Software Proposal Document for project Utilizing human genomes analysis to improve athletic performance

Mennat Allah Hisham, Donya Mohamed , Mazen Mohamed ,Dannel Maged

October 25, 2020

Proposal Version	Date	Reason for Change
1.0	24-October-2020	
2.0	27-October-2020	

Table 1: Document version history

GitHub: https://github.com/mazenmohammed/gradproject

Abstract

The main idea of this project is to combine the artificial intelligence and genomics to make a system using the genome analysis such as DNA sequences or gene expression to identify physical points of strength and points of weakness.This may help any one to understand his/her traits ,and how it affect his/her sport life .The system will recommend the suitable sport based on his/her points of strength And will recommend some exercises and a nutrition plan to enhance his/her points of weakness therefore his/her performance will be improved. In Addition ,We will also set a group of genes responsible for causing some disease using artificial intelligence algorithms to predict future diseases, so they can avoid risks,and increase their chances to reach the peak of the competition.Moreover, We are attempting to determine the muscle's tissues type whether it's Constriction or extension,(fast or slow) use machine learning techniques .Our challenge is to help athletes to improve their performance and the non athletes to choose the sport that will match their body features.

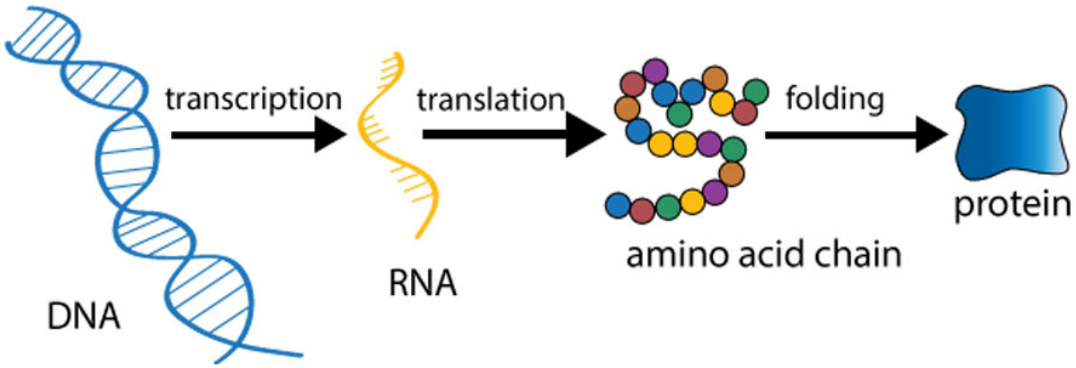
1 Introduction

1.1 Background

Its well known that the cells is the main building unit of the human body,and inside the nucleus of the cell there is the DNA. deoxyribonucleic acid (DNA) is a large molecule that contains the genetic code ,it also holds the instructions for building the proteins which are responsible for body functions. The genetic code is made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T)[3]. It store the information of DNA as the sequence of these bases determines the information for building an organism.As

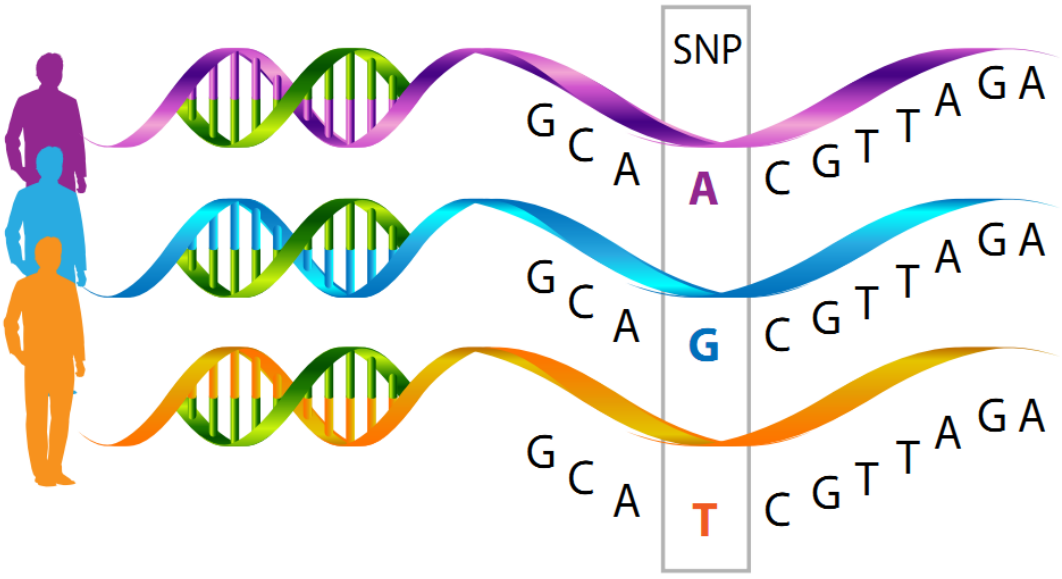
we mentioned before, The DNA is responsible for the making of protein , a there is special chemicals inside the nucleus make copies of DNA code called RNA. The RNA gets out of cytoplasm to ribosome which is the protein machine (where the proteins are made), the ribosome change the amino acid which made by RNA to proteins according to the code .Moreover, gene intensity could be measured by gene expression using microarrays. all humans share nearly 99% of genes, but what makes us different?

Figure 1: This is how DNA produces protein.



The difference in DNA is called mutation , there are types of mutation such as insertion,which is adding base pairs to the DNA, and there is also the deletion, it is the opposite of insertion as a section in DNA was deleted, and the last type is substitution(SNPs) which is exchange of one letter(A,T,G,C) in the codon with another. The substitution may lead to forming codon that encodes for another amino acid named "Missense mutation" , so it changes the protein. It may encodes for the same amino acid so it make no different in protein and its called "silent mutation" .finally , the change may lead to encode a stop codon "Nonsense mutation" which will stop forming the protein.
[8]

Figure 2: SNPs



SNPs are the most common type of genetic variation found among people , it averagely found in every 1,000 nucleotide ,it makes difference in shape and also not observed difference.So, there is an impact of

genetics on sporting performance and every person have different functionality due to his genes and also the environment. [7]

1.2 Motivation

1.2.1 Academic

Nowadays, many parents encourage their children to practice a specific sport for 2 or 3 days per week as practicing sports helps them build their bodies ,keep them fit and healthy and in every sport we find some children are doing very well in practicing a sport and others are weak or get tired quickly or even get injured, for this problem we decide trying to fix this problem by detecting the weakness and strength points of the child/player through his DNA and recommend the suitable sport for him/her.No doubt that every community believes that genetic factors contribute in the athletic performance. As in 2009, hundreds of genetic variants had been associated with physical performance, with more than 20 variants being in link with sport. As some studies examined the relation between genetic factors and sport performance in children or adults, this field of researches is more relevant to children.Prediction of future success of athletes in sports through genetic testing significantly increased recently . The history of our problem started with professional athletes who had some genetic diseases. Lou Gehrig was one of the best baseball players who has ended his sports career due to his infection with ALS (Amyotrophic lateral sclerosis).this is a fetal and generic disease. he was loved by many people so they called the disease with his name.Also Mohamed Ali Klai got infected with Parkinsons's disease . he was infected in 1984 that disease affects the central nervous system . He died in 2016 after experiencing a respiratory illnesses. Our project is interesting because by examining that player or athlete (getting his genes) that are responsible for speed,endurance and strength then get genes average expression for those athletes , by getting new data we compare between them and the average expression so by this way we can know what are the strength and weakness points of this player and we recommend some exercises,nutrition diet to improve his weaknesses.[1]

1.2.2 Business

We have done a survey to know the clients and users opinion about our project and we asked if they see it helpful .The results of the survey actually support us to do this project,mostly all the answers was that our project will help .we mentioned the survey and the analysis below in the supportive document.

1.3 Problem Statement

Many athletes start playing a sport but then they feel that it's not suitable for them , and they spend alot of money and time to play that sport.The problem is that the athletes don't know their abilities and don't know how to improve their skills and we aim that this project helps them.

2 Project Description

2.1 Objectives

• We will make it easier for all the people who are interested in sports to know their points of weakness and points of strength.
• Our system will recommend the suitable sport based on the over expressed genes , and recommend exercises to improve their points of weakness based on the under expressed genes.
• We will use machine learning techniques in helping Athletes improve in their fields.

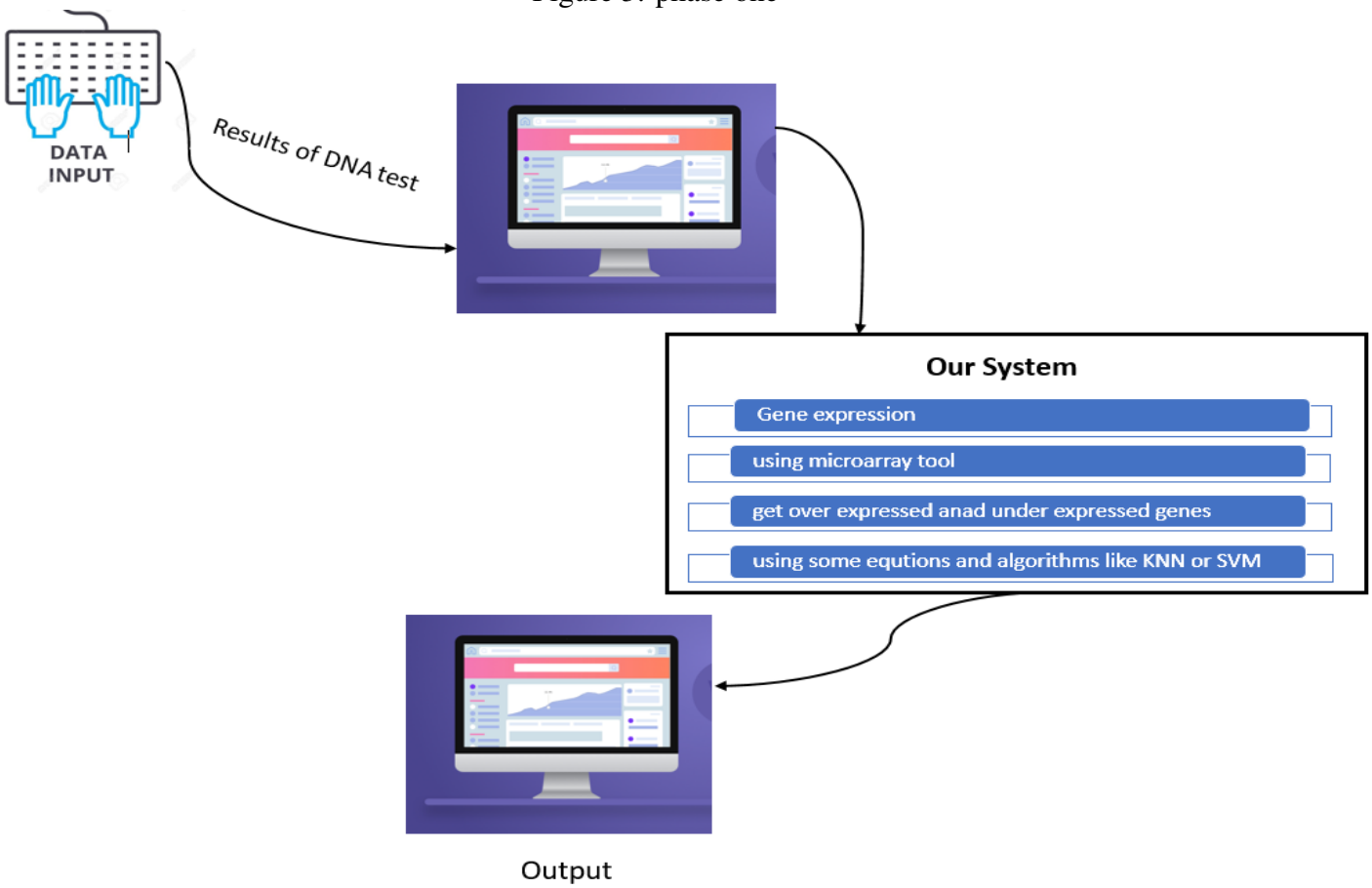
• We will early predict some of the diseases that can affect the players.

2.2 Scope

This system is mainly designed to help the athletes to know their abilities so they can choose the suitable sport and will not be needed to waste their time or money in the wrong place .Also, it will helps him/her know his/her points of weakness so,he would be able to improve himself. This system is developed to take the DNA from the players and push it in the system and by the machine learning techniques it will show him which sport is suitable for this person, also it will warn him about the predicted diseases if found. The Proposed system can:
1- It lets the user know his point of strength so , he will be able to choose the suitable sport
2- Will show him his points of weakness so he can improve his performance.
3- Can predict some of the diseases like lung cancer .

2.3 Project Overview

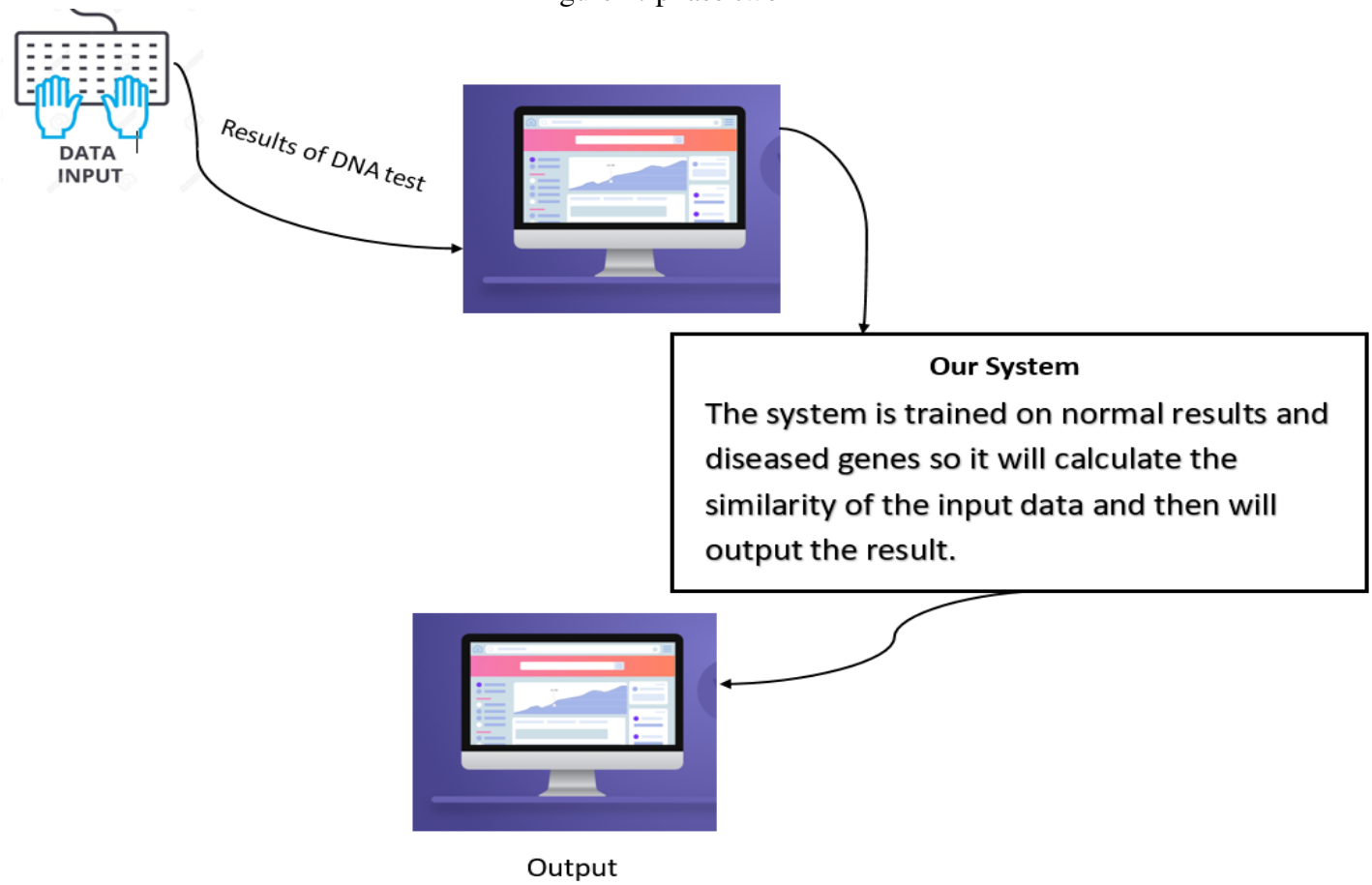
Figure 3: phase one



First of all our project consists of two main phases ,the first one is detecting points of weakness and point of strength.The system will work on 4 or 5 traits and from the DNA test will take the gene expression of the genes related to these traits ,and we will set an average standard of the gene's expression and by using tools

and some equations the entering data will divide into over expressed and under expressed. From knowing the over and under expressed genes,we can know the points of strength and points of weakness.Also the system will recommend some exercises to improve the point of weakness and can recommend a suitable sport based on the points of strength.The second phase is detecting whether the person carries a specific genes related to any disease. We will train our model on many data of diseased people and normal people ,so it can check whether this person is a carrier of the genes that may cause any of these disease.Our future work is to make our scope wider , to work on more traits and recommend more sports.

Figure 4: phase two



2.4 Stakeholder

2.4.1 Internal

- Mennat Allah Hisham is the Team leader
- Donya Mohamed Mohy ElDien
- Mazen Mohamed Fouad
- Dannel Maged

2.4.2 External

The End Users and clients will be doctors and hospitals. And mainly athletes and any one interested in sports. Also this system would be very helpful to many popular clubs as "El-Ahly" and "El-Zamalek" as

they will use this system to improve their players performance.

3 Similar System

3.1 Academic

1. **A Potential Endurance Algorithm Prediction in the Field of Sports Performance. [5]**
   The researchers wanted to design a genetic endurance prediction score (GES) of endurance performance and analyze its association with anthropometric, nutritional, and sport efficiency variables in a cross-sectional study within fifteen male cyclists. Their goal was to facilitate the finding of genetically talented athletes, improve their training and food habits, as well as help in the improvement of the physical condition of amateurs. In this study, the volunteers attended two different centers to complete the study: The Sports Medicine University Center to carry out a maximal incremental treadmill test and The Research Institute on Food and Health Sciences "IMDEA Food" (Madrid, Spain) for anthropometric measurements, body composition analysis, dietary records, DNA collection and genotyping. The data were analyzed using the R Statistical Software Version 3.4.1 The Mann–Whitney U test was used to check for significant differences in the continuous variables (not always normally distributed) for the different genotypes. The Spearman correlation coefficient was used for the association between the algorithm and the other variables. The Bonferroni correction was also applied to control against type-I errors for multiple tests. The most important result was the potential validation of an algorithm prediction of genetic susceptibility to endurance abilities. The outcomes of the present study confirm a positive relationship between an endurance prediction algorithm and the results of a cardiopulmonary exercise test.

2. **Muscle Gene Sets: a versatile methodological aid to functional genomics in the neuromuscular field. [4]**
   The approach of building large collections of gene sets and then systematically testing hypotheses across these collections is a powerful tool in functional genomics, both in the pathway analysis of omics data and to uncover the polygenic effects associated with complex diseases in genome-wide association study. The Molecular Signatures Database includes collections of oncogenic and immunologic signatures enabling researchers to compare transcriptional datasets across hundreds of previous studies and leading to important insights in these fields, but such a resource does not currently exist for neuromuscular research. In previous work, we have shown the utility of gene set approaches to understand muscle cell physiology and pathology.

3. **Sport and exercise genomics: the FIMS 2019 consensus statement update [6]**
   This SWOT analysis and the developed guiding reference highlight the need for scientists/clinicians to be well-versed in ethics and data protection policy to advance sport and exercise genomics without compromising the privacy of athletes and the efforts of international sports federations. Conducting research based on the present guiding reference will mitigate to a great extent the risks brought about by inappropriate use of genomic information and allow further development of sport and exercise genomics in accordance with best ethical standards and international data protection principles and policies. This guiding reference should regularly be updated on the basis of new information emerging from the area of sport and exercise medicine as well as from the developments and challenges in genomics of health and disease in general in order to best protect the athletes, patients and all other relevant stakeholders.

4. **Towards utilization of the human genome and microbiome for personalized nutrition. [2]**

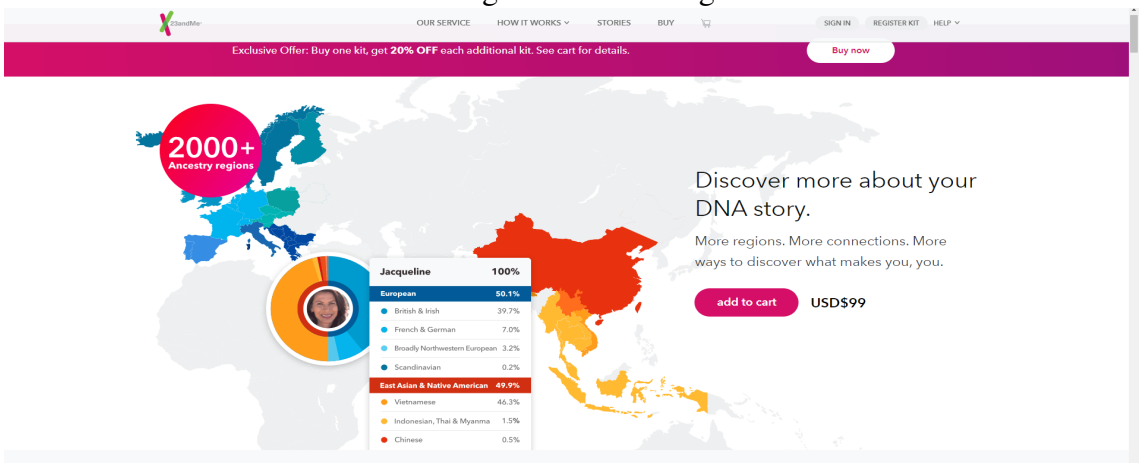
The researchers were trying to make personalized nutrition as it has become clear that the one-fit-for all diet approaches does not work and that there is an outstanding difference in the inter-individual responses to the diet and lifestyle interventions.

3.2 Business Applications

23andMe is a site that make a DNA test for users and it give them many results about their features , traits and family."https://www.23andme.com/en-int/"

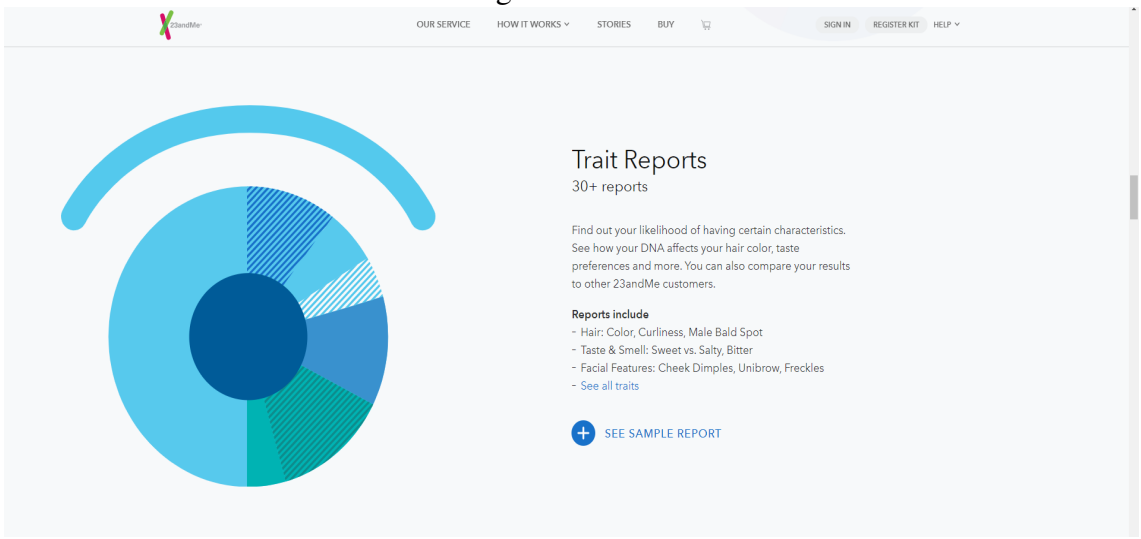
23andMe's mission is to help people access, understand and benefit from the human genome. Based in

Figure 5: Home Page of the website



Sunnyvale, California, the company currently employs well over 500 people and ships its product to more than 50 countries worldwide. The company uses data to revolutionize health, wellness and research. They want to improve healthcare. They want to prevent disease. And also to give individuals control over their health data. They want to dramatically accelerate the pace of research and also want to develop better drugs smarter and faster.

Figure 6: one of 23andMe service



4 What is new in the Proposed Project?

All the papers and researches similar to our project are mainly works on one trait, but we are attempting to work on many properties simultaneously. So by having the DNA test our system will work on genes of 4 or 5 property .

5 Proof of concept

We have done a trial using data set of DNA of group of people have covid 19,severe acute respiratory syndrome (SARS) and Zika virus disease ,just to proof that we can deal with biomedical data and by using machine learning with python. The out put was measuring the similarity between the diseases.Our demo is using python and an algorithm called SVM.

6 Project Management and Deliverable

6.1 Deliverables

Figure 7: Deliverables in brief details

Deliverable	Description	Deadline
Proposal Document and presentation	The main reason for the proposal is to explain why we choose the proposed system, what we added on the previous researches or projects and what we will do in details	24-Oct
Implementation	A 10% implemented from our system	27-Oct
Software Requirement Specification document (SRS)	it will include a detailed description of the developed software.	Third week of December
Implementation	A 35% implemented from our system	Third week of December
Software Design Description (SDD)	it mainly contains data design , architecture design and interface design .	Third week of February
Implementation	A 65% implemented from our system	Third week of February
System(Prototype)	80% of implementation	Last week of April
Contribution Paper		
Implementation	90% of implementation	Last week of May
Implementation	100% of implementation	Last 10 days in June
Thesis Document		Last 10 days in June

6.2 Budget and Resource Costs

There isn’t any Purchases.

6.3 Tasks and Time Plan

Figure 8: Project Timeline

Task	From	To	PERCENT COMPLETE
Discussion about the ideas	25-Jul	14-Sep	100%
Researches about the idea	05-Aug	01-Oct	100%
collecting of dataset	03-Sep	11-Oct	80%
Writing proposal	28-Sep	22-Oct	100%
Implementation	30-Sep	22-Oct	10%
Proposal Phase 1	24-Oct		100%
Proposal presentations	27-Oct		
collecting of dataset	30-Oct	10-Nov	20%
Writing SRS	10-Nov	14-Dec	100%
Implementation	01-Nov	14-Dec	35%
Pesentation of SRS	14-Dec	21-Dec	
Writing SDD	25-Dec	14-Feb	100%
Implementation	01-Jan	14-Feb	65%
Presentation of SDD	15-Feb	21-Feb	
Implementation	10-Mar	20-Apr	80%
Writing paper	29-Feb	02-May	100%
Deliver Paper			
Testing			
Implementation	21-Apr	21-May	90%
Implementation			100%
Thesis	01-May	15-Jun	100%
Final Presentation			

7 Supportive Documents

• users/survey

As we mentioned before , we have done a survey ,and after the analysis of answers we found that most of the answers support us to make this project. And here are the Questions and the Analysis of the answers.

Figure 9: Questions number 1,2 and 3 of the survey.

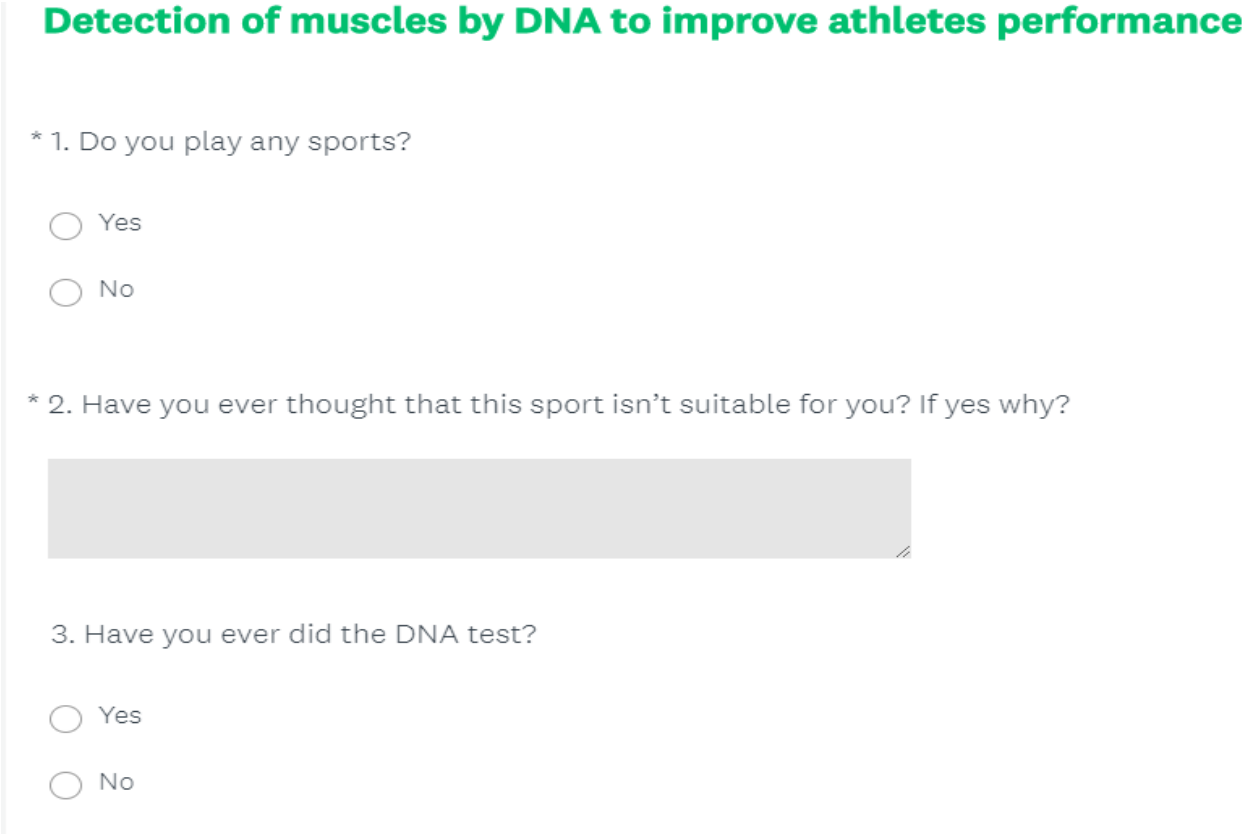


Figure 10: Questions number 4 and 5.

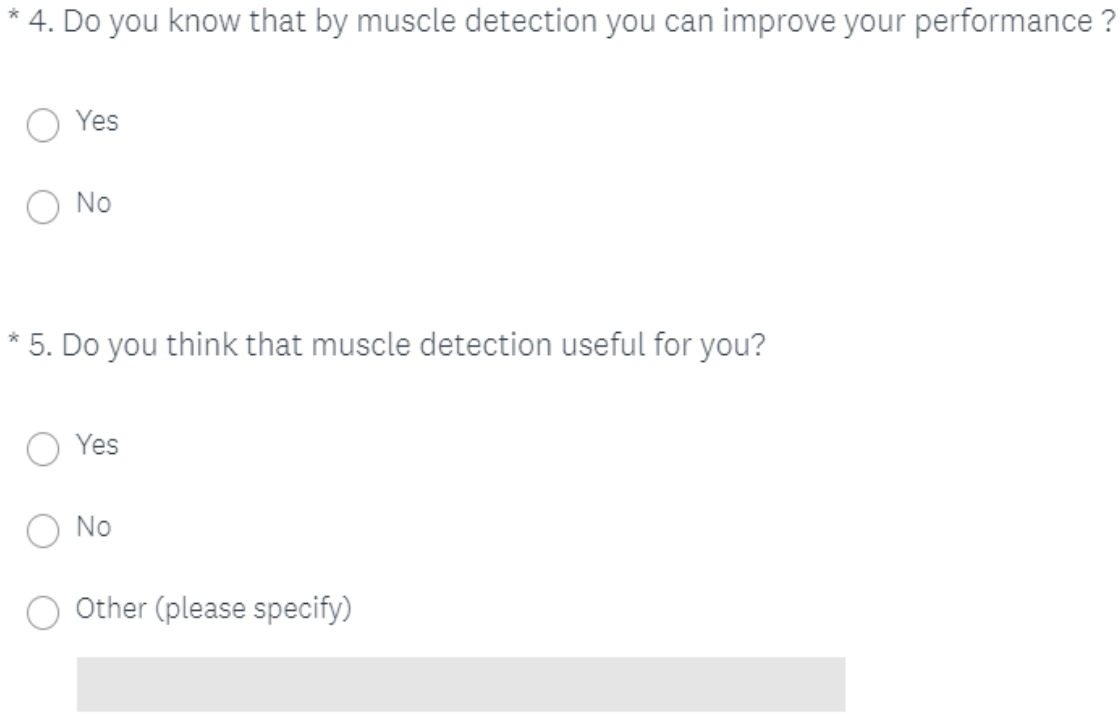


Figure 11: Questions number 6,7 and 8.

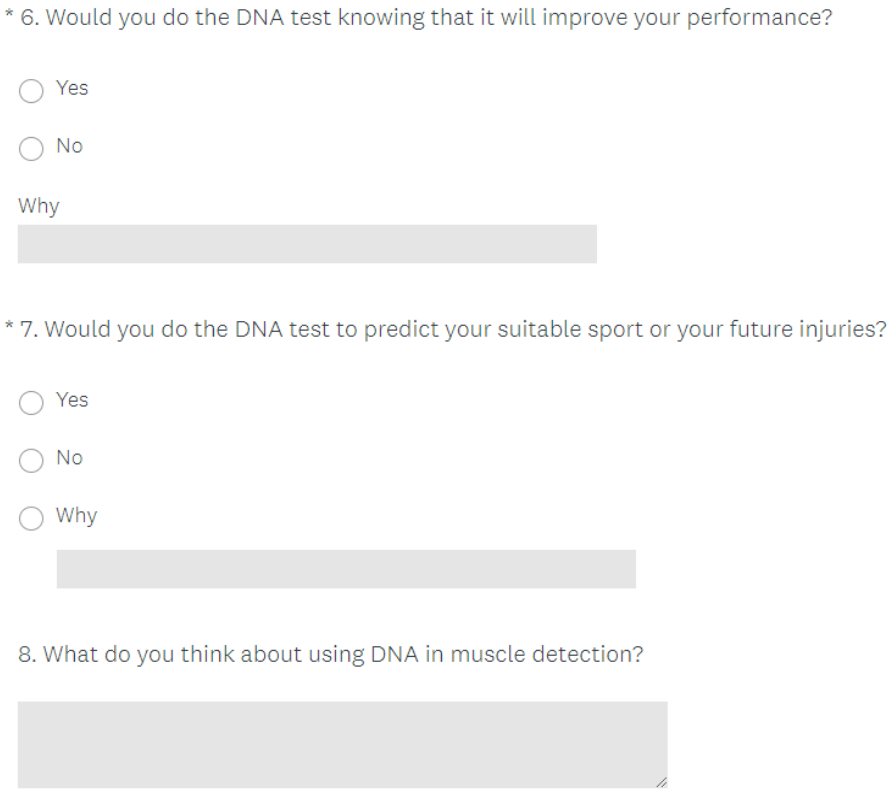
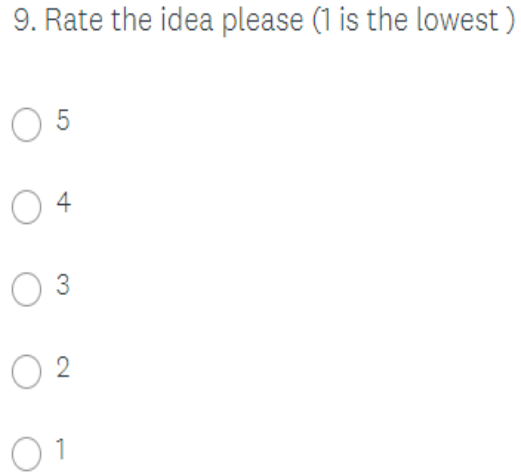


Figure 12: We asked the users to rate our idea.



In the second question many people answered that the sport they played wasn't suitable for them for not being sprint or for not being able to stand the exertion. The fourth Question they were asked if they know that the DNA detection can improve their performance , 62.5% answered that they didn't know .In the fifth question, 80% of the answers was that they think that the muscle detection will be useful for them and 81.25% answered that they would make the DNA test to make them choose the suitable sport. And for the 8th question , most of the answers was that they will do the DNA test to detect their muscles but the problem is the high cost of the DNA , but all the opinions was that if any famous athlete has the opportunity

will make the DNA to improve himself.We also asked about their opinion of the idea 50% of the answers gives the idea 5/5 , 31.25% give it 4/5 ,12.5% give it 3/5, 6.25% give it 2/5 and 0% give it 1/5, And the summary of the survey was that the system will be helpful for most of the athletes and if anyone has the ability to do the DNA test will use our project.

8 References

References

[1] São Paulo. *Scientific Electronic Library Online*. 2014. URL: https://www.scielo.br/scielo.php?Script=sci_arttext.

[2] Stavros Bashiardes et al. "Towards utilization of the human genome and microbiome for personalized nutrition". In: *Current opinion in biotechnology* 51 (2018), pp. 57–63.

[3] Tim Newman. *DNA explained: Structure and function*. Jan. 2018. URL: https://www.medicalnewstoday.com/articles/319818.

[4] Apostolos Malatras, Stephanie Duguez, and William Duddy. "Muscle Gene Sets: a versatile methodological aid to functional genomics in the neuromuscular field". In: *Skeletal muscle* 9.1 (2019), pp. 1–12.

[5] Rocio de la Iglesia et al. "A Potential Endurance Algorithm Prediction in the Field of Sports Performance". In: *Frontiers in Genetics* 11 (2020), p. 711.

[6] Kumpei Tanisawa et al. "Sport and exercise genomics: the FIMS 2019 consensus statement update". In: *British Journal of Sports Medicine* (2020).

[7] *What are single nucleotide polymorphisms (SNPs)?: MedlinePlus Genetics*. Sept. 2020. URL: https://medlineplus.gov/genetics/understanding/genomicresearch/snp/.

[8] *What kinds of gene mutations are possible?: MedlinePlus Genetics*. Sept. 2020. URL: https://medlineplus.gov/genetics/understanding/mutationsanddisorders/possiblemutations/.