

Tool for Visualisation the Gene Loci of Multiple Genes

Udayakumar Mani, M. Krishna Nishand , G. Raghu Ram

Department of Bioinformatics,
SASTRA University, Tanjore - 613401,
Tamil Nadu, India

Abstract:

NCBI contains the chromosome location for a particular gene of our interest. When we work on genes, causing a disease, it will be very cumbersome and time consuming to check the location of each and every gene to find out the effect of it. Not all disease causing gene loci can be represented at a single time. We consider the array express data set. Some of the chromosome gene loci are not been inputted in the data set. To address these problems, we have found out a visualization tool which can graphically represent all the disease causing genes for a particular disease. This tool retrieves gene loci information from the genbank database and the dataset is updated. The intuitive front end of the tool makes it convenient and easy for people studying that disease and gets their gene location at a single click.

Keywords: Chromosomes, Gene Locations, Ideogram, Graphical Representation.

INTRODUCTION

Human body cells have 46 chromosomes, made up of 23 pairs. There are 44 chromosomes numbered 1-22 called autosomes according to size from the smallest to the largest and two sex chromosomes: X and Y. The chromosomes consist of two very long thin strands of DNA chains twisted into the shape of a double helix and are located in the nucleus our body cells. The chromosomes are long strands of genes. Since the chromosomes come in pairs, the genes also come in pairs. Genes are also located in very small compartments called mitochondria that are randomly scattered in the cytoplasm of the cell outside the nucleus. Different genes are active in different cell types, tissues and organs, producing the necessary specific proteins. Some genes are ‘switched off’ and others are ‘switched on’. Changes to the genetic code can mean that a particular protein is not produced properly, produced in the wrong amounts or not produced at all.

Table 1: Nomenclature explanation

Component	Explanation
7	The chromosome number.
q	The position is on the chromosome's short arm (<i>p</i> for <i>petit</i> in French); q indicates the long arm (chosen as next letter in alphabet after p).
31.2	The numbers that follow the letter represent the position on the arm: region 3, band 1, and sub-band 2.

NCBI gene locator:

The Map Viewer [3] provides special browsing capabilities for a subset of organisms in Entrez Genomes. The organism subset is shown and also on the Map Viewer. Map Viewer allows you to view and search an organism's complete genome, display chromosome maps, and zoom into progressively greater levels of detail, down to the sequence data for a region of interest. The number and types of available maps vary by organism, and are described in the data and search tips file for each organism. The chromosome-specific view shows one or more detailed maps for a single chromosome. The results of a genome-wide search include a graphic of all the chromosomes (to scale), with red tick marks showing the location of markers that contained search term. The search terms can be clone name, gene symbol, gene name, marker name, aliases and text word

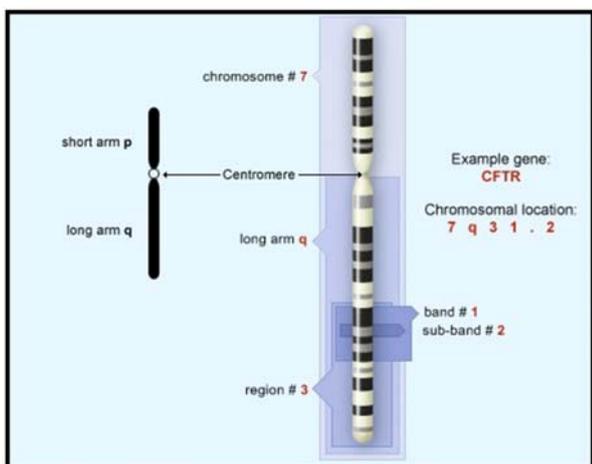


Figure 1: Nomenclature of chromosome

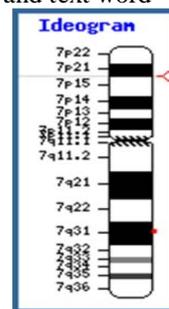


Figure 2: NCBI chromosomal representation at location: 7q31.2

NCBI [1] contains the chromosome location for a particular gene of our interest. When we work on genes, causing a disease, it will be very cumbersome and time consuming to check the location of each and every gene to find out the effect of it. Not all disease causing gene loci can be represented at a single time. To overcome this problem a tool has been developed which can graphically represent all the disease causing genes for a particular disease.

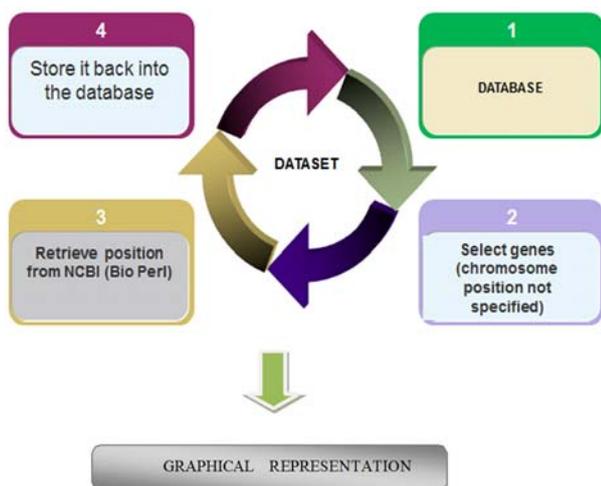


Figure 3: Dataflow diagram

Input Data:

Here we consider schizophrenia disease. The gene ids for schizophrenia are obtained from ArrayExpress [2]. Human-afly_HG-U133_Plus_2_2007-12-04. The annotation file is used as input for the tool. In general, to analyze we can use any input array express annotation file and further process can be done.

Storage:

The data is stored in MySQL database. The interface for storing data is done using php (hypertext pre-processor). The clone ids, genbank ids, Entrez ids, unigene id and chromosome positions are stored in the database. Each id is useful in retrieving the chromosome positions of the genes.

Processing:

Using the MySQL queries, the gene ids are retrieved from the database which does not have chromosomal location. The respective Entrez ids are used to obtain the missing gene loci. The retrieval is done using Perl programming. LWP module is used. This module retrieves the whole html web page which is obtained from the required URL. This LWP module is not an inbuilt package. It can be downloaded from perl.org website. Install this LWP module and location can be obtained. <http://www.ncbi.nlm.nih.gov/gene/> followed by the Entrez id downloads the whole html page. This page is in XML (Extensible Markup Language) format. This XML file is converted into text format. Using regular expression which is a very important function of the Perl language helps in recognizing the chromosome location. The gene position patter is used to identify the exact gene location. The identified gene location is updated back into the database using the php and MySQL queries.

Representation:

The gene location is displayed as a .png (portable network graphics) image using the graphics. The inbuilt graphics module GD module present in php is used for the representation. An image which exemplifies 23 chromosomes with all the gene loci is visualized simultaneously. First all 23 chromosomes are generated using php graphics functions. The retrieved gene loci values are represented on their respective chromosome region.

RESULTS AND DISCUSSION

The integrated tool can actually visualize the gene loci of multiple genes on a single platform or on a single shot. The aim of this project has been to help people who study a disease. While the drawback of NCBI has been that not all genes' chromosome location can be identified at the same time or to an extent visualized. This drawback in NCBI has been overcome using the project. We require files from ArrayExpress and the first file to be loaded is the annotation file from ArrayExpress which contains the CLONE_ID, GENBANK_ACC, REFSEQ_ACC, ENTREZ_ID, UNIGENE_ID, GENE_SYMBOL, GENE_TITLE and the CHROMOSOME_POSITION. These data are some of the parameters for the genes that are to be taken into consideration for the process below. First we load into the database which may or may not have their chromosome positions intact.

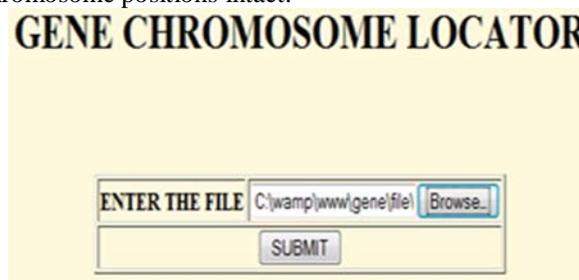


Figure 4: Input page

The file that is loaded into the database shows the following output is generated. Now all the contents are stored into the database in their respective attributes and the data needed for us is in an organized manner.

Clone ID	Entrez ID	Chrom Pos
1007_s_at	780	chr6:30960320-30975908(+)
1053_at	5982	chr7:73283770-73306674(-)
117_at	3310	chr1:159760964-159763203(+)
121_at	7849	chr2:113690046-113752968(-)
1255_g_at	2978	chr6:42231152-42255770(+)
1294_at	7318	chr3:49817643-49826395(-)
1316_at	7067	chr17:35472589-35499815(+)
1552256_a_at	949	NA
1552257_a_at	23170	chr22:41892573-41913045(-)
1552258_at	112597	NA
1552261_at	10406	chr20:43541899-43543585(+)
1552263_at	5594	NA
1552264_a_at	5594	NA
1552266_at	203102	NA

Figure 5: Tabulated output of the location of each unknown gene on a chromosome

Once the data is stored into the database, we get a confirmation saying that all the records have been safely stored in the database and hence we have gone for the next step. That will be to retrieve the chromosome positions of those genes which do not have their positions specified. The Bioperl file will run which will retrieve all the information from the XML file that is downloaded from NCBI. The above program is an illustration of how we exactly retrieve information from NCBI using an XML format. It actually first connects with the database that is present in the local server and takes those genes whose chromosome positions are not specified. It then goes to the URL of NCBI, i.e. www.ncbi.nlm.nih.gov/gene/GENE_ID. *GENE_ID* is the one the Gene id for a particular gene in NCBI. It fetches the information and stores in back to the database in the local server. Now we have all information required for building up the tool. So the pre- processes are over. Now that the file is ready, we will compile the above Bioperl file and the output generated will be like this:

GENE CHROMOSOME LOCATOR

Clone ID	Entrez ID	Chrom Pos
1552256_a_at	949	12q24.31
1552258_at	112597	2p11.2
1552263_at	5594	22q11.2
1552264_a_at	5594	22q11.2
1552266_at	203102	8p11.22
1552269_at	128153	1q41
1552276_a_at	57617	15q14
1552278_a_at	113235	17q11.2
1552286_at	90423	2p21
1552293_at	256130	7p21.1

Figure 6: Tabulated output of the location of each known gene on a chromosome

The generated output actually shows the location of all genes contained in the annotation file and hence we will have the all the inputs ready by our side. Now we need to plot all the diseased chromosomes on a plot which is a graphical representation which can be in any image file format namely, .jpg and .png formats. The representation will be like having all the 23 chromosomes on a single representation and hence we will have to plot all the genes on those 23 chromosomes to be plotted in a single image file or in a single shot representation.

Chromosome locus:

The location of a gene on a chromosome is called as locus. It is normally represented as 8p21.3

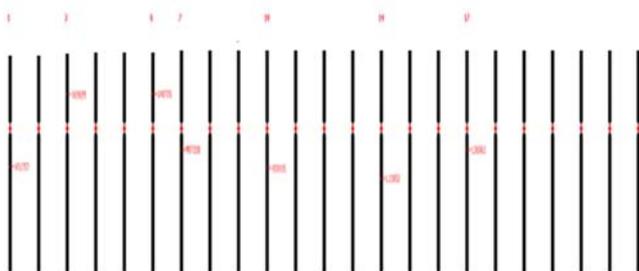


Figure 7: Pictorial image of the output

This is the exact representation of the genes in the chromosomes as we can clearly see all the genes are located on the 23 chromosomes available with the shorter hand that is the p hand at the top and the longer hand that is the q hand below. Now all the genes can be significantly represented without any glitches simultaneously and on a single representation comfortably. A magnified representation of the image file would be like Figure 8. Thus this is the representation of all the genes on the chromosome for first 8 chromosome location. This is a representation of the longer and the shorter hand and the gene id which has been plotted on the chromosome number. This is for all the disease causing genes and the innovation lies in the representation of the genes on a single platform or on a single image file.

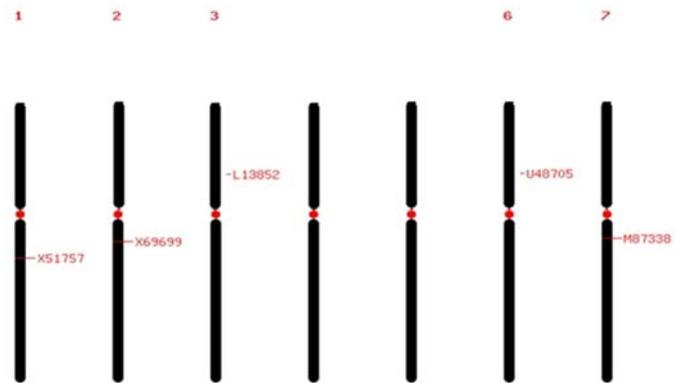


Figure 8: Magnified image of the output

SAMPLE CODE:

```
<?php
error_reporting(0);
echo "<body bgcolor='cornsilk'><center><h1><font
color=green>CHROMOSOME
LOCATOR</h1></font'";
$name=$_FILES['f']['name'];
$ploc=$_FILES['f']['tmp_name'];
if(move_uploaded_file($ploc,"file/data.txt"))
{echo "<h2>FILE UPLADEDED
SUCCESSFULLY!!!</h2><br>";}
else
{echo "Upload error";}
mysql_connect("localhost","root","");
mysql_select_db("gene1");
mysql_query("truncate data");
$k="load data local infile
'C:/wamp/www/Project/gene/file/data.txt' into table data";
mysql_query($k) or die(mysql_error());
$f=mysql_query("select CLONE_ID,ENTREZ_ID,CHROMPOS
from data limit 5");
echo "<form method=post action=/cgi-bin/loc.pl><input
type=submit value='Click here to
find the unknown gene location'><br><br></form></a>";
echo "<center><table border=1><tr><td><b>Clone
ID</td><td><b>Entrez
ID</td><td><b><center>Chromosome Position</td></tr>";
while($r=mysql_fetch_array($f))
{$a1=$r[0];$a2=$r[1];$a3=$r[2];
echo "<tr><td>$a1</td><td>$a2</td><td>$a3</td></tr>";}
?>
```

CONCLUSION

A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions

ACKNOWLEDGEMENT

We thank Department of Bioinformatics, SASTRA University for providing us the facilities to successfully carry forwards and complete the research project.

REFERENCES

- [1]. NCBI website: <http://www.ncbi.nlm.nih.gov/>. Accessed on November 12 2011
- [2]. ArrayExpress website <http://www.ebi.ac.uk/arrayexpress/>. Accessed on November 13 2011)
- [3]. NCBI Map Viewer: <http://www.ncbi.nlm.nih.gov/mapview/static/MapViewHelp.html>. Accessed on November 14 2011.