2012

Risk Prediction For A Fly Genome In A Clinical Context

Lakshmi Mamidi
Grand Valley State University

Risk Prediction for a fly genome in a clinical context

By

Lakshmi R Mamidi

Fall, 2012

Risk Prediction for a fly genome in a clinical context

By

Lakshmi R Mamidi

A project submitted in partial fulfillment of the requirements for the degree of

Master of Science in

Computer Information Systems

at

Grand Valley State University

December, 2012

____

Dr.Guenter Tusch **December 13, 2012**

Table of Contents

Abstract

This research project attempts to evaluate a fly genome from a clinical perspective. To study the clinical translation of genetic risk, Drosophila Melanogaster has been chosen as the model organism. A similar study to this, on the human genome, has been reported in [1]. Thanks to technological advancement, the cost of genome sequencing has significantly decreased in recent years, thus making genetic information potentially accessible for clinical use. However, the explanatory power and clinical implementation and utilization of risk estimates for common variants as found in genome-wide association studies still remain widely unclear [1]. Drosophila Melanogaster has been used in this model, because it is attractive to study as explained in [2]. Many basic biological, physiological, and neurological properties are conserved between mammals and D. melanogaster. Nearly 75% of human disease-causing genes are believed to have a functional homolog in the fly. The data source is the National Center for Biotechnology Information's (NCBI's) Sequence Read Archive (SRA) including the Bergman Lab [3] , which stores raw sequence data from the next generation of sequencing platforms ("Next-Generation Sequencing"). To quantify the genetic risk information available in different databases (NCBI SRA, Flybase, etc.) has been integrated as shown in the below diagram in the form of a pipeline.

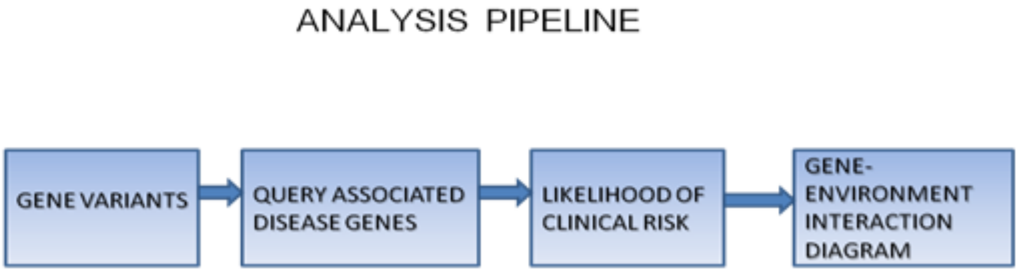

ANALYSIS PIPELINE
GENE VARIANTS
QUERY ASSOCIATED DISEASE GENES
LIKELIHOOD OF CLINICAL RISK
GENE-ENVIRONMENT INTERACTION DIAGRAM

Fig: 1

Introduction

Although the cost of sequencing has decreased rapidly, there has not been enough motivation on utilizing the data effectively. There has been increasing focus on precision medicine, which is multi-layered health data being used for diagnosis by providing better accuracy compared to the broader goals defined by personalized medicine. Although it sounds similar to personalized medicine, this multi-layered data is modeled similar to the geographical data which is also multi-layered, instead defines positions in google maps. Therefore utilizing the publicly available data and analyzing the clinical risk on a simpler organism e.g., Drosophila Melanogaster (D.Mel) is the aim of this project. Assuming that genetic likelihood (genotype) is directly associated to the phenotype and not any other causes which is not ideally the scenario; this project tries to understand the percentage importance of any sample exposed to a disease. This approach tries to apply the pipeline using the model organism as D.Mel; however the end goal would be to apply it to a human genome which is a quite large and complicated in nature.

Background and Related Work

This idea was initially available in the Lancet article—Clinical assessment incorporating a personal genome [1]. The analysis has been based out of a human genome which is complicated and huge to analyze. The paper fails to provide enough explanation to the variations that the genetic profile is currently exposed to. There has not been enough evidence to prove the reliability of information based on such evaluation. However, it has been able to demonstrate the availability of technology to lead to such analysis. It is very difficult to work with human genomes because of their unmanageable size having 23 chromosomes and insufficient information at any given reference point. Below is the gene environment diagram that has been used as a result to display the risk factors versus environmental factors that affect each of these clinical risks :-

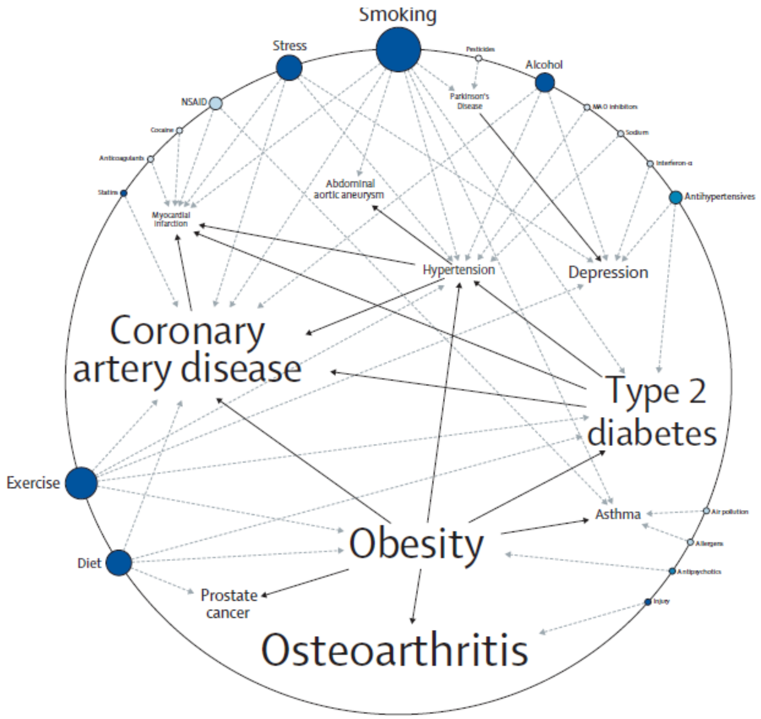


Fig: 2

About the model organism and data collection

The model organism is D.Mel with 14,000 genes and each gene occupying slightly less than 10,000 nucleotide sequences. Unlike the human genome having 23 pairs of chromosomes and billions of base pairs', D.Mel has around 6 different chromosomes. To acquire the datasets, SRAdb has been used for querying in R. Most of the search was manual due to lack of standards in the data in order to retrieve the desired results.

Role	Data source
Reference Genome	www.flybase.org[7]
Affected samples	SRA[8]
Test Samples	SRA, Bergman Lab[4]

Table: 1

Initial Idea

The initial idea is to use the affected sample, for each of the disease and then develop an individual model which at a later stage can be combined into a single model. The test samples then can be passed through the model to test for the various disease models. However, due to lack of availability of data, the test samples were derived only for single gene corresponding to a disease. Therefore, we arrive at a simpler evaluation technique.

Implementation

The pipeline of steps mentioned in Fig: 1, are divided into four steps. The pipeline shows the end to end process of the analysis, including the identification of the gene variants until the visualization of the gene-environment interaction.

Step 1-Identification of gene variants

The genome datasets for the 5 affected, 5 wild type and the 3 test samples were available in the SRA database and in [3] as NGS (Next Generation Sequencing) datasets. These datasets have been uploaded to GVSU-Quattro for further processing and were converted to fastq format. As per the steps mentioned in [6], each of the dataset was aligned and mapped to the reference genome available in flybase [7] using BWA (Burrows-Wheeler Aligner) aligner. This step helps in providing the dataset a reference with respect to the reference genome. It has been assumed that there are at least 4 matches for every read in the sequences for the datasets and the reference genome for any given read sequence. One of the test samples had data for the paired-end reads and therefore it has been mapped slightly differently using the BWA (sampe) instead of BWA (samse). The sorting and the indexing of the short reads was performed using the SAMTOOLS which produces a BAM file type as the output which is more usable for identifying SNPs'. Below is the Fig: 3 which is a BAM view of the bam file which has been viewed using BAM-Viewer [10]. The chromosomes listed in the figure are derived from the reference genome but also used to obtain the chromosomes that correspond to the aligned datasets. The gene coordinates as marked in the figure can be provided either as an input or can be dragged along the screen to view them by selecting the appropriate chromosome number. Each of the stacks represents a possible mapping of the reads to the reference genome. SAMTOOLS has also been helpful in identifying the differences among the datasets and the reference genome and for viewing them. These differences called SNPs' in the context of NGS such that these are only limited to the reference genome and help identify the potential differences after statistically calculating the likelihood of a specific nucleotide sequence being present at that location after several mapping alignments for that location. Each of the datasets consumed around 2 hours for converting and retrieving them to appropriate file. The more information available on the datasets, the more is the processing time.

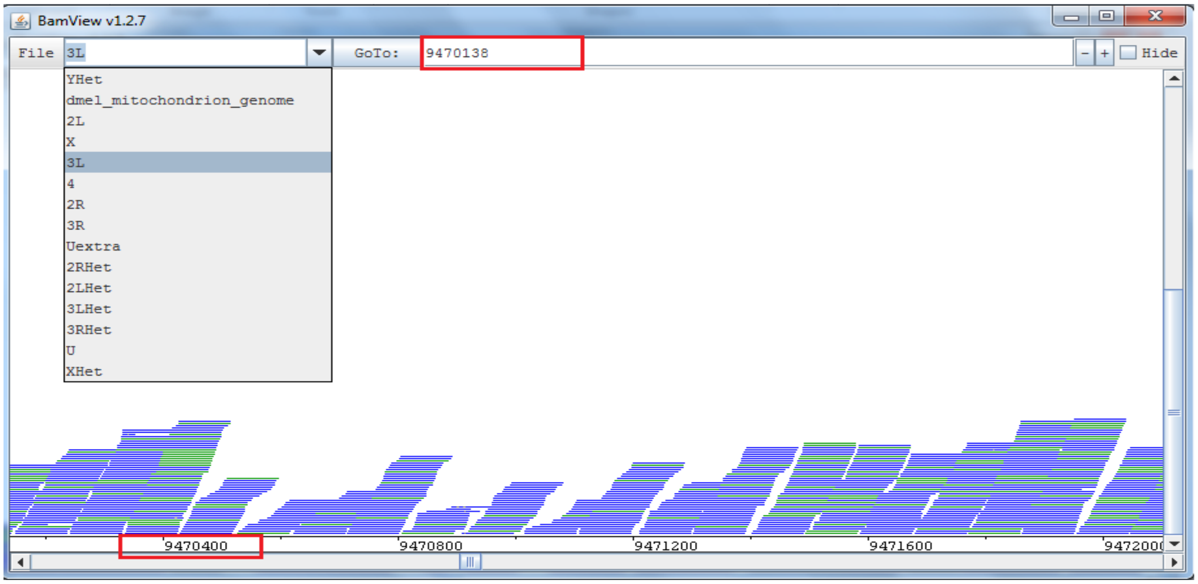


Fig: 3

Step 2: Query the associated genes

It has been observed that more than 75% of the human diseases causing genes have functional homologs in the fly. Therefore the following diagram shows that based on the diseases identified in the humans, the corresponding genes can be identified and then matched against the orthologous genes for D.Mel in order to obtain a similar function so that it can cause a similar effect on the phenotype. One such a gene that has been identified to cause autism namely Jarid2 in human beings has been identified to have caused autism in D.Mel. There has been experimental evidence that this protein coding gene is involved in the biological process for regulation of histone. Its gene sequence location in the reference genome is 3L: 9470138……9479630 according to flybase [7]. Since a fly genome is considered "isogenic", it has been assumed that every dataset belonging to the samples, start with the same position that contains the sequence for Jarid2.

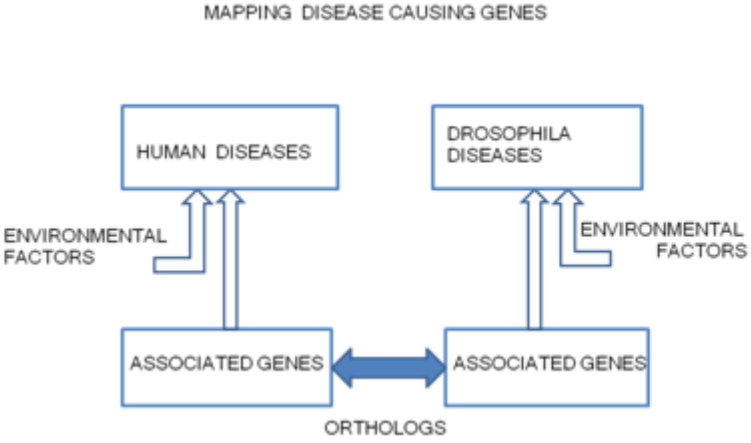


Fig: 4

Step 3:- Estimating the likelihood of clinical risk

Every dataset available has been calculated for the SNPs' (single nucleotide sequences') along with the sequence positions within the gene coordinates for Jarid2 using SAMTOOLS as mentioned in Step-1. The intersecting SNPs' among the 5 affected samples has been calculated using intersect_SNP.pl from biotoolbox [11] to identify the count and the location of them. There have been a total of 6 SNPs' that have been listed as the common intersecting ones. There was a total of 1 intersecting SNP among the 5 wild type samples that was also present in the intersecting list of the affected ones. Therefore this SNP was ignored. Below is a list of the SNPs that were in the mutant samples and not in the wild-type:-

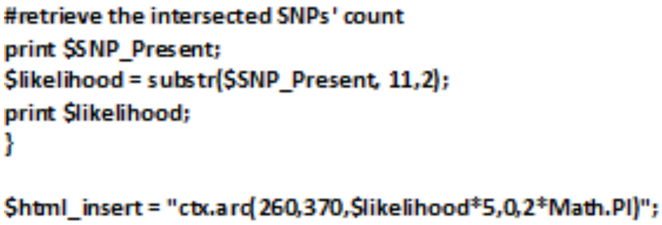
Chromosome	Relative Location of the change	Reference	Altered	Quality of the result
3L	9474046	C	T	210
3L	9474061	G	A	222
3L	9474142	C	T	128
3L	9474220	G	A	71
3L	9478656	G	A	77

Table: 2

This list has been used to test the three test samples for autism. Perl script, intersect_SNP.pl has been rerun to check for the intersecting SNPs among the existing list and the SNPs' for the test samples and as used for the next step to compute the likelihood.

Results

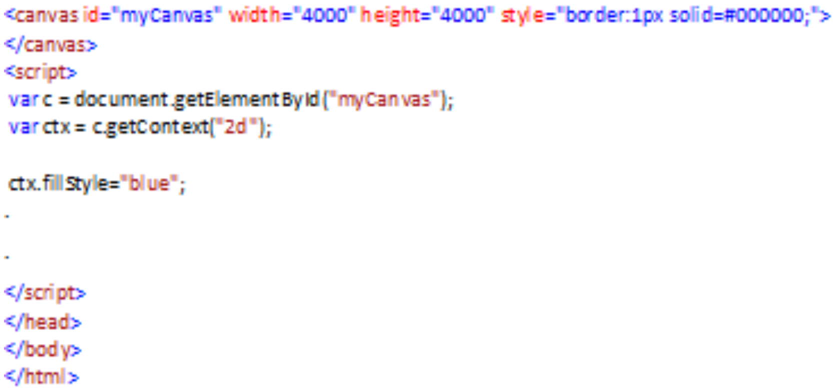
Results can also be considered as the step-4 of the pipeline which includes visualizing the likelihood of the clinical risk with respect to the environmental factors for any given test sample. Below is the diagram for the gene-environment interaction as in Fig 5. There has been an interface script implemented using PERL to generate this HTML file from the resulting SNPs from the intersect_SNP.pl available in biotoolbox. Each of the bubbles on the circle represents the exposure of an organism to the environmental factors. Information provided as in [2] states that the drosophila when exposed to the combined effect of chemicals and centrifugation had changes in the enriched gene-expressions of Jarid2 gene in Drosophila. This has been listed as the environmental factor for this example. Within the circle are the possible neurodegenerative and the neurological diseases that research experiments describe along with the gastric cancer in Drosophila. In case of the test sample having autism, a bubble appears against it to represent that it has been affected as shown in Fig: 5. This visualization has been implemented using HTML5 using canvas to draw the bubbles.



```
#retrieve the intersected SNPs' count
print $SNP_Present;
$likelihood = substr($SNP_Present, 11,2);
print $likelihood;
}

$html_insert = "ctx.arc(260,370,$likelihood*5,0,2*Math.PI)";
```

In the above piece of code, the number of intersections of the SNPs’ are extracted from the $SNP_Present variable using the “substr” command to locate them. This variable is stored as likelihood and is inserted as a string with the exception of likelihood passed as a number. The HTML5 <canvas> is used to draw graphics, on the fly, via scripting [12]. The bigger the bubble, the higher is the likelihood of risk in the sample.



```
<canvas id="myCanvas" width="4000" height="4000" style="border:1px solid=#000000;">
</canvas>
<script>
 var c = document.getElementById("myCanvas");
 var ctx = c.getContext("2d");

 ctx.fillStyle="blue";
.

.
</script>
</head>
</body>
</html>
```

The above code for canvas in HTML5 is embedded as a script and therefore all the graphics do not need other extensions or support to draw them. The whole code can be produced, without requiring much learning and the visualization is pretty appealing.

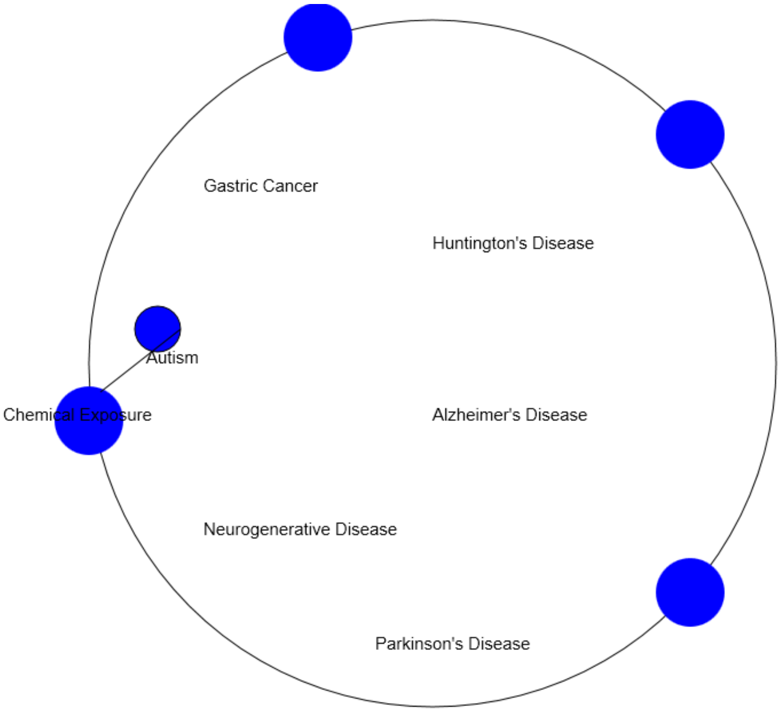


Fig: 5

Conclusions and Future Work

This demonstration has lead to interesting facts of the phenotype of a sample based on the genotype alone. Due to unavailability of data, the example demonstrates the likelihood of the test sample having a disease based on only a single gene that is likely to cause the disease. This work can be extended to using all the possible orthologs of human disease genes in flies to improve accuracy in obtaining the likelihoods. This project can also be extended to work with combination of environmental factors, genotype (for disease causing genes) and protein structures to demonstrate that it can be possible to experiment and bring the precision medicine to better implementation.

References

1) Ashley EA, et al. *Clinical assessment incorporating a personal genome.* Lancet. 2010 May 1; 375(9725):1525-35.
2) Pandey UB, Nichols CD. *Human Disease Models in Drosophila melanogaster and the Role of the Fly in Therapeutic Drug Discovery*, Pharmacol Rev. 2011 Jun; 63(2):411-36.
3) Haddrill, P. and C.M. Bergman (2012) *20 Drosophila melanogaster genomes from Montpellier, France*.

4) Hans-Martin Herz, Man Mohan. *Polycomb Repressive Complex 2-Dependent and –Independent, Functions of Jarid2 in Transcriptional Regulation in Drosophila,* Mol Cell Biol. 2012 May; 32(9): 1683–1693. doi: 10.1128/MCB.06503-11
5) http://flydiseasemodels.blogspot.com/search/label/Neurodegenerative%20Disease (Research articles on D. Mel)
6) http://ged.msu.edu/angus/tutorials-2011/snp_tutorial.html (tutorial for NGS)
7) www.flybase.org (information regarding Drosophila Melanogaster and the reference genome)
8) http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP009229 (affected samples and wild type samples)
9) http://samtools.sourceforge.net/samtools.shtml#1 (documentation for SAMTOOLS)
10) http://bamview.sourceforge.net/ (BAM view for viewing the .bam files)
11) http://code.google.com/p/biotoolbox/wiki/ProgramList (biotoolbox-intersect_SNPs)
12) http://www.w3schools.com/html/default.asp (HTML5 tutorial)

Acknowledgements

I would like to thank Dr.Guenter Tusch, for his immense support in helping me layout this project from the initial phase. I would also like to thank Dr.Greg Wolffe for sharing his ideas and the permission for accessing and sharing the access time with GVSU-Quattro machine. Dr. Martin Burg played a key role in helping me with the biology component for this project especially with the information regarding D.Mel and also helped me not to fall into ditches even before I can identify them. Ira Woodring had uploaded data for NGS datasets into Quattro allowing speeding up the process along with the installation of required software. Yuka Kutsumi volunteered to help me with some of the essential biology required to solve the analysis. I would finally thank GVSU for allowing me to work and get support to work with Quattro without any issues.