important traits and it is beneficial for genetic improvement of a crop. The use of CNVs has been documented in the literature as follows

i. In soybean (*Glycine max*), CNVs at the Rhg1 locus can mediate resistance to soybean cyst nematode (Cook *et al.,* 2012).
ii. In maize (*Zea mays*), CNV in a transporter gene (MATE1) was found to be the genetic basis for aluminum tolerance (Maron *et al.*, 2013).
iii. In barley (*Hordeum vulgare*), increased copy number of a Bot1 (boron transporter) gene provides tolerance to boron-toxicity (Sutton *et al*., 2007)
iv. In rice (*Oryza sativa*), a deletion in qPE9-1 gene is associated with the panicle erectness (Zhou *et al.,* 2009) and a deletion of the qSW5 gene caused the increase in grain size (Wang *et al.*, 2015).

Thus, there is a need to develop a solution that will help in retrieving the information associated with the CNVs.

**1.2 Need of Data Warehousing**

Information and communication technology (ICT) has been used in agriculture and allied fields for many years. The application of ICT usage start from conduct of experiment, data collection, analysis of experiment and their result in the different field of agriculture, the genome database development for different species, biological processes, decision support system, monitoring, evaluation and control, information management, knowledge dissemination to many others areas.

One of the most important and valuable assets of an organization is its information. This asset is almost always present in the organization in two forms, one is the operational system of record and the other is historical data storage. The current data is stored in operational systems while data warehouse is needed to preserve the historical data.

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management decision-making process (Kimball and Ross, 2011). It is developed as a central repository to provide the way of developing decision support system (DSS) and consists of integrated historical data both in summarized and detailed form. Data warehouse along with Online Analytical Processing (OLAP) tools are being increasingly developed for easy organization, retrieval and maintenance of large data in addition to provide the exploratory analysis data.

A data mart is a logical subset of the data warehouse of an organization. It deals with a specific domain or single subject which has an architectural foundation of a data warehouse. In a data mart, data is easily summarized, sorted or selected without any global considerations due to which it is more preferred by users and analysts. There is a limited number of sources and less processing overhead in data mart as compared to a data warehouse. The foremost need to integrate scattered information on a specific domain is to provide easy data availability and accessibility for researchers and agencies.

In agriculture sector, genomic phenomena generate complex data and need to be analysed in many ways. The detected CNVs are available in various forms and formats as per the detection mechanism and the size of data is becoming very big and available in disintegrated manner. There is a need to make them available on a single platform by collecting these different types of annotated and validated data by various researchers. The extraction and integration of this valuable information related to CNVs are of great challenge in current days due to the wide spread of these sources. In the present study, the main focus is the development of a data mart for Copy number variation (CNV) that will provide the available information in an analytical form.

### 1.3 Objectives

CNV information is available in different format which are difficult to be analysed or used for further analysis. In this thesis, an attempt was made to develop a data mart for copy number variation with the following objectives:

i. To design and develop schema for CNV mart.
ii. To implement ETL process for CNV mart.
iii. To test the solution with sample data.

### 1.4 Plan of Thesis

The present thesis consists of six various chapters.

**Chapter I** provides the brief introduction of the problems and accordingly objectives of the study has been set up.

**Chapter II** deals with the detailed description related to earlier work in this area in the form of review of literature.

**Chapter III** describes the materials and methods used in designing and developing the data mart.

**Chapter IV** mentions the steps requires in development of data mart for copy number variation and accordingly the results have been presented.

**Chapter V** discusses the working principle of data mart in the form of various types of reports which are helpful for easy to use in various ways.

**Chapter VI** concludes the work and summarizes the study. This is followed by abstract and references.

# CHAPTER II

# REVIEW OF LITERATURE

## 2.1 General Perspective

Variability is the most important factor for selection of plants for the breeding improvement programs. Copy number variation (CNV) is an important structure variants that helps in identifying the genetic and phenotypic variation in the breeding population. CNV is annotated by various researchers and published the annotated information in compiled formats and available in the literature as a part of their supplementary materials. This information is not available in ready to use formats and lacking in finding the association with many other parameters. The related work in this area has been summarized in the following section.

## 2.2 Related Work

- **From enterprise models to dimensional models: A methodology for data warehouse and data mart design (Moody, 2000)**

  This study revealed about different methods for developing dimensional models from traditional Entity Relationship (ER) models in the early days. These methods can be used to design data marts or data warehouses. Various schemas have been discussed to develop the data marts from enterprise data models.

- **Designing data marts for data warehouses (Bonifati *et al.*, 2001)**

  The data warehousing and OLAP technologies were discussed with an emphasis on incorporating upcoming new requirements in existing applications without affecting the running applications. Methods in data warehouse technology were also discussed with emphasis to identify and extract data marts out of an enterprise wide information system. Further, the specification of software and tools for data handling, storing data into data warehouse, building multidimensional data models for OLAP, front-end user interface for query and data analysis were also discuused.

- **Design of data marts for plantation crops (Kumar *et al.*, 2002)**

  The main objective of this study is to design and develop the data mart of plantation crops. Three data marts viz. statistics, agro-techniques, and research were identified and developed. The summary and detailed levels of data in this data mart were also documented.

- **Online analytical processing in agriculture using multidimensional cubes (Chaturvedi *et al.*, 2008)**

  An overview of online analytical processing in agriculture using multidimensional modeling have been discussed with emphasis on various types of schema. Different processes and techniques involved in designing and development of multidimensional cubes with reference to agricultural sector were discussed.

- **Dimensional issues in agricultural data warehouse designs (Nilakanta *et al.*, 2008)**

  Issues related to dimensional modeling in agriculture has been discussed with respect to data warehouse design. The study presented the challenges faced in designing the data warehouse, mainly in dimensional modeling and deployment of multidimensional cubes and also presented some early user judgment to the success of the warehouse.

- **Design and development of data mart for animal resources (Rai *et al.* 2008)**

  A data mart was developed for easy and fast access of animal resources which includes the complete process of building On-line Analytical Processing (OLAP) system for research managers dealing within an animal science domain. The capabilities of data quality and consistency checks were also implemented. The data has been provided in the form of web based OLAP cubes that is helpful in exploring the animal census data to develop the strategy towards animal resource planning and management.

- **Design and development of data marts for household amenities from census data of Maharashtra state (Suresh, 2008)**

  In this, a data mart was designed and developed for Household Amenities using Census Data (Maharashtra). The source data was available in excel sheets and MS-Access. It is very difficult to retrieve the desired data from multiple tables due to having improper coding schemes. Multidimensional cubes were designed and developed for census data of Maharashtra state and deployed over web based OLAP cubes.

- **Design and development of data mart for consumption expenditure survey data (Dutta, 2010)**

  The thesis described about design and development of data mart for consumption expenditure survey data. The survey was conducted by National Sample Survey Office (NSSO). Online Analytical Processing (OLAP) cubes based on multidimensional data models were designed and developed. The multidimensional cubes were deployed over web for easy and quick retrieval.

- **Development of an integrated Cropland and Soil Data Management system for cropping system applications (Yang *et al*., 2011)**

  Cropland and Soil Data Management system is capable of automatic data consolidation and integration. The Cropland Data Management component of the system is based on the Cropland Data Layer (CDL) products from the USDA National Agricultural Statistics Service. Management system implemented with seven program modules Map Cache Generator, Data Requester, Data Fetcher, Data Parser, Map Service Builder, Geo-database Builder, and Cropland Map Viewer. The Soil Data Management component is based on the Soil Survey Geographic (SSURGO) database from the USDA Natural Resources Conservation Service and is implemented with six program modules: Data Requester, Data Fetcher, Data Parser, Database Builder, Soil Map Generator, and Soil Map Viewer. This is a traditional data management software having GIS viewers capabilities.

- **Ensembl BioMarts: a hub for data retrieval across taxonomic space (Kinsella *et al*., 2011)**

  BioMart for genomic data was designed and developed in this study. The developed BioMart is useful for human, rat and chicken's genomic data. The BioMart is developed in Ensembl project. The Ensembl project was launched by a joint effort by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI).

- **Genome-wide copy number variations in *Oryza sativa* L. (Yu *et al*., 2013)**

  The research determined the CNV content using high-density array comparative genomic hybridization in a panel of 20 Asian cultivated rice comprising six indica, three asu, two aromatic, three tropical japonica, and four temperate japonica varieties. This study motivated to condct the current work and gave a direction to develop the data mart for CNVs.

- **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives (Zhao *et al*., 2013)**

  The research explored different strategies for CNV detection and also discussed about a CNV detection approach like Microarray-based comparative genome hybridization (array CGH) or genotyping arrays which were used for detection of large region repeats. The study also introduced newly and high resolution CNV detection method based on next-generation sequencing (NGS).

- **The impact and origin of copy number variations in the *Oryza species* (Bai *et al*., 2016)**

  Three out of five complementary approaches have been applied to identify CNVs in the *Oryza* species and these approaches are paired-end (PE) mapping, split-reed (SR) analysis, read-depth (RD) analysis. Full range of high-confidence CNVs were documented by these complementary sequence approaches (PE, RD, and SR) in the literature and the result was integrated with three CNV discovery

tools: BreakDancer, CNVnator, and Pindel. CNVs for 50 rice accessions (taken from rice3K repositories) covering various geographical areas of Asian region have been published.

# CHAPTER III

# MATERIAL AND METHODS

This chapter describes the requirements and design methodology for the development of the Data mart. Various tools and technologies were utilized in different phases of the Data mart development. The tools and technologies involved in development of this system are discussed. The step by step process has also been given in this chapter to develop the data mart and deploy using OLAP cubes, interactive reports and dashboards. The tools and software used in this study are available in public domain and under open source software license. This chapter will also discuss various phases of the data mart design, dimensional modeling and overall system design.

## 3.1 Data

Data word is originated from Latin word datum. A “datum” is a singular from and “data” is a plural form. Datum more commonly called data point means a single factual or a single entity which is the smallest measurable unit. Data can be qualitative or quantitative. The data need to be processed and acts as information that needed to be delivered to the end user for its use.

Copy number variation exists in most of the species and takes various forms that differ from one individual to another individual. Approximately 20% variations occur due to presence of multiple copies of CNVs in the genomes. The length of the CNVs is varying from few base pairs to mega base pairs. The variations in copy numbers helps in identifying various traits and stresses affecting that individual. Thus, CNVs identification and determination facilitates in development of diagnostic tools and treatments.

Initially, genome wide copy number variations in *Oryza sativa L*. was documented for 20 Asian rice accessions by using high-density array comparative genomic hybridization method (Yu *et al.*, 2013).

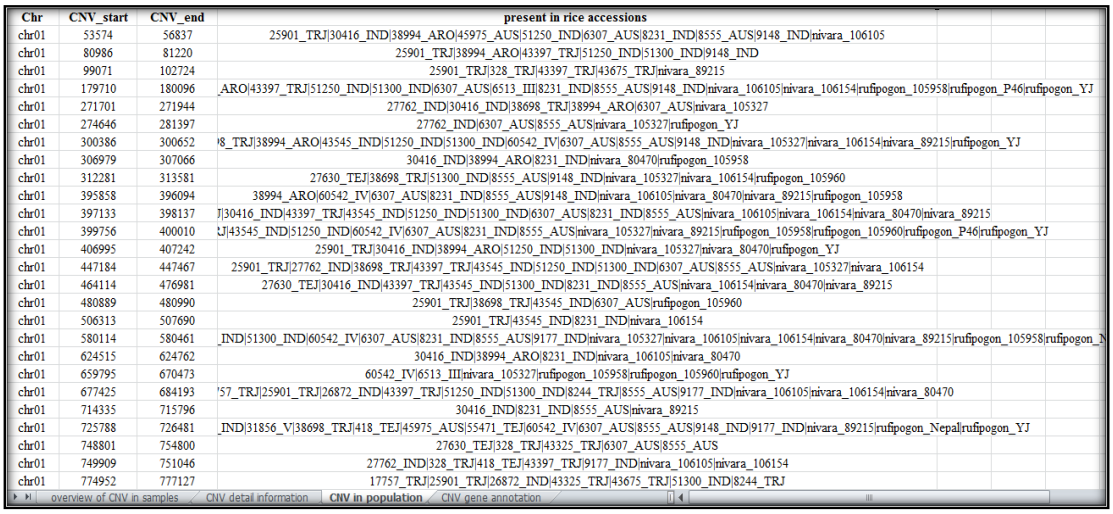
Another study reported the origin of copy number variations in Oryza species and documented full range of high-confidence CNVs by complementary sequence features like paired-end (PE) mapping, split-read (SR) analysis, and read-depth (RD) analysis (Bai *et al.*, 2016). They compiled the results for 50 rice accessions covering various geographical areas of the Asian region by using three CNV discovery tools i.e., BreakDancer, CNVnator, and Pindel. The compiled data is available as supplementary material that is used to develop the data mart.

In this supplementary material, the data is provided in the form of excel sheets. Table 3.1 describes the chromosome wise CNV presence in specific cultivars and sample data is shown in Figure 3.1. It is very difficult to find the CNV presence and position in the current format. Thus, this is needed to be transformed for easy to use and access the data required for further analysis. Once this data will be further extended for other cultivars/accessions, it will be very easy to incorporate into the existing schema.

*Table 3.1: Description of Cultivar Sample Data*

| **Attribute Name** | **Description** |
|---|---|
| Chr | Chromosome number |
| CNV_start | Starting position of CNV |
| CNV_end | Ending position of CNV |
| Present in rice accession | Cultivars name in which CNV is present |

Figure 3.2 specifies the classification of various cultivars. This classification contains groups and subgroups and their attirbutes description are given in Table 3.2.



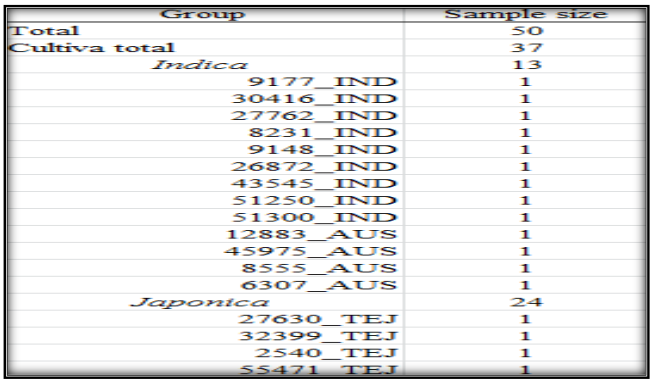
| Chr | CNV_start | CNV_end | present in rice accessions |
|---|---|---|---|
| chr01 | 53574 | 56837 | 25901_TRJ\|30416_IND\|38994_ARO\|45975_AUS\|51250_IND\|6307_AUS\|8231_IND\|8555_AUS\|9148_IND\|nivara_106105 |
| chr01 | 80986 | 81220 | 25901_TRJ\|38994_ARO\|43397_TRJ\|51250_IND\|51300_IND\|9148_IND |
| chr01 | 99071 | 102724 | 25901_TRJ\|328_TRJ\|43397_TRJ\|43675_TRJ\|nivara_89215 |
| chr01 | 179710 | 180096 | _ARO\|43397_TRJ\|51250_IND\|51300_IND\|6307_AUS\|6513_III\|8231_IND\|8555_AUS\|9148_IND\|nivara_106105\|nivara_106154\|rufipogon_105958\|rufipogon_P46\|rufipogon_YJ |
| chr01 | 271701 | 271944 | 27762_IND\|30416_IND\|38698_TRJ\|38994_ARO\|6307_AUS\|nivara_105327 |
| chr01 | 274646 | 281397 | 27762_IND\|6307_AUS\|8555_AUS\|nivara_105327\|rufipogon_YJ |
| chr01 | 300386 | 300652 | '8_TRJ\|38994_ARO\|43545_IND\|51250_IND\|51300_IND\|60542_IV\|6307_AUS\|8555_AUS\|9148_IND\|nivara_105327\|nivara_106154\|nivara_89215\|rufipogon_YJ |
| chr01 | 306979 | 307066 | 30416_IND\|38994_ARO\|8231_IND\|nivara_80470\|rufipogon_105958 |
| chr01 | 312281 | 313581 | 27630_TEJ\|38698_TRJ\|51300_IND\|8555_AUS\|9148_IND\|nivara_105327\|nivara_106154\|rufipogon_105960 |
| chr01 | 395858 | 396094 | 38994_ARO\|60542_IV\|6307_AUS\|8231_IND\|8555_AUS\|9148_IND\|nivara_106105\|nivara_80470\|nivara_89215\|rufipogon_105958 |
| chr01 | 397133 | 398137 | J\|30416_IND\|43397_TRJ\|43545_IND\|51250_IND\|51300_IND\|6307_AUS\|8231_IND\|8555_AUS\|nivara_106105\|nivara_106154\|nivara_80470\|nivara_89215 |
| chr01 | 399756 | 400010 | 'J\|43545_IND\|51250_IND\|60542_IV\|6307_AUS\|8231_IND\|8555_AUS\|nivara_105327\|nivara_89215\|rufipogon_105958\|rufipogon_105960\|rufipogon_P46\|rufipogon_YJ |
| chr01 | 406995 | 407242 | 25901_TRJ\|30416_IND\|38994_ARO\|51250_IND\|51300_IND\|nivara_105327\|nivara_80470\|rufipogon_YJ |
| chr01 | 447184 | 447467 | 25901_TRJ\|27762_IND\|38698_TRJ\|43397_TRJ\|43545_IND\|51250_IND\|51300_IND\|6307_AUS\|8555_AUS\|nivara_105327\|nivara_106154 |
| chr01 | 464114 | 476981 | 27630_TEJ\|30416_IND\|43397_TRJ\|43545_IND\|51300_IND\|8231_IND\|8555_AUS\|nivara_106154\|nivara_80470\|nivara_89215 |
| chr01 | 480889 | 480990 | 25901_TRJ\|38698_TRJ\|43545_IND\|6307_AUS\|rufipogon_105960 |
| chr01 | 506313 | 507690 | 25901_TRJ\|43545_IND\|8231_IND\|nivara_106154 |
| chr01 | 580114 | 580461 | _IND\|51300_IND\|60542_IV\|6307_AUS\|8231_IND\|8555_AUS\|9177_IND\|nivara_105327\|nivara_106105\|nivara_106154\|nivara_80470\|nivara_89215\|rufipogon_105958\|rufipogon_N |
| chr01 | 624515 | 624762 | 30416_IND\|38994_ARO\|8231_IND\|nivara_106105\|nivara_80470 |
| chr01 | 659795 | 670473 | 60542_IV\|6513_III\|nivara_105327\|rufipogon_105958\|rufipogon_105960\|rufipogon_YJ |
| chr01 | 677425 | 684193 | '57_TRJ\|25901_TRJ\|26872_IND\|43397_TRJ\|51250_IND\|51300_IND\|8244_TRJ\|8555_AUS\|9177_IND\|nivara_106105\|nivara_106154\|nivara_80470 |
| chr01 | 714335 | 715796 | 30416_IND\|8231_IND\|8555_AUS\|nivara_89215 |
| chr01 | 725788 | 726481 | _IND\|31856_V\|38698_TRJ\|418_TEJ\|45975_AUS\|55471_TEJ\|60542_IV\|6307_AUS\|8555_AUS\|9148_IND\|9177_IND\|nivara_89215\|rufipogon_Nepal\|rufipogon_YJ |
| chr01 | 748801 | 754800 | 27630_TEJ\|328_TRJ\|43325_TRJ\|6307_AUS\|8555_AUS |
| chr01 | 749909 | 751046 | 27762_IND\|328_TRJ\|418_TEJ\|43397_TRJ\|9177_IND\|nivara_106105\|nivara_106154 |
| chr01 | 774952 | 777127 | 17757_TRJ\|25901_TRJ\|26872_IND\|43325_TRJ\|43675_TRJ\|51300_IND\|8244_TRJ |

overview of CNV in samples | CNV detail information | **CNV in population** | CNV gene annotation

*Figure 3.1: Presence of CNV in Specific Cultivar*

*Table 3.2: Description of Cultivar Classification Sample Data*

| **Attribute Name** | **Description** |
|---|---|
| Group | Name of group and cultivar which is present in group |
| Sample size | Size of sample |



| Group | Sample size |
|---|---|
| Total | 50 |
| Cultiva total | 37 |
| *Indica* | 13 |
| 9177_IND | 1 |
| 30416_IND | 1 |
| 27762_IND | 1 |
| 8231_IND | 1 |
| 9148_IND | 1 |
| 26872_IND | 1 |
| 43545_IND | 1 |
| 51250_IND | 1 |
| 51300_IND | 1 |
| 12883_AUS | 1 |
| 45975_AUS | 1 |
| 8555_AUS | 1 |
| 6307_AUS | 1 |
| *Japonica* | 24 |
| 27630_TEJ | 1 |
| 32399_TEJ | 1 |
| 2540_TEJ | 1 |
| 55471_TEJ | 1 |

*Figure 3.2: Cultivar Classification*

Figure 3.3 mentions the information related to structural variation, CNV Start position, CNV end position, mechanism used, frequency, CDS-overlap gene and non-CDS overlap gene with respect to chromosome number and the description of attributes is given in Table 3.3.

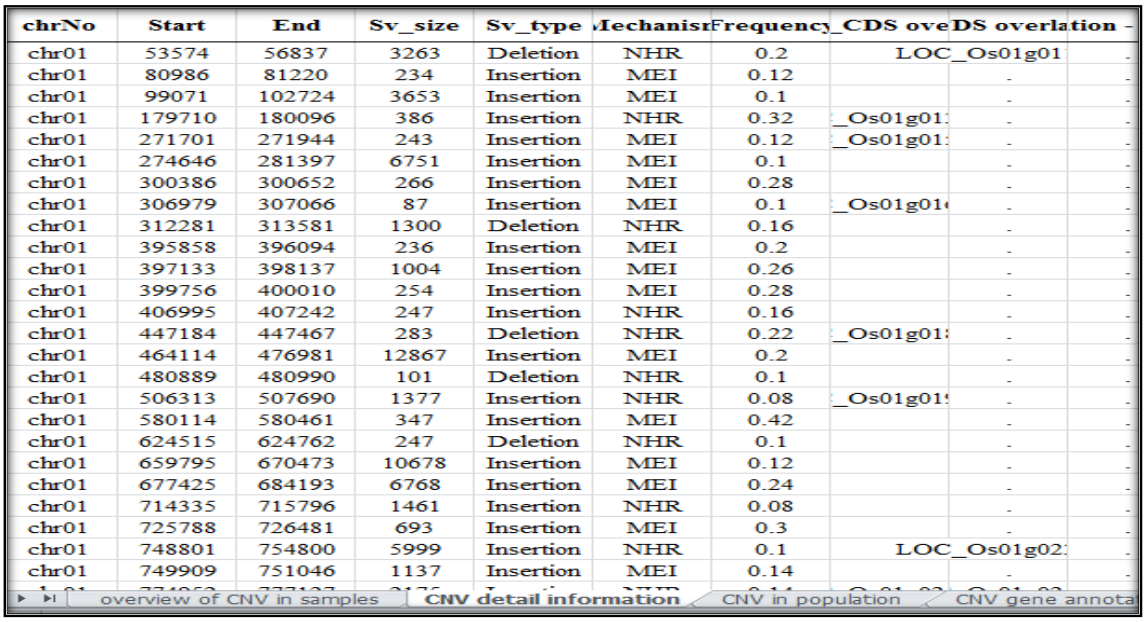
*Table 3.3: Description of Structural Variation and Frequency Sample Data*

| **Attribute Name** | **Description** |
|---|---|
| ChrNo | Chromosome number |
| Start | Starting position of CNV |
| End | Ending position of CNV |
| Sv_size | Size of structural variation |
| Sv_type | Type of structural variation |
| Mechanism | Name of Mechanism used in CNV detection |
| Frequency | Frequency of CNV |
| Non_CDS overlap | Non coding sequence overlap gene |
| CDS overlap | Coding sequence overlap gene |

Figure 3.4 mentioned the genes and their functions which can be integrated with the data mentioned in Figure 3.3. The description of attributes of gene sample data is shown in Table 3.4.

*Table 3.4: Description of Gene Sample Data*

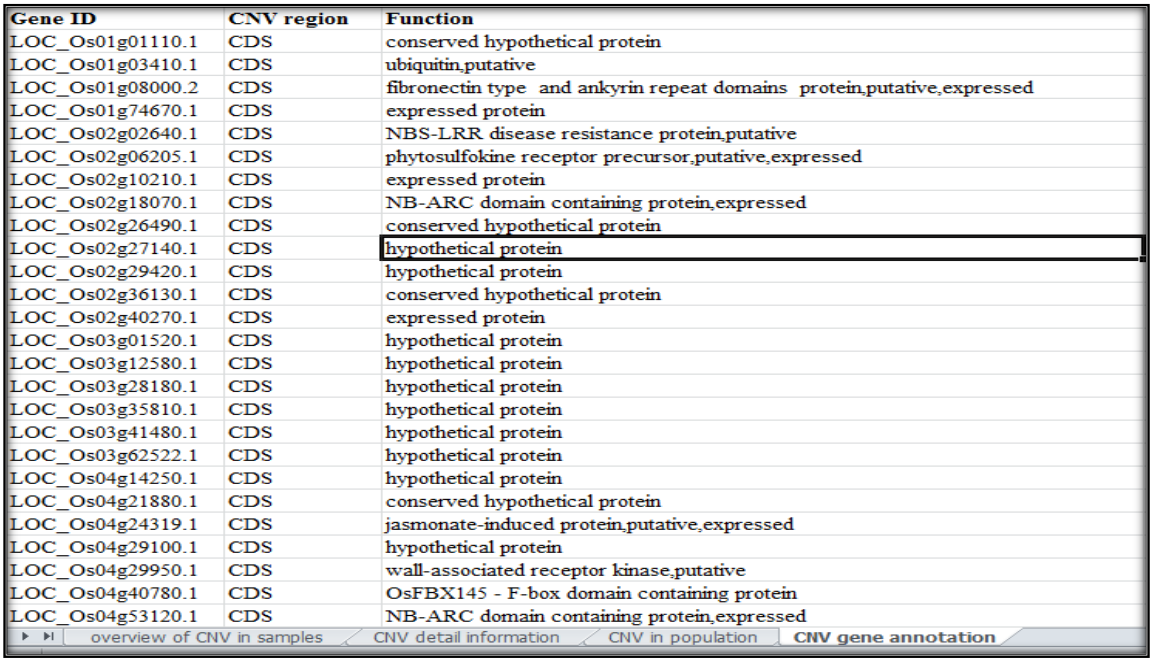
| **Attribute Name** | **Description** |
|---|---|
| Gene ID | Chromosome number |
| CNV region | Starting position of CNV |
| Function | Ending position of CNV |



| chrNo | Start | End | Sv_size | Sv_type | Mechanism | Frequency | _CDS ove | DS overlation |
|---|---|---|---|---|---|---|---|---|
| chr01 | 53574 | 56837 | 3263 | Deletion | NHR | 0.2 | LOC_Os01g01 | . |
| chr01 | 80986 | 81220 | 234 | Insertion | MEI | 0.12 | | . |
| chr01 | 99071 | 102724 | 3653 | Insertion | MEI | 0.1 | | . |
| chr01 | 179710 | 180096 | 386 | Insertion | NHR | 0.32 | _Os01g01 | . |
| chr01 | 271701 | 271944 | 243 | Insertion | MEI | 0.12 | _Os01g01 | . |
| chr01 | 274646 | 281397 | 6751 | Insertion | MEI | 0.1 | | . |
| chr01 | 300386 | 300652 | 266 | Insertion | MEI | 0.28 | | . |
| chr01 | 306979 | 307066 | 87 | Insertion | MEI | 0.1 | _Os01g01 | . |
| chr01 | 312281 | 313581 | 1300 | Deletion | NHR | 0.16 | | . |
| chr01 | 395858 | 396094 | 236 | Insertion | MEI | 0.2 | | . |
| chr01 | 397133 | 398137 | 1004 | Insertion | MEI | 0.26 | | . |
| chr01 | 399756 | 400010 | 254 | Insertion | MEI | 0.28 | | . |
| chr01 | 406995 | 407242 | 247 | Insertion | NHR | 0.16 | | . |
| chr01 | 447184 | 447467 | 283 | Deletion | NHR | 0.22 | _Os01g01 | . |
| chr01 | 464114 | 476981 | 12867 | Insertion | MEI | 0.2 | | . |
| chr01 | 480889 | 480990 | 101 | Deletion | NHR | 0.1 | | . |
| chr01 | 506313 | 507690 | 1377 | Insertion | NHR | 0.08 | _Os01g01 | . |
| chr01 | 580114 | 580461 | 347 | Insertion | MEI | 0.42 | | . |
| chr01 | 624515 | 624762 | 247 | Deletion | NHR | 0.1 | | . |
| chr01 | 659795 | 670473 | 10678 | Insertion | MEI | 0.12 | | . |
| chr01 | 677425 | 684193 | 6768 | Insertion | MEI | 0.24 | | . |
| chr01 | 714335 | 715796 | 1461 | Insertion | NHR | 0.08 | | . |
| chr01 | 725788 | 726481 | 693 | Insertion | MEI | 0.3 | | . |
| chr01 | 748801 | 754800 | 5999 | Insertion | NHR | 0.1 | LOC_Os01g02 | . |
| chr01 | 749909 | 751046 | 1137 | Insertion | MEI | 0.14 | | . |

overview of CNV in samples | CNV detail information | CNV in population | CNV gene annota

*Figure 3.3: Data of Structural Variation and CNV Frequency*

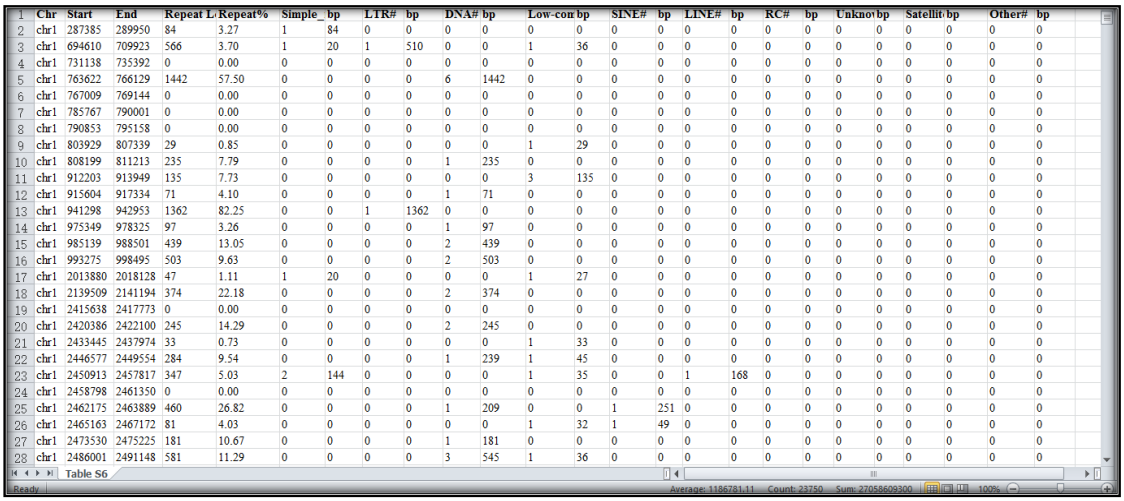


| Gene ID | CNV region | Function |
|---|---|---|
| LOC_Os01g01110.1 | CDS | conserved hypothetical protein |
| LOC_Os01g03410.1 | CDS | ubiquitin,putative |
| LOC_Os01g08000.2 | CDS | fibronectin type and ankyrin repeat domains protein,putative,expressed |
| LOC_Os01g74670.1 | CDS | expressed protein |
| LOC_Os02g02640.1 | CDS | NBS-LRR disease resistance protein,putative |
| LOC_Os02g06205.1 | CDS | phytosulfokine receptor precursor,putative,expressed |
| LOC_Os02g10210.1 | CDS | expressed protein |
| LOC_Os02g18070.1 | CDS | NB-ARC domain containing protein,expressed |
| LOC_Os02g26490.1 | CDS | conserved hypothetical protein |
| LOC_Os02g27140.1 | CDS | hypothetical protein |
| LOC_Os02g29420.1 | CDS | hypothetical protein |
| LOC_Os02g36130.1 | CDS | conserved hypothetical protein |
| LOC_Os02g40270.1 | CDS | expressed protein |
| LOC_Os03g01520.1 | CDS | hypothetical protein |
| LOC_Os03g12580.1 | CDS | hypothetical protein |
| LOC_Os03g28180.1 | CDS | hypothetical protein |
| LOC_Os03g35810.1 | CDS | hypothetical protein |
| LOC_Os03g41480.1 | CDS | hypothetical protein |
| LOC_Os03g62522.1 | CDS | hypothetical protein |
| LOC_Os04g14250.1 | CDS | hypothetical protein |
| LOC_Os04g21880.1 | CDS | conserved hypothetical protein |
| LOC_Os04g24319.1 | CDS | jasmonate-induced protein,putative,expressed |
| LOC_Os04g29100.1 | CDS | hypothetical protein |
| LOC_Os04g29950.1 | CDS | wall-associated receptor kinase,putative |
| LOC_Os04g40780.1 | CDS | OsFBX145 - F-box domain containing protein |
| LOC_Os04g53120.1 | CDS | NB-ARC domain containing protein,expressed |

overview of CNV in samples | CNV detail information | CNV in population | CNV gene annotation

*Figure 3.4: Data of Gene and Their Function*

In another data table related to chromosome, CNV start position, CNV end position, repeat type, number of repeats and their base pairs are presented in Figure 3.5. The detail of the attributes listed in Table 3.5.



| | Chr | Start | End | Repeat L | Repeat% | Simple_ | bp | LTR# | bp | DNA# | bp | Low-con | bp | SINE# | bp | LINE# | bp | RC# | bp | Unkno | bp | Satellit | bp | Other# | bp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | chr1 | 287385 | 289950 | 84 | 3.27 | 1 | 84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | chr1 | 694610 | 709923 | 566 | 3.70 | 1 | 20 | 1 | 510 | 0 | 0 | 1 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | chr1 | 731138 | 735392 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | chr1 | 763622 | 766129 | 1442 | 57.50 | 0 | 0 | 0 | 0 | 6 | 1442 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | chr1 | 767009 | 769144 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | chr1 | 785767 | 790001 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | chr1 | 790853 | 795158 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | chr1 | 803929 | 807339 | 29 | 0.85 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | chr1 | 808199 | 811213 | 235 | 7.79 | 0 | 0 | 0 | 0 | 1 | 235 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | chr1 | 912203 | 913949 | 135 | 7.73 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 135 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | chr1 | 915604 | 917334 | 71 | 4.10 | 0 | 0 | 0 | 0 | 1 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | chr1 | 941298 | 942953 | 1362 | 82.25 | 0 | 0 | 1 | 1362 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | chr1 | 975349 | 978325 | 97 | 3.26 | 0 | 0 | 0 | 0 | 1 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | chr1 | 985139 | 988501 | 439 | 13.05 | 0 | 0 | 0 | 0 | 2 | 439 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | chr1 | 993275 | 998495 | 503 | 9.63 | 0 | 0 | 0 | 0 | 2 | 503 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | chr1 | 2013880 | 2018128 | 47 | 1.11 | 1 | 20 | 0 | 0 | 0 | 0 | 1 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | chr1 | 2139509 | 2141194 | 374 | 22.18 | 0 | 0 | 0 | 0 | 2 | 374 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | chr1 | 2415638 | 2417773 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | chr1 | 2420386 | 2422100 | 245 | 14.29 | 0 | 0 | 0 | 0 | 2 | 245 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | chr1 | 2433445 | 2437974 | 33 | 0.73 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | chr1 | 2446577 | 2449554 | 284 | 9.54 | 0 | 0 | 0 | 0 | 1 | 239 | 1 | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | chr1 | 2450913 | 2457817 | 347 | 5.03 | 2 | 144 | 0 | 0 | 0 | 0 | 1 | 35 | 0 | 0 | 1 | 168 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | chr1 | 2458798 | 2461350 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | chr1 | 2462175 | 2463889 | 460 | 26.82 | 0 | 0 | 0 | 0 | 1 | 209 | 0 | 0 | 1 | 251 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | chr1 | 2465163 | 2467172 | 81 | 4.03 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 32 | 1 | 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | chr1 | 2473530 | 2475225 | 181 | 10.67 | 0 | 0 | 0 | 0 | 1 | 181 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | chr1 | 2486001 | 2491148 | 581 | 11.29 | 0 | 0 | 0 | 0 | 3 | 545 | 1 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table S6

Ready Average: 1186781.11 Count: 23750 Sum: 27058609300 100%

*Figure 3.5 Sample Data of Repeat Type in CNV.*

## 3.2 System Architecture

The Data Integration tool is used to create Data mart for CNV in the dimensional models. The Extract, Transform and Load (ETL) process is need to be implemented for CNV data. After designing the dimensional modeling, data are populated using ETL process and MySQL is used as storing place for data. The Pentaho Schema Workbench (PSW) is used for designing the schema of an OLAP cube. The output of the PSW is input for the Pentaho Business Intelligence (BI) Server. This BI server is used for OLAP visualization. The architecture is shown in Figure 3.6.

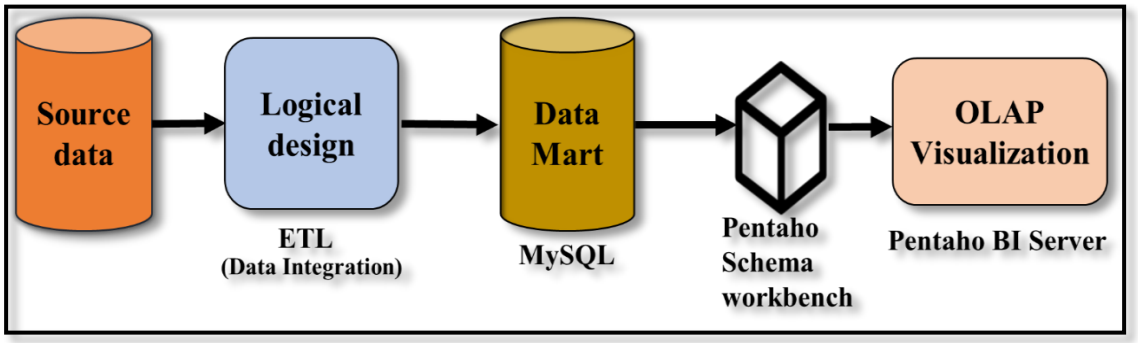



*Figure 3.6: System Architecture*

*Table 3.5: Description of Base pair Sample Data*

| **Attribute Name** | **Description** |
|---|---|
| Chr | Chromosome number |
| Start | Starting position of CNV |
| End | Ending position of CNV |
| Repeat length | Repeat length of CNV |
| Repeat % | Repeat percent |
| Simple_repeat# | Simple repeat number |
| LTR# | Long terminal repeat number |
| DNA# | DNA repeat number |
| Low_complexity | Low complexity repeat number |
| SINE# | Short interspersed nuclear element number |
| LINE# | Long interspersed nuclear element number |
| RC# | Read count repeat number |
| Unknown# | Unknown repeat number |
| Satellite# | Satellite repeat number |
| Other# | Other repeat number |
| BP | Base pair size |

## 3.3 Tools and Technologies

Several tools and technologies have been applied in the development of the complete Data mart. These are discussed as follows:

### 3.3.1 Pentaho Data Integration and Business Analytics Platform

This platform offers various set of open source products that are suitable to develop and implement data mart. This mainly provide data integration, OLAP cube development, OLAP exploration, interactive report generation, dashboards with advance capability of ETL (extraction, transformation, loading) process.

#### 3.3.1.1 Pentaho Data Integration (PDI)

Pentaho Data Integration (PDI) called kettle and now it is known as Spoon. PDI is responsible for creation of ETL processes. PDI offers many functionalities of data integration and cleansing. Some of these are listed below:

- Migrating data between different databases or applications
- Exporting data from various proprietary databases to and from flat files
- Loading data massively into databases
- Data cleansing
- Integrating applications

PDI is an easy to use and having a rich set of graphical tools. PDI is a meta data oriented tool. It refers to development of workflow or pipelines by using different subset of tools/components. This requires the select, drag, drop and connect through connecting arrows. These workflows are reusable and can be further customized with support of ever increasing user's demand.

PDI can be utilized as a standalone application as well, or it can be applied as part of the larger Pentaho Suite. PDI supports a big data array of input and output formats, including CSV file, text files, data sheets, and commercial and non-commercial database engines. The user can build different types of jobs and transformations processes. PDI offers three methods to save these workflows as follows:

- **Files**: The Jobs and transformation are saved as .kjb and .ktr extension respectively.

- **Database repository**: The execution of PDI workflows are required to be first saved in repository.

- **PDI Repository**: This is available in enterprise edition.

#### 3.3.1.2 Schema Workbench

The Schema Workbench is also recognized as a Mondrian Schema Workbench. It lets user to graphically create and test OLAP cube schema. The schema is saved as xml files. This xml will be send to the web server which further provide the generation of various kinds of reports through web browser.

#### 3.3.1.3 Business Intelligence (BI) Server

The BI Server is a set of programs that was invoked by Java servlets. This server will function as a web server and provide the web browser based interface in browsing and exploring various reports.
The BI server contains three layers:

- **The platform**: The assembling of components collectively known as the platform
- **BI components**: The platform forms the base for a number of ingredients that provide typical business intelligence functionality.
- **The presentation layer**: The presentation is available as web based interface and called the user console. This allows the users to interact with data resides in server. This layer will supports interactive reports, dashboard, OLAP exploration and many more reporting tools as BI content. The user console supports multiple dashboards, OLAP, and interactive reports simultaneously.

### 3.3.2 Apache Tomcat

Apache tomcat is an open source java servlet container that also called as tomcat web server. This web server is developed and maintained by the Apache Software Foundation (ASF) and it is available under open source license. Default port is used to run the web requests through HTTP connector through 8080.

### 3.3.3 MySQL and MySQL Workbench

MySQL is the most popular open source database software and now owned by Oracle. MySQL works support most of the operating systems. MySQL strongly supports data warehouse with major BI/ETL vendors. MySQL Workbench is a visual interface for MySQL database that integrates database creation, development, administration, design and maintenance.

# CHAPTER IV

# RESULTS

## 4.1 Introduction

Copy number variation is an important type of genetic variant. The information of CNV is used to develop improved and stress resistant varieties. To provide this data at the single platform, CNV mart has been developed. The ETL process of development of data mart has been implemented in this section.

## 4.2 Dimensional Modelling

Dimensional modelling helps to identify the fact and dimensions. Three fact tables have been identified based on the selected datasets. These fact tables are containing the occurrence of CNVs in various cultivars, frequency and associating genes. The dimensions are identified as classification hirerchy of cultivars consists of different groups/sub groups and types, CNV regions, CNV types, CDS overlaps types etc. These are further described in the subsequent sections of this chapter.

## 4.3 Cultivar Information

The cultivar information is shown in Figure 3.1 and 3.2 (chapter 3). The star schema for cultivar is shown in Figure 4.1. The tabular structure of these dimensions *viz.* chromosome, CNV and cultivar are described in Table 4.1, Table 4.2 and Table 4.3 respectively.

*Table 4.1: Description of Chromosome Dimension*

| **DIMENSION: CHROMOSOME** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Chromosome_id | Chromosome id |
| 2 | ChroNo | Chromosome number |
| 3 | Start_pos | Starting position of Chromosome |
| 4 | End_pos | Ending position of chromosome |

The source data related to cultivar contained in two tables. The first table represents main group, group, cultivar type and cultivar as single column in the excel sheet. The second table contains the cultivar name and CNV position. The ETL process is implemented to combine these two tables and a hierarchy has also been created to implement the cultivar types. The ETL process for development of Data mart for cultivar information requires following steps:

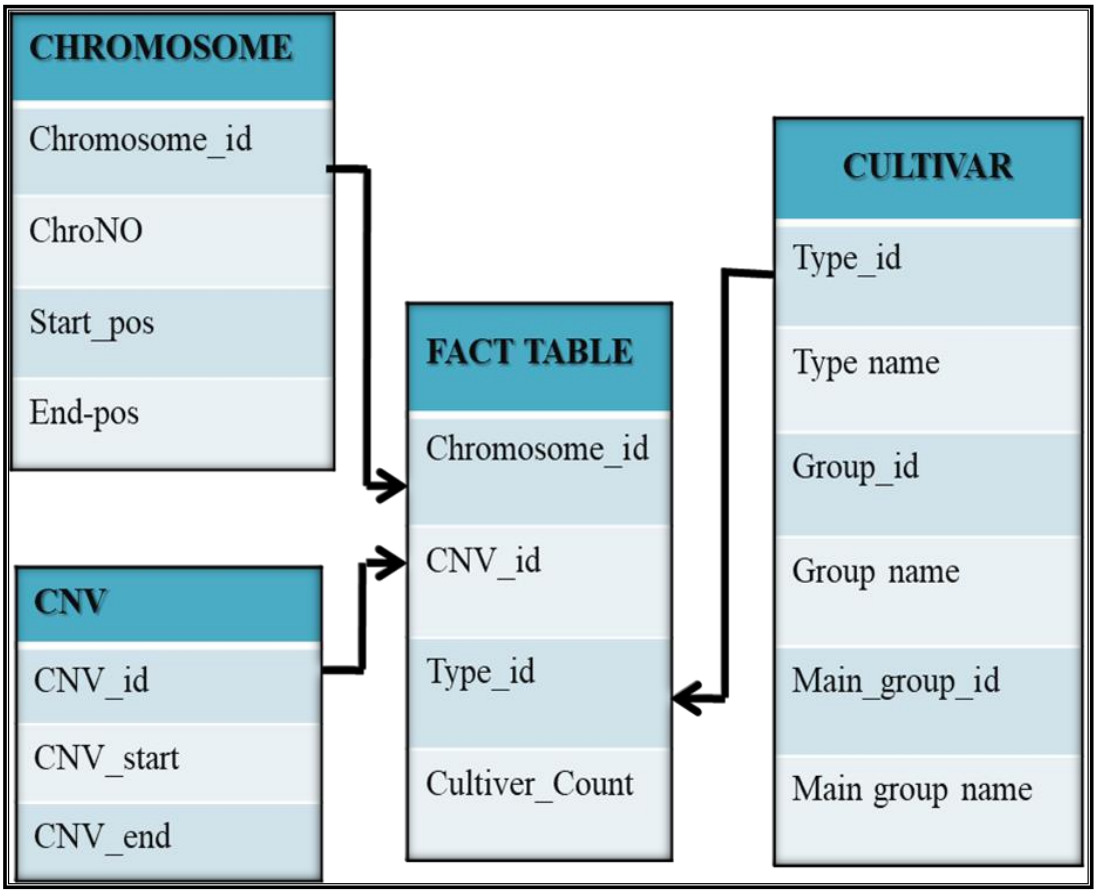



*Figure 4.1: Dimensional Modelling for Cultivar Information*

The fact cultivar count is presented in Table 4.4. The ETL process has been developed to implement this schema using Pendaho data integration tool.

*Table 4.2: Description of CNV Dimension*

| **DIMENSION: CNV** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | CNV_id | CNV id |
| 2 | CNV_start | Starting position of CNV |
| 3 | CNV_end | Ending position of CNV |

*Table 4.3: Description of Cultivar Dimension*

| **DIMENSION: CULTIVAR** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Type_id | Cultivar type id |
| 2 | Type name | Name of cultivar's type |
| 3 | Group_id | Group id |
| 4 | Group name | Name of cultivar's group |
| 5 | Main_group_id | Main group id |
| 6 | Main group name | Name of cultivar's main group |

*Table 4.4: Description of Fact Cultivar Count*

| **FACT: CULTIVAR_COUNT** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Chromosome_id | Chromosome id ( Primary key of chromosome dimension) |
| 2 | CNV_id | CNV id ( Primary key of CNV dimension) |
| 3 | Type_id | Type id ( Primary key of cultivar dimension) |
| 4 | Cultiver_Count | Count of cultivar (Fact) |

- Open MySQL and create schema for storage of mart as shown in Figure 4.2.

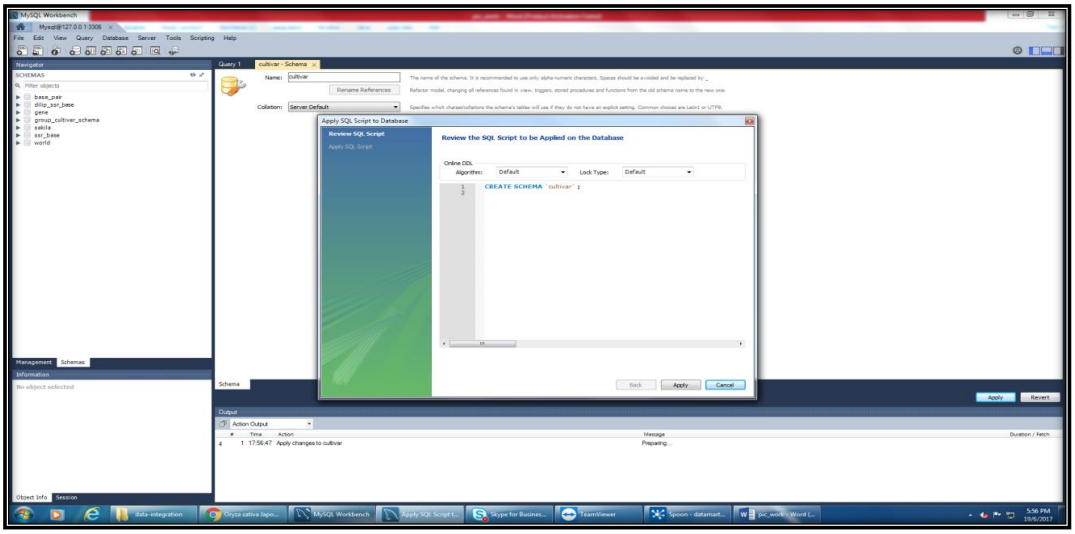


*Figure 4.2: Schema Creation*

- Open pentaho data integration tool by executing spoon.bat file.
- Open the new transformation to implement the ETL process.
- Create a connection with MySQL schema as shown in Figure 4.3.

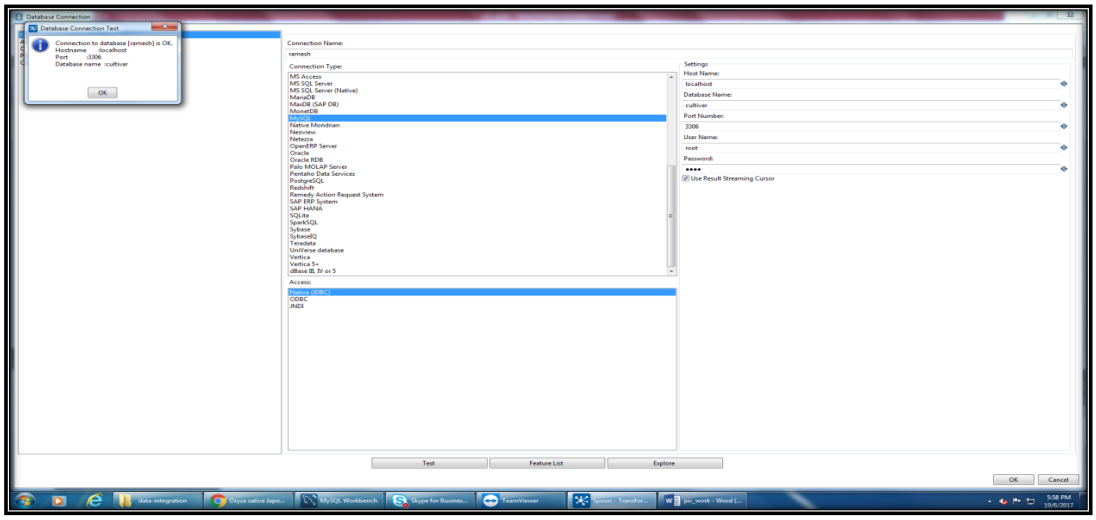


*Figure 4.3. Database Connection in Spoon*

- Take **Input** step to retrieve the contents from MS-Excel worksheet.
- Add a **Filter rows** to filter the rows. Cultivar types are extracted from input as presented in Figure 4.4.

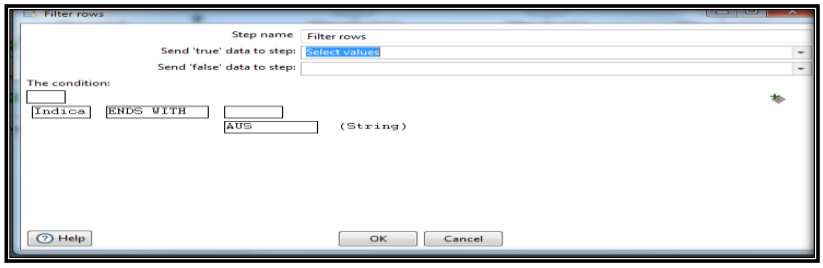



*Figure 4.4: Filter Rows*

- The step "**Select value**" is used to select a distinct field. The selected field i.e, *indica* is renamed to name_species as shown in Figure 4.5 as a part of transformation process.

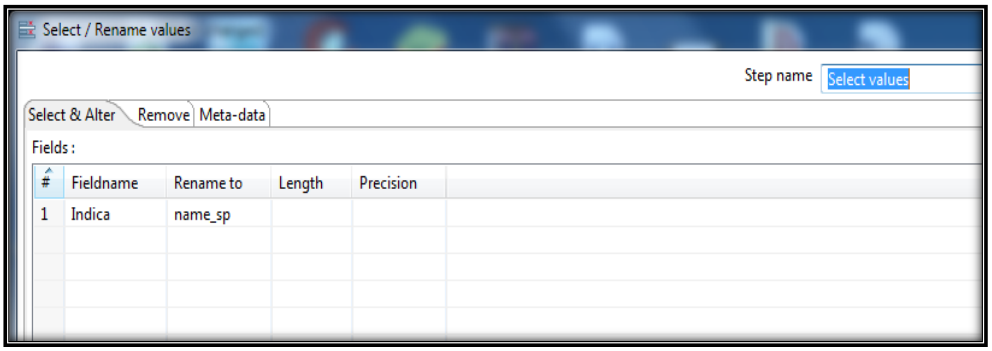



*Figure 4.5: Select/ Rename Values*

- The step "**Add constant**" used to create a new field *i.e.*, a field group_name. This field contained the values *viz.* indica, japonica, Oryza sativa group III, Oryza sativa group IV, Oryza sativa group V, O.rufipogon, and O.nivara for indica to create the classification as hirerchy groups and main groups (Figure 4.6).

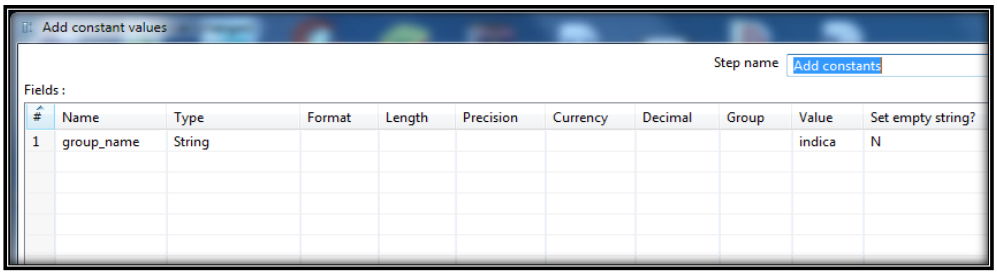



*Figure 4.6: Add Constant Values*

- To merge different source primary condition, data should be sorted. For sorting the data, **sort row** was used before merge join. Sorting of rows is done according to group name as shown in Figure 4.7.

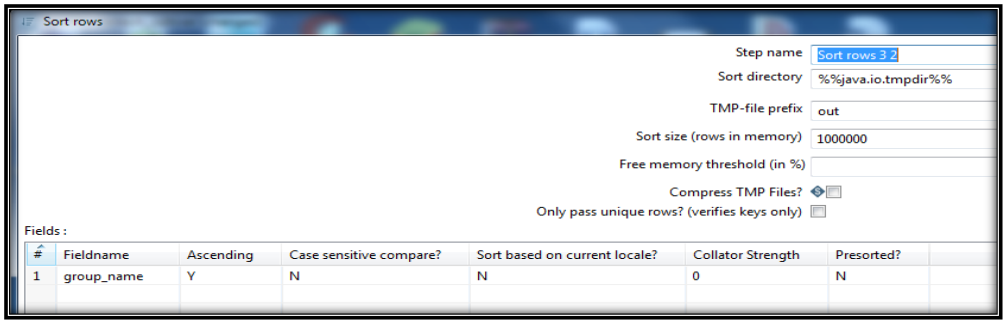



*Figure 4.7: Sort Rows*

- Add **Merge join** to join different extract field in single field i.e., extracted group names are merged in single field with name group_name as presented in Figure 4.8.

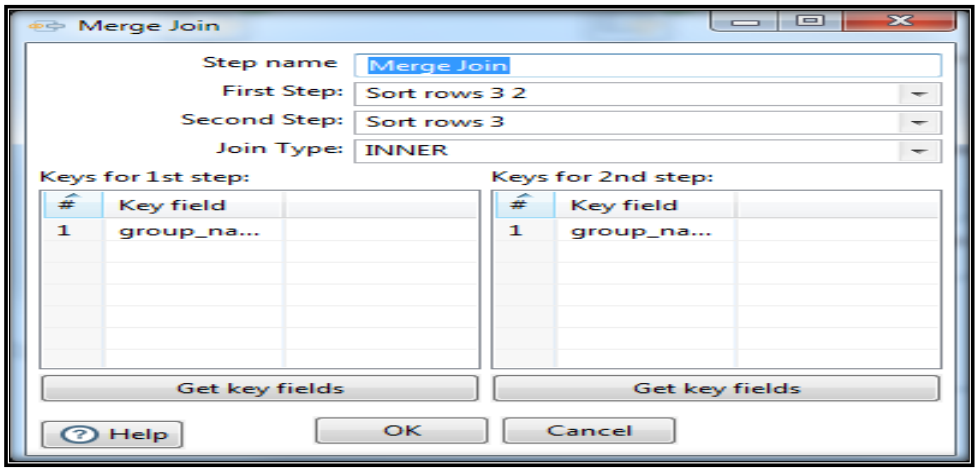



*Figure 4.8: Merge Join*

- The **Table output** step used to store the output of process in a single table which is shown in Figure 4.9.

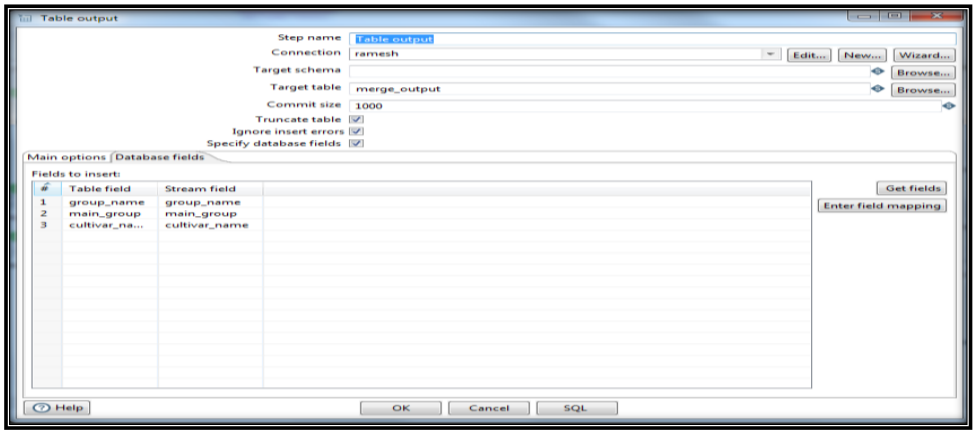



*Figure 4.9: Table Output*

- Above listed all step are shown in Figure 4.10 to understand the complete workflow.

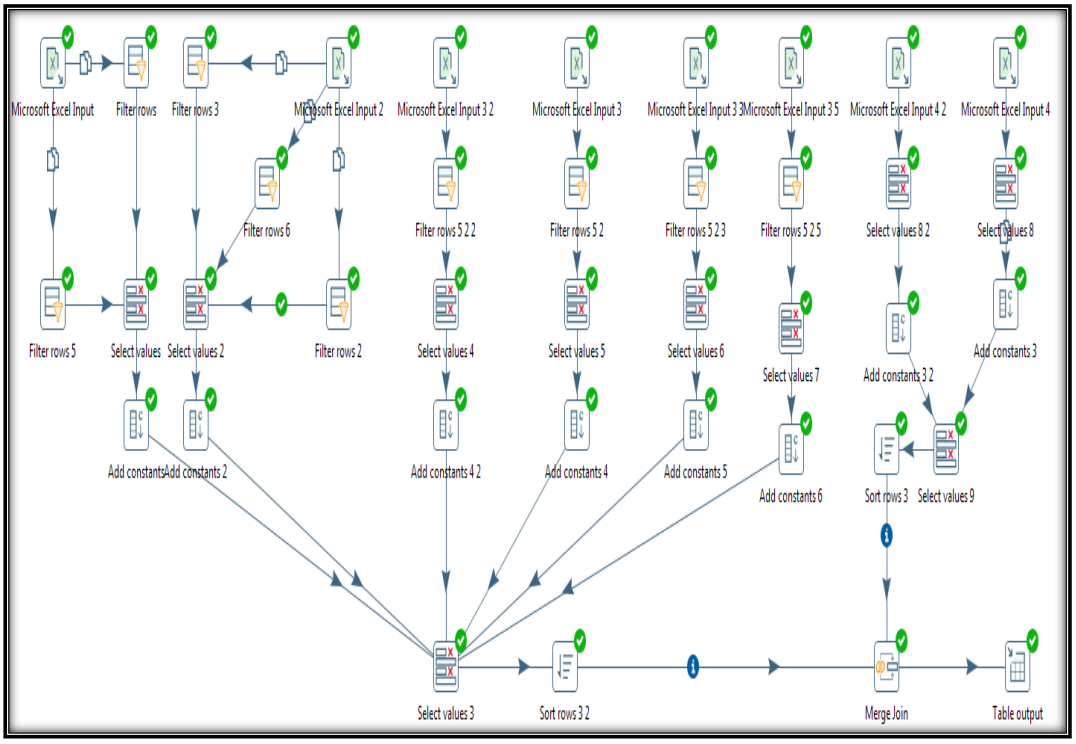



*Figure 4.10: Workflow for Creating The Hierarchy Of Cultivar*

- Take table output of the above menstioned transformation steps as input for further transformation.
- Add **Split field** to rows to split the field in more than one rows according to delimeter position as shown in Figure 4.11.

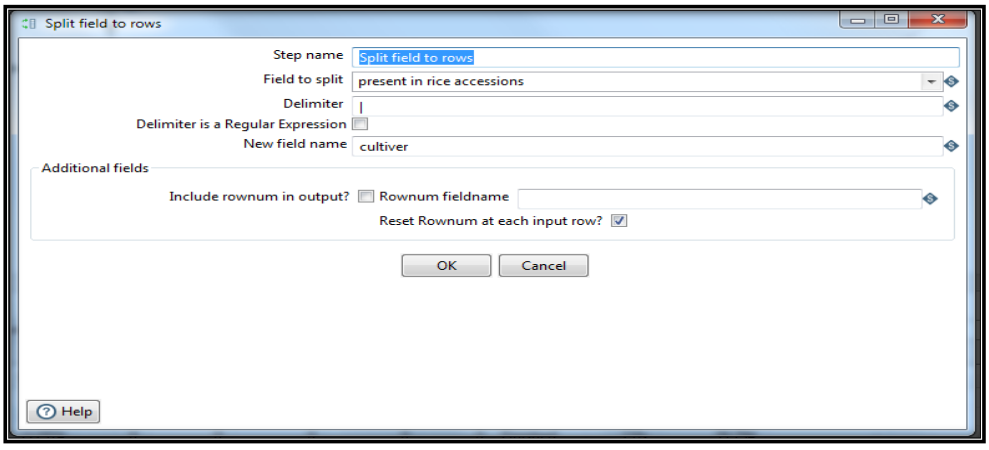



*Figure 4.11: Split Field*

- Add **Calculator** step to create a copy of field i.e., a new field caltivar copy is created which is replica of cultivar field as shown in Figure 4.12.

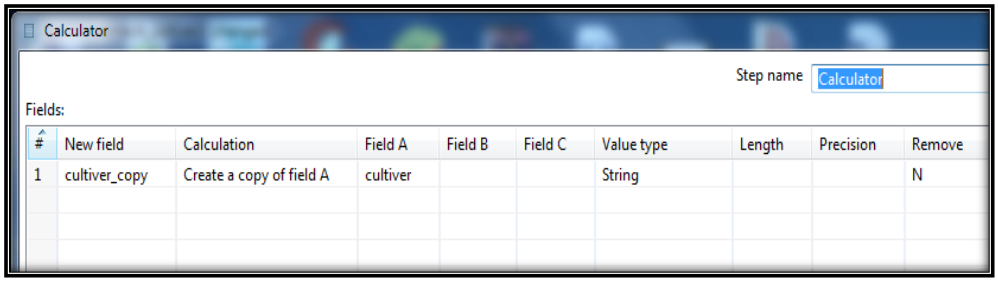



*Figure 4.12: Calculator*

- Add Split **field step** to split the cultivar copy field into cultivar type and cultivar number.
- **Combination lookup/update** is used to create chromosome dimension table as presented in Figure. Similarly, it also used for other dimension table like CNV and cultivar tables.

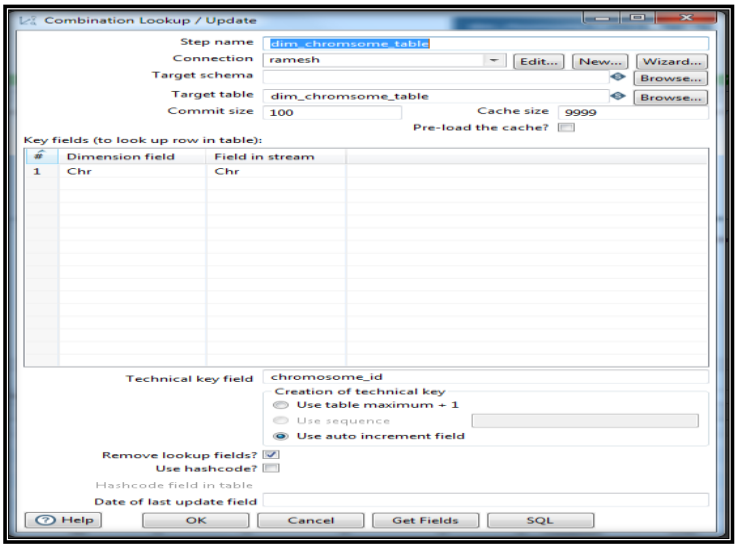



*Figure 4.13: Combination Lookup/ Update*

- Add **Table output** step for storing the output of fact table as in Figure 4.14.

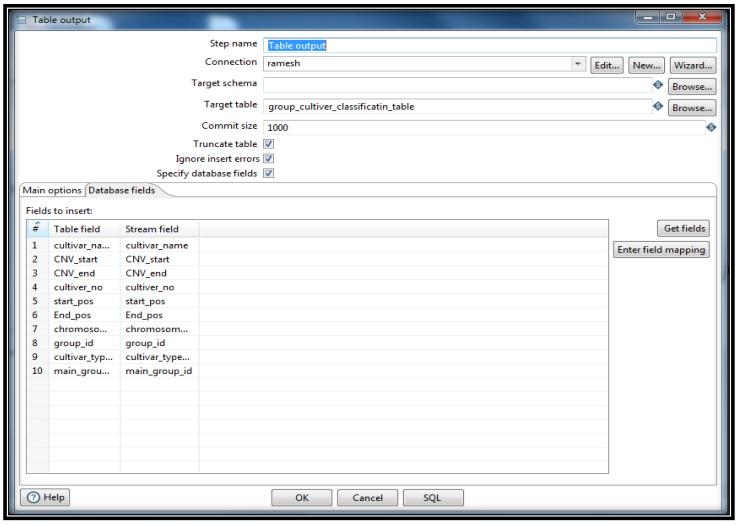



Figure 4.14: Table Output

- To load the Data into various dimension and fact tables, the workflow need to be executed.
- The final steps of transformation is summarized in Figure 4.15.

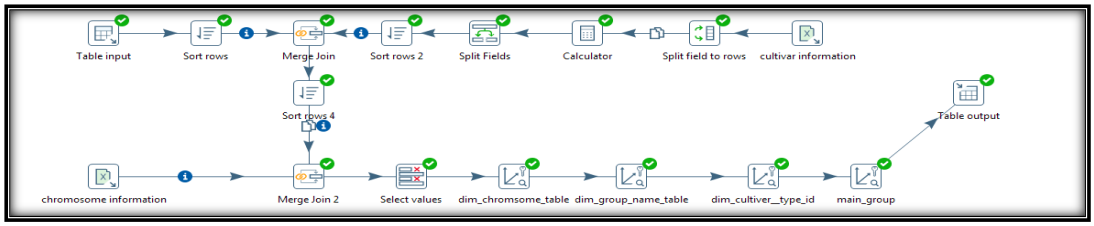



*Figure 4.15: ETL Process for Cultivar Information*

After executing this workflow, the fact table and dimension tables for cultivar information have been created. In the fact table i.e., cultivar information are having three dimensions viz. chromosome, CNV, and cultivar which are presented in Figure 4.16.

| cultivar_name | chromosome_id | group_id | cultivar_type_id | main_group_id | CNV_start | CNV_end | cultiver_no | start_pos | End_pos |
|---|---|---|---|---|---|---|---|---|---|
| 11010_TRJ | 1 | 1 | 1 | 1 | 725788 | 726481 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 870240 | 870851 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 936289 | 936625 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1032184 | 1032405 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1523116 | 1530026 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1573884 | 1574074 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1578832 | 1593637 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1602060 | 1602212 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1607752 | 1616733 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1645769 | 1646803 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1707765 | 1708919 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1793088 | 1793454 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1858583 | 1859764 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1871579 | 1871933 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1950744 | 1950969 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1960318 | 1960580 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1968186 | 1969800 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 1972017 | 1972644 | 11010 | 1 | 43270923 |
| 11010_TRJ | 1 | 1 | 1 | 1 | 2104495 | 2104868 | 11010 | 1 | 43270923 |

*Figure 4.16: Fcat Table of Cultivar Information*

### 4.4 CNV Association with Genes

For gene information in Figure 3.3 and 3.4 (Chapter 3), the star schema for gene is shown in Figure 4.17. The tabular structure of these dimensions *viz.* chromosome, gene id, CNV, CDS type, mechanism, structural variation, function, CNV region are described in Table 4.5, Table 4.6, Table 4.7, Table 4.8, Table 4.9, Table 4.10, Table 4.11 and Table 4.12 respectively. The facts count of gene and sum of frequency are presented in Table 4.13.

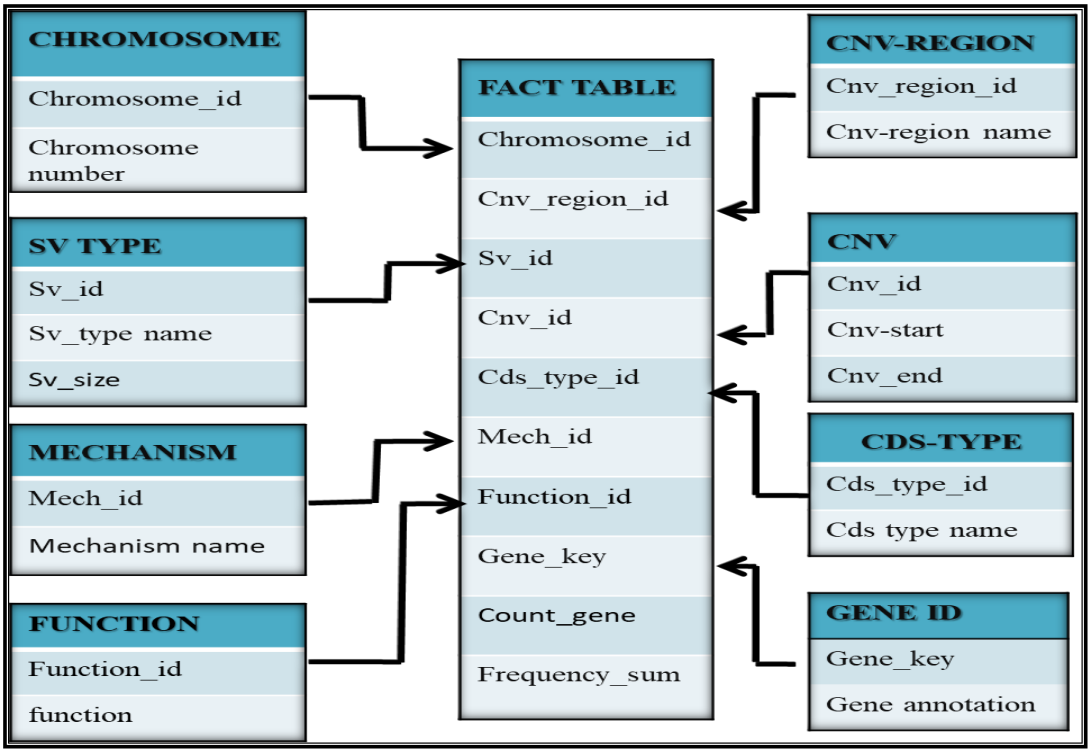



*Figure 4.17: Dimensional Modeling of Gene Information*

*Table 4.5: Description of Chromosome Dimension*

| **DIMENSION: CHROMOSOME** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Chromosome_id | Chromosome id |
| 2 | Chromosome number | Chromosome number |

*Table 4.6: Description of CNV Dimension*

| **DIMENSION: CNV** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | CNV_id | CNV id |
| 2 | CNV_start | Starting position of CNV |
| 3 | CNV_end | Ending position of CNV |

*Table 4.7: Description of Structural Variation Type Dimension*

| **DIMENSION: STRUCTURAL VARIATION TYPE** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Sv_id | Structural variation id |
| 2 | Sv_type | Name of structural variation |
| 3 | Sv_size | Size of sequence responsible for specific |

*Table 4.8: Description of Mechanism Dimension*

| **DIMENSION: MECHANISM** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Mech_id | Mechanism id |
| 2 | Mechanism name | CNV formation mechanisms |

*Table 4.9: Description of Gene Id Dimension*

| **DIMENSION: GENE ID** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Gene_key | Primary key of Gene |
| 2 | Gene_annotation | Gene annotation |

*Table 4.10: Description of Cds-type Dimension*

| **DIMENSION: CDS-TYPE** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Cds_type_id | Cds type id |
| 2 | Cds type name | Specific cds category name |

*Table 4.11: Description of CNV-region Dimension*

| **DIMENSION: CNV-REGION** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Cnv_region_id | CNV region id |
| 2 | Cnv-region name | Name of specific CNV region |

*Table 4.12: Description of Function Dimension*

| **DIMENSION: FUNCTION** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Function_id | Function id |
| 2 | Function name | Function of gene |

The ETL process for development of Data mart for gene information have following steps:

- **Microsoft execl input** step has taken as input for gene information which is shown in Figure 3.3.
- Now take **Split fields** into rows to split the CDS-overlap according to delimeter position into rows with new field name gene.
- **Table output** is used to store the output of process in MySQL.
- Above mentioned steps are shown in Figure 4.18. This flow helps the modeler to set this output as input for the remaining step of transformation process.

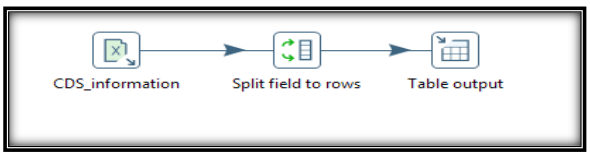



*Figure 4.18:Spliting the fields*

*Table 4.13: Description of Fact Count of Gene and Sum of Frequency*

| **FACT: COUNT_GENE & FREQUENCY_SUM** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Chromosome_id | Chromosome id ( Primary key of chromosome dimension) |
| 2 | CNV_id | CNV id ( Primary key of CNV dimension) |
| 3 | Sv_id | Sv id ( Primary key of structural variation dimension) |
| 4 | Cds_type_id | Cds type id CNV id ( Primary key of Cds type dimension) |
| 5 | Mech_id | Mechanism id ( Primary key of mechanism dimension) |
| 6 | Cnv_region_id | CNV region id ( Primary key of CNV region dimension) |
| 7 | Function_id | function id ( Primary key of function dimension) |
| 8 | Gene_key | Gene key ( Primary key of gene id dimension) |
| 9 | Count_gene | Count of gene (Fact) |
| 10 | Frequency_sum | Sum of frequency (Fact) |

- Add **Input table** step to access the output of previous step.
- Now add **Select values** to select value according to CDS or non-CDS overlap.
- Add **Add constant** to add new field cds_type and their values viz. CDS-overlap and CDS-non overlap.

- Add **Select values** to remove unwanted field of information and to rename the field gene as Gene_id.
- Now add **Filter rows** to remove the rows which have no gene information.
- Add **Sort rows** to sort the rows according to Gene_id.
- **Microsoft execl input** step has taken as input for gene information which is shown in Figure 3.4.
- Now, add **Calculator** to create a copy of Gene_id field with name gene_annotation.
- Add **Sort rows** to sort rows according to Gene_id.
- Add **Merge join** to merge the both tables based on Gene_id.
- Now add **Select value** to remove unwanted information.
- Add **Combination lookup/update** to create chromosome dimension Table and similarly it also used for other dimension Table like CNV, gene, cds_type, mechanism, sv_type, function, cnv_region Tables.
- Now add **Table output** step used to store the output of process in the Data mart.
- To load the output of transformation into Data mart, there is need to execute and run the developed workflow.
- Final transformation is summarized in Figure 4.19.

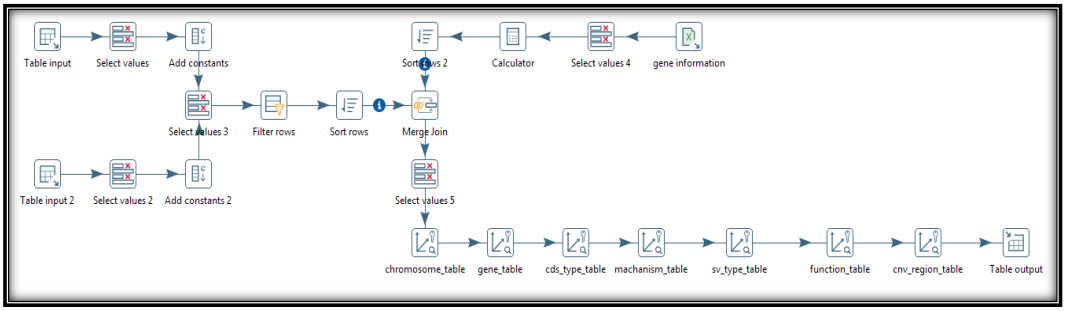



*Figure 4.19: ETL Process for Gene Information*

After performing the above steps, the fact and dimension tables related to gene information has been created and are available in MySQL for further use. The Data mart of gene information consists of eight dimensions viz. chromosome, CNV, gene, CDS type, mechanism, structural variation, function, CNV region and measures count of gene and sum of frequency as shown in Figure 4.20.

| gene_annotation | Start | End | Sv_size | Frequency | chromsome | gene_key | cds_type_id | mech_id | sv_id | function_id | cnv_region_id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LOC_Os01g01110.1 | 53574 | 56837 | 3263 | 0.2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| LOC_Os01g01110.1 | 53574 | 56837 | 3263 | 0.2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| LOC_Os01g01360.1 | 179710 | 180096 | 386 | 0.32 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| LOC_Os01g01360.1 | 179710 | 180096 | 386 | 0.32 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| LOC_Os01g01520.1 | 271701 | 271944 | 243 | 0.12 | 1 | 3 | 2 | 2 | 2 | 3 | 3 |
| LOC_Os01g01520.1 | 271701 | 271944 | 243 | 0.12 | 1 | 3 | 2 | 2 | 2 | 3 | 3 |
| LOC_Os01g01610.1 | 306979 | 307066 | 87 | 0.1 | 1 | 4 | 2 | 2 | 2 | 4 | 3 |
| LOC_Os01g01610.1 | 306979 | 307066 | 87 | 0.1 | 1 | 4 | 2 | 2 | 2 | 4 | 3 |
| LOC_Os01g01830.1 | 447184 | 447467 | 283 | 0.22 | 1 | 5 | 2 | 1 | 1 | 5 | 3 |
| LOC_Os01g01830.1 | 447184 | 447467 | 283 | 0.22 | 1 | 5 | 2 | 1 | 1 | 5 | 3 |
| LOC_Os01g01925.1 | 506313 | 507690 | 1377 | 0.08 | 1 | 6 | 2 | 1 | 2 | 6 | 3 |
| LOC_Os01g01925.1 | 506313 | 507690 | 1377 | 0.08 | 1 | 6 | 2 | 1 | 2 | 6 | 3 |
| LOC_Os01g02370.1 | 748801 | 754800 | 5999 | 0.1 | 1 | 7 | 1 | 1 | 2 | 7 | 4 |
| LOC_Os01g02370.1 | 748801 | 754800 | 5999 | 0.1 | 1 | 7 | 1 | 1 | 2 | 7 | 4 |
| LOC_Os01g02410.1 | 774952 | 777127 | 2175 | 0.14 | 1 | 8 | 1 | 1 | 2 | 8 | 1 |
| LOC_Os01g02410.1 | 774952 | 777127 | 2175 | 0.14 | 1 | 8 | 2 | 1 | 2 | 8 | 1 |
| LOC_Os01g02410.1 | 774952 | 777127 | 2175 | 0.14 | 1 | 8 | 1 | 1 | 2 | 8 | 1 |

*Figure 4.20: Data Mart of Gene Information*

### 4.5 CNV Types with Repect to Bp Frequency

For base-pair (bp) information shown in Figure 3.5 (chapter 3). The star schema for gene is shown in Figure 4.21. The tabular structure of these dimensions *viz.* chromosome, CNV and repeat are described in Table 4.14, Table 4.15 and Table 4.16 respectively. The fact basepair sum is presented in Table 4.17.

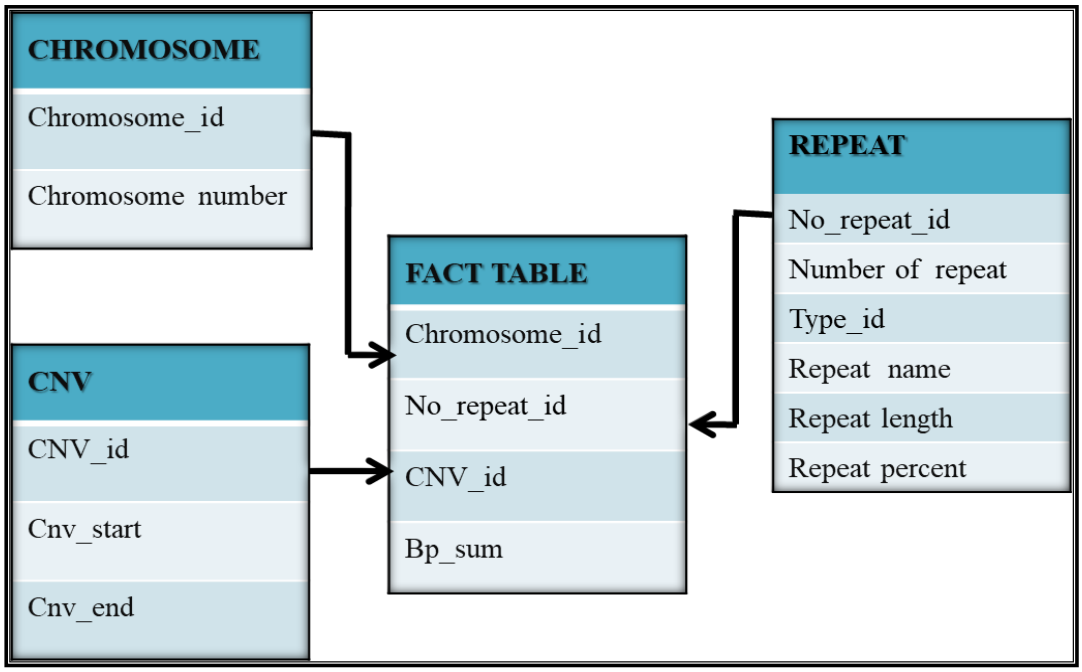



*Figure 4.21: Dimensional Modeling for Base-Pair Size*

*Table 4.14: Description of Chromosome Dimension*

| **DIMENSION: CHROMOSOME** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Chromosome_id | Chromosome id |
| 2 | Chromosome number | Chromosome number |

*Table 4.15: Description of CNV Dimension*

| **DIMENSION: CNV** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | CNV_id | CNV id |
| 2 | CNV_start | Starting position of CNV |
| 3 | CNV_end | Ending position of CNV |

*Table 4.16: Description of Repeat Dimension*

| **DIMENSION: REPEAT** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | No_repeat_id | Repeat number id |
| 2 | Number of repeat | Number of repeats in CNV |
| 3 | Type_id | Type of repeats |
| 4 | Repeat name | Name of repeat type |
| 5 | Repeat length | Repeat length in basepair |
| 6 | Repeat percent | Percentage of repeat in CNV |

*Table 4.17: Description of Fact Sum Of Basepair*

| **FACT: SUM OF BASEPAIR** | | |
|---|---|---|
| S.No. | **Dimension Attribute** | **Description** |
| 1 | Chromosome_id | Chromosome id ( Primary key of chromosome dimension) |
| 2 | CNV_id | CNV id ( Primary key of CNV dimension) |
| 3 | No_repeat_id | Number of repeats id ( Primary key of repeat dimension) |
| 4 | Bp_sum | Sum of basepair (Fact) |

The base-pair sizes corresponding to individual repeat type like simple repeat, LTR repeat, RC repeat etc and repeat number are present in sample data. The ETL process for development of Data mart for base-pair information have following steps:

- **Table input** step is used for taking the input as base-pair information which is presented in Figure 3.5.
- Add **Select values** to select individual repeat type like simple repeat and to rename the field of repeat number of individual repeat type as no_repeat.
- Now add **Replace in string** to replace the no_repeat's numeric value in string values i.e., in simple repeat no_repeat's value 2 is replaced with sr2.
- Add **Filter rows** to remove rows who have no information.
- All above mentioned steps are also performed for all other remaining repeat types.
- Add **Table output** to store the outcome of every repeat types transformation in single table of MySQL.
- Now add **Combination lookup/update** to creates chromosome, CNV and repeat dimension tables.
- Add **Table output** has store the dimension and fact table in data mart.
- To populate the data mart execute the run.
- All the above mentioned step is shown in Figure 4.22.

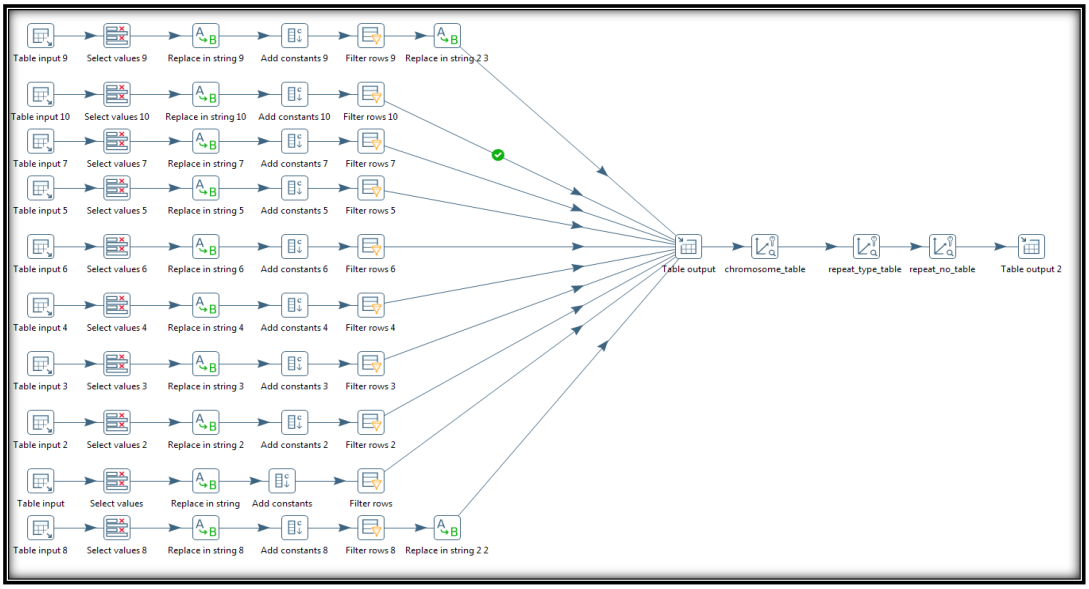



*Figure 4.22: ETL Process for Base-pair Size*

In the Data mart of base-pair dimensions viz. chromosome, CNV and repeat and fact basepair sum are stored which is shown in Figure 4.23.

| Start_point | End_point | Repeat Length | Repeat_percent | bp | chro_id | repeat_type_id | no_repeat_id |
|---|---|---|---|---|---|---|---|
| 694610 | 709923 | 566 | 3.69596447695 | 36 | 1 | 1 | 1 |
| 803929 | 807339 | 29 | 0.850190559953 | 29 | 1 | 1 | 1 |
| 912203 | 913949 | 135 | 7.72753291357 | 135 | 1 | 1 | 2 |
| 2013880 | 2018128 | 47 | 1.10614262179 | 27 | 1 | 1 | 1 |
| 2433445 | 2437974 | 33 | 0.728476821192 | 33 | 1 | 1 | 1 |
| 2446577 | 2449554 | 284 | 9.53660174614 | 45 | 1 | 1 | 1 |
| 2450913 | 2457817 | 347 | 5.02534395366 | 35 | 1 | 1 | 1 |
| 2465163 | 2467172 | 81 | 4.02985074627 | 32 | 1 | 1 | 1 |
| 2486001 | 2491148 | 581 | 11.2859362859 | 36 | 1 | 1 | 1 |
| 2616184 | 2622200 | 1323 | 21.9877015124 | 86 | 1 | 1 | 3 |
| 2631678 | 2633847 | 139 | 6.40552995392 | 38 | 1 | 1 | 1 |
| 2815623 | 2819045 | 375 | 10.9553023663 | 21 | 1 | 1 | 1 |
| 3440915 | 3444348 | 499 | 14.5311589983 | 71 | 1 | 1 | 1 |
| 10324347 | 10332011 | 690 | 9.00195694716 | 28 | 1 | 1 | 1 |
| 12348082 | 12352760 | 630 | 13.4644154734 | 133 | 1 | 1 | 3 |
| 13540603 | 13543140 | 973 | 38.3372734437 | 101 | 1 | 1 | 3 |
| 13548918 | 13550651 | 53 | 3.05651672434 | 53 | 1 | 1 | 1 |
| 13565670 | 13567415 | 44 | 2.52004581901 | 44 | 1 | 1 | 3 |

*Figure 4.23: Data Mart of Base-pair Information*

## 4.6 OLAP Cube Schema Development

After creating the Data mart, there is need to create schema for OLAP cube. This schema can be created by using the design tool i.e., schema workbench. The steps of OLAP cube schema development for listed below:

- Start the design tool schema workbench by executing workbench.bat file.
- Go to menu bar and click on option tab
- Select connection
- Fill the database details for establishing the connection with the database

- Test the connectivity.
- Go to file click on new and select schema
- Fill the value of schema's name attribute.
- Right click on schema and select Add cube.
- Fill the value of cube's name attribute and provide the description of cube in description attribute.
- Right click on cube and select Add Table. Here, add the table of data mart. This table is work as fact table for cube development.
- Right click on cube and select Add Dimension.
- Fill the value of Dimension's name attribute.
- Choose column for foreign key attribute. The foregin key link the dimension table with fact table.
- Right click on dimension name and select add hierarchy and give a name to it.
- Add the primary key in the hierarchy of dimension.
- Right click on dimension name and select add table and choose name attribute's value.
- Right click on hierarchy name and select add level and choose the value of Name, Column, type and level type attributes.
- Right click on cube and select Add Measure.
- Fill the Name, Aggregator, and Column attribute's values.
- Save cube.
- Publish cube on BI server.

Based on above mentioned steps, the required OLAP cube schema have been created. These are described in the next section.

### 4.6.1 OLAP Cube Schema for Cultivar Information

In OLAP cube for Cultivar information, three dimensions are presented namely chromosome, CNV and cultivar. The measures of this cube is count of cultvar. The cube is shown in Figure 4.24.

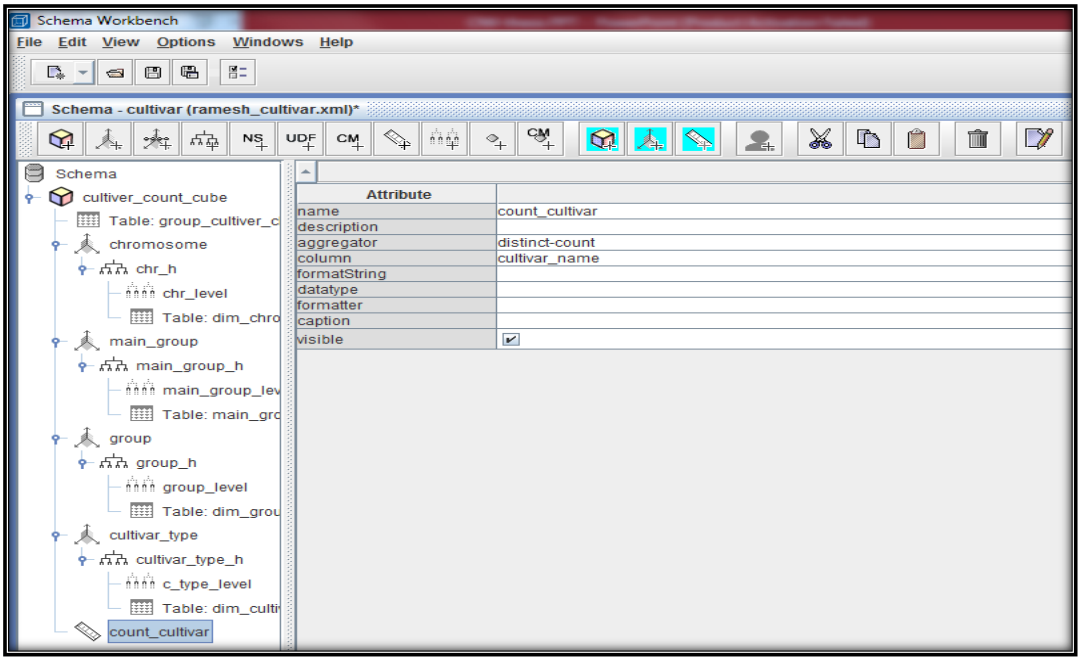



*Figure 4.24: OLAP Cube for Cultivar Information*

### 4.6.2 OLAP Cube Schema for Gene Information

In OLAP cube for gene information eight dimension are present namely chromosome,CNV, CDS type, sv type, mechanism, function, gene, and CNV_region. The measures of this cube is sum of frequency and count of gene. This cube schema is presented in Figure 4.25.

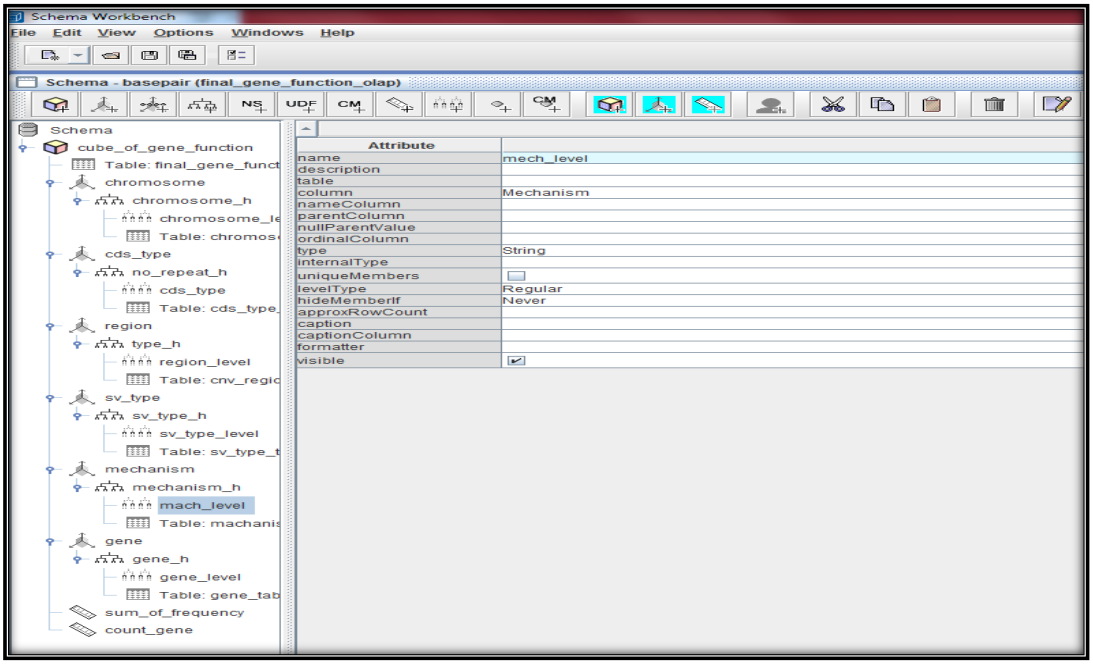



*Figure 4.25: OLAP Cube for Gene Information*

### 4.6.3 OLAP Cube Schema for Base-pair Information

In OLAP cube schema for base-pair information, three dimension namely chromosome, CNV and repeat are identified. The measure of this cube is base-pair sum. The schema for this cube is shown in Figure 4.26.

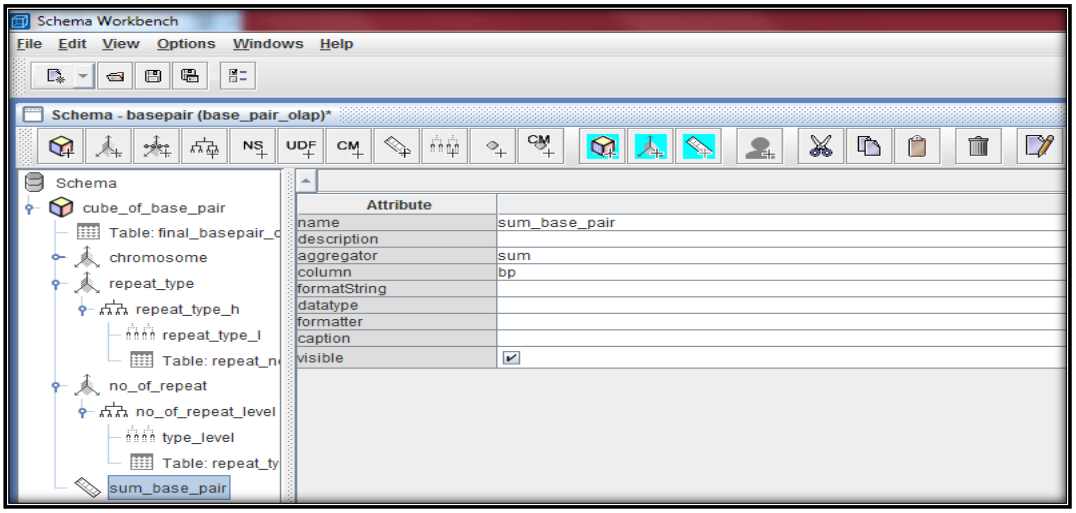



*Figure 4.26: OLAP Cube for Base-Pair Size*

# CHAPTER V

# GENERAL DISCUSSION

The table and graph give eye bird view of information. Data mart provides information in a tabular format and can be reused in desired format. BI server supports the output in tables as well as graphs. The information in the form of cube is summarized and easily understandable. From these cubes and schema, researcher and scientist can easily access and find the required CNVs in their desired format for required task. This will serve as an exploratory analysis tool. The reports are prepared in tow ways i.e., OLAP reports and interactive reports.

## 5.1 OLAP Cube Exploration

OLAP Cubes are useful in customizing the reports in various diensions as per the need of users. The admininstrator will create the sample reports by mentioning various dimensions and aggregation rules which will be useful in exploring the information using slicing/dicing and drill dwon/ roll-up across various dimensions. The customized information will be useful to other users by visualizing/exporting the information/data in tabular and graphical format according to selected dimensions.

### 5.1.1 OLAP Report of Cultivar Information

The count of cultivar corresponding to the various dimensions viz chromosome number, CNV and cultivar of cultivar information is shown in Figure 5.1 and the graphical representation of count of cultivar is shown in Figure 5.2.

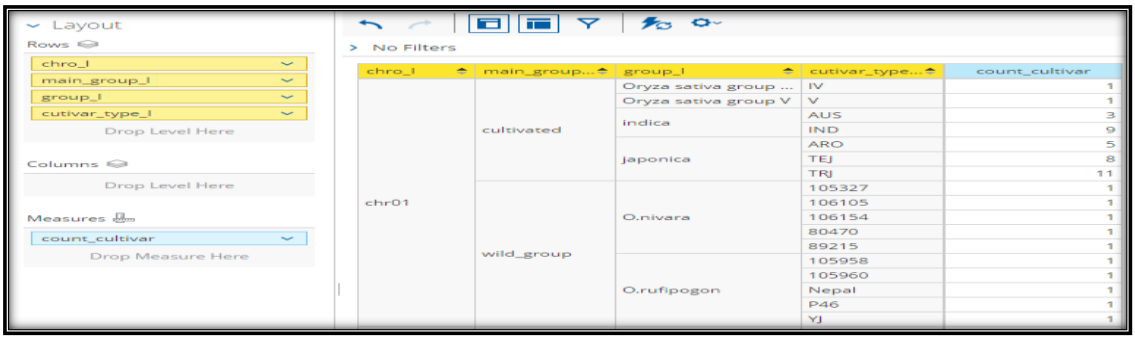



*Figure 5.1: Count Of Cultivar*

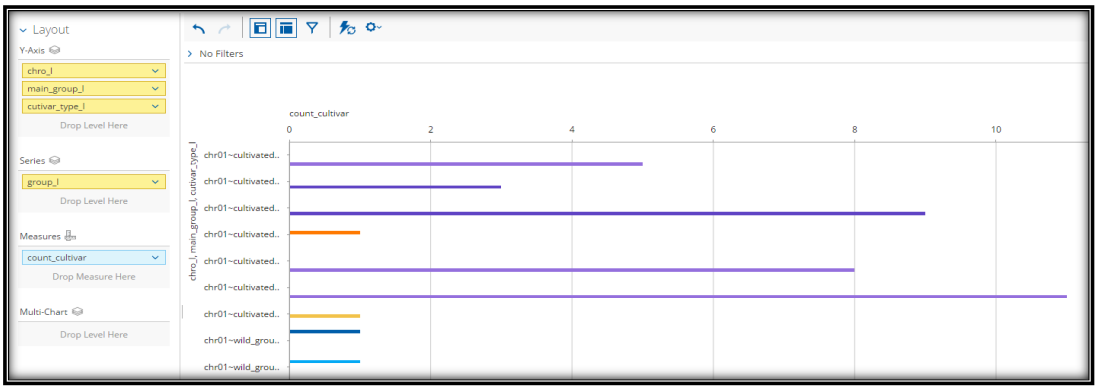



*Figure 5.2: Graphical view of cultivar count*

Finally on can also see group wise count of cultivar is shown in Figure 5.3. This will help in identifiy the contribution of CNVs presence in particular group.

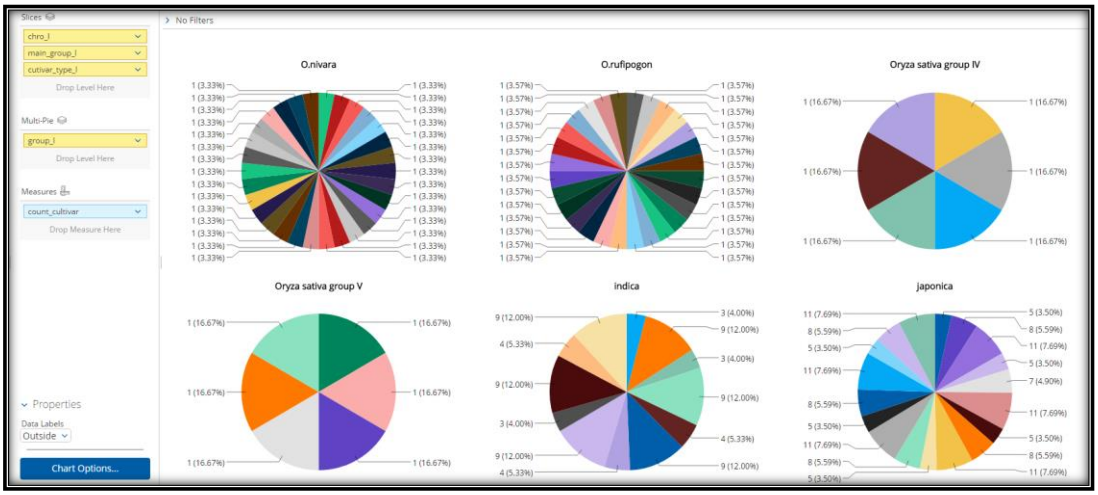
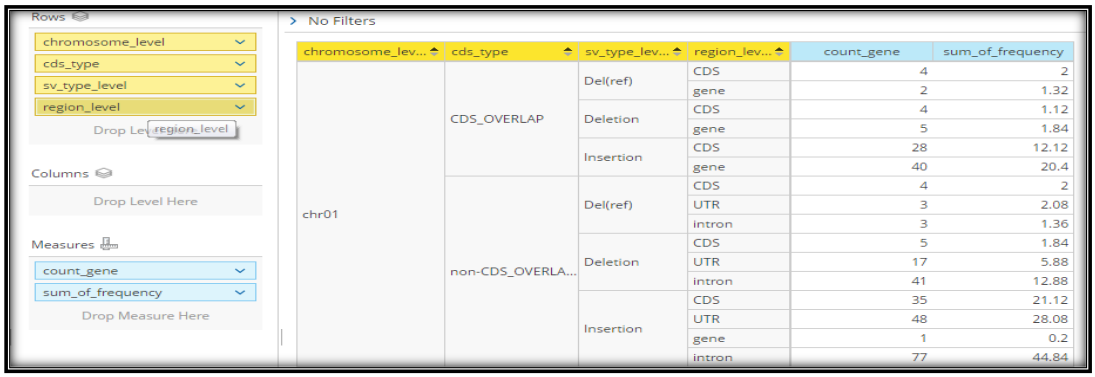



*Figure 5.3: Graphical Representation of Cultivar Count in Different Group*

### 5.1.2 OLAP Report for Gene Information

The count of genes and sum of frequency corresponding to the various dimensions of gene information viz chromosome number, CDS type, structural variation type and region of CNVis shown in Figure 5.4. The graphical represntation of count of gene and sum of frequency are presented in Figure 5.5 and Figure 5.6 respectively.




| chromosome_lev... | cds_type | sv_type_lev... | region_lev... | count_gene | sum_of_frequency |
|---|---|---|---|---|---|
| chr01 | CDS_OVERLAP | Del(ref) | CDS | 4 | 2 |
| | | | gene | 2 | 1.32 |
| | | Deletion | CDS | 4 | 1.12 |
| | | | gene | 5 | 1.84 |
| | | Insertion | CDS | 28 | 12.12 |
| | | | gene | 40 | 20.4 |
| | non-CDS_OVERLA... | Del(ref) | CDS | 4 | 2 |
| | | | UTR | 3 | 2.08 |
| | | | intron | 3 | 1.36 |
| | | Deletion | CDS | 5 | 1.84 |
| | | | UTR | 17 | 5.88 |
| | | | intron | 41 | 12.88 |
| | | Insertion | CDS | 35 | 21.12 |
| | | | UTR | 48 | 28.08 |
| | | | gene | 1 | 0.2 |
| | | | intron | 77 | 44.84 |

*Figure 5.4: Count of Gene and sum of base-pair*

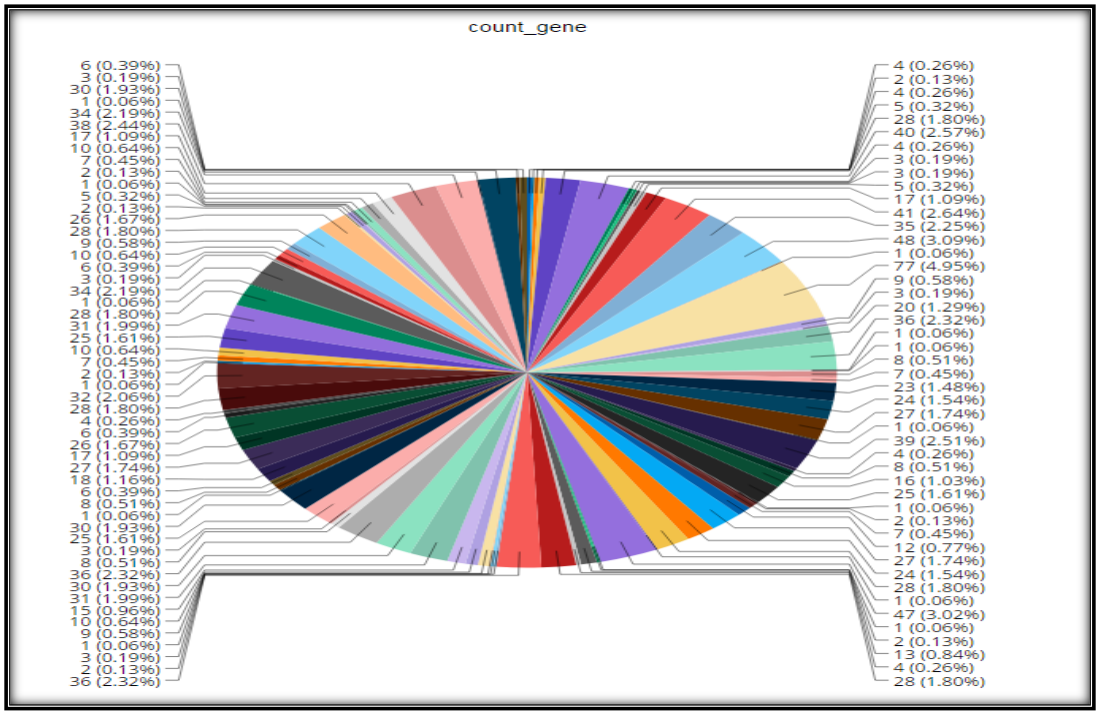



*Figure 5.5: Graphical Representation Of Count Of Gene*

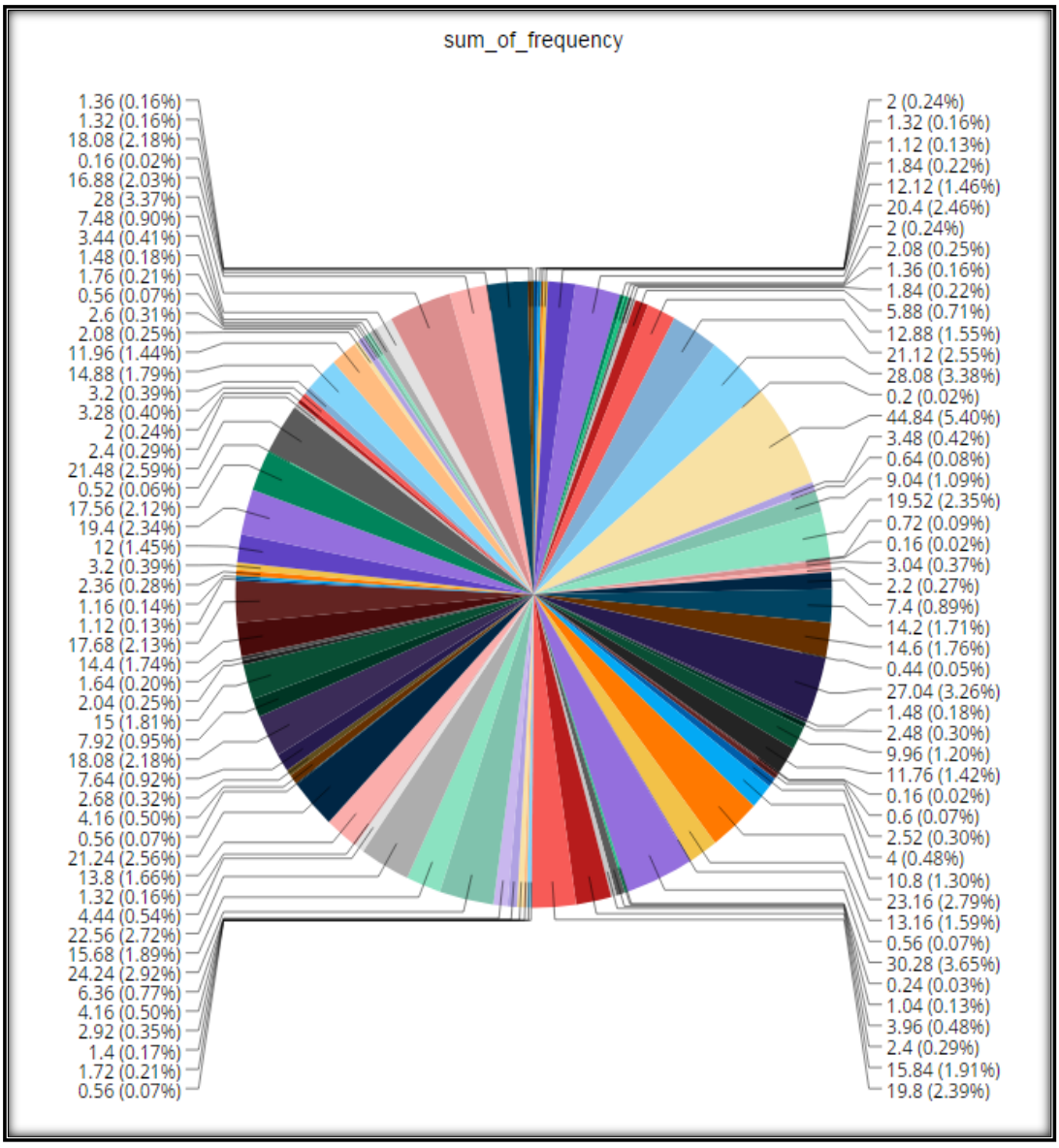
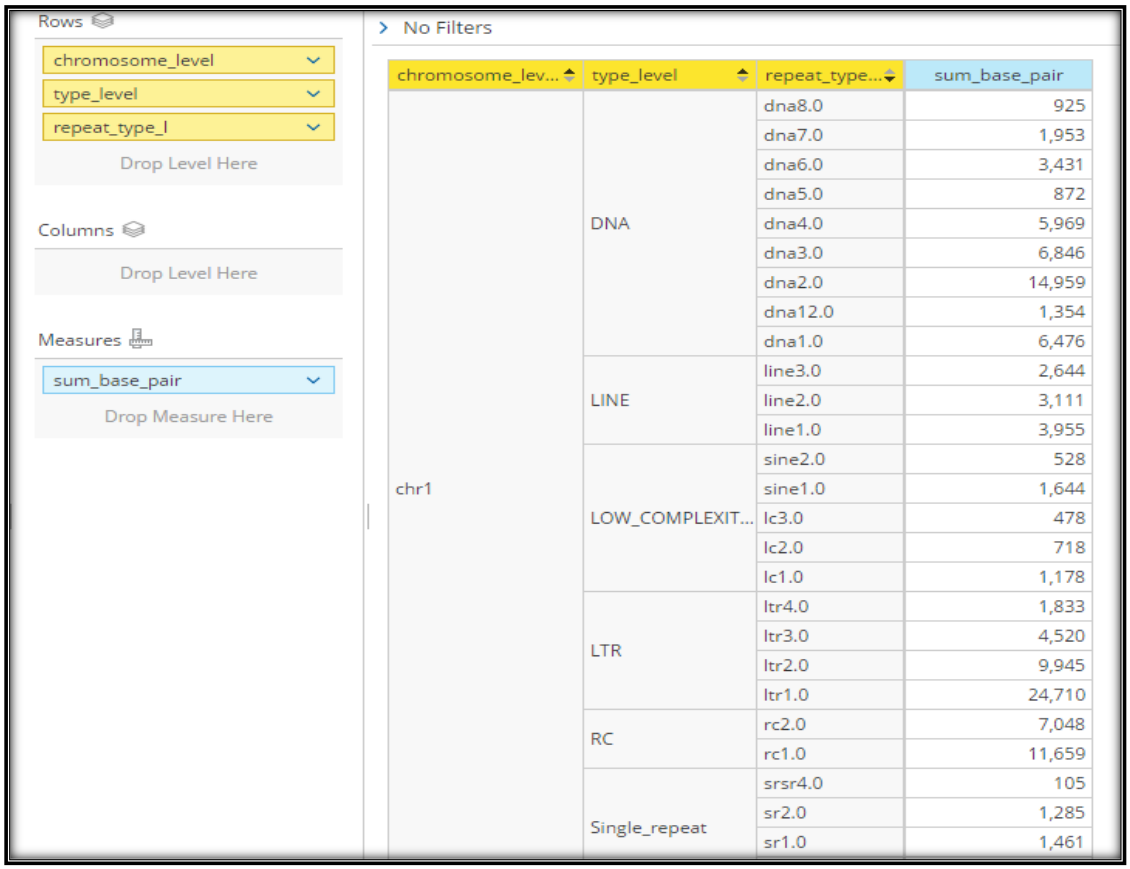



*Figure 5.6: Graphical Representation of Sum of Frequency*

### 5.1.3 OLAP Report of Base-pair Information

The sum of base-pair corresponding to the various dimensions of basepair information viz chromosome number, CNV, repeat is shown in Figure 5.7 and graphical representation is presented in Figure 5.8.




| chromosome_lev... | type_level | repeat_type... | sum_base_pair |
|---|---|---|---|
| chr1 | DNA | dna8.0 | 925 |
| | | dna7.0 | 1,953 |
| | | dna6.0 | 3,431 |
| | | dna5.0 | 872 |
| | | dna4.0 | 5,969 |
| | | dna3.0 | 6,846 |
| | | dna2.0 | 14,959 |
| | | dna12.0 | 1,354 |
| | | dna1.0 | 6,476 |
| | LINE | line3.0 | 2,644 |
| | | line2.0 | 3,111 |
| | | line1.0 | 3,955 |
| | LOW_COMPLEXIT... | sine2.0 | 528 |
| | | sine1.0 | 1,644 |
| | | lc3.0 | 478 |
| | | lc2.0 | 718 |
| | | lc1.0 | 1,178 |
| | LTR | ltr4.0 | 1,833 |
| | | ltr3.0 | 4,520 |
| | | ltr2.0 | 9,945 |
| | | ltr1.0 | 24,710 |
| | RC | rc2.0 | 7,048 |
| | | rc1.0 | 11,659 |
| | Single_repeat | srsr4.0 | 105 |
| | | sr2.0 | 1,285 |
| | | sr1.0 | 1,461 |

*Table 5.7: Sum of Base-Pair*

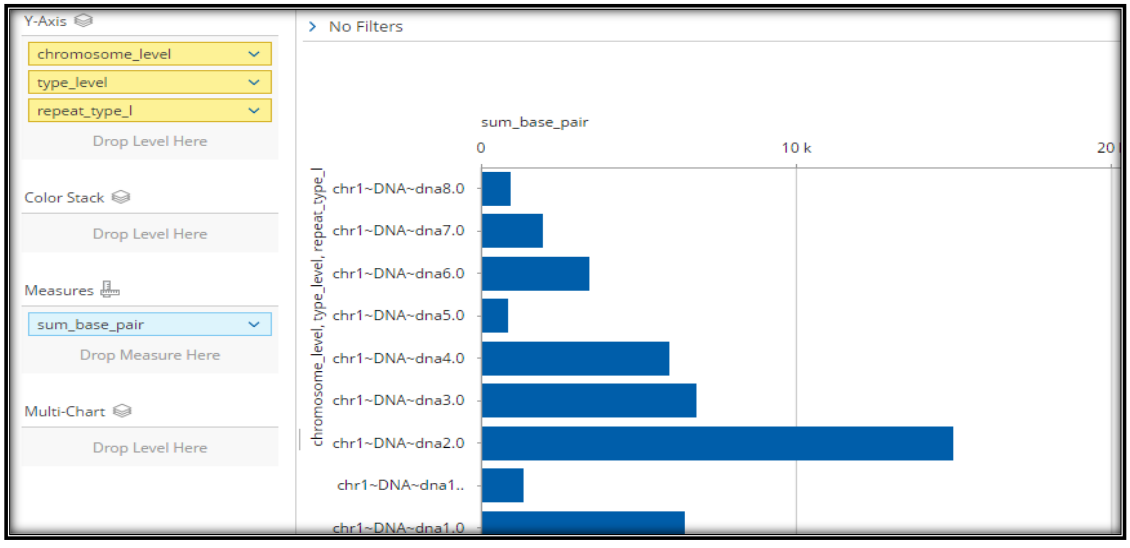



*Figure 5.8: Graphical Representation of Sum Of Base-pair*

## 5.2 Interactive Report

The Interactive report provides an operational on-demand web based report format which can be customised as per the requirements.

### 5.2.1 Interactive Report of Cultivar Classification

In the interactive report of cultivar classification, chromosome number, main group, and group name and cultivar type are used as prompt for filtering the records. Interactive report gives only those data which is selected in column and fulfils the condition of these four prompts. The output with column CNV start, CNV end and cultivar name is shown in Figure 5.9.

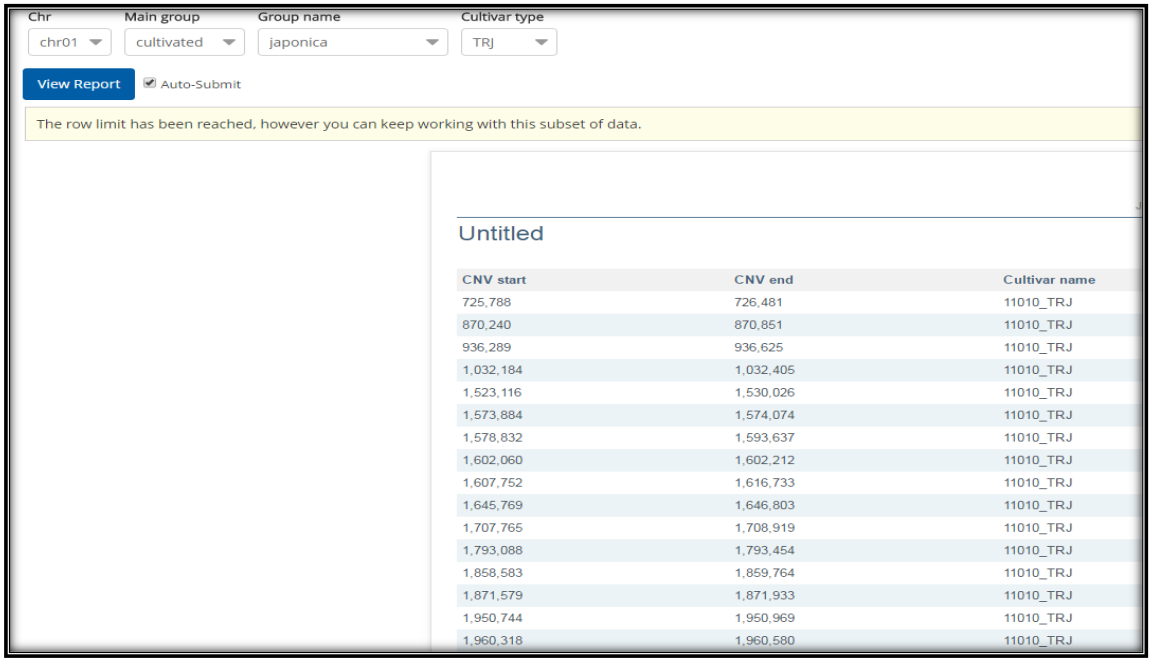



*Figure 5.9: CNV Position and Cultivar Name*

### 5.2.2 Interactive Report of Gene Information

The output with column name start, end, gene id, frequency and function on the bases of five prompts condition namely Chromosome number, CDS type, mechanism, structural variation type, and CNV region is shown in *Figure* 5.10.

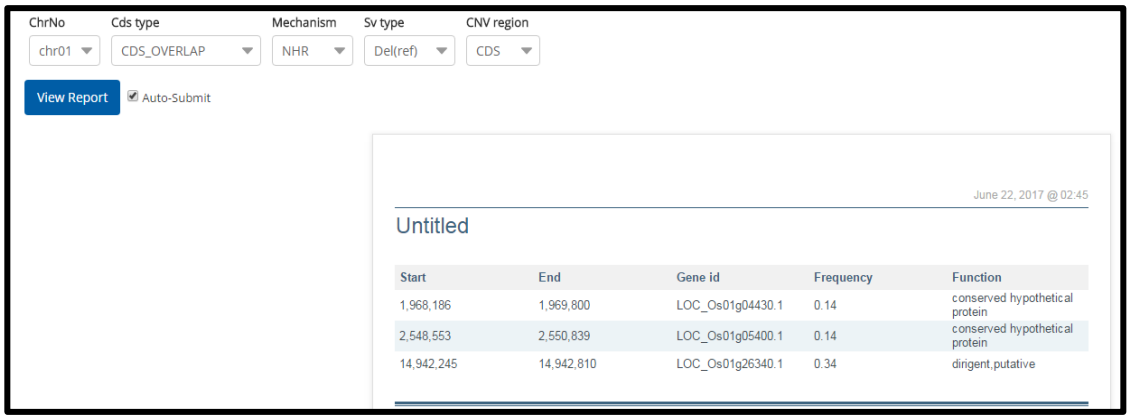



*Figure 5.10: CNV Position, Gene Id, Frequency and Function*

### 5.2.3 Interactive Report of Base-pair Information

The output with column name start point, end point repeat length and repeat percent is generated by using chromosome number, type name, and repeat type prompt conditions. The output is presented in Figure 5.11.

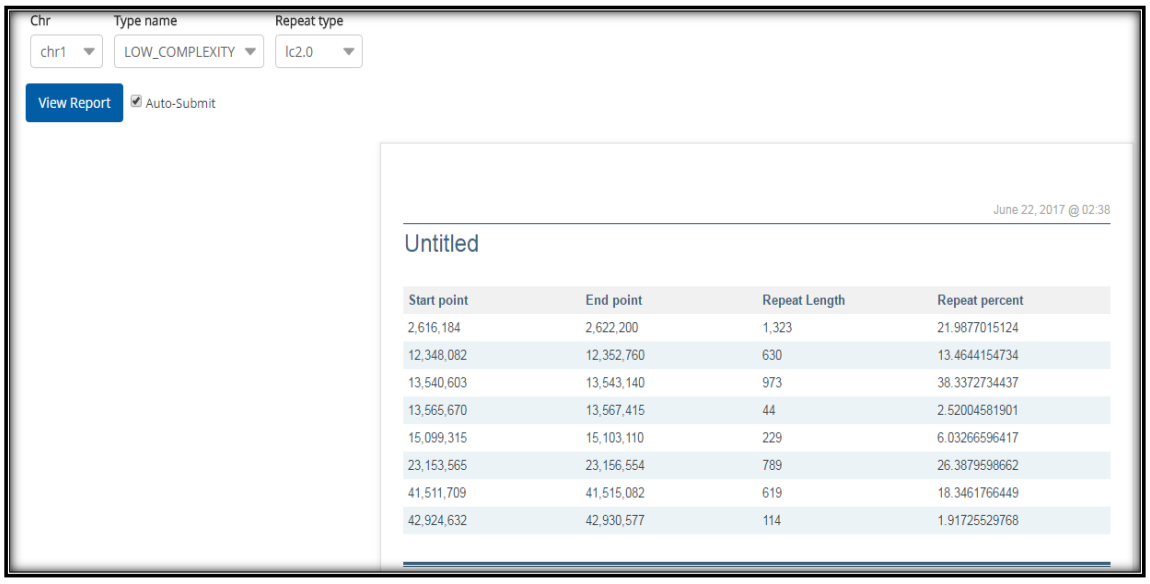



*Figure 5.11: CNV Position, Repeat Length, and Repeat Percent*

# CHAPTER VI

# SUMMARY AND CONCLUSIONS

Agriculture is backbone of India and plays significant role in Indian economy. Most of Indian rural family depends on agriculture for meeting their daily livelihood. The growth rate of Indian population is dramatically increasing and area under agriculture is continuously decreasing. Further, various biotic and abiotic stresses also reduced the production of crop. To feed the population of our country, traditional farming methods are time consumable and not sufficient. There is a need to implement advance methods of breeding involving latest molecular breeding, markers selection and development techniques to help in developing improved high yielding and stress resistant varieties. The study of variability is one of most important factor for selection of vigorous and resistant plant in breeding population. Copy number variation is one type of variability. CNV is a genomic variant in which a sequence of genome identifies the multiple copies of markers and genes in the genome. Basically CNV are of two types i.e., short repeats and long repeats. In short repeats, two or three nucleotide is repeated but in long repeats entire gene sequence is repeated several times. The CNV helps in studying genomic and structural variation in a crop which is helpful in identification of variability and diversity. The information related to different aspects of CNV is not readily available and is being reported by researchers in different ways available at varied places. Since, these genomic resources are scattered and also providing a great insights to the researchers but similarly they are finding it difficult to get the collective information of CNVs at one place. Development of data mart will be helpful in integrating the CNV related information at central place by providing the CNV data in various dimensions. Data mart provides the related information at single platform by extending the summary, sort or select in various dimension without making any change in the metadata structure.

In this study, the data Mart for CNV has been tested and implemented using data of 50 rice accessions covering various geographical areas of the Asian region. The data mart can be further updated, once the information for other accessions will be available in same format. This source data is provided as supplementary material for 50 rice

accessions (Bai *et al.*, 2016). Different open source set of tools i.e., Pentaho data integration tool, schema workbench and BI sever were used during the development process of Data mart. MySQL was also used to create and store database. Pentaho data integration tool spoon was use to implement the extraction, transformation, and loading of data. Spoon was used to create different dimensions. Schema workbench was used to define the dimension and measures to create OLAP cube schema. The schema workbench produces output in the form of metadata as xml file. This xml file was used in BI sever as input to develop the schema for OLAP cube and other types of reports. BI server was use to visualize the OLAP cube and for generating other reports like interactive reports using different dimensions and parameters. The graphical reports are very useful to the researchers.

The developed mart will be useful to the researchers in providing collective information of CNV. The mart will help in easy accessibility of the data through its user friendly web interface. Information from CNV mart can be accessed across various dimensions such as chromosome, gene, structural variation, CNV detection mechanism, cultivar, and repeat types. The developed mart has the flexibility of providing reports such as:

i. Retrieval of CNV information for the study of varietal development and improvement.
ii. Querying of CNVs based on user requirement.
iii. Visualization of data through various dimensions.
iv. Exploring the CNV data for the purpose of diversity analysis.

Currently, the CNV mart provides the limited information of CNV contents along with their exact location in genome including traits. Single nucleotide polymorphisms (SNPs), SNP haplotypes, insertion and deletions (InDels) and simple sequence repeats (SSRs) will be further integrated. CNV Data mart can also join other genomic marts to form an integrated information Warehouse.

# ABSTRACT

The biggest challenge facing Indian agriculture is to develop high yielding varieties to feed the vast increasing population of the country. Seed is critical and basic input for attaining high crop yields and sustainable growth in agricultural production. The advances in genome sequencing technologies are helpful in identification of different types of markers which can help in development of high yielding varieties. Single reference genome is not able to provide the representation of genetic diversity in a given species. The diversity can be identified and discovered using the study of structural variation in the form of copy number variants (CNVs) by studying the sequences of different accessions. The CNV will account for complete value of genetic information that is present in individual species. Copy number variation (CNV) plays an important role in identifying the genetic and phenotypic variation in the breeding population. Data mart of CNV has been developed and enabled end users in identifying the association between CNV and cultivar with respect to various types of traits. Pentaho Business Analytics platform is an open source set of software tools for data mart development and MySQL database has been used to store the data. Different kind of reports such as interactive and OLAP reports has been generated and can be viewed through the web interface using BI server which was hosted on Apache web server. The exploration of data from data mart and OLAP cube visualizations will provide the CNV related information in various dimensions viz. chromosome wise, gene wise, structural variation wise, CNV detection mechanism wise, cultivar wise and repeat type wise through a single window access. The CNV mart can further be used by researchers for retrieval of CNV information to identify disorder and helpful in developing diagnostic kits and treatments, varietal development and improvement by genome wide association of different cultivars.

# सार

वर्तमान समय में भारतीय कृषि विशाल जनसंख्या के भोजनापूर्ति के लिए एक बड़ी चुनौती का सामना कर रही है, जो की उच्च पैदावार वाली किस्मों को विकसित करना है। फसल की पैदावार बढ़ाने और कृषि उत्पादन में सतत विकास के लिए बीज महत्वपूर्ण और बुनियादी उत्पादक सामग्री है। जीनोम अनुक्रमण तकनीकों में हुए प्रगति विभिन्न प्रकार के मार्करों की पहचान करने में सहायक होती है जो उच्च पैदावार वाली किस्मों के विकास में मदद कर सकती हैं। केवल एक संदर्भ जीनोम किसी प्रजाति में आनुवंशिक विविधता का प्रतिनिधित्व करने में सक्षम नहीं होता है। विभिन्न परिग्रहणों के अनुक्रमों का अध्ययन करके "कॉपी संख्या विविधता" (सीएनवी) के रूप में संरचनात्मक भिन्नता के अध्ययन का उपयोग करते हुए विविधता को पहचाना एवं उसका पता लगाया जा सकता है। सीएनवी आनुवांशिक जानकारी के पूर्ण महत्त्व के लिए उत्तरदायी है जो व्यक्तिगत प्रजातियों में मौजूद होता है। कॉपी संख्या विविधता (सीएनवी) प्रजनन आबादी में आनुवंशिकी और प्ररूपी भिन्नता की पहचान करने में एक महत्वपूर्ण भूमिका निभाती है। उपयोगकर्ताओं को विभिन्न प्रकार के फसलों के गुणों एवं सीएनवी के बीच के सहयोग की पहचान करने में सक्षम करने हेतु सीएनवी डाटा मार्ट विकसित किया गया है। "पेंटाहो व्यावसायिक विश्लेषिकी मंच" सॉफ़्टवेयर टूल का समुच्चय है जो डाटा मार्ट के विकास के लिए एक खुला श्रोत है और डेटा के संग्रह के लिए MySQL डाटाबेस का इस्तेमाल किया गया है। विभिन्न प्रकार के विवरण जैसे इंटरैक्टिव और ओ.एल.ए.पी. विवरण बी.आई. सर्वर का उपयोग कर वेब इंटरफेस के माध्यम से बनाये और देखे जा सकते हैं जो की अपाचे वेब सर्वर पर होस्ट की गई है। डेटा मार्ट और ओएलएपी क्यूब विज़ुअलाइजेशंस से डेटा की अन्वेषण विभिन्न आयामों जैसे गुणसूत्र के अनुसार,संरचनात्मक भिन्नता के अनुसार, जीन के अनुसार, सीएनवी पहचान तंत्र के अनुसार, फसल के अनुसार और दोहराने की प्रकार के अनुसार एकल खिड़की अभिगम के माध्यम से सभी जानकारी प्रदान करेगा। सीएनवी मार्ट को शोधकर्ताओं द्वारा सीएनवी जानकारी की पुनर्प्राप्ति के लिए विकार की पहचान करने और नैदानिक किटों और उपचार, विविधता के विकास और विभिन्न किस्मों के जीनोम व्यापक सहयोग के द्वारा सुधार के विकास में मददगार हो सकता है।

# BIBLIOGRAPHY


Bai, Zetao, Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R.A., & Chen, M. (2016). The impact and origin of copy number variations in the Oryza species. *BMC genomics*, **17(1)**:261.

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., & Paraboschi, S. (2001). Designing data marts for data warehouses. *ACM transactions on software engineering and methodology*, ***10*(4)**: 452-483.

Chaturvedi, K., Rai, A., Dubey, V., & Malhotra, P. (2008). On-line analytical processing in agriculture using multidimensional cubes. *Journal of the Indian Society of Agricultural Statistics*, ***62*(1)**, 56-64.

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., & Diers, B. W. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*, **338(6111)**:1206-1209.

Dutta, D., (2010). Design and Development of Data Mart for Consumption Expenditure Survey Data. M.Sc. Thesis, IARI, New Delhi.

Kimball, R., & Ross, M. (2011). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 3rd Kindle Edition, John Wiley & Sons.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., & Kersey, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, ***2011***, bar030.

Kumar, N.R., Muralidharan, K., Sairam, C.V., Palaniswamy, C., Arulraj, R.D.S., Rai, A., Dubey, V., Chaturvedi, K.K., Sreedharan, K., Vinod Kumar, P.K. & Chulaki, B.M. (2002). Design of data marts for plantation crops. In Proceedings of the *15th Plantation Crops Symposium Placrosym XV, Mysore, India, 10-13 December, 2002.* pp. 784-790.

Maron, L. G., Guimaraes, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., & Magalhaes, J. V. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proceedings of the National Academy of Sciences,* **110(13)**:5241-5246.

McCarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics,* **39(Supp 7)**:S37-42.

Moody, D. L., & Kortink, M. A. (2000). From enterprise models to dimensional models: a methsodology for data warehouse and data mart design. *DMDW*, 5-17.

Nilakanta, S., Scheibe, K., & Rai, A. (2008). Dimensional issues in agricultural data warehouse Designs. *Computers and Electronics in Agriculture*, **60(2)**:263-278.

Rai, A., Dubey, V., Chaturvedi, K. K., & Malhotra, P. K. (2008). Design and development of data mart for animal resources. *Computers and Electronics in Agriculture*, **64(2)**:111-119.

Suresh, R.S. (2008). Design and Development of Data Mart for Household Amenities from Census Data (Maharashtra). M.Sc. Thesis, IARI, New Delhi.

Sutton, T., Baumann, U., Hayes, J., Collins, N. C., Shi, B. J., Schnurbusch, T., & Langridge, P. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science,* **318(5855)**:1446-1449.

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., & Xu, J. (2015). Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nature Genetics,* **47**:244-248.

Yang, Y., Wilson, L. T., Wang, J., & Li, X. (2011). Development of an integrated cropland and soil data management system for cropping system applications. *Computers and Electronics in Agriculture*, ***76*(1)**:105-118.

Yu, P., Wang, C. H., Xu, Q., Feng, Y., Yuan, X. P., Yu, H. Y., & Wei, X. H.(2013). Genome-wide copy number variations in Oryza sativa L. *BMC genomics,* **14**:649.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, **14(Suppli 11)**:S1.

Zhou, Y., Zhu, J., Li, Z., Yi, C., Liu, J., Zhang, H., & Liang, G. (2009). Deletion in a quantitative trait gene qPE9-1 associated with panicle erectness improves plant architecture during rice domestication. *Genetics*, **183(1)**:315-324.