IN SILICO DIAGNOSIS OF TUBERCULOSIS IN MULTIPLE SPECIES USING SINGLE NUCLEOTIDE POLYMORPHISM AS BIOLOGICAL MARKER

Meetali Sinha[1], Neha Jain[2], Kabani Sudevan[3]

[1, 2] Department of Bioengineering, Faculty of Engineering, Integral university, Lucknow, (India)

[2]Department of Bioinformatics, School of Bioengineering, SRM University, Chennai, (India)

ABSTRACT

Since the initial sequencing of the human genome, many projects are underway to understand the effects of genetic variation between individuals. Predicting and understanding the effects of genetic variation using computational methods are becoming increasingly important for Single Nucleotide Polymorphism (SNP) selection in genetic studies and understanding the molecular basis of disease. SNP analysis allows us to find the disease-causing gene much more quickly. Tuberculosis is a very deadly disease and also to some extent it can be genetic. With the help of dbSNP, SMART, SNPS3D databases the clinically important SNP causing Tuberculosis is found. These clinically important genes can be set as markers to find Tuberculosis in multiple species, thus making detection of disease much easier.

Keywords-Biological Markers, Genetic Variations, Mutations, SNP, Tuberculosis.

I. INTRODUCTION

A Single Nucleotide Polymorphism, or SNP, is a small genetic variation, that can occur within a individual's DNA sequence. More than 99% of human DNA sequences are the same across the population, small variations or changes in DNA sequence, such as SNPs, can have a major influence on how humans respond to disease, environmental factors, and medicines. While SNPs do not cause disease, they can help to determine the probability that someone will develop a particular disease. These SNPs can be used as a marker to find disease in multiple species. In our investigation, if a mutation is more frequent across multiple-species and if the mutation can be matched with its phenotype across species, it can be clinically very important.

1.2 Genetic Variations and Mutations

Genetic variation describes naturally occurring genetic differences among individuals of the same species **[1]**. Variation is studied for the better understanding of ourselves as a species and uses this knowledge to improve our health and well-being. Variations are simply differences in genetic sequence. Major deletions, insertions, and genetic rearrangements can affect several genes or large areas of a chromosome at once. Genetic variation is brought about by mutation. A mutation is a permanent change in the DNA sequence of a gene. Mutations in DNA can alter the amino acid sequence of the protein encoded by the gene. If a mutation changes a protein produced by a gene, it will be harmful. In order to function properly, each cell depends on number of proteins to function in the right places at the right times **[2]**. When a mutation alters a protein that plays a critical role in the body, a disease is generated. A condition caused by mutations in one or more genes is called a genetic disorder. Mutations can be of two types-around a whole genome or at a particular point.

TABLE 1: Base pair substitution types and their functions

Base pair substitution types	**Functions**
Silent mutation	Results in a no new amino acid in the protein sequence.
Missense mutation	Results in an amino acid substitution.
Nonsense mutation	Results from substitutions in a protein coding region which may mutate an amino acid codon to a termination codon.
Frame shift mutation	Results from the insertion or deletion of one or more nucleotides in the coding region of a gene.
Inversion mutation	An entire section of DNA is reversed.
Deletion mutations	Results in missing DNA.

1.3 Single Nucleotide Polymorphism

A single-nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a biological species. SNPs are stable, highly abundant, and distributed throughout the genome. SNPs usually occur more in non-coding regions than in coding regions of a gene **[3]**. SNPs which are found within a coding sequence are of particular interest to researchers because they are more likely to alter the biological function of a protein. A synonymous polymorphism is a SNP in which both alleles produce the same polypeptide sequence whereas a replacement polymorphism produces a different polypeptide sequence. A replacement polymorphism may be either missense or nonsense. Replacement polymorphisms are usually the cause of over half of all known disease mutations. Most of the SNPs are not responsible for a disease state **[4].** Instead, they serve as biological markers for a disease on the human genome map, because they are generally located near a gene found to be associated with a certain disease. SNP may sometimes be the cause of a disease and, therefore, can be used to search and isolate the disease-causing gene. Associating SNPs with human disease phenotypes has great potential for direct clinical application by providing new and more accurate genetic markers for diagnostic and prognostic purposes and, possibly, novel therapeutic targets even in multiple species also. SNPs are largely the easiest to ascertain, and the most useful and widely applied biological markers in genetic studies in the modern age. All humans have almost the same sequence of 3 billion DNA bases and the variation is 0.1% only **[5].** Majority of these variations are contributed by SNPs. SNPs serve as excellent biological markers because they occur frequently throughout the genome and tend to be relatively genetically stable. Many common diseases in humans are not caused by a genetic variation within a single gene but are influenced by complex interactions among multiple genes. Environmental and lifestyle factors also play a vital role to add tremendously to the uncertainty of developing a disease, though it is still remain a difficult task to measure and evaluate their overall effect on a disease process. SNPs can provide more accurate results. Recently, it has been suggested that SNPs can be used for homogeneity testing and pharmacogenetic studies. SNP can also be helpful to identify and map various complex and common diseases such as high blood pressure, diabetes, and heart disease **[6].** Although genome-wide association studies (GWAS) have identified many disease-susceptibility single-nucleotide polymorphisms (SNPs), these findings can only explain a small portion of genetic contributions to complex diseases, which is known as the missing heritability **[7]**. SNPs are not only more common than other types of polymorphisms but also occur at a frequency of approximately 1 in 1000 base pairs throughout the human genome and researchers report that more than ten million SNPs have already been identified in the human genome. SNPs that fall into promoter regions of proteins (coding region) comprise only a small fraction of the presently annotated SNPs. Nearly four thousand SNPs have been mapped till date which had the disease/disorder status. SNPs have found to have a greater impact to immune responses to chemicals, pathogens, drugs and vaccines. SNPs that fall into protein domains in the human genome potentially contribute to disease **[8**].

1.4 Tuberculosis

Tuberculosis, MTB or TB, is an infectious disease caused by various strains of *Mycobacterium (Mycobacterium tuberculosis and Mycobacterium bovis)* usually *Mycobacterium tuberculosis.* Tuberculosis is a major public health problem in India. *M.tuberculosis* is a rod-shaped, slow-growing bacterium. Their cell wall has high acid content, which makes it hydrophobic, resistant to oral fluids. Due to these properties, this bacterium has many harmful effects. Tuberculosis typically attacks the lungs but can also affect other parts of the body. The disease is spread through the air when an active TB infected person cough, sneeze, or transmit their saliva through the air. Once infectious particles reach the alveoli, another cell, called the macrophage, engulfs the TB bacteria **[9]**. Then the bacteria are transmitted to the lymphatic system and bloodstream and spread to other organs occurs. About 90% of those infected with *M. tuberculosis* have asymptomatic, latent TB infections (sometimes called LTBI), with only a 10% lifetime chance that a latent infection will progress to TB disease **[10].** However, if untreated, the death rate for these active TB cases is more than 50%. A number of factors make people more susceptible to TB infections. The most important of these is HIV, with co infection present in about 13% of cases **[11]**. The symptoms are loss of weight, loss of energy, poor appetite, fever, a productive cough, and night sweats. Although most initial infections have no symptoms and people overcome them, they may develop fever, dry cough, and abnormalities that may be seen on a chest X-ray. This is known as pulmonary tuberculosis. Pulmonary tuberculosis frequently goes away by itself, but in 50%-60% of cases, the disease can return. About 15% of people may develop tuberculosis in an organ other than their lungs. About 25% of these people usually had known TB with inadequate treatment. The most common sites include the following- lymph nodes, genitourinary tract, bone and joint sites, meninges. In the beginning TB can be diagnosed by X-ray that leads to the suspicion of infection. The Mantoux skin test also known as tuberculin skin test (TST or PPD test). This test helps identify people infected with *M. tuberculosis* but who have no symptoms **[12]**. QuantiFERON-TB Gold test can help detect active and latent tuberculosis. Sputum testing for acid-fast bacilli is the only test that confirms a TB diagnosis. Sputum testing in lab may confirm up to 30% of people with active disease. The only currently available vaccine is Bacillus Calmette-Guérin (BCG) **[13]**, which, while effective against disseminated disease in childhood, confers inconsistent protection against pulmonary disease. It is the most widely used vaccine worldwide, with more than 90% of children vaccinated.

II. METHODOLOGY

The information about tuberculosis was searched in **OMIM database** and the limit was given as allelic variants. Thirteen genes were identified and the respective SNPs were collected. From the OMIM database, the link-gene view in **dbSNP** was selected. The SNP: GeneView display tabulates SNPs mapped to transcript variants of a particular gene. Each entry in database is given as reference sequence id. These are arranged from different species or from different chromosome position if they have same SNPs. Next step was to find out whether the

SNP has an impact on protein structure and function by going to the **SNPs3D** database. The result shows the SNP analysis on the particular gene which we have given, along with the information including refseq accession number, snps, snp_ids, molecular effect, etc. If the particular SNP is shown in red color, then we can say that it has an effect on the protein structure i.e., it will cause mutation. **UniProt database** was searched from protein sequences of diseased SNP. FASTA sequence was submitted in **SMART Database** (Simple Modular Architecture Research Tool) which allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. The result showed the protein domains where the mutation has occured. Here the protein domains that were present are SAND,PHD and BROMO. Similar diseased SNPs were found in multiple species by searchung SNP gene in dbSNP. The allelic change and the residue change in multiple species which were found similar to that of the *Homo sapiens* were selected. The allelic change as well as the protein domain that were similar in Homo sapiens and *Rattus norvegicus* for the gene SP110.The allelic change and protein domain that were similar in Homo sapiens and *Bos Taurus* for the gene CISH. Homology modeling was done for those proteins without structures. The **prime software** is used for homology modeling.

III. RESULTS AND DISCUSSIONS

The OMIM database was searched for disease tuberculosis and all the related genes were collected. The identified genes are **TIRAP, IL12RB1, SLC11A1, CD209, SP110, IFNGR1, STAT1, IFNG, CISH, CCL2, CYBB, IKBKG, and MBL2.** Each of these was again searched in dbSNP for SNP analysis. From the database, all exons under missense mutation were tabulated. Exons under missense mutation had more diseased SNPs.

TABLE 2: All Diseased SNPS With Respective Ids Were Obtained From SNPS3D Database.

GENES	SNP	SNP ID
TIRAP	D96N	8177400
IL12RB1	P47S	17887176
	A525T	11575935
CD209	K223N	11465379
	L242V	11465380
SP110	A128V	11556887
	M523T	1135791
STAT1	P27T	11549696
	I30T	34255470
	P538L	1803838
CISH	H162Q	419160
	P185H	419346
MBL2	R52C	5030737
	G54D	1800450
	G57E	1800451

The protein domain where mutation was found was obtained from SMART tool. Of the all genes from SNPS3D database, only CD209, SP110, STAT1, CISH and MBL2 have clinically important SNPs. Protein domain of CD209 gene in humans was **CLECT** with a starting position 256 and ending position 378. The diseased SNP was found to be in exon3c. Mainly this domain functions as calcium-depended carbohydrate binding molecule.

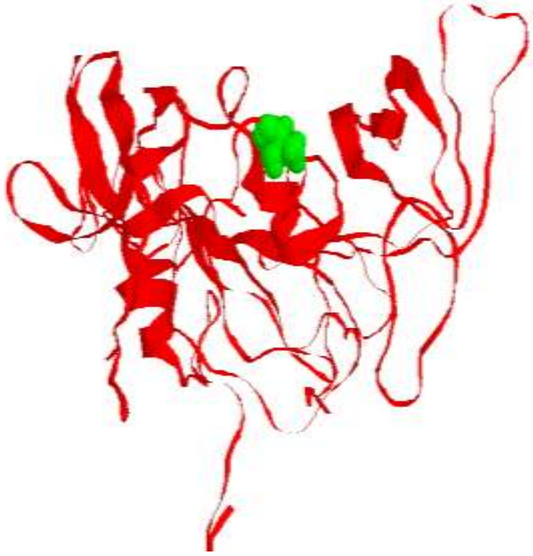
The resultant SNP id is **rs187439891** with mRNA position 293, observed allele change from **G->A** and protein position at 76 with a change in residue **A[Ala]->T[Thr]**. In case of SP110 gene, this has three domains: **SAND** with starting position 462 and ending at 535, **PHP** with start at 536 and end at 578 and lastly **BROMO** with a start at 581 and ending at 676. Of these three **SAND** domain is having clinically important SNP, **rs113453818** with an allele change of **A->G** at mRNA position 544 and having a possible residue change of **S[Ser]->G[Gly]** at 102 position. The SAND domain (named after Sp100, AIRE-1, NucP41/75, DEAF-1) is a conserved ~80 residue region found in a number of nuclear proteins, many of which function in chromatin-dependent transcriptional control. These include proteins linked to various human diseases, such as the Sp100 (Speckled protein 100 kDa), NUDR (Nuclear DEAF-1 related), GMEB (Glucocorticoid Modulatory Element Binding) proteins and AIRE-1 (Autoimmune regulator 1) proteins. STAT1 gene has **STAT-int** protein domain with starting position at2 and ending position at 122. This gene does not have a clinically important SNP domain. In case of CISH gene, **SOCS** is the domain with a start at 214 and end at 255.With an allele change from **A->C** and SNPid- rs79987458 .They have mRNA position at 303 and a residue change at position 6 with a change from T[Thr]->P[Pro].They are suppressors of cytokine signaling. It has been proposed that this could couple bound proteins to the ubiquitination or proteasomal compartments. SOCS proteins interact with multiple tyrosine kinases activated signaling proteins and inhibit cytokine signal transduction by binding to tyrosine kinase receptors. SOCS proteins regulate cell growth and survival, placental development, hematopoiesis, and T-cell allergic responses.MBL2 gene also have **CELT** protein domain from position 127 to position 245. Four SNP ids were identified:

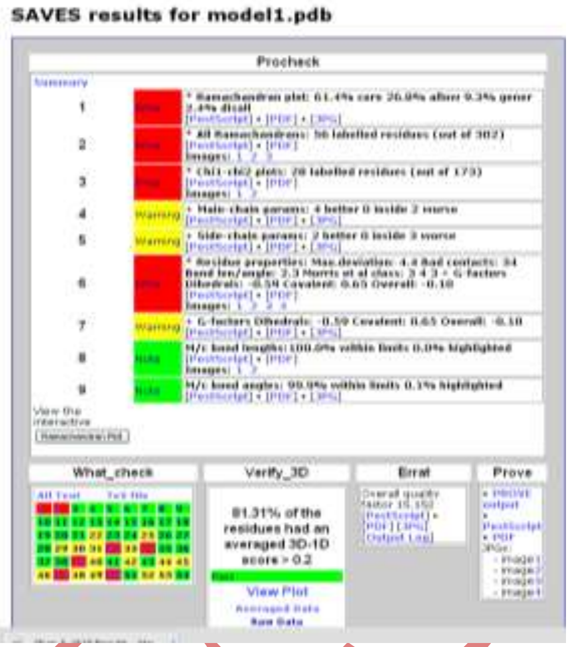
rs id	mRNA position	Allele change	Protein position	Residue change
rs72661128	139	G->C	25	C[Cys]->S[Ser]
rs146004726	181	G-.A	39	S[Ser]->N[Asn]
rs148483303	183	T->G	40	S[Ser]->A[Ala]
rs148078249	220	G->A	52	R[Arg]->H[His]

With the help of diseased SNP diagnosed in humans, tuberculosis was identified in multiple species. Only two genes SP110 and CISH were having same clinical importance as that of humans. These two genes were again searched in dbSNP to find disease spots in multiple species.

SP110 has rs175495167, rs175495166, rs8146428, rs198014247, rs108886247, rs108811220, rs108261062, rs108058919, rs49894818, rs13474185, rs13474187, rs108624819, rs108207432, rs108145509, rs107807772, rs107678013, rs50143038, rs47718372, rs36238733 as defective SNPs from these 3 species: *Pongo abelii* , *Rattus norvegicus* , *Mus musculus.* Similarly, in case of CISH: rs137311656, rs134059616, rs133317693, rs107024373, rs8170068 were identified from the species: *Bos taurus*, *Rattus norvegicus.*

SAND domain was found to be common to both humans as well as *Rattus norvegicus.* They has same clinical importance with allele change **T->G.** The diseased SNP id is **rs198014247**. Observed protein residue change is at **N[Asn]->K[Lyn]** in 232 position. SOCS domain was found to be common to both humans as well as *Bos Taurus.* They have same clinical importance with allele change A -> C**.** The diseased SNP id is rs137311656. Observed protein residue change is at **N[Asn]->T[Thr]** in 228 position.





SAVES results for model1.pdb

Procheck

Summary

1 Error * Ramachandran plot: 61.4% core 26.0% allow 9.3% gener 3.4% disall
[PostScript] • [PDF] • [JPG]

2 Error * All Ramachandrans: 56 labelled residues (out of 302)
[PostScript] • [PDF]
Images: 1 2 3

3 Error * Chi1-chi2 plots: 20 labelled residues (out of 173)
[PostScript] • [PDF]
Images: 1 2

4 Warning + Main-chain params: 4 better 0 inside 2 worse
[PostScript] • [PDF] • [JPG]

5 Warning + Side-chain params: 2 better 0 inside 3 worse
[PostScript] • [PDF] • [JPG]

6 Error * Residue properties: Max.deviation: 4.4 Bad contacts: 34
Bond len/angle: 2.3 Morris et al class: 3 4 3 + G-factors
Dihedrals: -0.59 Covalent: 0.65 Overall: -0.10
[PostScript] • [PDF]
Images: 1 2 3 4

7 Warning + G-factors Dihedrals: -0.59 Covalent: 0.65 Overall: -0.10
[PostScript] • [PDF] • [JPG]

8 Note M/c bond lengths: 100.0% within limits 0.0% highlighted
[PostScript] • [PDF]
Images: 1 2

9 Note M/c bond angles: 99.9% within limits 0.1% highlighted
[PostScript] • [PDF] • [JPG]

View the interactive
Ramachandran Plot

What_check
All Text Txt file

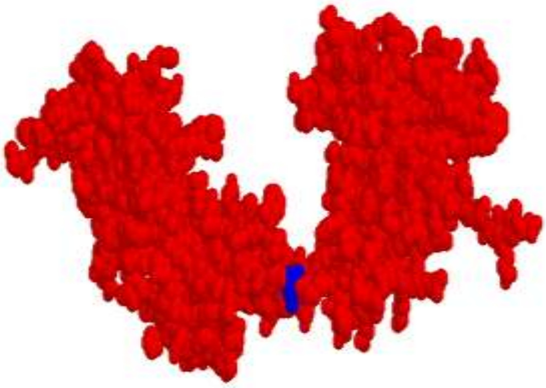
Verify_3D
81.31% of the residues had an averaged 3D-1D score > 0.2
Pass
View Plot
Averaged Data
Raw Data

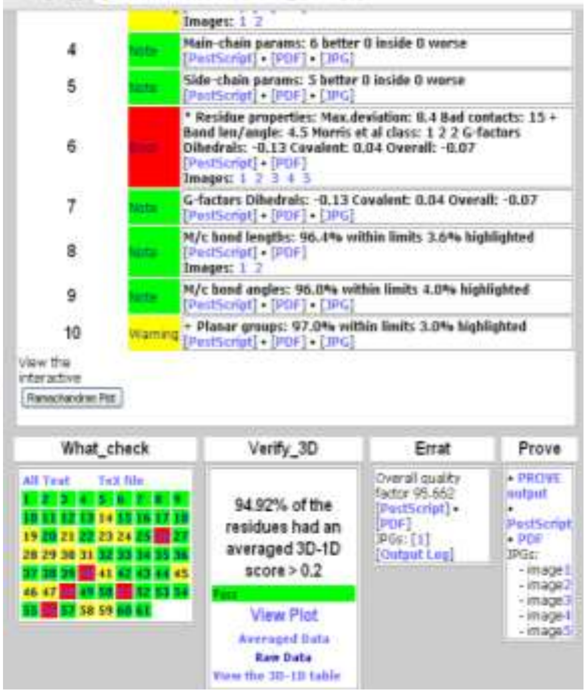
Errat
Overall quality factor 15.152
[PostScript] • [PDF] [JPG]
[Output Log]

Prove
• PROVE output
• PostScript
• PDF
JPGs:
- image1
- image2
- image3
- image4

Figure(left) **Modeled structure of CISH protein in *Bos Taurus.*** Figure(right) **Saves server result for modelled structure of CISH protein in *Bos Taurus***

Homology modeling was done for *Rattus norvegicus* and *Bos Taurus.* The green region shows the amino acid change of **Thr** residue at position 228 in protein sequence. Homology modeling was done using PRIME software and was validated using SAVES server. Green colored residues indicates clinically diseased SNP.





Images: 1 2

4 Note Main-chain params: 6 better 0 inside 0 worse
[PostScript] • [PDF] • [JPG]

5 Note Side-chain params: 5 better 0 inside 0 worse
[PostScript] • [PDF] • [JPG]

6 Error * Residue properties: Max.deviation: 8.4 Bad contacts: 15 +
Bond len/angle: 4.5 Morris et al class: 1 2 2 G-factors
Dihedrals: -0.13 Covalent: 0.04 Overall: -0.07
[PostScript] • [PDF]
Images: 1 2 3 4 5

7 Note G-factors Dihedrals: -0.13 Covalent: 0.04 Overall: -0.07
[PostScript] • [PDF] • [JPG]

8 Note M/c bond lengths: 96.4% within limits 3.6% highlighted
[PostScript] • [PDF]
Images: 1 2

9 Note M/c bond angles: 96.0% within limits 4.0% highlighted
[PostScript] • [PDF] • [JPG]

10 Warning + Planar groups: 97.0% within limits 3.0% highlighted
[PostScript] • [PDF] • [JPG]

View the interactive
Ramachandran Plot

What_check
All Text Txt file

Verify_3D
94.92% of the residues had an averaged 3D-1D score > 0.2
Pass
View Plot
Averaged Data
Raw Data
View the 3D-1D table

Errat
Overall quality factor 95.662
[PostScript] • [PDF]
JPGs: [1]
[Output Log]

Prove
• PROVE output
• PostScript
• PDF
JPGs:
- image1
- image2
- image3
- image4
- image5

Figure (left): Modeled structure of SP110 protein in *Rattus norvegicus.* Figure (right): Saves server result for SP110 protein in *Rattus norvegicus*

Since the sequence length of *Rattus norvegicus* was greater and there were no results from BLAST to do homology modeling, we used a protein threading tool called MUSTER. The resultant protein structure was viewed in RASMOL viewer. In the above Fig blue colored atoms shows ASN residue at position 232 which is diseased in humans as well as *Rattus norvegicus.*The blue colored residues indicate the diseased SNP.

IV. CONCLUSION

From the above mentioned databases and methodology, we found that genes **SP110** and **CISH** are having damaging effects in genome of humans and also in multiple species. Talking about multiple species, *Bos taurus* have **rs137311656** as a damaging SNP with an amino acid change of **N[Asn]->T[Thr]** at position 228 and *Rattus norvegicus* have **rs198014247** as damaging SNP with a change of **N[Asn]->K[Lyn]** at position 232. Tuberculosis is a very deadly disease and the methods existing now to detect this disease will take more time. With the help of these SNP markers, it is easier to identify the disease and cure as soon as possible. SNP analysis method can be used anywhere with good validation results. From our study we can conclude that genes SP110 and CISH are SNPs with clinical importance and any species with these genes are prone to this deadly disease. Tuberculosis remains in the body for long time without causing any disease until and unless a favorable mutation is occurred. These SNP markers can be used as a kit to detect mutation, determine the disease and prevent it as early as possible.

REFERENCES

[1] Arthur M.Lesk. *Introduction to Genomics* (New York, Oxford, 2008)

[2] Beth A. Montelone, *Human Genetics* (BIOL400, 1998).

[3] Rune Andreassen, Sigbjørn Lunner and Bjørn Høyheim , Targeted SNP discovery in Atlantic salmon (*Salmo salar*) genes using a 3'UTR-primed SNP detection approach . *BMC Genomics . 11:706.2010.*

[4] Jeremy W.Dale and Malcolm von Schantz. *From Genes to Genomes* (Wiley,Blackwell,2008)

[5] Guoliang Zhang, Xinchun Chen, Long Chan, Mingxia Zhang, Baohua Zhu, Lantian Wang, Xiuyun Zhu, Jieyun Zhang, Boping Zhou & Junwen Wang. An SNP selection strategy identified IL-22 associating with susceptibility to tuberculosis in Chinese**.** *Scientific Reports 1.* (2011). *20: 10.1038*

[6] Barkur S. Shastry.. SNP alleles in human disease and evolution. *Jpn Soc Hum Genet and Springer-Ver4l6a0g0.* (2002).47:561-566

[7] Can Yang, Xiaowei Zhou, Xiang Wan, Qiang Yang, Hong Xue and Weichuan Yu. Identifying disease-associated SNP clusters via contiguous outlier detection. *Bioinformatics. 2011 .27: 2578-2585*

[8] Theresa A. Sergel, Lori W. McGinnes, and Trudy G. Morrison. A Single Amino Acid Change in the Newcastle Disease Virus Fusion Protein Alters the Requirement for HN Protein in Fusion. *Journal of Virology*. 2000.74:5101-5107

[9] Houben E, Nguyen L, Pieters J. "Interaction of pathogenic mycobacteria with the host immune system". *Curr Opin Microbiol* ***9*** *(1): 76–85. (2006)*

[10] Skolnik, Richard (2011). *Global health 101* (2nd ed.). Burlington, MA: Jones & Bartlett Learning. p. 253. ISBN 978-0-7637-9751-5.

[11] Arch G. Mainous III, Claire Pomeroy, (2009). *Management of antimicrobials in infectious diseases : impact of antibiotic resistance.* (2nd rev. ed.). Totowa, N.J.: Humana. p. 74. ISBN 978-1-60327-238-4.

[12] Escalante, P (2 June 2009). "In the clinic. Tuberculosis". *Annals of internal medicine* **150** (11): ITC61–614; quiz ITV616. PMID 19487708.

[13] McShane, H (12 October 2011). "Tuberculosis vaccines: beyond bacilli Calmette–Guérin". *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **366** (1579): 2782–9. doi:10.1098/rstb.2011.0097. PMC 3146779..