Curation of Genetic Variants inComplex Diseases And Traits

Dissertation submitted in partial fulfilment for the degree of

Master of Science in Biotechnology

Submitted By

Shrestha Sen Gupta

1661024

KIIT School of Biotechnology, Campus- 11

KIIT deemed to be University

Bhubaneswar, Odisha, India

Under the Supervision of

Dr.Thenral S. G.

Associate Scientist

Medgenome Labs Ltd.,

Bommasandra, Bangalore - 560099,

Karnataka, India

MedGenomes Lab Ltd.

CERTIFICATE

This is to certify the dissertation entitled " ***Curation Of Genetic Variants In Complex Diseases And Phenotypic Traits***"Submitted by***Shrestha Sen Gupta***in partial fulfilment of the requirement for the degree of Master of Science in Biotechnology, KIIT School of Biotechnology, KIIT deemed to be University, Bhubaneswar bearing Roll No. 1661024 &Registration No.16666351680 is a bonafide research work carried out by his/her under my guidance and supervision from ***08-01-2018***to ***11-05-2018 .***

Date: 11.05.2018
Place: Bengaluru

Dr. Sakthivel Murugan S.M.
(Associate Director, Operation)

MedGenome Labs Ltd.

CERTIFICATE

This is to certify that the dissertation entitled "***Curation Of Genetic Variants In Complex Diseases And Phenotypic Traits***" submitted by *Shrestha Sen GuptaRoll No. 1661024 Registration No.* 16666351680to the KIIT School of Biotechnology, KIIT deemed to be University, Bhubaneswar-751024, for the degree of Master of Science in Biotechnology is his original work, based on the results of the experiments and investigations carried out independently by his/her during the period from***08-01-2018*** to***11-05-2018*** of study under my guidance.

This is also to certify that the above said work has not previously submitted for the award of any degree, diploma, fellowship in any Indian or foreign University.

Date: 11.05.2018
Place: Bengaluru

Dr.Thenral S.G.
(Associate Scientist, Operation)

DECLARATION

I hereby declare that the dissertation entitled "***Curation Of Genetic Variants In Complex Diseases And Phenotypic Traits***" submitted by me, for the degree of Master of Science to KIITdeemed to be University is a record of bonafide work carried by me under the supervision of, ***Dr.Thenral S.G., Associate Scientist, Medgenome, Bangalore, Karnataka, India***.

Date: 11.05.2018

Place: BengaluruShrestha Sen Gupta

1. Abstract

Curation is a systematic process of extracting information from scientific texts, such as research articles. It is a broad term which is generally used to facilitate processes and activities related to an organization and combination of data collected from variety of sources, annotation of the data, publication and presentation of data, in such a way that the value of the data is maintained and available for further use. Structured data enables access to comprehensive information which is efficient to extract and less prone to errors. The technological revolution in the recent years in sequencing and information technology has led to enormous amount of genomic data published asbiomedical literature, relevant to the study and understanding of human diseases.

The curation of genetic variants based on standard nomenclature with evidence necessary to support an association between a gene and a particular disease is a valuable resource for interpretation for both complex diseases and Mendelian diseases.Comprehensive curation including the frequency, prevalence, ethnicity, clinical features, segregation studies, inheritance patterns, in silico prediction, and functional studies contribute to ascertaining the variant's significance. Around 143 publication related to phenotype traits were curated in accordance to the structured format. In addition founder mutations specific to Indian population was also curated. It was observed that most of the data available for complex traits were based on studies in the western population compared to those available for Indian population. The variants were documented based on standard nomenclature and all available evidence documented.

2. Acknowledgements

I would like to express my sincere gratitude to my advisor and reporting manager Dr.Thenral S. G. for her continuous support in my work and dissertation project, patience, motivation, enthusiasm and immense knowledge. Her guidance helped me in all the time of my project and writing my thesis.

Beside that, I would also like to thanks my team members, Udita Mahadevia, Aishwarya Subramaniyan and Bhuvandeep Narangfor their encouragement, insightful comments.

My sincere thanks also goes to Dr. Sakthivel Murugan S. M., Associate Director, MedGenome, for offering me the opportunity to do my dissertation project from this organization.

I would also Dr.MrutyunjaySuar, Director, Kiit School Of Biotechnology, for allowing me to do my project from this organization.

Last, I would like to thanks my friends and family members for supporting me in every step of my life.

Date: 11.05.2018
Place: BengaluruShrestha Sen Gupta

Table of Contents

Abbreviations

DNA	Deoxyribonucleic acid
CNV	Copy Number Variation
SNP	Single Nucleotide Polymorphism
ACMG	American College of Medical Genetics and Genomics
VEGFA	Vascular Endothelial Growth Factor A
OMIM	Online Mendelian Inheritance in Man
ACL	Anterior Cruciate Ligament
LOVD	Leiden Open Variation Databse
SIFT	Sorting Intolerant From Tolerant
LRT	Likelihood ratio Test
UCSC	University of California, Santa Cruz
MAF	Minor Allele Frequency
NGS	Next Generation Sequencing
FISH	Fluorescence In Situ Hybridization
PCR	Polymerase Chain Reaction
cDNA	Complementory Deoxyribonucleic acid
HGMD	Human Gene Mutation Database
NCBI	National Center of Biotechnology Information
DCN	Decorin
BGN	Biglycan
KDR	Kinase Insert Domain Receptor
ATP7B	ATPase Copper Transporting Beta
G6PD	Glucose 6 Phosphate Dehydrogenase

List of Figures

Similarly you can automatically generate a list of figures from paragraphs of style. This contains the list of all figures with a short legend and page no in the right margin.

1. Aim

Retrospective comprehensive curation of genetic variant data in a structured format for both complex traits as well as Mendelian diseases from biomedical literature for Genotype-phenotype correlation.

2. Objectives

2.1 Objective 1

1. Literature mining using keywords for specific complex traits / Mendelian diseases using multiple search tools to acquire relevant publications.
2. Selection of the literature for curation based relevance and authenticity
3. Curation of data for genetic variants, demographic details and disease phenotype in a structured format.
4. Standardizing with the curated variants with respect to GRCh37 and HGVS nomenclature.

2.2 Objective 2

1. Detail documentation about patient, history, consanguinity, inheritance, and demographic details
2. Detailed variant annotation (gene, cDNA, protein change) and *in silico* predictions and population frequency forthe variant.
3. Classification of clinical variant according to ACMG guidelines
4. Quality check of the curated data by identifying inconsistencies in the data (pathogenic variants at high frequencies), conflicting variant interpretations and annotations.

3. INTRODUCTION:

The biomedical research articles are increasing at a rapid rate, which is making the task of knowledge retrieval and extraction more difficult. Tools that give means to search and mine full text of literatures or research articles thus represent an important way by which the efficiency of these process can be improved.

Although procedures, institutions and ability for preserving and circulate digital information have been known and in some disciplines for several decades, for recognizing that assemblage of practices is very important to fully establish the field of digital curation.

Curation is a broad term which is generally used to facilitate processes and activities related to an organization and integration of data collected from variety of sources, annotation of the data, publication and presentation of data, in such a way that the value of the data is maintained and should be available for further use and preservation. Curation refers to "all the procedure needed for principled and controlled data creation, maintenance and management, together with the capacity to add value to data."

Basically, it is a process of extraction of essential information from scientific texts, such as research articles by specialists to be transformed into an electronic format such as entry of biological databases.

The first examples of curated content emerged in Renaissance Europe, five centuries ago, in the form of newsletters. Handwritten newsletters circulated privately among merchants handwritten newsletters were circulated confidentially, for passing information about everything from economic conditions to social customs.

Curated data can be used for both clinical and research purposes. Curation can help an analyst to interpret result byweighing the evidence to determine the significance. All the information mentioned in a scientific paper is not necessarily compiled. Thus, only necessary information of relevance to the objective and required by analysts are documented in a structured format which can be used as standard during analysis. Availability of comprehensive information facilitates the downstream process of variant interpretation of the analyst by reducing time and effort to collate data for individual cases.

The below listed are a few basic rules followed for scientific literaturecuration

- Selection of specific key words for the topic of interest and search using in online database tools to obtainrelevant literatures

- Excerpt all information related to the variants of interest including the authors' conclusion on the significance of the variant with respect to the phenotype studied. The required information is documented in standard format.

A variant is any change in the gene sequence. A variant can be a mutation, can be SNP or can be CNV. On an average variant occurs once after each 300 base pairs in DNA. This change can either be harmful or disease causing or polymorphic. If it is disease causing it is considered as pathogenic. A benign / polymorphic variant is may not be responsible for effect on the phenotype but may contribute to the phenotypic traits such as -eye colour, skin pigmentation, vitamin level etc.

Phenotypic characteristics of an individual may be inherited or influenced by environmental conditionsor occurs as a combination of both.Traits controlled by genes, the observable trait depends on the normal expression of gene. Thus even in complex phenotypic traits the genetic variants may affect the overall outcome of the phenotype. Thus risk variants in complex traits can help us determine the role of genetics in determining that particular trait. For exampleeye colour, skin pigmentation, behaviour, vitamin level, taste, smell, deep sleep, alcohol flush reaction etc. For some phenotypic traits there may be major modifier of the phenotype in the gene sequence responsible. Thus particular variation in the gene may result in variation in the trait. For example, HERC2 is one of the gene involved in eye colour. Expression of this gene results in the production of P protein. Now individuals with A allele in the desired gene has increased production of P protein which lead to brown coloured eye in an individual in almost 80% of the individuals with this genotype. The alternate allele which results in base change from A to G causes reduced production of P protein, thus resulting in blue coloured eyes. Hence, mutation in a particular gene associated with any trait of an individual, may lead to variation in in the trait.

Founder mutations are a specific mutation arisen in an individual of a closely knit community and then pass down to the subsequent generation in the same population. Thus these variants are specific to a sub population or in a community. This founder effect is caused due to genetic drift. As this small population gets separated from the main population due to inbreeding/ consanguinitythe frequency of this mutation will be higher compared to the other populations thus the chances the phenotype caused due to the mutation. This situation is unlikely in the main population due to random mating of alleles. This reproductive isolation of communities results in higher frequency of the mutant allele and thus increase chance of observing the phenotype. For example – in Amish population, as

outbreeding is uncommon, they generally marry within the community. They carry one of the rare phenomena of polydactyly in which individuals tend to have extra finger and toes. Specific communities in Indian population do practise endogamy resulting in increased risk of the disease causing alleles these specific communities thereby increasing the risk of the disease. Systematic curation of founder mutations for Indian population can help in the development of molecular diagnostic tests for specific high risk communities at affordable costand can also develop in population screening strategies.

In this work I have curated both complex phenotypes and Mendelian phenotypes. A brief overview has been presented below on the Mendelian diseases. The inheritance of traits by the offspring from the parents exhibit calculable segregation ratios according to the Mendelian laws. According to Mendel, allele pairs get separated or segregate randomly from each other during the gamete production. Two gametes from each parent unite during fertilization, each of them contributes one allele to maintain the paired condition in the offspring. Now the movement of two alleles in offspring from each parent follows three mendelian laws. And the laws are -

Laws of segregation :- According to this law, during gamete formation, the alleles from each gene segregate from each other so that each gamete can carry only one allele for each gene.

Laws of independent assortment :- Genes for different traits can segregate independently during the formation of gametes.

Laws of dominance :- Some alleles are dominant while others are recessive. Organism with at least one dominant allele will display the effect of the dominant allele. And this effect of the dominant allele is observable.

Dominant allele refers to that particular allele, in which only one allele among the two is enough to show its effect as an observable trait. Whereas, in the case of the recessive allele, both the alleles are required to show their effect as phenotypic trait. One recessive allele fails to display its effect.

The mendelian trait are regulated by a single gene locus in the a particular inheritance pattern. If there is a mutation taking place in a single gene that can cause a disease and the mutation can be passed on to the next generation according to the Mendel's Law. This inheritance patterns are of four types. They are as follows –

Autosomal Dominant:- An individual carrying one copy of a mutated gene and one normal gene on a pair of chromosome. This mutated gene is inherited from each one of the parent as

well as the normal gene also comes from the other parent. There is 50% chance for the progeny to inherit the phenotype.

Autosomal Recessive:- Individuals carrying two mutated gene inherited from each of the parents. So, in that individual the two mutated genes are able to show their effect together only. The progeny of individuals who are both carriers have a 25% of inheriting the phenotype.

X – Linked Dominant:- the mutated dominant gene if is present in the X chromosome will display the effect in the offspring. X linked dominant mutation generally arise *de no vo.*

X – Linked Recessive:- mode of inheritance in which a mutation in a gene on the X chromosome cause the phenotype to be expressed in males hemizygous and females who are heterozygous or carrier.

The clinical disease variants curated here follows one of the above mentioned inheritance pattern. For example DMD – is a severe form of muscular dystrophy that causes muscle weakness, loss of ability to walk etc. Dystrophin is basically the largest gene of human that provides the information for the production of dystrophin protein. The gene in located in X chromosome. This protein helps in muscle movement, strengthening of muscle fibres, protect them from any injury. Any mutation taking place in dystrophin gene will hamper the protein production as well as its function. As the gene is located in the X chromosome, during fertilization it passes on to offspring. If the offspring is a female then no disease is observed because of the presence one more X chromosome in the female which compensate the effect of the mutated allele. Whereas, if offspring is a male then the mutated gene will show its effect as males do not have an extra X chromosome. So it is classified as a X – linked recessive disorder.

The clinical variant which is thought to be matching the clinical phenotype of the individual is annotated and supporting functional evidence collected to ascertain the significance to the relevant variant according to ACMG guidelines. For example – c.827G>A – this variant is associated with albinism. Thus relevant information pertaining to the variant of interest is compiled from databases and literature to attribute pathogenicity to the variant. There are several of databases and *in silico* tools are available publicly which is helpful in the interpretation of sequence variants. Some important databases are dbSNP, SNPedia, OMIM, ensemble etc. the in – silico tools that a curator can use during curation are Mutalyzer, variation repoter etc. which is also publicly available. Pubmed, Sci - hub and SNPedia is generally referred to pull out publications. SNPedia basically display the Pubmed id link for

the literatures available for a particular variant. The above mentioned databases and the is silico tools are explained as follows –

- ***dbSNP:***

The **Single Nucleotide Polymorphism Database** [1] (dbSNP) is a free public archive for genetic variation within and acrossdifferent species developed and hosted by the National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI). Although the name of the database implies a collection of one class of polymorphisms only (i.e., single nucleotide polymorphisms (SNPs)), it in fact contains a range of molecular variation: (1) SNPs, (2) short deletion and insertion polymorphisms (indels/DIPs), (3) microsatellite markers or short tandem repeats (STRs), (4) multinucleotide polymorphisms (MNPs), (5) heterozygous sequences, and (6) named variants [2]. The dbSNP accepts apparently neutral polymorphisms, polymorphisms corresponding to known phenotypes, and regions of no variation.

dbSNP is an online resource implemented to aid biology researchers. Its goal is to act as a single database that contains all identified genetic variation, which can be used to investigate a wide variety of genetically based natural phenomenon. Specifically, access to the molecular variation catalogued within dbSNP aids basic research such as physical mapping, population genetics, investigations into evolutionary relationships, as well as being able to quickly and easily quantify the amount of variation at a given site of interest. In addition, dbSNP guides applied research in pharmacogenomics and the association of genetic variation with phenotypic traits [3].

- ***SNPedia:***SNPedia is wiki - based bioinformatics database for SNPs (Single Neucleotide Polymorphism) for a disease or trait. It is. There are articles on each SNPs, which provide short descriptions, links for the scientific articles and microarray information about that SNP.. Since August 2007, the number of SNPs in SNPedia doubles roughly once in every 14 months and have 107,123 SNPs till 4 September, 2017.
- ***OMIM:***Online Mendelian Inheritance in Man(OMIM) is a constantly upgraded catalogue of human genes, genetic disorders and traits, along with a particular focus on the gene-phenotype relationship. Phenotypes represents approximately 8425 entries among 23,00 in OMIM. The rest entries include gene which are related to the phenotypes.

Johns Hopkins School of Medicine made and curated the content of OMIM. Based on the selection and review of the published biomedical literature the content of OMIM was made. Updating of content is done by a team of science writers and curators of Johns Hopkins University. It was designed primarily for the use by the physicians and other health care professionals dealing with genetic disorders, by genetics researchers, and by advanced students in science and medicine [4]. The database can also be used as a resource for placing literature relevant to inherited conditions [5].

During report curation or analysis, OMIM is used to get the inheritance pattern and gene phenotype relationship. It is also preferred to know whether any functional studies were done on the query gene or disease, which is important for analysis purpose.

- ***Ensembl:***Ensemble was launched in the year 1999 in response to the close completion of HGP (Human Genome Project). It was collaboration between the European Bioinformatics Institute and Wellcome trust Sanger Institute. Ensembl provides a compact resource for molecular biologists , geneticists and other researchers studying the genomes of human species, other vertebrates and model organisms. Ensembl is one of well known genome browsers for fetching genomic information.

During curation Ensemble is referred to get the NMID and the NPID along with transcript ID. It shows the length (either by number of bases or number of amino acids) of each transcript which helps to chosoe the longest transcript. Else it displays the gene details like chromosome number and location.

- ***Mutalyzer:*** It is a free web - based software tool. During genetic testing it is required to check the descriptions of a sequence variant in a particular gene. for this purpose, this software was developed [6]. This tool uses the rules of the standard human sequence variant nomenclature and according to that it can correct the description. Mutalyzer needs a such a record of DNA sequence which carries the transcript and protein feature annotations used as reference. Mutalyzer 2 agreed the record of GeneBank and Locus Reference Genomic. To generate DNA and protein variant descriptions for any organism the above mentioned annotations are used which can be applied to correct the codon translation table.

Leiden Open Variation Database (LOVD) uses Mutalyzer, which reserves sequence variant informations for numerous human genes, to check the variant descriptions before submission of new data [7]. This helps data sharing, display and integration with other genetic resources.

In curation mutalyzer is used to check the exon / intron / UTR postion of the query vatiant. NMID and cDNA position of the query variant is used as input in mutalyzer. NPID and protein change can also be obtained from this tool.

- ***Variation reporter:***Variation reporter is a free tool. To acquire the content of human variation resources at NCBI, variation reported is used. By using variety of formats the data query can be generated. The databases match them to the reserved data to produce a report that draws on dbSNP, DBVAR, ClinVar and NCBI's own human genomic annotation.

The variant can typed / pasted directly into a text box, or upload them in VCF, HGVS, GVF or BED format files. To help interpret the data, a assembly is chosen must (especially important if you identify chromosome 1 as "chr1" or just "1").

GRCh37 and GRCh38 these are the two builds which is used.

During curation, the preferred build is chosen. Then the user should input the NMID and cDNA position in the text box and upload the file. It will show some relevant result regarding that particular NMID and cDNA position.

Basically, to search for a rsID for a particular cDNA position, variation reporter is used. In the result section the rsID will be provided regarding the query position, which links directly to dbSNP to get all the details for the user regarding the rsID.

- **NCBI-**

NCBI is the main hub of the publications. It is a part of NLM (United National Library of Medicine). NCBI keeps a series of database on biotechnology, biomedicine, bioinformatics tools and services. Under NCBI there are two major databases Genebank and Pubmed.

1. **Gene Bank :** - The DNA sequence database available by the GeneBank which is kept by NCBI since 1992 [8]. This GeneBankmakes this DNA sequence available in coordination with laboratories and other sequence databases. [9]
2. **Pubmed:-**Pubmed is a bibliographic database for all the biomedical literatures. It is a free search engine, which accesses primarily the MEDLINE database of references, abstract on life science and biochemical topics. The United States National Library of Medicine (NLM) maintains this database as a part of the Entrez system of information retrieval. Pubmed was first released, which became the guide of primary and secondary scale MEDLINE searching [10]. In 1997 it was free to all.
3. **NCBI bookshelf**– it is a collection of online version of selected biomedical books which are freely accessible. It houses wide range of topics which includes molecular biology,

microbiology, genetics, cell biology, biochemistry, virology, research methods, disease states from the cellular and molecular point of view etc.

- **Sci – Hub**– Sci-Hub is a reservoir of over 64.5 million academic literatures and articles. It bypasses the paywalls.

If any publication is not freely available directly in Pubmed, Sci – Hub is used to download the paper free of cost.

Reports are also curated in a standard format. For report curation requires databases are referred. Like 1000 genome, Clinvar, Swissvar, LOVD, NCBI Gene etc are used. MedVar is a in – house database serves the same purpose. To cross check the result of the report, in silico tools such as Like mutation taster, LRT, SIFT, UCSC browser, Uniprot etc. are used. This curated report's database will be useful for reanalysing cases. These databases are useful to get the MAF of the variant. If the variant is reported in any of the said databases it will show the minor allele frequency of the minor allele or the risk allele (the allele which is responsible for a disease or trait is considered as risk allele). This is helpful to predict the type of the variant that whether it is harmful or not. From the prediction tools users can interpret the significance of the data (Damaging or disease causing or polymorphism). All the above mentioned databases and tools are described below –

- ***1000 genome project:*** it was an international research effort to establish by far the most detailed catalogue of human genetic variation. Scientists planned to sequence the genome of at 1000 anonymous participants from a number of different ethnic group using newly developed technologies which were faster and less expensive.

The goal of 1000 genome project was to find most genetic variants with frequencies of at least 1% in population studied. It took advantage of sequence technology, which sharply reduces the cost of sequencing. The project was planned to sequence each sample to 4X genome coverage, at this depth sequencing can not discover all variant in each sample, but can allow the detection of most variants with frequencies as low as 1%.

- ***ExAC***: The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

After giving the gene name in the search box, it shows the result in a tabulated form. First, the variant (chromosome no. : chromosome position reference allele / alternate allele) then the chromosome number and position in two separate column. If there is any protein change reported then that is also mentioned in consequence column. Then it shows the annotation that is whether the variant is intronic, missense, nonsense or 5'UTR or ibn splice region or not followed by allele count and allele frequency in two separate columns. For analysis or report curation, if the query variant is reported in he database then the MAF value is taken from it.

- ***ClinVar:***Clinvar is an open source prediction tool used to predict pathogenicity of the human gene variants with supportive evidences. It's a data base which contains reported variants and phenotypes and its relationships with the health status. The data base is generated by reported variants found in patient samples, its clinical significance, its refence sequence and with other supporting data. The alleles which are searched in this database are mapped according the to reference sequences and will display the result according to the HGVS standard.

User should input the gene name in the search box and will get all relevant variants in that gene. It is used during analysis to know the clinical significance of the variant.

- ***SwissVar:***To search for a variant in SwissProt of UnitProt Knowledgebase (UniProtKB), SwissVar is the gateway for that purpose. It also provides a direct access of Swiss – Prot Variant pages.

The Swiss-Prot Variant pages summarize all the information related to a particular variant and containThe variant page of SwissProt combined all the information relevant to a variant and contains:

- **manual annotation on the genotype-phenotype relationship of each specific variant based on literature;**
- **pre-computed information (such as conservation scores and a list of structural features when available) to help assess the effect of the variant.**

Three main search categories are provided:

Search categories	Functionalities
Disease	Enable search using disease names, OMIM identifier, as well as MeSH terms or identifiers of the disease category
Protein	Enable search using protein or gene names, UniProt accession number or identifier
Functional/structural features	Enable search using a list of functional and structural parameters of the variant

The combination of the above three categories is possible, and results can be downloaded in xml or tab-delimited format.

Swiss National Science Foundation and the European Community's Seventh Framework Programmecreated SwissVar in the frame of UniMed project.

- ***LOVD:***Leiden `University Medical Centre in the Netherlands designed LOVD as a freely available open source database. This database was createdfor the collection and display of variants in the DNA sequence.*[11][12]*The main focus of LOVD is to get the relation between gene and genetically inherited diseases. All the sequence variants found in human are collected in this database, along with the information about whether they could be normally connected to the disease or not. LOVD is also used by specialized doctors to

diagnose and advise patients who are carrying a genetic disease. Ideally, if a mutation is found in the information present in LOVD can predict the progress of the disease.

LOVD is used as a prediction tool for pathogenicity. Gene name is used as the input in the search box. Along with other details like gene symbol, gene name, chromosomal location, chromosome band number, genomic reference, transcript reference (NMID). It shows the classification of the variant het whether it is disease causing or polymorphism or not.

- ***Mutation Taster :*** Mutation taster is a free web based in – silico tool to evaluate DNA sequence variants for their disease – causing potential. Hence mutation tester is not used to single amino acids substitutions, but it also handles synonymous or intronic variants [13][14].

During analysis, for a particular mutationuser either need to input the chromosome number, position of the variant in the chromosome. The reference allele and alternative allele are used as a input quires. After this it will show that whether the said variant is reported or not and if reported then whether it is disease causing or polymorphism or not. Thedatabase along with the pathogenicity gives an insight about the gene, ensemble transcript id, uniport number, type of alteration, region of alteration, the cDNA change, amino acid change and its position, splice sites, prediction of PhyloP and PhatsCons, distance form splice site, protein is affected or not, reference sequence and alternate sequence and more other genomic details. It's a very helpful tool for analysing or reanalysing a variant.

- ***LRT:*** Likelihood ratio test (LR test) is a statistical test used to compare the goodness of fit of two statistical models, one which is the null model and another is a special case which is the alternative model. The test is based on the likelihood ratio, which explain how many times more likely a data is under null model than the other. This likelihood ratio, or its logarithm (both are equivalent), can be used to calculate a p-value, or compare to a critical values to judge whether the null model should be rejected or not.

The likelihood ratio test rejects the null hypothesis if the value of the statistic is too small. How small is too small depends on the significance level of the test, i.e., on what probability of Type I error is considered tolerable ("Type I" error consist of the rejection of a null hypothesis that is true).

The likelihood ratio lies between 0 to 1, lower the value means that the observed outcome will most likely to lie under null hypothesis , which can not be rejected.

Nested models are required for likelihood – ratio test (in nested model complex models are transferred into simpler one by applying some limitations on the parameter). If models are not nested, then instead of using relative likelihood, generalized form of likelihood – ratio test can be used.

- ***SIFT:*** Single nucleotide polymorphism (SNP)are used as markers which helps in linkage and association studies. These studies detects specific regions in human genome which is responsible for any disease. SNPS that leads to amino acid changes in proteins are of main interest causes this will lead to substitution of amino acids which results in some human diseases. SIFT (Sorting Intolerant From Tolerant) is a program which uses sequence homology to predict whether an amino acids substitution affects protein function or not so that users can prioritize substitutions for further study. It is shown that SIFT can distinguish between functionally neutral and deleterious amino acid changes in mutagenesis studies and on disease causing human polymorphisms.

Users can simply provide their protein sequences and amino acid substitutions. SIFT will automatically search for protein sequences which is homologous to the query protein sequences and based on these sequences, SIFT calculates the probabilities for each possible amino acid change.

SIFT predicts the substitution less than a score of 0.05 is considered as deleterious. In some cases, it is observed that substitution score less than 0.1 provides better sensitivity for detecting deleterious Single Nucleotide Polymorphism (SNP). Score permits the users to prioritize their amino acid changes by ranking them from the lowest to the highest score.
rsIDor query protein sequence is as an input for SIFT. SIFT will show the result in a tabulated format, after comparing the query sequence with the orthologues protein alignment along with scale of that prediction.

- ***UCSC browser:*** It is a freely available, online, downloadable genome browser which is organized by the University of California, Santa Cruz (UCSC). This website offers the entry to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms, which is assimilated with a large collection of aligned annotations. The browser database, browsing tools, downloadable data files, and documentation are available on the UCSC genome Bioinformatics website. It shows a graphical display of the first full chromosome draft assembly of human genome sequence.

For analysis purpose user uses this UCSC Browser to know whether a particular region of a specific chromosome is conserved or across the species. For this user should enter the

chromosome number with the position. It shows the reference sequence along with the position and the amino acid sequence. It also display the gene expression from different tissue in a graphical form.

- ***Uniprot:-*** It is another freely available tool for protein sequence and functional analysis of domain. It holds huge amount of data about biological function of protein taken from different research literaures and other databases. It is the Universal protein resource which is a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases.

During analysis, gene name is used as an input ranging from different animal species to human. Only human variant should be chosen.

The "view protein in Pfam" is selected and a table containing on the start and stop position of the amino acid based on protein domains pops out the protein family domain 'Pfam" for the gene is scaled based on the protein change position.

By analysing the protein change position domain can be predicted. This domain finding is an important part of analysis because this will also help to decide the ACMG classification of a particular variant.

- ***Polyphen2:*** It is an *in silico* tool which predicts possible effect of amino acid substitutions, both structural and functional mutation, of human proteins using physical and evolutionary comparative considerations.

Polyphen – 2 is a modification of the Polyphen tool for annotating coding and nonsynonymous SNPs. The highlighted of the new version are:

- **High quality multiple sequence alignment pipeline**
- **Probabilistic classifier based on machine - learning method**
- **Optimized for high throughput analysis of the next generation sequencing data.**

In this tool, position of the mutation is used as input in the protein identifier column. Then the position of substitution. Then the alphabet should be chosen of the corresponding amino acid that takes part in the substitution.

In the result it shows that whether the protein change is disease causing or polymorphism.

The American College of Medical Genetics and Genomics (ACMG) was developed for the guidance, for interpreting of sequence technology. This guideline is primarily applied to broadness of various genetic tests used in many clinical laboratories or research laboratories, which includes genotyping, gene panels, single gene, exomes and whole genomes. According

to this ACMG guideline some specific terminology is used to classify a variant that cause Mendelian disorders. They are – "pathogenic", "likely pathogenic", "benign", "likely benign" and "uncertain significance". So this guideline explains the process for classifying a variant in one of the five above mentioned categories and this classification is based on several criteria using typical types of variant evidences, like – population data, computational data, segregation data, functional data etc. to classify the variants in the above mentioned categories, some parameters are followed, an example - if a disease causing variant is found in a patient with complete matching symptoms, then according to ACMG guideline the variant is classified as "pathogenic". A variant will be considered as "Likely Pathogenic" when the variant is likely to contribute to the development of a disease, but evidence are not sufficient to prove the same, additional evidence are required. If a variant is detected but the variant is not responsible for any harmful disease, then it is considered as "Benign". If a variant has been detected in a patient, but based on available evidences it is difficult to conclude that whether the variant is pathogenic or benign, then the variant is classified as "Variant of Uncertain Significance". These rules are followed during report curation or analysis to classify the reported variant in a patient.

3.1 Background:

- **About Medgenome:-**

I joined Medgenome as a scientific curator trainee in operation department. Medgenome is a genomic based diagnostic and research company. The research solutions apply cutting-edge genomic technologies, computing, bioinformaticsand big data analytics to the genetically diverse population to know the genetic basis of cancer, eye disorders, metabolic disordersand other rare diseases. Medgenome offers a wide variety of genetic tests such as – Cardiology, Haematology, Oncology, Neurology, Nephrology, Immunology, Metabolic disorder, eye disorders, connective tissue disorder, prenatal screening etc. For these tests company uses NGS, FISH, PCR, Sanger, Microarray, Flow cytometer technologies form the basis of this company by providing a wide range of sequencing, such as - whole exome sequencing, RNA sequencing, miRNA sequencing, single cell sequencing.

- **NGS: -**

The next-generation sequencing (NGS) has revolutionised the field of genomics by its ultra-high throughput, scalability, and speed. An entire genome or a specific areas of interest can be sequenced by NGS, including all 22 000 coding genes (a whole exome) or few

numbers of individual genes. An entire human genome can be sequenced by NGS within days. The DNA is fragmented before sequencing. Because NGS platforms can sequence of small fragments of DNA in parallel. The data generated from NGS is analysed by Bioinformatics, which is used to piece together those DNA fragments by mapping the individual reads to the human reference genome. NGS sequence about three billion bases of human genome multiple times for providing high depth to provide accurate data and unexpected DNA variation. From the NGS data it is easy to discover entirely novel mutations and disease - causing genes. It provides more sensitive read-outs and thus can be used to identify variants which present in just a few per cent of the cells, including mosaic variation. The sensitivity of NGS sequencing can be increased further, simply by increasing sequencing depth. The main disadvantage of NGS is the proper clinical infrastructure, such as computer capacity and its storage, the personnel expertise required to handle the NGS and analyse and interpret the NGS data.

3.2 Achivements:

After the prerequisite training and guidance in the short span of time, I have been able learn and work in the field of genetic variant curation and literature mining and contribute to the overall objective of the project.

- Curation is very interesting area of work. It helps to enhance the knowledge about the disease, gene and the related information.
- For curation (both of publications and reports) I have got hands on experience on using various genomic/population/disease databases and in silico prediction tools. Curation provides the opportunity to use these database and tools as required to search for a query variant. It also helps to interpret a result displayed by a database or in *in silico* tool about a query variant. So, the hand on experience of using those data bases and tools was very helpful.
- I was able to learn the genetic basis of both complex and Medelian genetic diseases for the Indian Population.
- Constructing and maintaining a huge amount of information with less amount of error, was a difficult task but it enhances my capacity in organizing the data.

In addition to the technical knowledge, I have been able to contribute to the team for attaining the overall objective in the set timelines and as a resourceful team player

3.3Overview of dissertation:

The main aim of my work as a scientific curator trainee was to compile genetic variant data form human diseases. Initially was guided through the objective and overall workflow including sample processing, bioinformatics applications, working principles of the genetic variation prediction tools with frequent inductions by the other teams to brief their work. Then gradually were trained to get acquainted with the annotation of the variants and using various databases to obtain the required information.

As a part of the work, I had to curate variants associated with Indian founder mutation, several and phenotypic traits. Around 113 relevant scientific papers were curated and documented in a systematic format. In the process I learnt to use the *in silico* tools to predict the effect of these risk variant for several phenotypic traits (such as – vitamin B9 level, eye colour, skin and hair pigmentation, ability to taste bitter food etc.). In addition clinically relevant variants were curated in accordance with the ACMG guidelines with strict quality control and proof reading.

4. Methodology:

The process and the format of curating variants from biomedical literature and clinical report were different based on the content curated. So there are two different methods followed during paper and report curation. The methods are described below –

4.1 Curation of Literatures in Complex Traits -

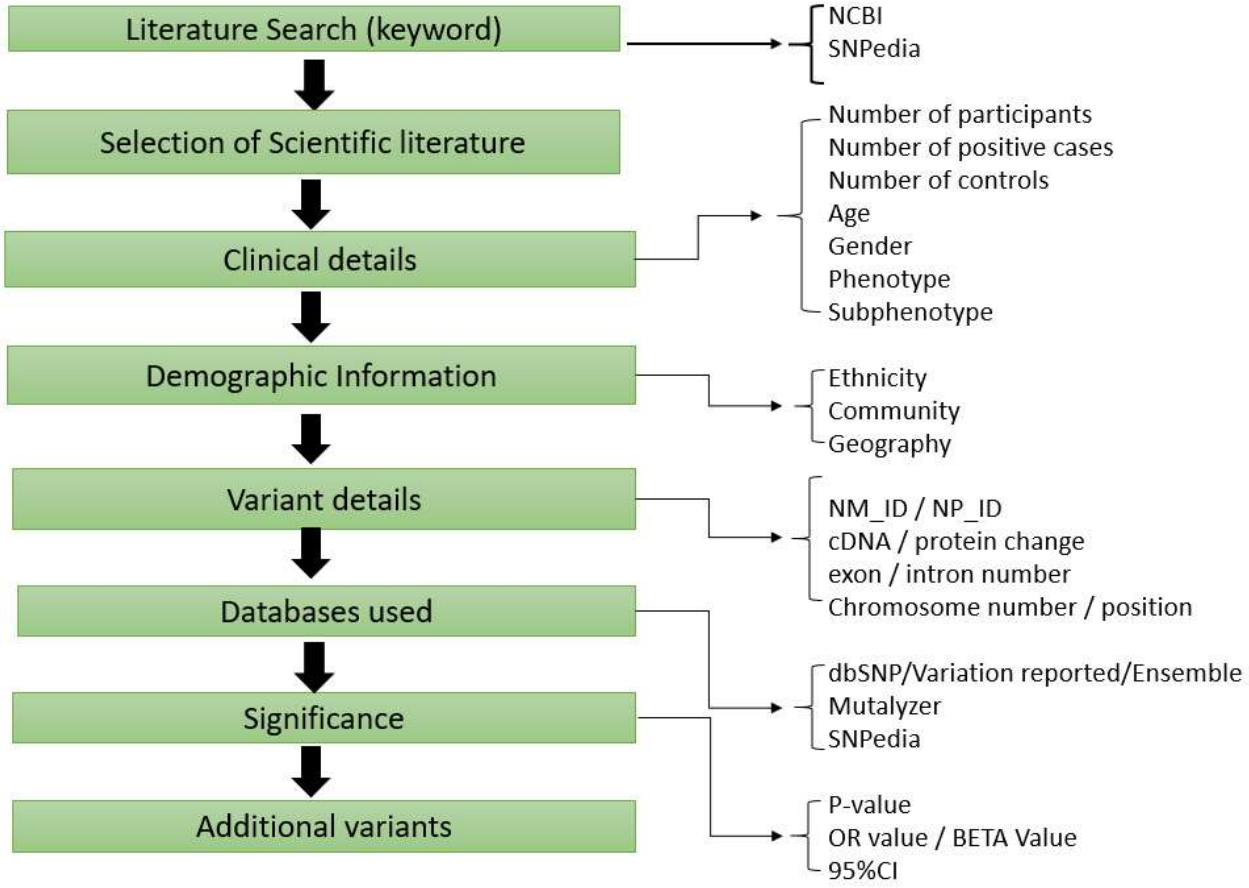


Figure1 : Flow Chart of Literature Curation in ComplexTraits

- First curator needs to choose a particular variant associated with a disease or phenotype or injury risk.
- The relevant literatures were searched in NCBI – Pubmed and SNPedia. HGMD (in house data base) can also be used. HGMD provides the PMID (Pubmed) link of the query variant.SNPedia also provides the PMID link of the publication relevant to the query variant. If full texts are not available in Pubmed, the PMID is use as an input in another database that is Sci – hub. Sci – hub is preferred for the full text.
- Then all the relevant informations were curated one after one.
- Then the curator needs to find whether the paper describes about the variant and the variant is responsible for the disease or phenotype only.
- The variant in the paper can be rsID / cDNA position / protein position.
- Th rsID found in the paper used as an input of dbSNP database.
- The database shows information about the variant.

- In the allele section it shows the class of the variant whether it is a SNP or CNV / reference allele or ancestral allele / clinical significance of the variant whether it's a pathogenic variant or disease - causing variant.
- Some MAF values are also given in this context which are taken from some other databases like 1000 genome or ExAC etc. In the next section HGVS names are there. There we get know about the NMID, NPID, NCID of the variant. The NMID refers to the cDNA position of the variant. NPID refers to the protein change. The genomic position of the variant is assigned by the NCID. There is a box below which shows the genomic build, chromosome number, position of the variant in the chromosome. Below the box the gene name is mentioned.
- In dbSNP first need to check the HGVS column for NMID and NPID. There if more than one NMID AND NPID is present then should go with the longest transcript. For that need to select the link of gene name, which shows a wide description in about the gene. There need to search with longest transcript of that rsID. Then should take that NMID and NPID as our required evidence. For genomic position we need chromosome position first then the position of the variant in chromosome and the reference and alternate allele.

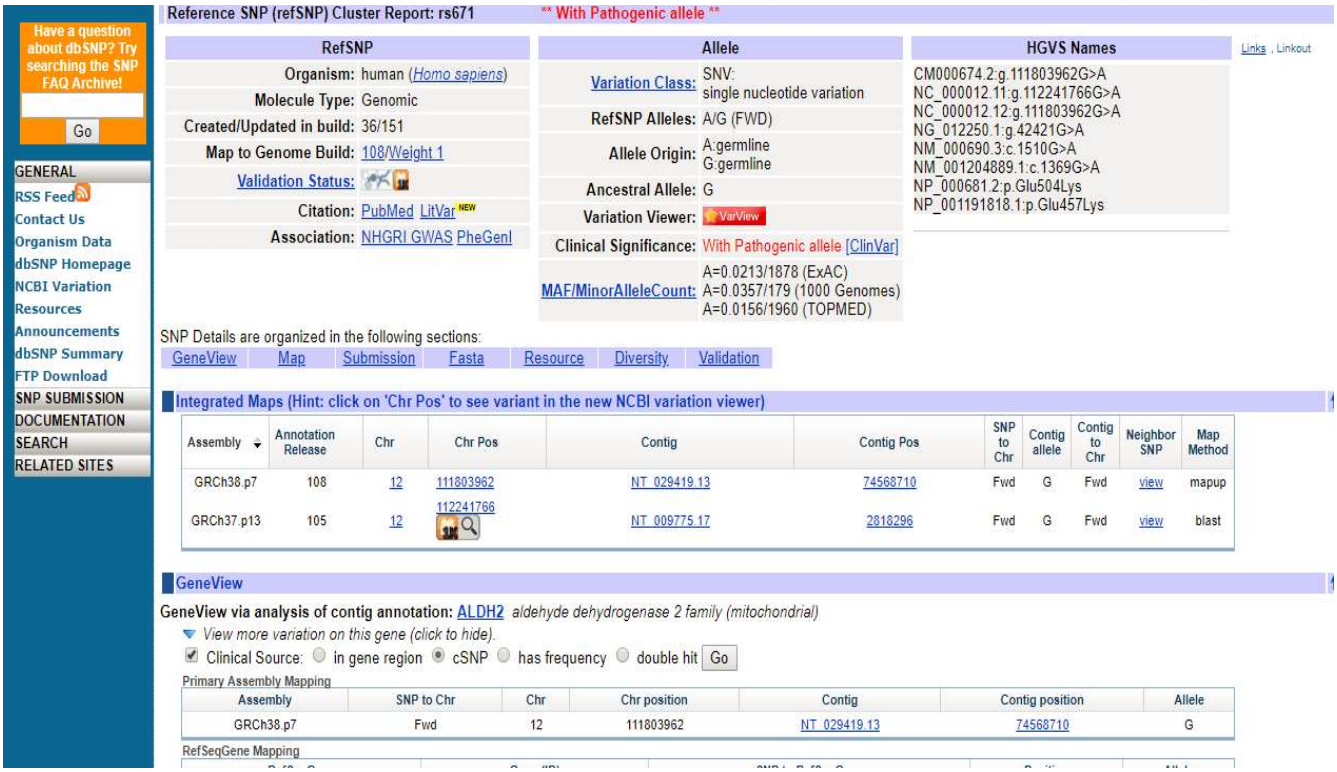


Figure2: Image of dbSNP

- If no rsID is mentioned in the paper but the gene is provided, then the longest transcript can be checked from ensemble database. Gene name is used as input and by comparing base pair or no of amino acid the longest transcript is chosen.

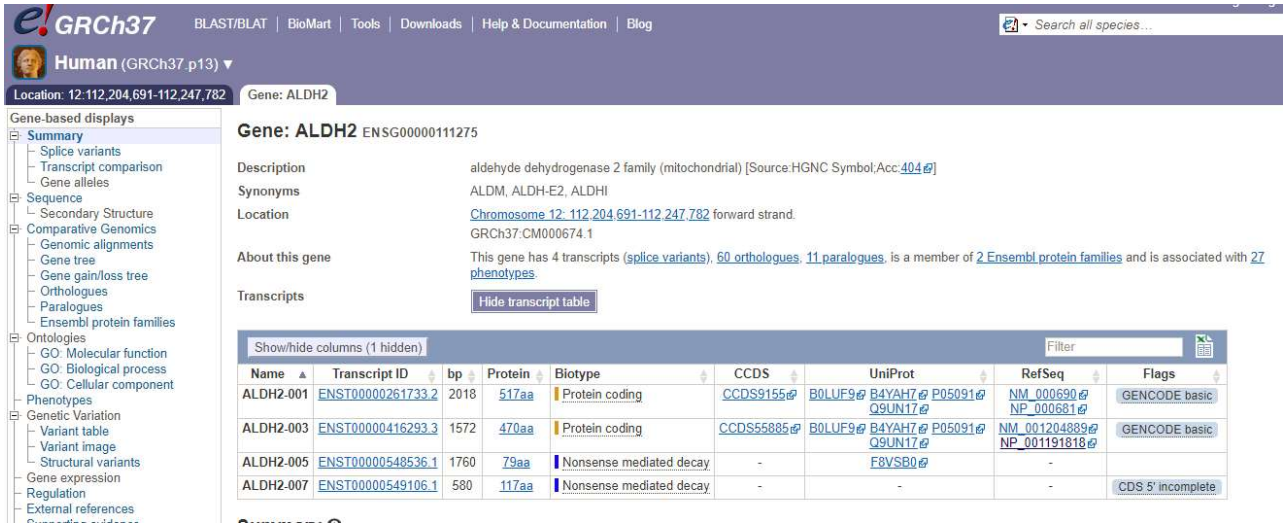
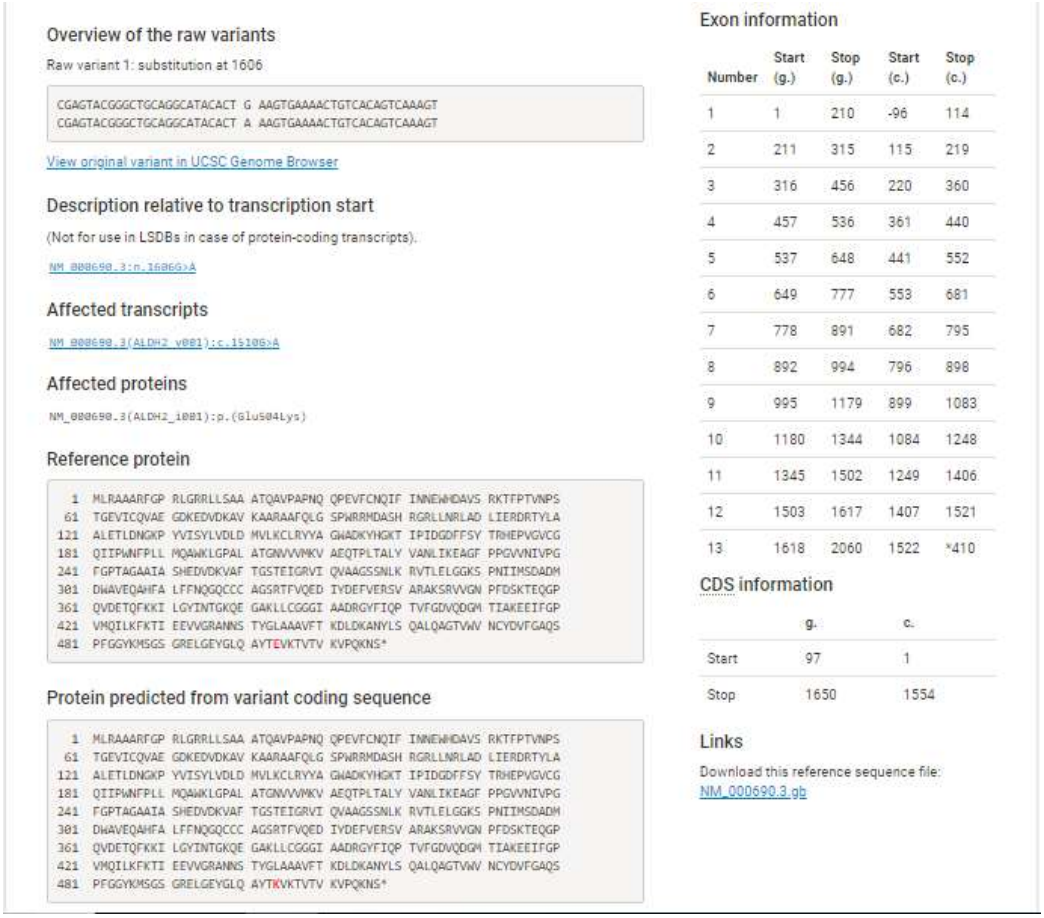


Fig 3: Image of Ensembl

- All the informations(cDNA position and its corresponding NMID and protein change with its corresponding NPID) from dbSNP is curated in its respective column of the curation sheet.
- To find the position of the variant (exonic / intronic / UTR) Mutalyzer database is used. In the Name Checker of mutalyzerdatabase the NMID and cDNA position or NPID and its corresponding protein change is used as an input and vice versa. As a result the whole exon information of the gene ispresented from which the variant class and its position is curated.



Overview of the raw variants

Raw variant 1: substitution at 1606

```
CGAGTACGGGCTGCAGGCATACACT G AAGTGAAAACTGTCACAGTCAAAGT
CGAGTACGGGCTGCAGGCATACACT A AAGTGAAAACTGTCACAGTCAAAGT
```

View original variant in UCSC Genome Browser

Description relative to transcription start

(Not for use in LSDBs in case of protein-coding transcripts).

NM_000690.3:n.1606G>A

Affected transcripts

NM_000690.3(ALDH2_v001):c.1510G>A

Affected proteins

NM_000690.3(ALDH2_i001):p.(Glu504Lys)

Reference protein

```
  1 MLRAAARFGP RLGRRLLSAA ATQAVPAPNQ QPEVFCNQIF INNEWHDAVS RKTFPTVNPS
 61 TGEVICQVAE GDKEDVDKAV KAARAAFQLG SPWRRMDASH RGRLLNRLAD LIERDRTYLA
121 ALETLDNGKP YVISYLVDLD MVLKCLRYYA GWADKYHGKT IPIDGDFFSY TRHEPVGVCG
181 QIIPWNFPLL MQAWKLGPAL ATGNVVVMKV AEQTPLTALY VANLIKEAGF PPGVVNIVPG
241 FGPTAGAAIA SHEDVDKVAF TGSTEIGRVI QVAAGSSNLK RVTLELGGKS PNIIMSDADM
301 DWAVEQAHFA LFFNQGQCCC AGSRTFVQED IYDEFVERSV ARAKSRVVGN PFDSKTEQGP
361 QVDETQFKKI LGYINTGKQE GAKLLCGGGI AADRGYFIQP TVFGDVQDGM TIAKEEIFGP
421 VMQILKFKTI EEVVGRANNS TYGLAAAVFT KDLDKANYLS QALQAGTVWV NCYDVFGAQS
481 PFGGYKMSGS GRELGEYGLQ AYTEVKTVTV KVPQKNS*
```

Protein predicted from variant coding sequence

```
  1 MLRAAARFGP RLGRRLLSAA ATQAVPAPNQ QPEVFCNQIF INNEWHDAVS RKTFPTVNPS
 61 TGEVICQVAE GDKEDVDKAV KAARAAFQLG SPWRRMDASH RGRLLNRLAD LIERDRTYLA
121 ALETLDNGKP YVISYLVDLD MVLKCLRYYA GWADKYHGKT IPIDGDFFSY TRHEPVGVCG
181 QIIPWNFPLL MQAWKLGPAL ATGNVVVMKV AEQTPLTALY VANLIKEAGF PPGVVNIVPG
241 FGPTAGAAIA SHEDVDKVAF TGSTEIGRVI QVAAGSSNLK RVTLELGGKS PNIIMSDADM
301 DWAVEQAHFA LFFNQGQCCC AGSRTFVQED IYDEFVERSV ARAKSRVVGN PFDSKTEQGP
361 QVDETQFKKI LGYINTGKQE GAKLLCGGGI AADRGYFIQP TVFGDVQDGM TIAKEEIFGP
421 VMQILKFKTI EEVVGRANNS TYGLAAAVFT KDLDKANYLS QALQAGTVWV NCYDVFGAQS
481 PFGGYKMSGS GRELGEYGLQ AYTKVKTVTV KVPQKNS*
```

Exon information

Number	Start (g.)	Stop (g.)	Start (c.)	Stop (c.)
1	1	210	-96	114
2	211	315	115	219
3	316	456	220	360
4	457	536	361	440
5	537	648	441	552
6	649	777	553	681
7	778	891	682	795
8	892	994	796	898
9	995	1179	899	1083
10	1180	1344	1084	1248
11	1345	1502	1249	1406
12	1503	1617	1407	1521
13	1618	2060	1522	*410

CDS information

	g.	c.
Start	97	1
Stop	1650	1554

Links

Download this reference sequence file:
NM_000690.3.gb

Figure4 : Name checker of mutalyzer

- From the paper curator searches for the number of of individuals (cases / controls), ethnicity of the population / individuals are.
- The genotype column is curated depending upon the risk associated with the disease which can be cross checked from the SNPedia by using rsID as an input. The risk allele in that particular case should also be mentioned in the curation sheet. To cross check this result of the paper, SNPedia can be referred. rsID is used as an input in SNPedia for which it shows genotype as well as some paper on that particular variant. For an example – for this particular rsID which is associated with alcohol flush reaction AA and AG genotypic individual have increased risk of alcohol flush reaction and esophageal cancer whereas GG genotype carrying individual has normal flush reaction.
- Along with this the risk allele and the risk allele frequency was also curated in its respective place in the curation sheet.
- Each variant has a particular p value, ORor BETA value and 95% CI value which are curated according to the values mentioned in the paper and cross checked fromSNPedia. If

these values are not mentioned, then the column should be left blank. But if it is there in the paper, then we need to cross check the values with SNPedia.

- Additional variants found in the literature responsible for the similar disease or phenotype or injury is also curated.
- The reference of the paper against which the curation was done is also provided in the curation sheet along with all the above - mentioned details.
- During the curation process while checking the data of the paper with the online databases, if any discrepancy was observed that is also mentioned in the comment section of the same sheet.

4.2Curation of literatures for IndianFounder mutation

The process is similar tophenotypic traits curation, however additional databases are referred to collect relevant information. The extra information includes – allele frequency, functional study, gender of participants, zygosity and number of individuals under that zygosity, segregation in the family.

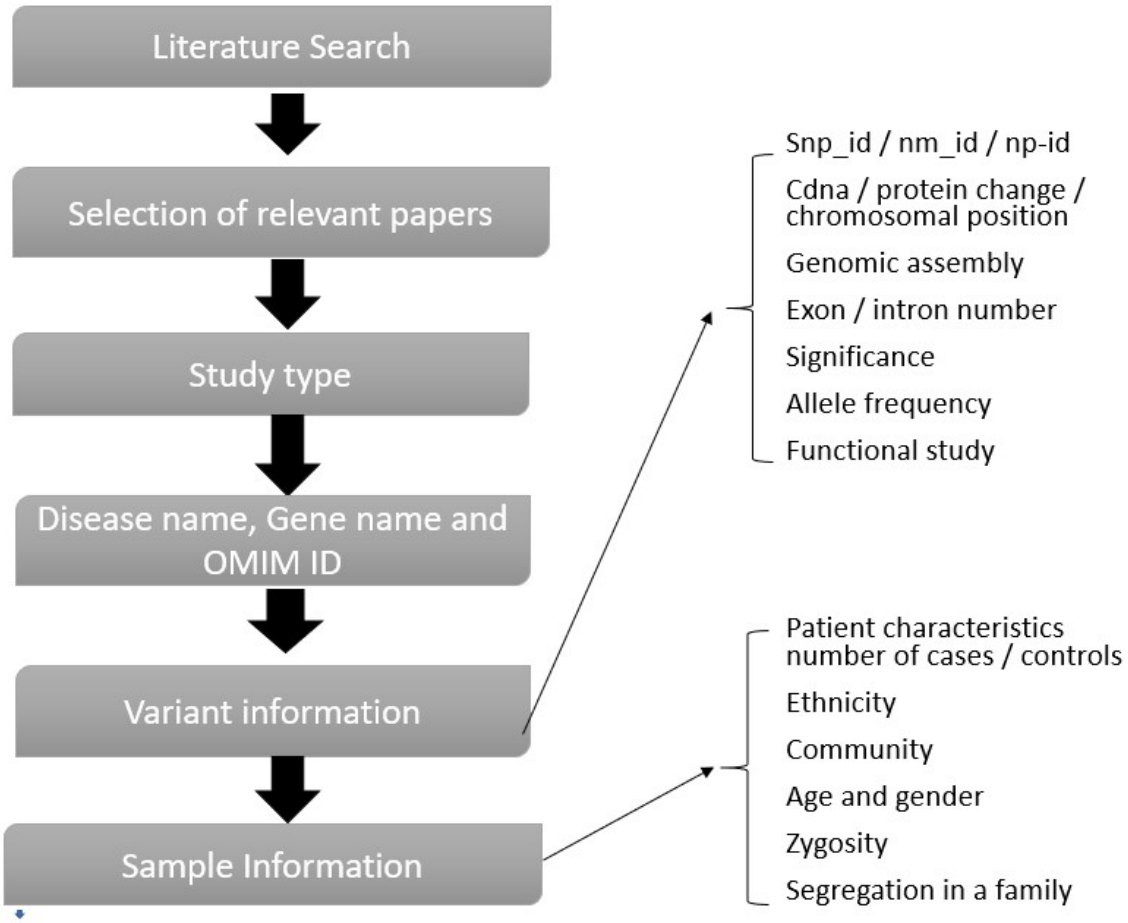


Figure 5: Curation Work Flow in Complex Diseases

4.3Report Curation –

The process of curating a report is different from curating a publication. In this case information from a patient report is compiled in a standard format. More number of databases and in silico tools are used (mentioned above) than paper curation. The report curation is done to keep a back - up data for any further query or for reanalysis of sample. The process is mentioned in the form of a flow diagram below -

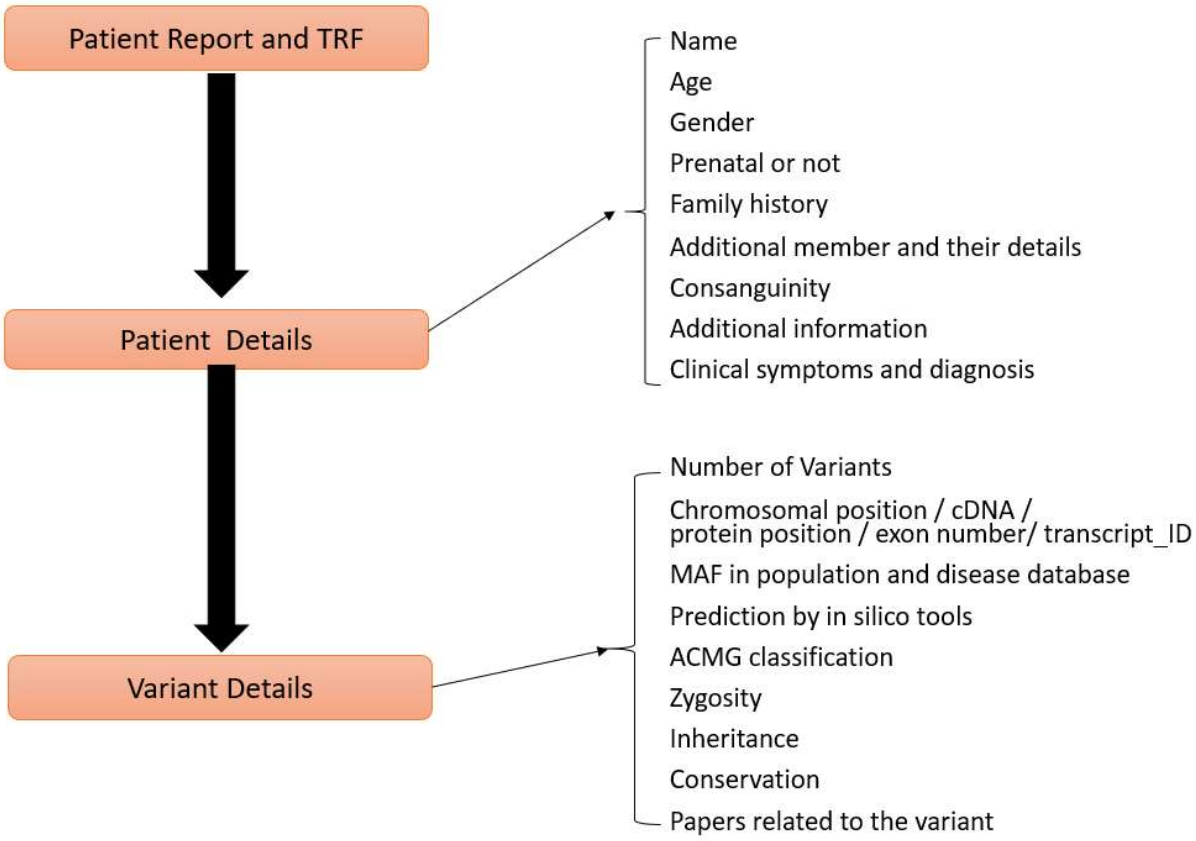


Figure 6: Flow chart for report curation

- The clinical report of a patient along with the TRF was received.
- Each patient has a separate family ID, sample ID and order ID. Additional family members were in the same family ID was curated referring both the TRF and the clinical report.
- The panel was curated in the respective sample IDs depending upon the disease or phenotype desired or whether the samples are required for whole genome sequencing / whole exome sequencing.
- The Clinician name and details about the proband (like – name, age, date of birth and gender) are curated from the clinical report. Along with this consanguinity / degree of consanguinity about the parents was also curated from the TRF.
- If any additional information was mentioned such as – place of birth, ethnicity, mother tongue etc. was also curated.
- The details about the family members if mentioned are curated accordingly followed by the relationship details about the proband.

- Additional family members' sample IDs and relationship with the proband should be mentioned after that.
- If the proband sample is a prenatal / fetal sample, it should be curated with caution.
- Sometimes family history is important in the analysis of a genetic disease. So, the clinician provides details of the family history along with the other informations in the TRF. That is also captured during curation.
- The clinical symptoms which were observed in the proband, should be mentioned, based on that diagnosis could be done. All this information (clinical symptoms / diagnosis) was curated from the TRF.
- After sequencing if variants were found, number of the variants should be mentioned in the curation sheet.
- The details about the variants found and (chromosomal position / cDNA position / protein change / exon or intron number / zygosity / transcript ID / variant class) was mentioned in the report was captured from the clinical report.
- Whether the reported gene has any pseudogene or not, that was checked by using "Genecard" and this information was also captured.
- Using the online databases (mentioned in the introduction section) pathogenicity about the variant, the phenotype details, conservation, the population frequencies are captured.
- If the variant found in the patient, is reported in the paper previously, then the PMID of the variant should be captured.
- ACMG classification was used to interpret the significance of the variant and curated accordingly. Here, curator should use the in – house – database (varclass) to check what is the prediction of the variant according to ACMG classification. Some query questions were used as an input to get the prediction.

5. Results:

A total of 113 publications on phenotypic traits, 13 literatures on founder mutation and 17 publications on genetic risk variants for ligament injury were curated. Studies onACL injury and phenotypic traits were case control study based and founder mutation studies were predominantly on Mendelian diseases. The different for which the variants were compiled

were deep sleep, smell, taste, eye colour, hair pigmentation, skin pigmentation, lactose deficiency, Vitamin B12 level, Alcohol flush reaction, caffeine consumption, Folic acid, muscle consumption etc.

Indian founder mutation was performed for Wilson disease, Congenital Disorder of Glycosylation, Type Iq, Progressive Pseudorheumatoid, HemolyticAnemia, Hyperoxaluria, primary type – I, Trichohepatoenteric syndrome 1 respectively. For all the required data points including segregation of the variants, information on functional studies, and prevalence in population databases. Detailed information in controlled vocabulary (HPO terms) was systematically captured. For ACL injury gene specific curation was done (*COL1A1, COL12A1, COL5A1, DCN, COL3A1, VEGFA*). In addition to genetic variant, the process of the study, i.e. background of the study, the setup, criteria of patient selection, the statistical data, analysis of the result and genetic analysis statistical method used were also compiled.

In my curation study association of more than one gene variant for a single disease or phenotype was also noted. Example –. It was observed that various mutations in different locus change the expression *HERC2* gene resulting in different eye colour; hair pigmentation is influenced by multiple genes *MC1R, IRF4, SLC24A4* and *ALDH2* gene which is associated with alcohol flush reaction, three different variants were observed in the same gene in different exonic and intronic positions. In case of vitamin D level observation studies, it was stated that one variant in GC gene is responsible for that. But in association with the variant on GC gene another variant with IVL gene also effects the vitamin D level in body.

In ACL injury curation studies, VEGFA gene was associated with the increase the risk of injury. Two different variants were reported in two different papers in this gene responsible for the risk of injury.

In Wilson disease (founder mutation study), four different variants were reported in the *ATP7B* gene in Indian population. On the other hand, in case of haemolyticanaemia, two different variants were studied in different locations in G6PD gene. These variants were mostly found in Indian population.

Figure 7: Pie chart showing the percentage of scientific literature curated for the specific diseases in Indian population

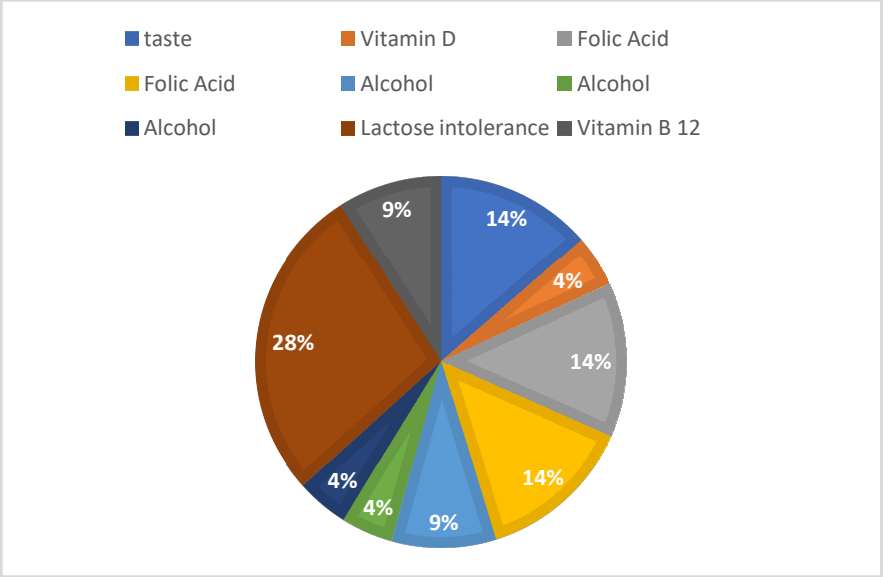
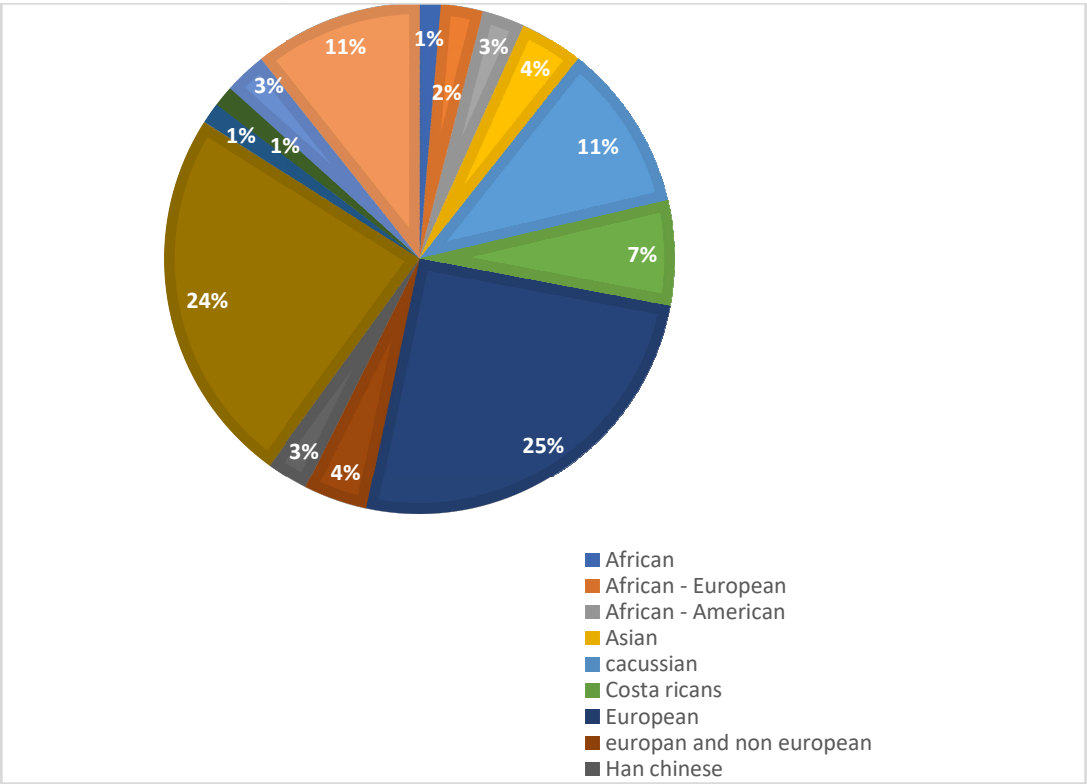


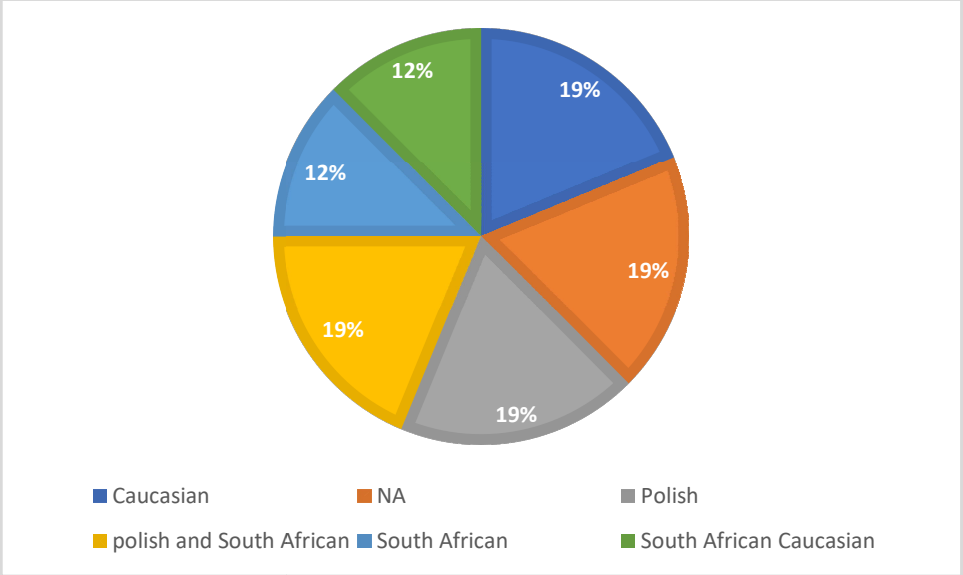
Figure 8: Pie chart showing percentage of scientific literature curated from different ethnic background for traits' of interest



In my study, I found six different traits which are very common in Indian population. The trait list of Indian population includes taste, vitamin D, folic acid, alcohol flush reaction, lactose intolerance and vitamin B12. Among this two different variants of folic acid were found to be predominant in Indian population. At the same time 3 different variants of alcohol flush reaction was found to be predominant in Indian population. The above pie chart shows percentages of different traits which are found to be common in Indian population study. It displays lactose intolerance occupies 28% of all traits in Indians. Then taste, two variants of folic acid are of 14% each, alcohol flush reaction and vitamin B12 level occupies 9% each. At last, each of vitamin D and two different variants of alcohol flush reaction takes 4% of the total traits. From this it is concluded that the most predominant trait among the six most common traits is lactose intolerance.

The publications were from various populations which includes African, Asian, Costa Ricans, Europeans, Han Chinese, Indian and Japanese. Some of the publications were performed in more than one population.And also it was obsered that most of the publication were from the European population (25%) and very few in Indian population (represented in figure 8). When a comparative study was done among different population considering all the traits, a different result appeared. In this case, I tried to observe the percentage of each population considering all the traits.

Figure 9: Pie chart showing different percentages for different ethnicity of ACL injury risk



After curating 17 different literatures on ACL injury risk, it was very clear that the risk of ACL injury is high in those individuals who are associated with any kind of sports. This study was also done on different sports participants from different ethnic background. The study was based on football players, tennis players and mostly skiers. South Africans, Cacussians and Polish participants were mostly referred for these studies. The literature collected were predominantly on Polish, Caucasian and south African population(19%).

6. Conclusion:

The variants associated with deep sleep, smell, eye colour, skin and hair pigmentation, caffeine consumption, muscle consumption etc. are mostly observed in European population compared to the other ethnic groups. Variants associated with alcohol flush reaction, taste, Vitamin B12, Vitamin B9 (Folic acid), Vitamin D, Lactose intolerance etc traits were also found in Indian population. Genetic risk of Ligament injury studies have been done in participants who were associated with sports confirming from Polish, Caucasian and South African ethnicity. Based on the curation it was evident that there is paucity of literature on Indian population.

The sequencing technologies has enabled opportunities to design and address questions to identify gene mutations responsible for different rare and common genetic disorders. According to joint World Health Organization birth defects account for 7% of all neonatal mortality and 3.3 million under five deaths, where in India birth defects prevalence varies from 61 to 69.9/1000 live births and congenital malformations are the second commonest cause (9.9%) of mortality among stillbirths. The structured documentation of these mutations specific to Indian population can be used for both research and clinical purposes and to design population specific diagnostic tests cost effectively.

7. Challenges Faced

Curation of publication / reports take significant time, because the curator required to find, verify and organize high quality information from multiple sources, so as to present in an effective way.

- It requires experience in the field being curated, as well as great familiarity with the tools and databases that is required during curation.
- The curated data should be in anunaltered format, so that the meaning of the information should be the same as it was mentioned in the paper.
- Biomedical literature especially those published a few years back may not have used the current annotations and may also not necessarily have used standard names and nomenclature which makes curation difficult

8. How to complement in corporate?

The most important is to be able to understand the objective and application of the assigned task both immediate and long term. Based on observation, I have listed a few traits below.

- Time management, punctuality and to be able to meet the timeline for completing the work
- Detailed documentation
- Confidentiality should be maintained.
- Given the situation, the ability to manage multiple tasks, if required.One should know how to deal with the situation and should be capable to fulfil the requirements.
- Ability to cooperate as a team and to be able to respect decisions which are based solely on the work outcome
- Good ethics
- To be able to build in-depth work knowledge to evaluate data and identify discrepancies
- To be able to communicate with clarity
- Continuous learning to be able to build work related skill and knowledge

9. References:

[1] Mizrachi, Ilene "GenBank: The Nucleotide Sequence Database" (22 August 2007).

[2] Sherry ST, Ward M; Sirotkin, K. "dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation". Genome Research (1999)..

[3] Kitts A; Sherry S, "The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation"(2009).

[4]Amberger, J.; Bocchini, C.; Hamosh, A. "A new face and new challenges for Online Mendelian Inheritance in Man (2011).

[5] Gitomer, W.; Pak, C. "Recent advances in the biochemical and molecular biological basis of cystinuria". The Journal of Urology(1996).

[6] Wildeman, Martin; Van Ophuizen, Ernest; Den Dunnen, Johan T.; Taschner, Peter E.M. "Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker". Human Mutation(2008).

[7]Fokkema, Ivo F.A.C.; Den Dunnen, Johan T.; Taschner, Peter E.M. "LOVD: Easy creation of a locus-specific sequence variation database using an 'LSDB-in-a-box' approach"(2005).

[8] Mizrachi, Ilene "GenBank: The Nucleotide Sequence Database"(22 August 2007).

[9] Mizrachi, Ilene "GenBank: The Nucleotide Sequence Database"(22 August 2007).

[10]"PubMed Celebrates its 10th Anniversary". Technical Bulletin. United States National Library of Medicine.

[11]Fokkema, Ivo F.A.C.; Den Dunnen, Johan T.; Taschner, Peter E.M. "LOVD: Easy creation of a locus-specific sequence variation database using an 'LSDB-in-a-box' approach"(2005).

[12] Fokkema, IF; Taschner, PE; Schaafsma, GC; Celli, J; Laros, JF; den Dunnen, JT "LOVD v.2.0: the next generation in gene variant databases"(May 2011).

[`13] Schwarz, Jana Marie; Rödelsperger, Christian; Schuelke, Markus; Seelow, Dominik "MutationTaster evaluates disease-causing potential of sequence alterations"(2010-08-01).

[14] Schwarz, Jana Marie; Cooper, David N; Schuelke, Markus; Seelow, Dominik "MutationTaster2: mutation prediction for the deep-sequencing age"(2014-03-28).