

***IN-SILICO* MOLECULAR ANALYSIS AND APPLICATION OF REVERSE
VACCINOLOGY APPROACH TO IDENTIFY PROTEIN SUBUNITS FOR
VACCINE DEVELOPMENT AGAINST *TROPHERYMA WHIPPLEI***

**A
THESIS
SUBMITTED
TO**

For the award of

DOCTOR OF PHILOSOPHY (Ph.D.)

**In
Biochemistry**

**By
AMIT JOSHI
ENROLLMENT NO: 41800810**

**Supervised By:
Dr. Vikas Kaushik**



Transforming Education Transforming India

**LOVELY FACULTY OF TECHNOLOGY AND SCIENCES
LOVELY PROFESSIONAL UNIVERSITY
PUNJAB
2022**



DECLARATION

The thesis entitled “*In-Silico* Molecular Analysis and Application of Reverse Vaccinology Approach to Identify Protein Subunits for Vaccine Development Against *Tropheryma whipplei*” is conducted under the supervision of **Dr. Vikas Kaushik**, Professor, School of Bioengineering and Biosciences, Lovely Professional University, Phagwara, Punjab, India.

Herewith I declare that the information reported is the result of my research work. The thesis has not been accepted for any degree and is not concurrently submitted to any candidature for any other degree or diploma.



AMIT JOSHI

DATE 15-07-2022

41800810



LOVELY
PROFESSIONAL
UNIVERSITY

Transforming Education Transforming India

CERTIFICATE

This is to certify that thesis entitled “***In-Silico* Molecular Analysis and Application of Reverse Vaccinology Approach to Identify Protein Subunits for Vaccine Development Against *Tropheryma whipplei***” submitted by Mr. Amit Joshi (41800810) in the partial fulfillment of the requirement for the degree of Doctor of Philosophy in Biotechnology is a record of research work done by him during the period of his research study in Lovely Professional University under my guidance and supervision. This thesis has not been previously formed the basis of the award for the degree, diploma, associate ship, fellowship, or any other similar titles.

Vikas Kaushik

Supervisor

DATE 15-07-2022

Dr. Vikas Kaushik

Professor, School of Bioengineering and Biosciences

Lovely Professional University, Phagwara Punjab (India)

Acknowledgements

This PhD has been a completely life-changing experience for me, and I would not have been able to complete it without the help and advice of many people. First and foremost, I'd like to express my gratitude to my guide sir, Dr. Vikas Kaushik, for all of his support and motivation.

My research work is a continuous effort of all godly powers working behind me, including my University, Guide sir, Family, Colleagues and off course nature/ environment. Every day I woke up and look in to my mobile for routine email check, there is always at least 2-3 mails of guide sir providing new guidelines, discussions, training platforms, research problems and latest article links, and by the end of the day around 12 o'clock I was able to understand the concept that he tried to impart on my cognitive domain.

I am extremely thankful towards my guide Dr. Vikas Kaushik, who not only understood my problems but also provided ways to resolve them. My brain dictionary has very less words to define his passionate interest on research, but still I would like to say he is one of the best human being I ever found in my career, because to hold a set of degrees not makes a person good human being, this thing comes from experience of real ups and downs and a continuous conscious struggle to hold such a great positive psychological foundation in dealing research candidates.

Lovely professional university is one of the great battle ground for academicians, researchers and intellectual scholars those who want to bring philosophy back to the research, as this degree PhD holds secret ingredient of philosophy which seems to be lost in current scenario. It is the only University in the Globe that open doors not only for experts and rich intellectuals but also for socially backward, financially challenged or economically deprived learners. My deep appreciations for respected fellow professors and whole research fraternity for conducting my end term presentations smoothly, and provide useful suggestions for improving the quality of my research work and thesis.

I am always thankful to all universities where I worked during my research tenure, Kumauni University, Invertis University, these Universities not only provided better computational facility to me but also intellectuals interaction for framing my research in unidirectional way of success. I am also thankful for research support of my friends, Teachers and colleagues Mr. Sunil Krishnan (LPU), Mr. Nahid Akhtar (LPU), Dr. Dinesh Chandra Pathak (IVRI), Dr. Inderpal

Singh (Sri Mata Vaishno Devi University, Jammu), Dr. Pankaj Rai (Invertis University, Bareilly), Dr. Shashank Upadhyay (Invertis University, Bareilly), Dr. Mohd Amin-ul Manan (LPU), Dr. Keshwananda Tripathi (IARI), Dr. Vandana (AIIMS), Dr. Parabhjeet Singh (GADVASU, Ludhiana), Dr. Joginder Pal Singh(LPU) and Dr. Ashutosh Dubey (GBPUA&T, Pantnagar).

Very special thanks to my father (Mr. Rajendra Joshi), mother (Mrs. Kamla Joshi), my wife (Mrs. Preeti Joshi), my brother (Mr. Sumit Joshi), and my whole family who are always in support of my educational development, and provide emotional support and stability.

Table of Contents

SI. No.	Chapter Title	Page No.
1.	Abstract	1-3
2.	Chapter 1: Introduction	4-10
3.	Chapter 2: Review and Literature	11-20
4.	Chapter 3: Objectives & Scope of Study	21-23
	Chapter 4: Material and Methods	24-33
	4.1. Genomic trends analysis in <i>Tropheryma whipplei</i> and Phylogenetic analysis	25
	4.2. Transcriptomic data analysis of model organism affected by <i>Tropheryma whipplei</i>	25-30
5.	4.2.1. Retrieval of <i>Whipplei</i> microarray gene expression profile datasets	
	4.2.2. Differentially expressed genes comparison analysis	
	4.2.3 Gene Functional Classification and Identification of functional related gene	
	4.3. T-cell Epitope prediction for peptide vaccine crafting	30-32
	4.3.1. Retrieval of Proteins for <i>T. whipplei</i>	
	4.3.2. Allergenicity Prediction for Proteins	
	4.3.3.T-Cell Epitope Prediction	
	4.3.4. Epitope's Molecular Docking with HLA Alleles	
	4.3.5. Analysis of Population Coverage	
	4.3.6. Molecular Dynamics and Simulations	
	4.4. <i>In-silico</i> drug discovery against <i>Tropheryma whipplei</i>	32-33
	4.4.1. Enzyme Selection Bias	
	4.4.2. Enzyme Annotation studies	
	4.4.3. Structural analysis: Docking & Simulation	
6.	Chapter 5: Results and Discussion	34-121
	5.1. Transcriptomic data analysis for model organism to reveal host-pathogen interaction	35-85
	5.1.1 <i>T. whipplei</i> microarray profile data	
	5.1.2 Identification of differentially expressed genes	
	5.1.3 Gene Functional Classification and identification	
	5.1.4 Core functionality and interpretation	
	5.2. Codon usage and Amino acid usage patterns in <i>Tropheryma whipplei</i>	86-91
	5.3.T-Cell Epitope based vaccine prediction	91-103
	5.3.1 Proteins sequence retrieval and non-Allergen	

	determination	
	5.3.2 T-Cell Epitope prediction	
	5.3.3 Molecular 3D modeling of epitopes and HLA alleles	
	5.3.4 Molecular Docking of epitopes and HLA alleles	
	5.3.5 Toxicity prognostication of putative vaccine targets	
	5.3.6 Population coverage analysis of epitopes	
	5.3.7 Molecular Dynamic and Simulation Studies	
	5.4. <i>In-silico</i> drug prediction against <i>Tropheryma whipplei</i>	103-108
	5.5. Codon usage studies and epitope-based peptide vaccine prediction	109-121
	5.5.1. <i>Tropheryma whipplei str. Twist</i> codon usage	
	5.5.2. Rare and very rare codons	
	5.5.3. Codon usage measurements	
	5.5.4. Epitope based vaccine prediction: Application of Codon usage studies	
7.	Chapter 6: Summary and Conclusion	122-127
8.	Chapter 7: Future Aspects	128-130
9.	Bibliography	131-143
10.	Appendices	144-209

List of Tables

Table No.	Table Titles	Page No.
1.	List of microarray experiments accessed from NCBI Geo-Datasets for microarray analysis	29-30
2.	Multiple Parameters for encoded proteins	91
3.	Allergenicity results for analyzed proteins of <i>T. whipplei</i>	92
4.	Predicted epitopes based on NetMHCII 3.2 server and VaxiJen score	92-94
5.	HLA Template model based on Pdb Id derived for MHC Class II alleles Structure from RCSB-PDB	94
6.	Molecular docking outcomes of epitopes with HLA alleles using patch dock	95-96
7.	Selected multi-target epitopes and toxicity score based on TOXIN-PRED tool	98
8.	MHCPRED results for selected epitopes with their desired HLA DRB1 allele's binders	98
9.	Selected proteins information for <i>Tropheryma whipplei</i> , that are found to be drug targets	105
10.	AutoDock-vina docked results: Binding energies of best docked complexes	105
11.	Drug characteristics analyzed by PubChem database and SwissADME	108
12.	Effective Number of Codon Pairs for each <i>T. Whipplei</i>	111
13.	<i>Tropheryma whipplei</i> RefSeq codon table contains 88597 CDSs	112
14.	<i>Tropheryma whipplei</i> str. Twist 808 CDS	113
15.	<i>Tropheryma whipplei</i> TW08/27 783 CDS	114
16.	<i>Tropheryma whipplei</i> strain Twist 23S ribosomal RNA gene	115
17.	<i>Tropheryma whipplei</i> str. Twist 16S ribosomal RNA	115
18.	AllergenFP score and proteins considered for <i>Tropheryma whipplei</i>	117-119
19.	Peptides showing interaction to HLA-DRB0101, NETMHCII PAN 4.0 server results and VaxiJen score	119
20.	ProtParam results: Biochemical properties of epitopes	119
21.	ACE Value, Global energy, and Binding energy for selected docked complexes	120

List of figures

Fig No.	Figure legends	P. No.
1	Whipple's disease histological changes	6
2	Vaccination advancement in modern world	9
3	Spread of Whipple's' disease in USA in different races	12
4	Whipple's disease aetiology from a clinical standpoint	13
5	<i>T. whipplei</i> increases HLA-G expression and in turn reduces TNF that assist in promoting its own replication in monocytes	15
6	Microbial pathogenesis and vaccine development interaction	16
7	Schematic process of vaccine immunization cellular-physiology	18
8	Microarray analysis plots for group Cold shock temperature 4C Vs 37C under series GSE3693	36
9	Microarray analysis plots for group Cold shock temperature 28C Vs 37C under series GSE3693	37
10	Microarray analysis plots for group Heat shock temperature 43C Vs 37C under series GSE3693	38
11	Microarray analysis plots for group Doxycycline @ 0.5 mg/l vs <i>Tropheryma whipplei</i> Twist strain	40
12	Microarray analysis plots for group Doxycycline @ 5 mg/l vs <i>Tropheryma whipplei</i> Twist strain	41
13	Microarray analysis plots for group ART1 Vs <i>Tropheryma whipplei</i> Twist strain	43
14	Microarray analysis plots for group DigNeuro14 Vs <i>Tropheryma whipplei</i> Twist strain	44
15	Microarray analysis plots for group Dig7 Vs <i>Tropheryma whipplei</i> Twist strain	45
16	Microarray analysis plots for group Dig9 Vs <i>Tropheryma whipplei</i> Twist strain	46
17	Microarray analysis plots for group Dig10 Vs <i>Tropheryma whipplei</i> Twist strain	47
18	Microarray analysis plots for group Dig15 Vs <i>Tropheryma whipplei</i> Twist strain	48
19	Microarray analysis plots for group DigADP11 Vs <i>Tropheryma whipplei</i> Twist strain	49
20	Microarray analysis plots for group DigMusc17 Vs <i>Tropheryma whipplei</i> Twist strain	50
21	Microarray analysis plots for group DigNeuro18 Vs <i>Tropheryma whipplei</i> Twist strain	51
22	Microarray analysis plots for group Endo5 Vs <i>Tropheryma whipplei</i> Twist strain	52
23	Microarray analysis plots for group Endo7 Vs <i>Tropheryma whipplei</i> Twist strain	53
24	Microarray analysis plots for group Neuro1 Vs <i>Tropheryma whipplei</i> Twist strain	54
25	Microarray analysis plots for group Neuro2 Vs <i>Tropheryma whipplei</i> Twist strain	55
26	Microarray analysis plots for group Slow1B Vs <i>Tropheryma whipplei</i> Twist strain	56
27	Microarray analysis plots for group Slow2 Vs <i>Tropheryma whipplei</i> Twist strain	57
28	Microarray analysis plots for group BMDM_ <i>Tropheryma whipplei</i> Twist strain Vs BMDM_Control	59

29	Microarray analysis plots for group IL-16 Knock out Vs BMDM_ <i>Tropheryma whipplei</i> Twist strain	61
30	Microarray analysis plots for group BMDM_ LPS Vs BMDM_ Control	63
31	Microarray analysis plots for group IL-16 Knockout BMDM_ LPS Vs IL-16 Knockout BMDM_ Control	64
32	Microarray analysis plots for group <i>T. whipplei</i> infected DC's Vs Unstimulated DC's	66
33	Microarray analysis plots for BCG Treated group (Patient 1,2,3 Vs Control1)	68
34	Microarray analysis plots for non-stimulated group (Patient 1,2,3 Vs Control1)	69
35	Microarray analysis plots for <i>Tropheryma whipplei</i> infected group (Patient 1,2,3 Vs Control1)	70
36	Microarray analysis plots for BCG Treated group (HET1,2,3 Vs Control1)	71
37	Microarray analysis plots for non-stimulated group (HET1,2,3 Vs Control1)	72
38	Microarray analysis plots for <i>Tropheryma whipplei</i> infected group (HET1,2,3 Vs Control1)	73
39	Microarray analysis plots for BCG Treated group (WT1,2,3,4 Vs Control1)	74
40	Microarray analysis plots for non-stimulated group (WT1,2,3,4 Vs Control1)	75
41	Microarray analysis plots for <i>Tropheryma whipplei</i> infected group (WT1,2,3,4 Vs Control1)	76
42	Microarray analysis plots for BCG treated group (Patient 1,2,3 Vs Control2)	77
43	Microarray analysis plots for non-stimulated group (Patient 1,2,3 Vs Control2)	78
44	Microarray analysis plots for <i>Tropheryma whipplei</i> infected group (Patient 1,2,3 Vs Control2)	79
45	Microarray analysis plots for BCG Treated group (HET1,2,3 Vs Control2)	80
46	Microarray analysis plots for non-stimulated group (HET1,2,3 Vs Control2)	81
47	Microarray analysis plots for <i>Tropheryma whipplei</i> infected group (HET1,2,3 Vs Control2)	82
48	Microarray analysis plots for BCG Treated group (WT1,2,3,4 Vs Control2)	83
49	Microarray analysis plots for non-stimulated group (WT1,2,3,4 Vs Control2)	84
50	Microarray analysis plots for <i>Tropheryma whipplei</i> infected group (WT1,2,3,4 Vs Control2)	85
51	Dendrogram based on genomic blast	86
52	BLAST results for Twist and TW 08/27 Strains	87
53	BLAST score per cent identity for Twist and TW08/27 strains	87
54	Results of Phylogenetic analysis among 4 strains of <i>Tropheryma whipplei</i>	88
55	Correspondence study on synonymous transcriptome codon usage	89
56	Correspondence study on Relative amino acid usage	90
57	Molecular docking between epitope IRYLAALHL and HLA	96
58	Molecular docking of epitope VLMVSAFPL and HLA	97
59	Graphical depiction of docked and selected Epitopes- HLA Alleles corresponding to their atomic contact energies.	97
60	Graphical representation along of IEDB Population coverage for VLMVSAFPL	99
61	Graphical representation of IEDB Population coverage for IRYLAALHL	99

62	RMSD Value Vs Time (ps) for Epitope IRYLAALHL	100
63	RMSD Value vs Time (ps) plot for epitope VLMVSFAPL	100
64	Molprobit Ramachandran Plot analysis results for Epitopes Chemical structure of drugs obtained from PubChem database and analyzed in	101
65	Pymol	106
66	Molecular interactions between drugs and receptor	107
67	Molecular simulation analysis of drugs complexed with enzymes of <i>T. whipplei</i>	108
68	Taxonomy and strains of <i>Tropheryma whipplei</i>	110
69	Heatmap of log-transformed codon usage	111
70	Codon frequencies of <i>Tropheryma whipplei</i>	112
71	Dinucleotide frequencies of <i>Tropheryma whipplei</i>	113
72	rRNA length & nucleotide composition	116
73	Percentage of rRNA nucleotide composition	116
74	Percentage of rRNA synonymous codons	117
75	Codon measurement values plot	117
76	Molecular Docking results of epitopes with HLA-DRB-0101	121

Abbreviations

HLA	Human Leuckocyte Antigen
MHC	Major Histocompatibility factor
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
MD	Molecular Dynamics
RSCU	Relative Synonymous Codon Usage
RAAU	Relative Amino Acid Usage
IL	InterLeukin
TNF	Tumor Necrosis Factor
CAI	Codon Adaptive Index
BMDM	Bone Marrow Derived Macrophages
GEO	Gene Expression Omnibus
NCBI	National Center for Biotechnology Information Database for Annotation, Visualization and Integrated
DAVID	Discovery
KEGG	Kyoto Encyclopedia of Genes and Genomes
TW	Tropheryma whipplei
CADD	Computer Aided Drug Discovery

ABSTRACT

Abstract

Tropheryma whipplei bacterium is harmful as it is associated with many diseases, not only in humans, but also in animals. It is involved in multi-systemic disorders like gastroenteritis, lipodystrophy, neuronal damages and also affects immune system. Currently very few drugs are available to treat Whipple's disease caused by this bacterium in humans. So, our study aims to develop novel therapeutics by deploying modern computational biology approaches against Whipple's disease. Proteomic analysis of different proteins of bacterium *Tropheryma whipplei* leads to select antigenic, non-allergenic epitopes that can elicit immune response in human population. Computer based drug designing after screening various pharmacoeactive molecules, and metabolic enzymes of bacterium. Molecular dynamics and simulation analysis use for validation for interaction of drugs and vaccine candidates. In this study we analyze microarray transcriptomic data related to model organism affected by strains of *Tropheryma whipplei*. Also, Codon usage and amino acid usage was analyzed to determine target antigens. The core objective of the study was to employ reverse vaccinology approach for the identification and prediction of *Tropheryma whipplei* candidate antigens. This was useful in predicting epitope-based vaccine prediction to control Whipple's disease. Also we conducted designing of potent and effective drug against this pathogen. NCBI-GenBank database was used to retrieve genomic as well as proteomic data sets of *Tropheryma whipplei* bacterium. *In-silico* tools was used like codonW for codon usage analysis, NetMHC servers for epitope screening, ADMET analysis for pharmacoeactive compounds screening, AutoDock, PatchDock and DINC server for Molecular docking, and Gromacs, MD-Web server, Desmond like tools for molecular dynamics and simulation studies for validation of results were used. Many other computational tools like PepFold, Molprobit, ProtParam, ExPASy tools was deployed for structural and physio-chemical properties assessment for various interacting macromolecules. Many other databases like PubChem, chembridge, and RCSB-PDB were also used to study chemical properties during drug assessment. For structural visualization PyMOL was

used. IL-16 up-regulation in macrophages results in under-expression of defensive genes (IFN- γ , Interferons, ICAM, and Complement C3etc) and enhance bacterial multiplication in macrophages. In this study we obtained VLMVSAFPL and IRYLAALHL as predicted epitopes for vaccine crafting. *In-silico* investigation discovers that 2-APC, NMN, and RFMP as possible medications to treat Whipple's disease. These modern computational approaches are still very new, but provide easy predictions of epitope-based vaccines and drugs against *Tropheryma whipplei*. These epitope and drugs can be used after wet-lab validations and animal testing's. This study opens doors for fast and efficient vaccine and drug discovery against many other viral and bacterial diseases. In our current research we included social welfare concept and conducted successful epitope-based vaccine prediction for SARS-CoV2, and Dengue like viruses also.

CHAPTER 1

INTRODUCTION

INTRODUCTION

Tropheryma whipplei causes an infectious disease that can be extremely fatal if not cured by using good drugs and vaccines. In modern scenario, it is present mostly in Caucasian populations of Northern America and Europe. This bacterium shows extreme infection in gastrointestinal tract and valve sites of heart primarily responsible for gastroenteritis and endocarditis like problems in humans, which is termed as Whipple's disease in respect of G.H Whipple who first described role of this bacteria in lipodystrophy, a kind of malfunction in lipid metabolic reactions and absorption in patients at 1907. *Tropheryma whipplei* also infect animals, where it can cause colitis in dogs. Currently, hydroxychloroquine and doxycycline drugs are used for treatment of this disease which is effective but required long time treatment up to 2 years with life time follows up by gastroenterologist. Currently geographical stretch of *Tropheryma whipplei* bacterium is Canada, France, Germany, and Portugal; more specifically in Caucasian race but it is a main causative bacterium to study as it was observed in different pathological diseases likely classical Whipple's disease, gastroenteritis, and in some neuropathies and endocarditis etc.

George Hoyt Whipple in 1907 explained Whipple's disease, as a multisystemic chronic infectious disease. He identified silver stained rod-shaped bacterium in vacuoles associated with macrophages of patients, but initially he did not think of them as the cause for the disease rather he thought that intestinal lipodystrophy (Whipple's disease) was caused due to some novel disturbances in fat metabolic schemes (**Whipple GH, 1907**). When the first successful treatment started by using antibiotics in 1952, determined that this bacterium might be the major causative agent of this disease (**Paulley JW, 1952**). An electron microscopic study in 1960's provided additional support for this hypothesis (**Cohen AS, et al., 1960 and Yardley JH. 1961**). But finally in 1990's specific segments of 16S r RNA multiplied by PCR amplification and this confirmation leads to developing bacterium tentative name *Tropheryma whipplei* (Etymology-Greek *trophe* means nourishment and *eryma* means obstacle or blockade due to malabsorption) (**Wilson KH.1991 and Relman DA. 1992**). Officially approved name

was given to this bacterium in 2001 as *Tropheryma whipplei* (La Scola B. 2001). In 2003, 1st synthesized-strain of Gram-positive *T. whipplei* named as Twist Marseille strain got established and its complete set of DNAs was partially sequenced with GenBank accession no. NC_004572 (Raoult D. 2003). The complete genome of *T. whipplei* reference strain TW08/27 was completely sequenced and has GC content of 46% and size of genome was determined to be 925,938 base pair (Bentley SD. 2003). Intracellular bacterium can be visualized by using periodic-acid-Schiff (PAS) staining with this method bacterium is visible within vacuoles and mostly preferred (La Scola B. 2001).

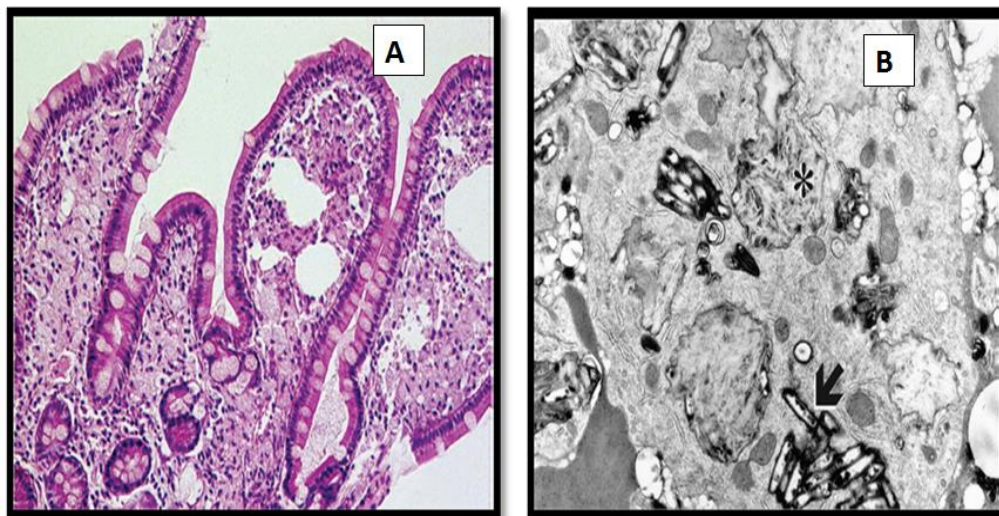


Figure1. Whipple's disease histological changes: **A.** Whipple's ailment: aggrandizement of the villi repleted by frothy macrophages or histiocyte, by Hematoxylin-eosin stain method. **B.** clusters of bacterium species (arrow) and distinctive lysosomes (asterisk) observed in Gastrointestinal tissue macrophage in Whipple's ailment (Bureš et al., 2013).

The vaccination concept was proposed by Edward Jenner in 1796 by constructing vaccine for curing smallpox and removing the infection by extracting some biological substances from cattle. He pioneered the term vaccine. Extensive and progressive use of vaccines globally for multiple infectious disease-causing microbes has been the turning point in medical world. Louis Pasteur proposed key vaccinology laws once it was discovered that

microorganisms were the principal pathogenic organisms for infectious illnesses. Salk and Sabin employed Pasteur's directives to develop a Polio vaccine that paved the path for the administration of dead and attenuated (pacified) live Polio virus, respectively. Measles and Rubella are two more contagious illnesses, which primarily affect youngsters. Vaccination strategies developed by Hilleman against measles, mumps and rubella by using attenuated viruses lead to the new avenue for development of vaccines against pathogens such as Diphtheria, Tetanus, and *S. pneumonia*. With reference to Hepatitis virus, the vaccines were designed by using inactivated viral antigenic peptides. Therefore, vaccine development with the use of directives proposed by Pasteur has been very effective approach in the pharmaceutical history regarding drug development.

Vaccine specialists and scientists have followed Pasteur's criteria for over a century, and this traditional technique has resulted in the development of several innovative vaccines. Their methodology has been efficient in many cases, but is always dilatory or detained and fails when the pathogens cannot be cultured by *in vitro* methods, or when many antigens are not consistent in peptide sequence. By the end of the twentieth century, several vaccines were developed using the traditional methods. Therefore, scientific as well as medical fraternity required new approaches to get success over remaining microbial disease-causing organisms. A scientific revolution approached after Craig Venter published genomic literatures for the first time for biological organisms. The availability of entire genome sequences, together with the advancement of computer server-based technologies that are likely to be functional and structural genomics, has produced a new environment and ease in the creation of harmful organism vaccines. This breakthrough technique allowed scientists to think beyond Pasteur's recommendations, allowing them to use databases and computer algorithms to rationally design vaccines using data from genome sequences, eliminating the need for wet lab procedures. This novel method of vaccine study and designing was termed as “reverse vaccinology”. It was entitled so, because the process of vaccine design and investigation initially requires *in silico* approaches like database analysis of expressed genomic sequences to determine new potential antigenic components instead of analyzing the microbe itself.

Reverse vaccinology defines the process of computational biology to design antigenic epitope usually starting from genomic database involvement. Rappuoli initially presented this method in the year 2000 that, illustrated a genome DNA sequence-based vaccine formulation methodology. Dr. Rappuoli was the first to successfully apply reverse vaccination against Serogroup B. *Meningococcus* (MenB), the organism that causes meningococcal meningitis across the world. As bacterial capsular polysaccharides were almost comparable to a human self-antigenic determinant, standard techniques were challenged in this circumstance, even though bacterial surface proteins were highly variable due to their sequential nature. Accessibility of the *Meningococcus* B strain complete DNA sequence provided a new hope for formulating a vaccine against it. The complete genomic DNA sequence of *Neisseria meningitidis* was used for bioinformatics analysis and computational approaches were successfully used to select 600 novel vaccine candidates that were expressed in to proteins in *Escherichia coli* later. From them, 350 were finally selected and purified, and used for antibody production in mice and testing of mice serum for *in vitro* anti-meningococcal activity (complement system based *in vitro* disintegration of the bacterium). Finally, 29 antigenic surface-exposed proteins of *Meningococci* were examined, and several of them induced antibacterial antibodies, despite the fact that many of these vaccine candidates were missed by all prior vaccination techniques. The antigens exhibiting the effective anti-bacterial activity were screened out and allowed to develop vaccine prototypes that were capable to generate active immunization in selected mice against most of the *Meningococcus* B strains. After development of this vaccine successfully reverse vaccination has been used in many new vaccine development programs. So, one can easily interpret that after successful accomplishment of human genome project in 2003 new computational analytical approaches were designed with multiple algorithms to identify genomic and proteomic parameters that can be used for better prediction of vaccine candidates and ease the cumbersome process of wet lab vaccine design by hit and trial approaches.

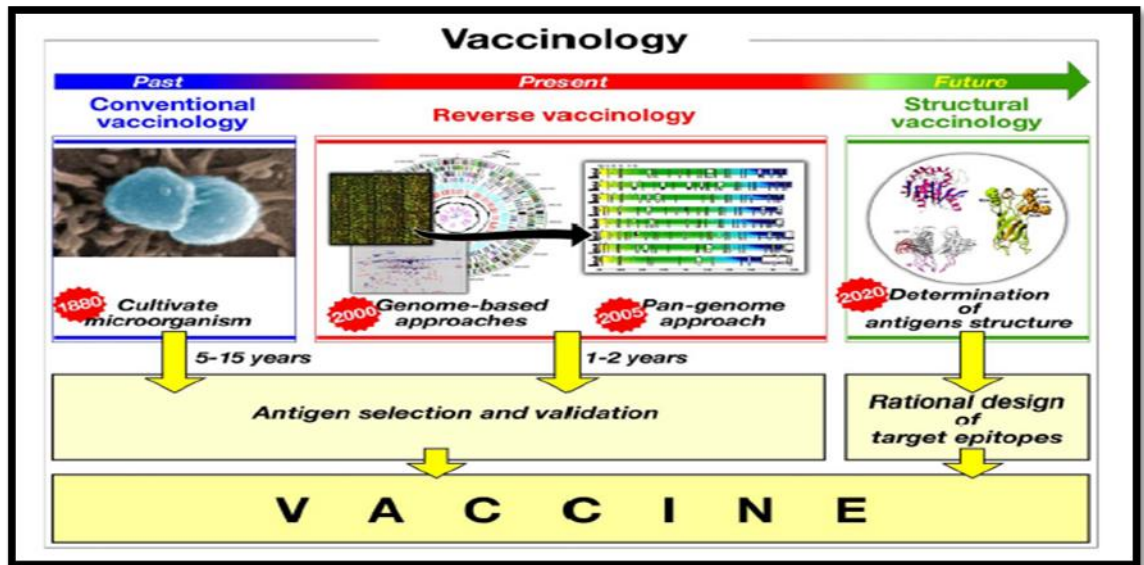


Figure 2. Vaccination advancement in modern world (Serruto & Rappuoli., 2006)

This opened up new directions in vaccine prediction and designing with economically efficient approaches, as all the tools are available online and genomic as well as proteomic data can be easily retrieved from NCBI (National Center for Biotechnology information) database of US NIH (United states National Institute of Health). In this study, we have moved from genome to vaccinome direction so that multiple epitopes can be designed for *T. whipplei* twist strain and its interaction with MHC class II alleles was identified and analyzed by using various immunological bioinformatics approaches to finally predict a useful vaccine against this bacterium. So far, we have deployed *in-silico* immunoinformatics approaches and predicted many immunogenic epitopes for diseases like HCMV, Dengue virus, SARS-CoV-2, *Candida aureus*, and *T. whipplei*.

After determining epitopes from bacterial proteome, they need to be synthesized. Primarily, solid-phase epitope synthesis and liquid phase epitope synthesis are used to synthesize peptides to design vaccine constructs. Rapid solid phase synthesis or a mixture of solid phase synthesis of different method exists to extract components like condensation reactions into a single peptide in solution is preferred for B and T-helper epitopes comprising fairly lengthy peptides (Moisa et al., 2012). Solid phase synthesis procedure involves covalent binding of peptides or epitopes to be synthesized on polymeric base, the good thing is that all activators, enzymes get washed by deploying

solvent wash (Lloyd-Williams et al., 2020). Amino acid derivatives protected by a fluorenyl-methoxy-carbonyl group containing 9 fluorenyl methoxycarbonyl groups have become the most widely used peptide synthesis method. Unlike the production of tertbutyloxycarbonyl-Boc-amino acid derivatives, the produced peptide does not need the application of a strong acid like HCl or HF to cleave it off the polymeric basal-support. After synthesis of peptides adjuvants like Freund's adjuvant (containing lipopolysaccharides with lanolin and oil), QS21 saponin fraction adjuvant, poly-lysine and poly arginine is joined with synthesized epitopes (Moisa et al., 2012). Also, currently scientist prefer adjuvants and synthesize them as per pattern recognizing receptors like TLR binding Pam2CYS in case of Hepatitis C infection in many recent studies. Liposomes and virosomes containing epitope linked to adjuvants are deployed as mobile agents to transfer such immunogenic epitope directly to antigen presenting cells (Moisa et al., 2012). These vaccines need to be inserted in animal models that have xenotransplantation of human tissues. In case the selected organism is already immune to disease under consideration, then we can go with human cell lines study (Moisa et al., 2012). The last evaluation step is to check cytokines concentration along with antibodies concentration in serum levels of model organism, which directly correlates with immunogenic activity against considered pathogen after injecting vaccine. In this study we analyzed microarray transcriptomic data related to model organism affected by strains of *Tropheryma whipplei*, for which genomic and proteomic data was retrieved from NCBI. Also, Codon usage and amino acid usage was analyzed to determine target antigens. The core objective of the study was to employ reverse vaccinology approach for the identification and prediction of *Tropheryma whipplei* candidate antigens. This was useful in predicting epitope-based vaccine prediction to control Whipple's disease. Also, we conducted designing of potent and effective drug against this pathogen.

CHAPTER 2

REVIEW & LITERATURE

George Hoyt Whipple in 1907 explained the Whipple's disease, as a multisystemic chronic infectious disease. He identified silver stained rod-shaped bacterium in vacuoles associated with macrophages of patients and he initially did not think of them as the cause for the disease. Rather he thought that intestinal lipodystrophy (Whipple's disease) was caused due to some novel disturbances in fat metabolic schemes. *Tropheryma whipplei* is oxygen-dependent, rod-shaped, gram-positive, non-acidic, periodic acid Schiff-positive bacterium, found both intracellularly and extracellularly, and grows slowly in acidic vacuoles of cells. This bacterium mostly infects the intestine's lamina propria and the vacuoles of foamy macrophages. *In situ* hybridization testing revealed that the bacterium is primarily present around the villi tips of the intestinal walls in patients' deep mucosa and lamina propria. The bacterium then enters mucosal macrophages, but they are unable to kill it because it decreases CD11b expression in the macrophages (CD11b on macrophages actually mediates the intracellular degradation of bacteria), which leads to inappropriate antigen presentation by the macrophages and dendritic cells.

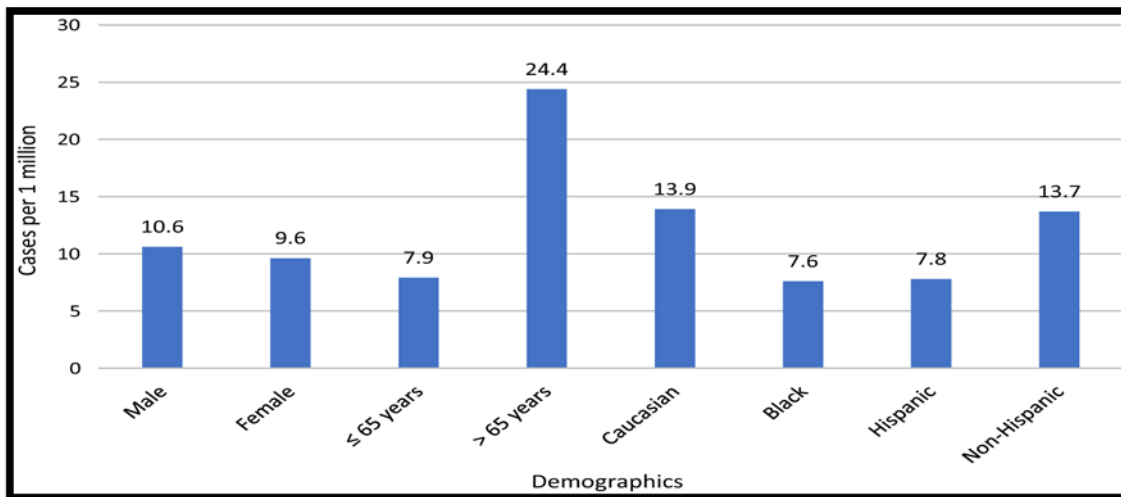


Figure 3. Spread of Whipple's' disease in USA in different races (Elchert JA. 2019) (Values are in percentage).

This results in an increase of IL-10, TGF- α , and CCL-18 expression and a reduction in IFN- γ , which causes phagosome maturation to be disrupted and thioredoxin production to be reduced, leaving them unable to kill bacteria and present antigens (Moos V. 2010). Along with these modulations, the bacterium also interferes with the development of CD4+ cells into Th2 cells, and as a result of the bacterium's modifications in the immune system, disease elimination becomes insufficient, and the bacterium takes the advantage of this situation to multiply in macrophages (Moos V. 2006).

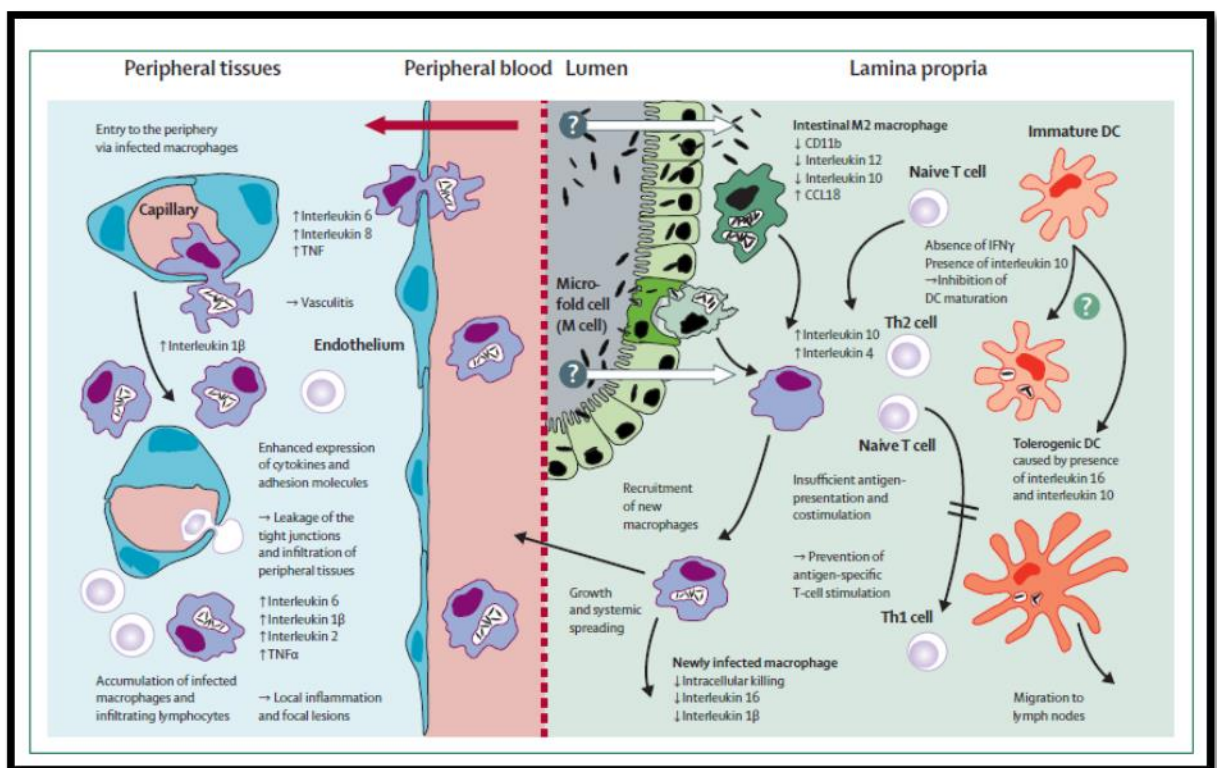


Figure 4. Whipple's disease aetiology from a clinical standpoint: The presence of interleukin 10 and interleukin 16, as well as the absence of interferon- γ and interleukin 12, may cause improper maturation of professional antigen-presenting cells, resulting in insufficient antigen presentation and inhibiting the stimulation of antigen-specific T-helper type 1 cells, allowing *T whippelii* to grow and spread systemically. Even in immunologically protected organs like joints or the brain, local generation of inflammatory cytokines by macrophages and endothelial cells in the periphery may cause

lymphocyte infiltration through a leaky endothelial barrier, followed by localized inflammation=chemokine (C-C motif) ligand. DC=dendritic cell. TNF α =tumor necrosis factor α . **(Schneider T.2008).**

Human monocytes can kill *T. whipplei*, but they can multiply in monocyte-derived macrophages by initiating a unique activation programme. Thioredoxin is suggested to be bacterial death through its upregulation in infected monocytes, because adding thioredoxin to infected macrophages reduced bacterial proliferation. *T. whipplei* can multiply in monocytes because of IL-16, which was upregulated in macrophages. This boosted bacterial replication in macrophages. Antibodies against IL-16 also stopped *T. whipplei* reproduction in Macrophages. The expression of thioredoxin was inhibited by IL-16, whereas the expression of IL-16 and proapoptotic genes was increased. *T. whipplei* replication was greater in Whipple's Disease patients than in healthy people, and it was linked to high levels of circulating IL-16 **(Desnues B. 2005).**

Whipple's disease patient of Indian origin also shows the typical and unique oculomastigatory myorhythmia along with neurological systemic disorder **(Chandra SR. 2018)**. Patients with Whipple's disease are mostly observed in Caucasian male of nearly 50-year age who have initial arthralgia, chronic diarrhea or suffering from weight loss **(Lagier JC. 2014)**. Gastrointestinal tract (duodenum, jejunum, and ileum) is primarily affected in patients with classic Whipple's disease. Gastric organ, hepatic organ and oesophagus can also be affected in this disease. Major symptoms like diarrhea, abdominal pain, steatorrhea, hepatosplenomegaly, anorexia, cachexia, hematochezia and malabsorption are common. Bone and joint manifestation include arthralgia, arthritis and spondylodiscitis along with polyarthritis in Whipple's disease patients **(Puechal X.2001)**. *Tropheryma whipplei* causes neurological problems (choreiform movements, myoclonus, occulomasticatory myorhythmia), hypersomnia, cognitive impairment, cerebral ataxia on the basis of location of lesions formation **(Compain C. 2013)**. *Tropheryma whipplei* also causes- cardiac symptoms like heart failure, destruction of heart valves, acute ischemic stroke **(Fournier PE. 2010)**. While the major disease in *T. whipplei* infected patients is

gastroenteritis causing liquid diarrhea and colicky abdominal pain (Maizel H. 1970). The fact that Whipple's illness is more common in persons who have the HLA-B27 antigen suggests that there may be a genetic predisposition in those who have the disease, leading in an aberrant host response to a microbe that is common in humans.

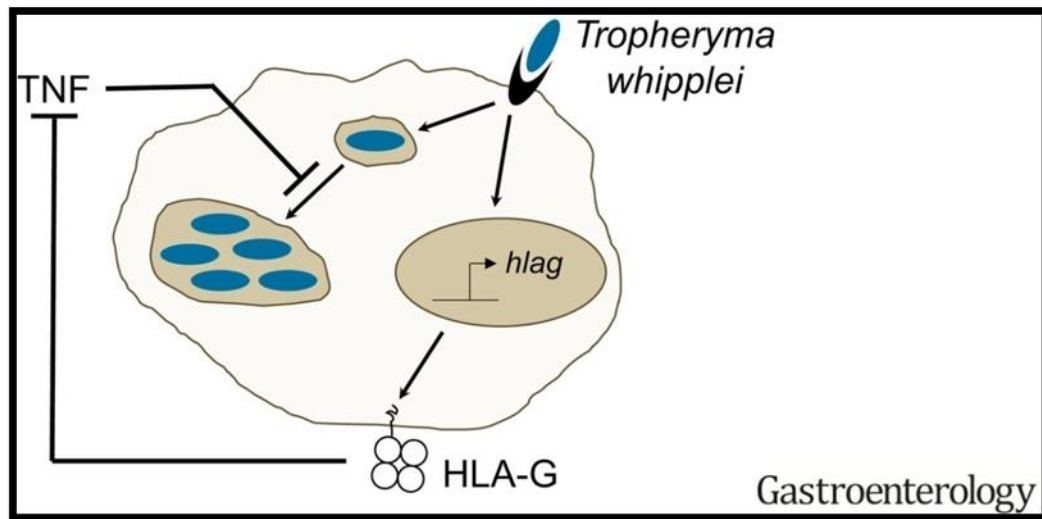


Figure 5. *T. whipplei* increases HLA-G expression and in turn reduces TNF that assist in promoting its own replication in monocytes (Azzouz EB.2018).

Reverse vaccinology is a suitable strategy as well as a unique scientific method that uses genetic data in conjunction with the usage of a computer to assemble antibodies without cultivating bacteria species (Kampalliwar .2013; Tang et al. 2012). It allows the choice in hands of human interface for selecting antigens from pathogenic set of DNA and most antigenic areas could be used to synthesize potential immunization to initiate defensive responses against such pathogenic species (Kazi A. 2018). Epitopes based antibodies selection and production is explicitly less time consuming, economical and considered safest approach in vaccine designing.

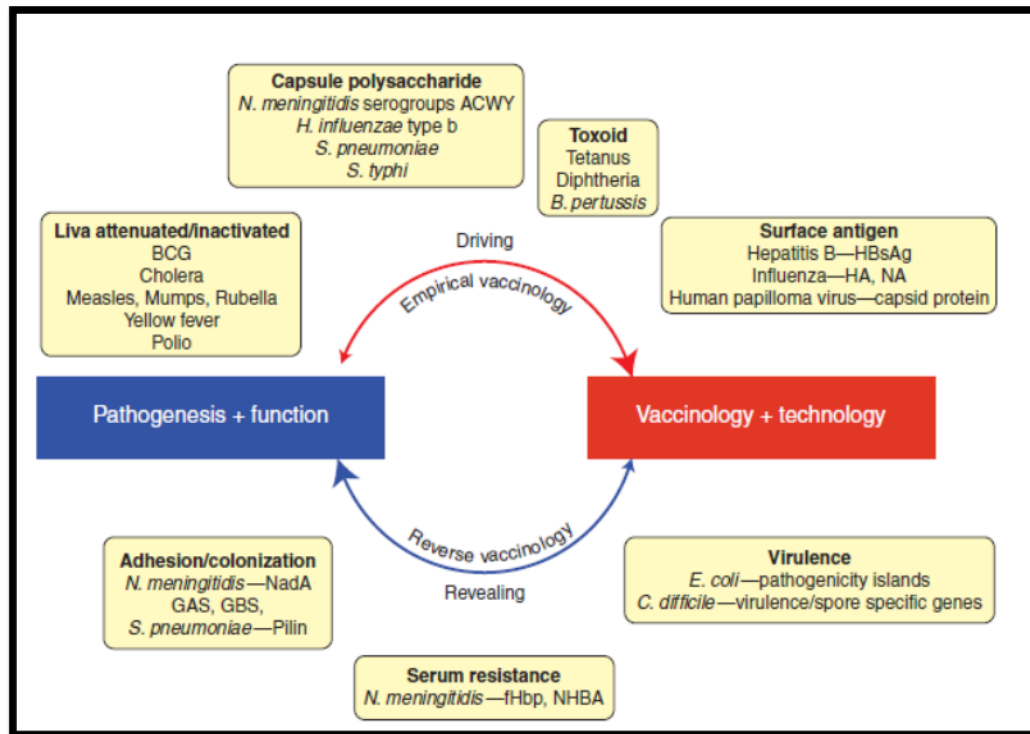


Figure 6. Microbial pathogenesis and vaccine development interaction (Del Tordello E. 2017.)

Improvements in Reverse Vaccination technology

A. Pan Genomic Reverse Vaccination technology

Pan Genomic approach implies to the whole genomic sequences shared by every primarily sequenced microbial strains and also the different allocated genome presents only in few subsets of main domain of microbial species. It is applicable to achieve the genomic size and its diversity that one can access (Tettelin H. 2009). The genomes of multiple isolates of the same species are analyzed using computational approach in Pan genomic reverse vaccinology. The pan genomic reverse vaccination technology was applied for the first time in *Streptococcus agalactiae* (Lefébure T, et al.2007). This approach has also been applied in present study of *Whipplei* vaccine development.

B. Comparative Reverse Vaccination technology

This technology deals with comparison of pathogenic as well as non-pathogenic microbial strains on the basis of genetic sequences. It is also associated with the analysis of proteins from various biological entities having different structures.

C. T cell epitope study

T cell epitopes are amino acid fragments of antigens that have capacity to interact with MHC (Major Histocompatibility Complex) molecules and induce immunogenicity. One can simply say these epitopes have capacity to stimulate CD4 or CD8 T-cells that in turn initiate immune response against such antigenic parts which can be helpful in vaccine prediction (**Ahmed RK. 2009**).

D. B cell epitope study

B cell epitope study focuses on identification that is important for structure functions prediction in practical analysis. In linear B-cell epitopes prediction amino acid propensity scales along with hydrophilicity calculations (**Hopp TP. 1983**) are commonly used as hydrophilic peptide portions are present on surface and capable to have antigenic properties.

E. Clinical aspects and Wet-lab strategies for Epitope based vaccine

The stimulation of T-helper subtypes and the production of type-specific cytokines are some of the most essential aspects of immune function. The Th1 and Th2 helper cells can be activated by antigens bound on MHC-II (**Figure 7**). Th2 cells primarily stimulate production of antibodies in response to an external bacterium, whereas Th1 cells stimulate cell - mediated immunity in reaction to intracellular infections (viruses, cancer). Th1 and Th2 immunity, on the other hand, are not strictly comparable in terms of cell-mediated and humoral immunity (**Skwarczynski et al., 2016**). The Th1 system, for example, may trigger low levels of antibody-based reactions. Th1 cytokines are linked to pro-inflammatory actions, whereas Th2 cytokines are linked to anti-inflammatory responses. Immuno-pathophysiological consequences such as cellular injury from acute inflammation or powerful allergy reactions might result from unbalanced Th1/Th2

sensitivities. As a result, in vaccine construct planning, a well-balanced Th1 and Th2 activation must be considered.

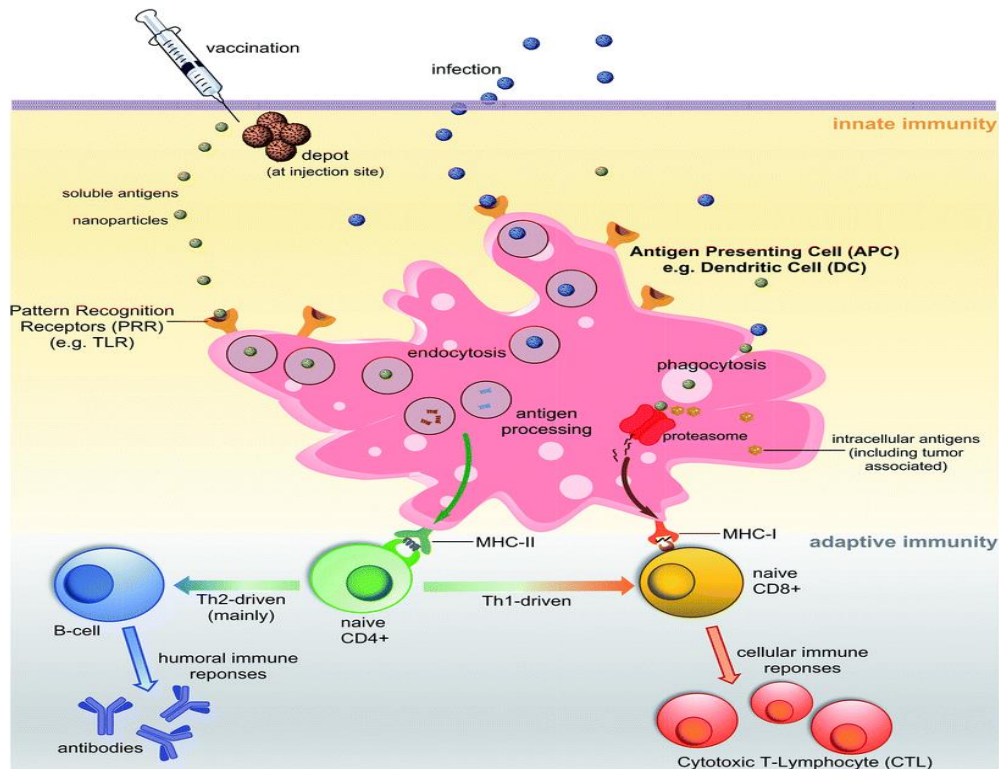


Figure 7. Schematic process of vaccine immunization cellular-physiology (Özcan et al., 2019)

In many studies epitope-based vaccines were primarily tested on cell lines, and viability was checked by MTT assay. The cytokine assays were conducted by ELISA (Pouriayevali et al., 2011). In case of Alzheimer’s disease, the use of epitope or peptide-based vaccines was found to be effective (Evans et al., 2014). In one of the studies, it was observed chimeric epitope-based vaccine construct was used to treat three breast cancer mice model organisms. This indicates immense scope of immunoinformatics in predicting epitopes that can be used against pathogenic infections and cancer related problems (Mahdevi et al., 2014).

Recent studies on Epitope based vaccine designing and Reverse vaccinology

So far, many studies on epitope-based vaccine designing have been conducted on variety of pathogens. The study by **Joshi et al., 2020**, and **Akhtar et al., 2021** designed epitope-based vaccine candidates against SARS-CoV2, **Krishnan et al 2021** also designed epitope-based vaccine against dengue. The epitopes prediction in these studies followed similar approach that we have discussed here in our study against *Tropheryma whipplei*. In a recent study on codon usage patterns our group also found amino acids such as valine, aspartate, leucine, and phenylalanine had a high codon use frequency and detected epitopes KPSYLSALSAHLNDK and FKSFNYNVAIGVRQP, which were screened from the proteins excinuclease ABC subunit UvrC and 3-oxoacyl-ACP reductase FabG, respectively (**Joshi et al., 2022**). In two more studies, the immunogenic epitopes were predicted against (Human Cytomegalovirus) HCMV virus and *Candida aurius*. Then, the predicted epitopes were joined with adjuvants and PADRE sequence to design vaccine candidates which were further subjected to molecular docking, molecular dynamics simulation and in silico cloning (**Akhtar et al., 2021**).

Also, on few other studies on Canine circovirus (**Jain et al., 2021**) and Nervous necrosis virus that affect dogs and fishes respectively the immunogenic epitopes were predicted by similar approach that we have followed in current research. This assisted us in drawing conclusion that “to predict vaccine any model organism can be selected if its pathogen genomic or proteomic information is available in primary databases like NCBI, EMBL, and DDBJ”. In one of the recent studies on *Mycobacterium tuberculosis*, the predicted epitope-based vaccine was docked with TLR3, for better internalization. Such studies show the importance of bioinformatics in the field of immunogenic vaccine construction (**Bibi et al., 2021**). **Ashfaq et al., 2021** designed epitope-based vaccine construct against MERS-COV by deploying the viral proteomic analysis by reverse vaccinology methods. **Kumar et al., 2020** also designed epitope-based vaccine construct against bacterium *Campylobacterium jejuni* by proteomic analysis.

In one of the recent research projects Chikungunya viruses (**Sharma et al., 2022**) epitope vaccine development was conducted by reverse vaccinology approach that followed similar patterns of our current study. This suggests that various recent studies are following the same approach with different target pathogens that would lead to develop immunity in host organisms. Furthermore, the pandemic caused by SARS-CoV2 has increased the demand for robust and quick approaches of vaccine design. To full-fill such demands immunoinformatics is the only approach that can assist in rapid vaccine construction and provide better hope of treatments for patients.

CHAPTER 3

OBJECTIVES & SCOPE OF STUDY

Currently hydroxychloroquine (600mg/day) and doxycycline (200mg/day) combinedly used for treatment of Whipple's disease for 12 to 18 months, but life time follow up is required (**Lagier JC. 2014**), so it is time consuming treatment process and only few handful trials have been conducted (**Feurle GE. 2013**). Due to bacterium evolution and horizontal gene transfer that promotes gain of new antibiotic resistance genetic sequences, it become very crucial to implement new strategies to fight infections (**Mondal SI. 2015**). Based on the clinical interest, there are few research utilizing *T. Whipple* proteome analysis, and most of the approaches do not include switch or reverse vaccinology or molecular docking for vaccine or medication prediction, which will open up new opportunities in comparative proteomic investigations.

Whipple disease is associated with neurological disorders (**Gerard A. 2002**). Whipple's disease is also associated with arthritis (**Pu  chal X. 2001**). A case of dementia has been observed in Indian individual due to Whipple's disease (**Chandra SR. 2018**). Case of Whipple's disease has been observed in Japanese population (**Chandra SR. 2018**). Whipple disease is responsible for cerebellar ataxia (**Matthews BR. 2005**). Granulomatous colitis in dogs is also associated with *Tropheryma whipplei* (**Craven M.2011**)

Proteomic analysis of different proteins of bacterium *Tropheryma whipplei* leads to select antigenic, non-allergenic epitopes to elicit immune response in human population. Computer based drug designing after screening various pharmacoactive molecules, and metabolic enzymes of bacterium. These epitopes and drug molecules binding to various receptors and enzymes were validated by deploying molecular dynamic simulation studies. In this study microarray transcriptomic data analysis related to model organism affected by strains of *Tropheryma whipplei* was conducted along with Codon usage and amino acid usage was analyzed to determine target antigens. The core objective of the study was to use reverse vaccinology approach for the identification and prediction of *Tropheryma whipplei* antigens that are non-allergenic and used to predict epitope-based vaccine prediction to control Whipple's disease. Here we also used CADD approach to predict potent drug against this pathogen.

In present study a novel epitope vaccine designing was achieved by using reverse vaccination approach, this is going to help the medical world in many aspects. It will provide better cure for Whipple's disease and specifically target *T. whipplei* antigenic non-allergic epitopes. This investigation will provide economically (time as well as money) efficient strategy development for elimination of *T. whipplei* from human by using reverse vaccination. This study opens more dimensions in the research domain for understanding least studied *T. whipplei* structure loop holes by molecular modeling of selected antigenic epitopes and their interaction with MHC class II alleles.

CHAPTER 4

MATERIAL & METHODS

METHODOLOGY

4.1 Genomic trends analysis in *Tropheryma whipplei* and Phylogenetic analysis

All sequences of genomic and proteomic interest related to *Tropheryma whipplei* were retrieved from NCBI GenBank. For phylogenetic establishment between considered strains BLAST-n tool was deployed. To reduce sampling related errors, annotated coding sequences up to 100 codons was excluded out. Duplicate sequences, transposable elements with internal stop codons, no translational fragments were not considered during the analysis to reduce the error. Genes of template and non-template strands were retrieved, Oriloc was employed for this investigation based on oriC in *T. whipplei* chosen strains. (Frank CA.2000). RSCU (Relative_synonymous codon usage) and RAAU (Relativeaminoacidusage) among *T. whipplei* strains was analyzed by software CODONW 1.4.2, which assisted in applying COA (corresponding analyses).

GC content analysis for each codon positions was calculated for all selected coding sequences in both strains of *T. whipplei* by deploying GC content calculator (<https://jamiemcgowan.ie/bioinf/gc.html>). Other important parameters such as CAI (Codon adaptive index) (Sharp PM.1987), RSCU, RAAU, GRAVY Score (average hydrophobicity), total number of repeatable features of each and every selected codon and also average complexity quotient was analyzed for translated proteins to set up a systematic *in-silico* assay to determine factors influencing amino acid usage. MEGA program was deployed to study pairwise synonymous divergences (dS) as well as non-synonymous divergence divergences were calculated (dN) (Nei.1986). BLAST-n used to determine comparative phylogenetic relationship between various strains of *Tropheryma whipplei*.

4.2 Transcriptomic data analysis of model organism affected by *Tropheryma whipplei*

4.2.1. Retrieval of *T. whipplei* microarray gene expression profile datasets:

T. whipplei microarray gene expression profile GEO dataset was retrieved from GEO database (Barrett et al., 2012).

Differential Gene Expression Analysis (Microarray GEO Datasets)

The Gene Expression profiling analysis was performed by using GEO2R which compares two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Force Normalization were applied for the array normalization which uses quantile normalization to the expression data making all selected samples have identical value distribution. Limma package were used for the pvalue calculation between different groups. The Benjamini & Hochberg false discovery rate method were selected by default which is the most commonly used adjustment for microarray data and provides a good balance between discovery of statistically significant genes and limitation of false positives.

The differential expression was calculated based on biological analysis plan. The significant differentially expressed genes were filtered with less than or equal to pvalue ≤ 0.05 . Upregulated genes ≥ 1 -fold change and downregulated genes ≤ -1 -fold change were applied for significant two-fold difference in treated vs control.

Plots Interpretation

1. Volcano plot

A volcano plot displays statistical significance ($-\log_{10}$ P value) versus magnitude of change (\log_2 fold change) and is useful for visualizing differentially expressed genes. Highlighted genes are significantly differentially expressed at a default adjusted p-value cutoff of 0.05 (red = upregulated, blue = downregulated). A volcano plot displays the test results for a single contrast (a contrast is one Sample group compared to another Sample group). Volcano plot were generated using Limma (volcano plot).

2. Mean difference (MD) plot

A mean difference (MD) plot displays \log_2 fold change versus average \log_2 expression values and is useful for visualizing differentially expressed genes. There, similar to

volcano plot. Highlighted genes are significantly differentially expressed at a default adjusted p-value cutoff of 0.05 (red = upregulated, blue = downregulated). A mean difference plot displays the test results for a single contrast (a contrast is one Sample group compared to another Sample group). Mean difference (MD) plot was generated using Limma (plotMD).

3. UMAP

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique useful for visualizing how Samples are related to each other. The number of nearest neighbors used in the calculation is indicated in the plot. UMAP were generated using umap.

4. Boxplot

Boxplot is used to view the distribution of the values of the selected Samples. The Samples are colored according to groups. Viewing the distribution can be useful for determining if your selected Samples are suitable for differential expression analysis. Median-centered values are indicative that the data are normalized and cross-comparable. The plot shows data after log transform and normalization. It is generated using R boxplot.

5. Expression density

Expression density is used to view the distribution of the values of the selected Samples. The Samples are colored according to groups. This plot complements boxplot in checking for data normalization before differential expression analysis. The plot shows data after log transform and normalization. It is generated using R plot Densities.

6. Adjusted P-value histogram

Adjusted P-value histogram is used to view the distribution of the P-values in the analysis results. The P-value here is the same as in the Top differentially expressed genes table and computed using all selected contrasts. It is generated using R hist.

7. Moderated t-statistic quantile-quantile (q-q) plot

Moderated t-statistic quantile-quantile plot is the quantiles of a data sample against the theoretical quantiles of a student's t distribution. This plot helps to assess the quality of the limma test results. Ideally the points should lie along a straight line, meaning that the values for moderated t-statistic computed during the test follow their theoretically predicted distribution. It is generated using R limma (qqt).

8. Mean Variance trend

Mean Variance trend plot is used to check the mean-variance relationship of the expression data, after fitting a linear model. It can help show if there is a lot of variation in the data. This plot can help assess whether applying the precision weights option to take mean-variance trend into account is recommended. Precision weights improve accuracy of test results when a strong mean-variance trend is present. The plot does not require group selection. Each point represents a gene. The red line is mean-variance trend approximation that can be taken into account during differential gene expression analysis. The blue line is constant variance approximation. It is generated using R limma (plot SA, vooma).

9. Venn Diagram

Mathematical set-based representation plot, it represents no. of genes group wise. Microarray analysis was conducted by using GEO-2R tool of NCBI, latest microarray analysis schemes that were constructed to study various 8 Geo series data that was retrieved from Geo-Datasets were represented in **Table1**.

10. Differentially expressed genes comparison analysis

GEO2R web-tool permits to compare GEO series sample groups and identify differentially expressed genes (DEGs) (**Barrett et al., 2012**). From the GEO series samples were compared and significant DEGs across experimental conditions were identified using Bioconductor supported limma R package. The Benjamini & Hochberg (BH) standard operating procedure was used to reduce false *P* value discovery rate (**Tolfvenstam et al., 2011**). The median-centered distribution values of samples selected

for optimum cross comparability and DEG identification. In DEG analysis, LogFC (fold-change) value greater than 1 indicate up-regulation of genes while logFC value less than -1 was considered down-regulated genes (Shi et al., 2021).

11. Gene Functional Classification and Identification of functional related gene

The differentially expressed genes of *T. whipplei* were classified into functional related gene groups using gene_functional_classification tool of Database_for_Annotation, Visualization and Integrated Discovery (DAVID) from <https://david.ncifcrf.gov/tools.jsp>. Microarray analysis was conducted by using GEO-2R tool of NCBI, latest microarray analysis schemes that were constructed to study various 8 Geoseries data that was retrieved from Geo-Datasets were represented in **table** below.

Table1: List of microarray experiments accessed from NCBI Geo-Datasets for microarray analysis

Geoseries Accession	No.of sample	Organism name	Detailed Interpretation of Experiment	Groups Designed for Biological analysis	Source
GSE10286 2	36	Human	PBMC(Peripheral blood mononuclear cells like T Cells, B Cells, NK cells, Monocytes) were obtained from infected and carrier for IRF-4 (Interferon regulatory factor 4)mutation, R98W from french population and microarray experiments was conducted in <i>Tropheryma whipplei</i> infected(multiplicity of infection:1) and non-infected individuals.	6	Guérin et al., 2018
GSE20210	18	Mouse	IL-16(Interleukin 16) promotes <i>Tropheryma whipplei</i> replication by inhibiting phagosome and macrophage modulation	4	Ghigo et al., 2010
GSE20209	6	Mouse	Whole genome microarray analysis was conducted in mouse bone marrow derived macrophages(BMDM) by knocking out IL-16 gene and infected with	1	Ghigo et al., 2010

			Tropheryma whipplei and control samples.		
GSE16180	6	Mouse	BMDM infected with <i>Tropheryma whipplei</i> (MOI: 50:1) and control samples	1	Ghigo et al., 2010
GSE7453	8	<i>Tropheryma whipplei</i>	In this work genomic diversity of 15 <i>Tropheryma whipplei</i> strains were analyzed by comparative genomic hybridization, here all 14 strains were compared with Twist strain to indicate evolutionary diversity	15	La et al., 2007
GSE3693	6	<i>Tropheryma whipplei</i>	Transcriptomic analysis of <i>Tropheryma whipplei</i> in response to temperature stress	3	Crapoulet et al., 2006
GSE5717	9	<i>Tropheryma whipplei</i>	Susceptibility of drug doxycycline was checked on <i>Tropheryma whipplei</i> at 0.5 mg and 5mg concentration by microarray experiments	2	Van La et al., 2007
GSE49016	24	Human	Different bacterium including <i>Tropheryma whipplei</i> effect on human dendritic cells maturation was observed by microarray experiment	1	Gorvel et al., 2014

4.3. T-cell Epitope prediction for peptide vaccine crafting

4.3.1. Retrieval of Proteins for *T. whipplei*

The NCBI-GenBank and UniProtKB databases were used to get proteomes in FASTA format. The following accession numbers were chosen to represent five proteins with varied functions: WP 042507409. WP 033800049, DNA-directed RNA polymerase subunit beta (RPO-B). GroES, WP 038104819, is a co-chaperone. WP 042505650 is a metal homeostasis membrane protein from the TerC/Alx family. WP 042505746 is a membrane protein insertase YidC-integral membrane protein involved in murein production. This selection demonstrates the variety in pathogenic domain proteins that must be included.

4.3.2. Allergenicity Prediction for Proteins

The protein sequences were then deployed for further analysis based on Allergen FP V 1.0 for predicting allergenicity. Allergenicity was calculated on the basis of TSI (Tanimoto similarity index) > 0.81 by deploying AllergenFP tool. This score was based on molecular fingerprint analysis or similarity indexing (**Bajusz et al., 2015, Dimitrov et al., 2014**).

4.3.3. T-Cell Epitope Prediction

Net MHCII PAN 3.2 server was used to find and screen out HLA alleles which have good interaction with selected non-allergens of pathogenic origin (**Jensen et al. 2018**). To bring higher confidence in selecting epitope, VaxiJen server was deployed to determine antigenicity with threshold ≥ 0.7 for selected rare bacterium (**Doytchinova & Flower., 2007**). By subjecting proteomic sequences to Net MHCII PAN 3.2 server we obtained 1147 epitopes for WP_042507409.1, 90 epitopes for WP_033800049.1, 309 epitopes for WP_038104819.1, 302 epitopes for WP_042505650.1, and 510 epitopes for WP_011096746.1. This server was used because of its neural networking algorithm-based approaches for fine predictions. $1 - \log_{50k}(\text{affinity score}) \leq 0.6$ is used to screen out possible epitopes.

4.3.4. Epitope's Molecular Docking with HLA Alleles

The docking experiment was conducted using Patch-Dock tool (**Schneidman-Duhovny et al. 2005**), The predicted docked models of putative epitope and HLA alleles was selected on the basis of score, which relies on highest geometric shape complementarities and atomic contact energy (**Zhang et al. 1997**). This allows the best selection of epitope and HLA allele interaction. This tool is easy to deploy for all life science domains.

4.3.5. Analysis of Population Coverage

Immune Epitope Database (IEDB) analysis Resource tool of population coverage was used to predict population coverage of the putative epitopes that are exhibiting interaction to HLA alleles and based on MHC-II restriction data (**Bui et al. 2007**). MHCpred tool

was deployed for quantitative prediction of selected epitopes interacting to major histocompatibility complexes (Guan et al. 2003).

4.3.6. Molecular Dynamics and Simulations

Epitope-HLA allele docked sets were then used for simulation and dynamics analysis by deploying NAMD (Phillips et al. 2005) associated with VMD (Visual Molecular Dynamics) tool (Humphrey et al. 1996). In this study OPLS-AA force field was used, for 10ps. Using the psfgen package of VMD, a protein structure file was created using the topology files and preliminary PDB files of the HLA II allele- epitope docking complex. The trajectory DCD file was generated by NAMD. The root mean square deviation (RMSD) of the docked complex was calculated to examine the simulation outcome. The value of RMSD obtained from rmsd.dat file, which was then examined using Excel sheet for graphical representation (Joshi et al., 2021). Best RMSD values lies under 10 angstrom for docked complexes (Akhtar et al., 2021).

4.4. In-silico drug discovery against *Tropheryma whipplei*

4.4.1. Enzyme Selection Bias

Proteomic sequences of *Tropheryma whipplei* were retrieved from NCBI-GenBank for two significant compounds to be specific, DNA Ligase (AAO44511) and Chorismate synthase (WP_011096348), and these enzymatic edifices are engaged with DNA replication, biosynthesis of amino acids individually. The choice of these fundamental proteins depends on DEG (database of basic qualities) server investigation.

4.4.2 Enzyme Annotation studies

After this CD-HIT server was utilized to distinguish paralogs. It depends on fast heuristic examination approach, and accommodating in deciding likeness investigation between peptide stretches. Basic local search alignment was utilized for deciding homology between considered proteins of *Tropheryma whipplei* and proteomic spaces of *Homo sapiens*. This gives more approval to the determination from escalated proteomic sets of bacteria. To examine pharmacogenecity or medication capacity of considered proteins

drug bank web-server (<http://www.drugbank.ca/>) was applied. Two online services, conserved domain architecture retrieval tool (CDART) and Pfam, were used to detect space homology in enigmatic protein groupings. KEGG automatic annotation server (KAAS) assisted in distinguish metabolic pathway for selected biocatalysts and here *T. whipplei* Twist and *T. whipplei* TW08/27 strains was browsed from the organism list at the time of investigation.

4.4.3 Structural analysis: Docking & Simulation

The selected proteins after intensive investigation and KEGG annotation was exposed to homology displaying by means of Phyre2 (**Kelly., 2015**), which is a Hidden Markov model-based server for structural predictions of catalytic enzymes. Quick overview of medications acting on chosen proteins and their 3D structure was acquired by utilizing PubChem web server and RCSB-PDB databank. Sub-atomic docking was led by means of AutoDock-vina (**Trott., 2010**) assembly to examine interaction energies of ligand-protein docked structures. SwissADME tool (**Daina., 2017**) was deployed to examine biochemical properties like pharmo-kinetics, drug-likeness, and inhibitory action on cytochrome P450 isoforms, structural properties, bioavailability and synthetic accessibility. SwissADME tool assisted in observing Lipinski violations for the current research (**Lipinski., 2004**). Molecular simulation studies were performed for 40ns by using Gromacs suite (**Van Der Spoel., 2005**).

CHAPTER 5

RESULTS & DISCUSSIONS

RESULTS AND DISCUSSION

5.1 Transcriptomic data analysis for model organisms to reveal host-pathogen interaction

5.1.1 *T. whipplei* microarray profile data

The NCBI-GEO database search was deployed for expression profiling related to model organisms. Microarray analysis was conducted by using GEO-2R tool of NCBI, latest microarray analysis schemes that were constructed to study various 8 Geoseries data that was retrieved from Geo-Datasets were represented in (Table1).

5.1.2 Identification of differentially expressed genes

Differential expression analyses for various genes of different groups that were considered from Geoseries were reported and are as follows:

a. GSE3693:

For this series 3 groups were identified for bacterium on the basis of temperature treatment. *Tropheryma whipplei* genomic patterns in exposure to heat stressors revealed distinct transcription patterns, as per microarray analysis. The dnaK regulon, which includes grpE, hspR, dnaK, clpB, and cbpA, as well as the TWT745 ORF, that also translate for heat-shock polypeptide, were all up-regulated after 15 minutes at 43°C. RibC and IspDF proteins, which were thought to be virulent, were similarly up-regulated in Heat shock. *T. whipplei* transcript was extensively changed after cold shock for 4°C, in comparison to the heat-shock activity. Nine regulons were discovered among the 149 genes differentially expressed, one of which was made up of five genes having ABC transporter commonalities that suggests enhanced uptake of nutrients. The overexpression of heat-shock molecules such as GroEL2 and ClpP1, and also numerous genes associated with metabolism, was seen after *T. whipplei* in exposure to cold. These findings reveal that *T. whipplei* has a distinct adaptive control to thermal stressors, which is consistent with its presumed environment origins.

1. Cold shock temperature 4°C Vs 37°C

Here 57 genes were found to be down regulated and 75 genes were found to be up regulated. The down regulated genes include peptidase, ribonuclease, deoxy ribonuclease and 3-dehydroquininate synthase; while up regulated genes include cell division protein fts-W, 4-aminobutyrate amino transferase gab-T, Chaperonin groEL and Glucokinase glkA. **Figure 8** shows microarray analysis plots for this group.

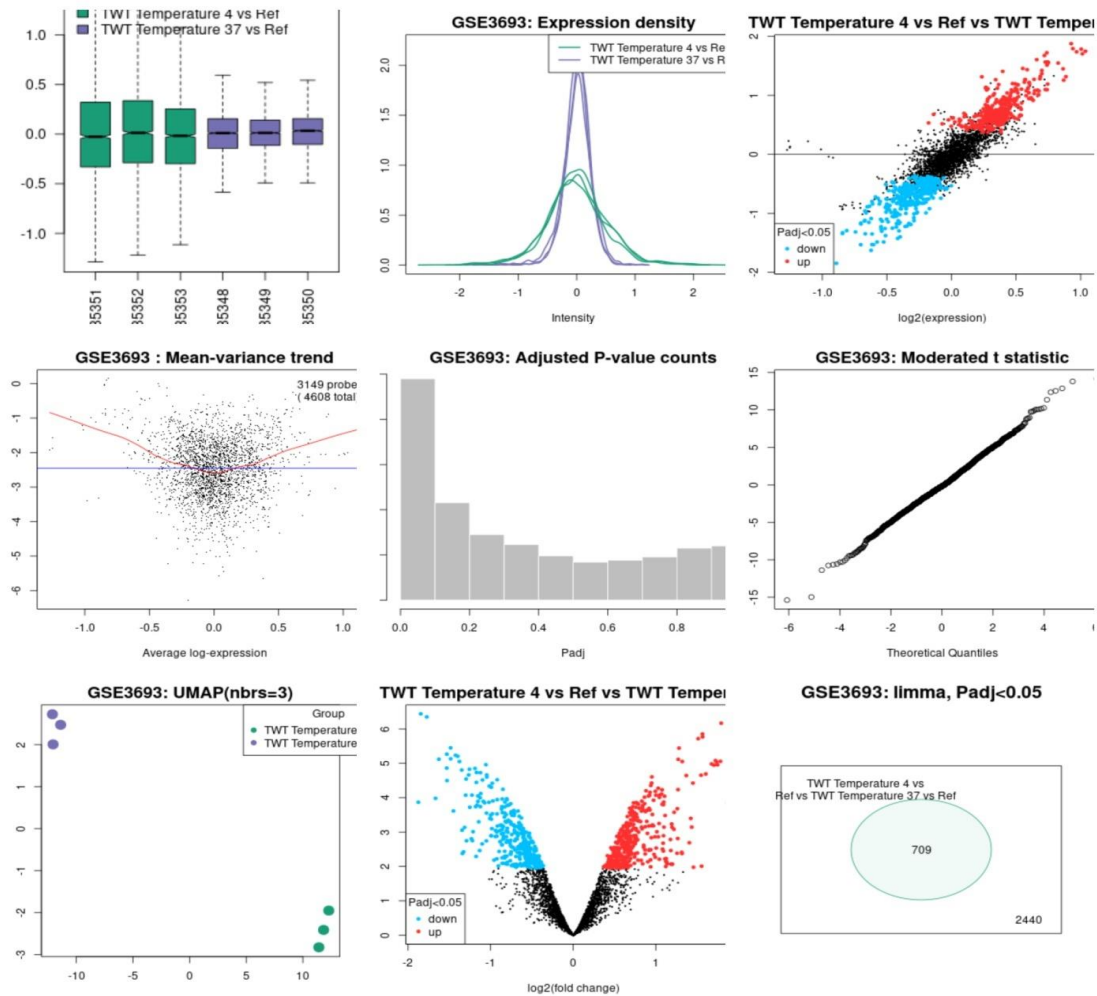


Figure 8. Microarray analysis plots for group Cold shock temperature 4°C Vs 37°C under series GSE3693

2. Cold shock temperature 28°C Vs 37°C

Here 4 genes were found to be down regulated and 3 genes were found to be up regulated. The down regulated gene was curved DNA binding protein (cbpA), and the significant up regulated gene was found to be 4-aminobutyrate amino transferase (gab-T). **Figure9** shows microarray analysis plots for this group.

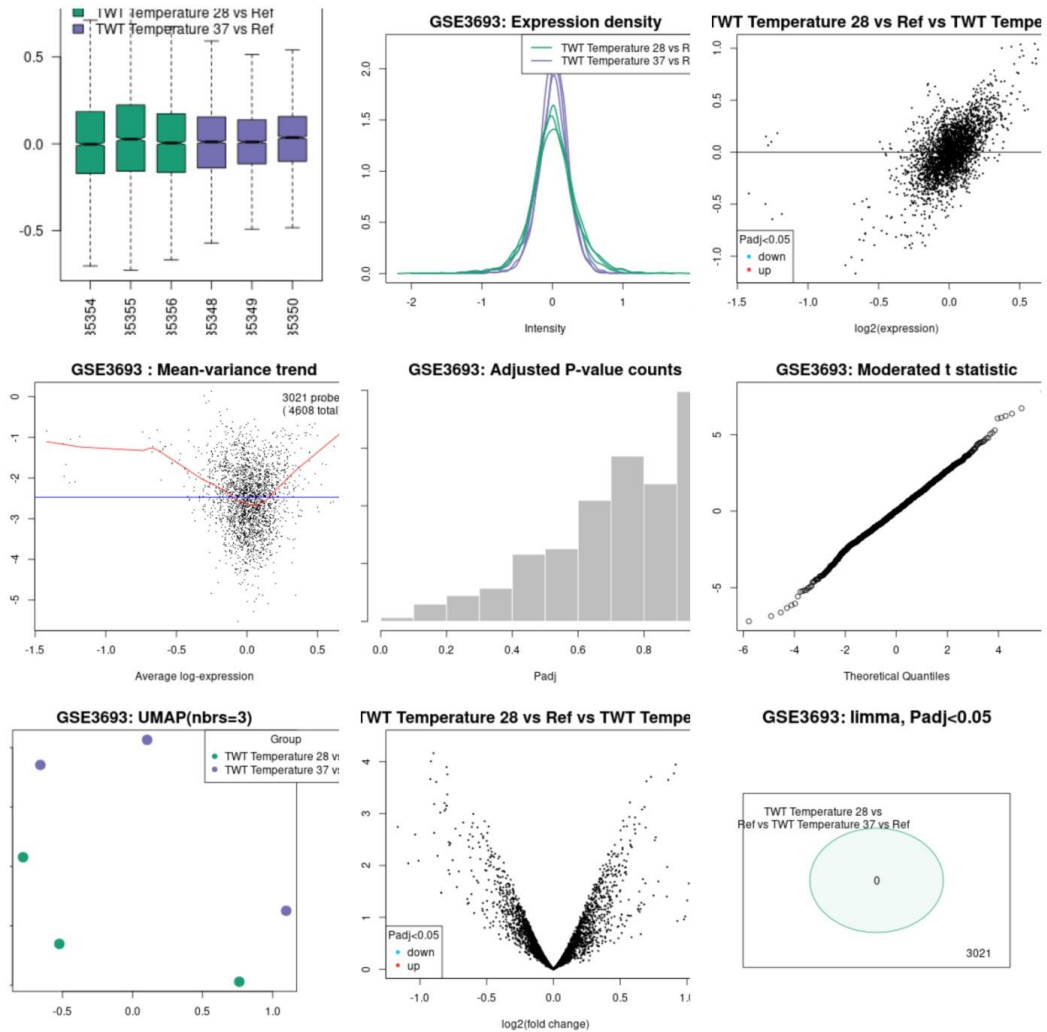


Figure 9. Microarray analysis plots for group Cold shock temperature 28°C Vs 37°C under series GSE3693

3. Heat shock temperature 43°C Vs 37°C

Here 43 up regulated genes were found and no significant down regulated genes were found. Fe-ABC transporter gene, HSP-70 cofactor, cbpA gene, riboflavin synthase, and GTP cyclohydrolase I was few common up- regulated genes. **Figure10** shows microarray analysis plots for this group.

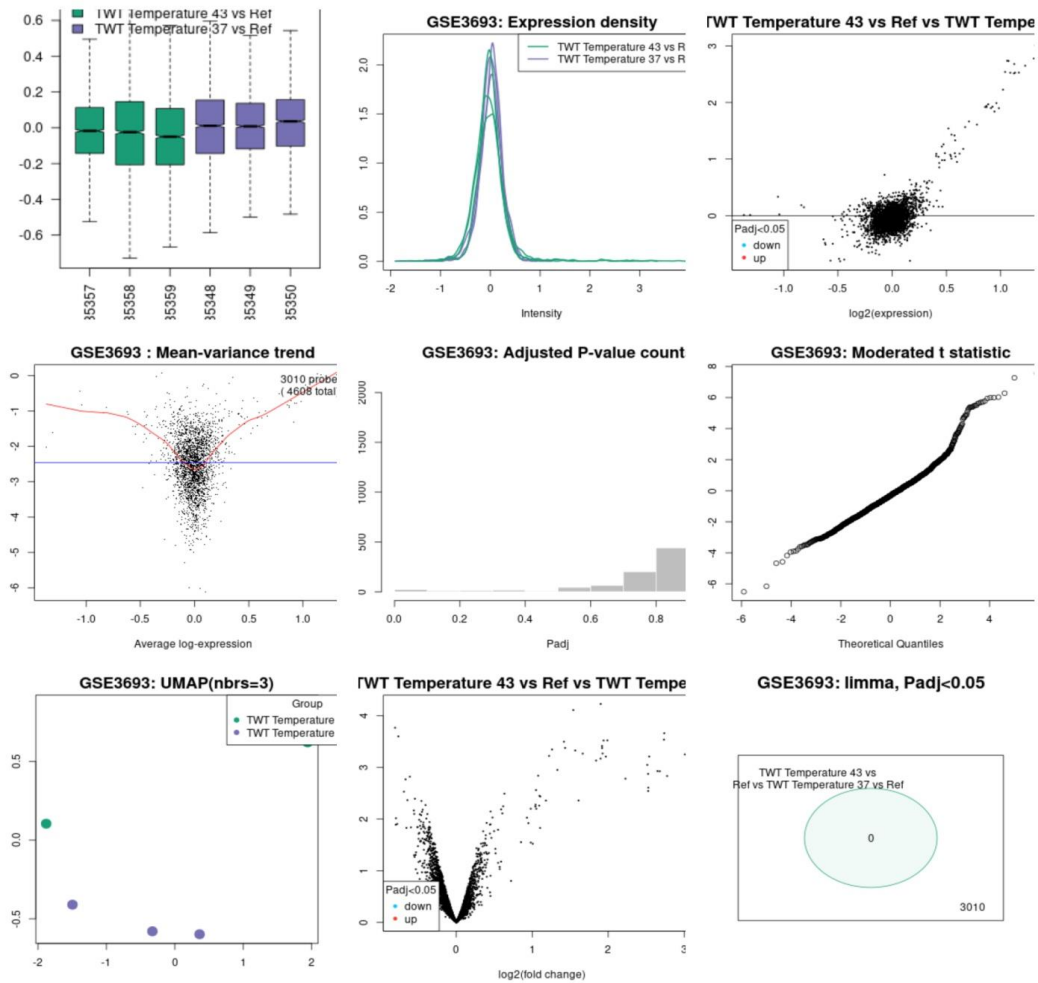


Figure 10. Microarray analysis plots for group Heat shock temperature 43°C Vs 37°C under series GSE3693

b. GSE5717:

For this series two groups were identified, based on drug doxycycline treatment with varied proportions on bacterium *Tropheryma whipplei*. Using a whole-genome DNA microarray data, the susceptibility of *T. whipplei* to doxycycline was studied at the genetic expression level. Antibiotic related primary transcriptional patterns were observed when *T. whipplei* was exposed to a minimum inhibitory dosage of 0.5mg doxycycline, but indirect effects were identified with a 5mg dosage of doxycycline. At low doses of 0.5 mg translation inhibition of bacterial ribosome proteins and transcription factors were observed, that can be linked with arrest of cellular growth. In higher conc. of Doxycycline, up-regulation of Membrane proteins like ATP Binding cassette transporters, which may create export and detoxify systems by which *T. whipplei* may restrict the impact of the bactericidal chemical.

1. Doxycycline @ 0.5 mg/l vs *Tropheryma whipplei* Twist strain

Here we found 9 genes (especially TWT685 gene that expresses DNA ribonuclease) to be up regulated while 4 genes to be down regulated (especially rpo-B gene that expresses DNA directed RNA-polymerase). **Figure11** shows microarray analysis plots for this group.

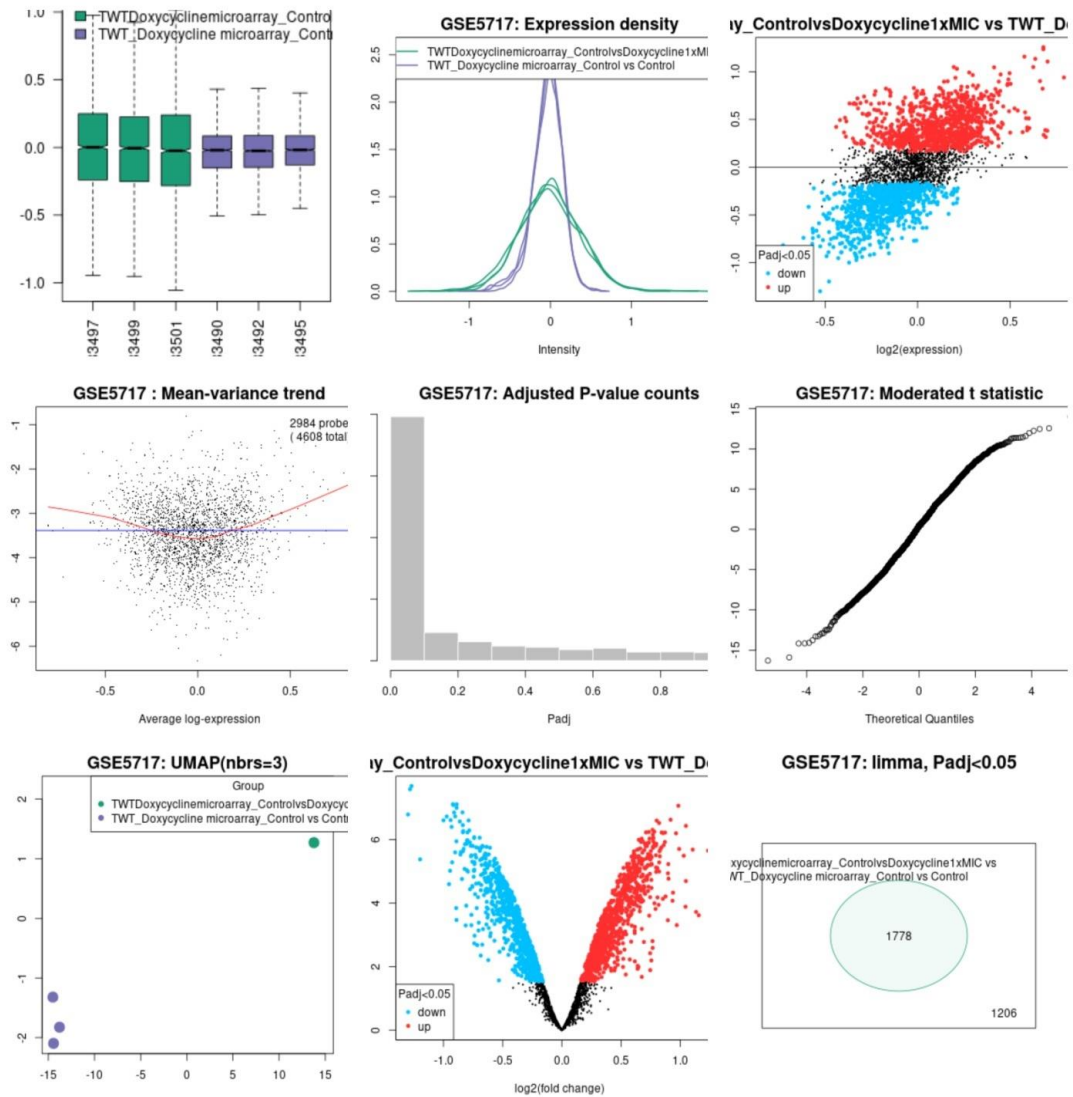


Figure 11. Microarray analysis plots for group Doxycycline @ 0.5 mg/l vs *Tropheryma whipplei* Twist strain

2. Doxycycline @ 5 mg/l vs *Tropheryma whipplei* Twist strain

Here we found 42 genes were up regulated while 3 genes were down regulated. Here down regulated genes include ribosomal protein genes. Up-regulated genes were found related to protease, aminotransferase, and nucleotidyl transferases. **Figure12** shows microarray analysis plots for this group.

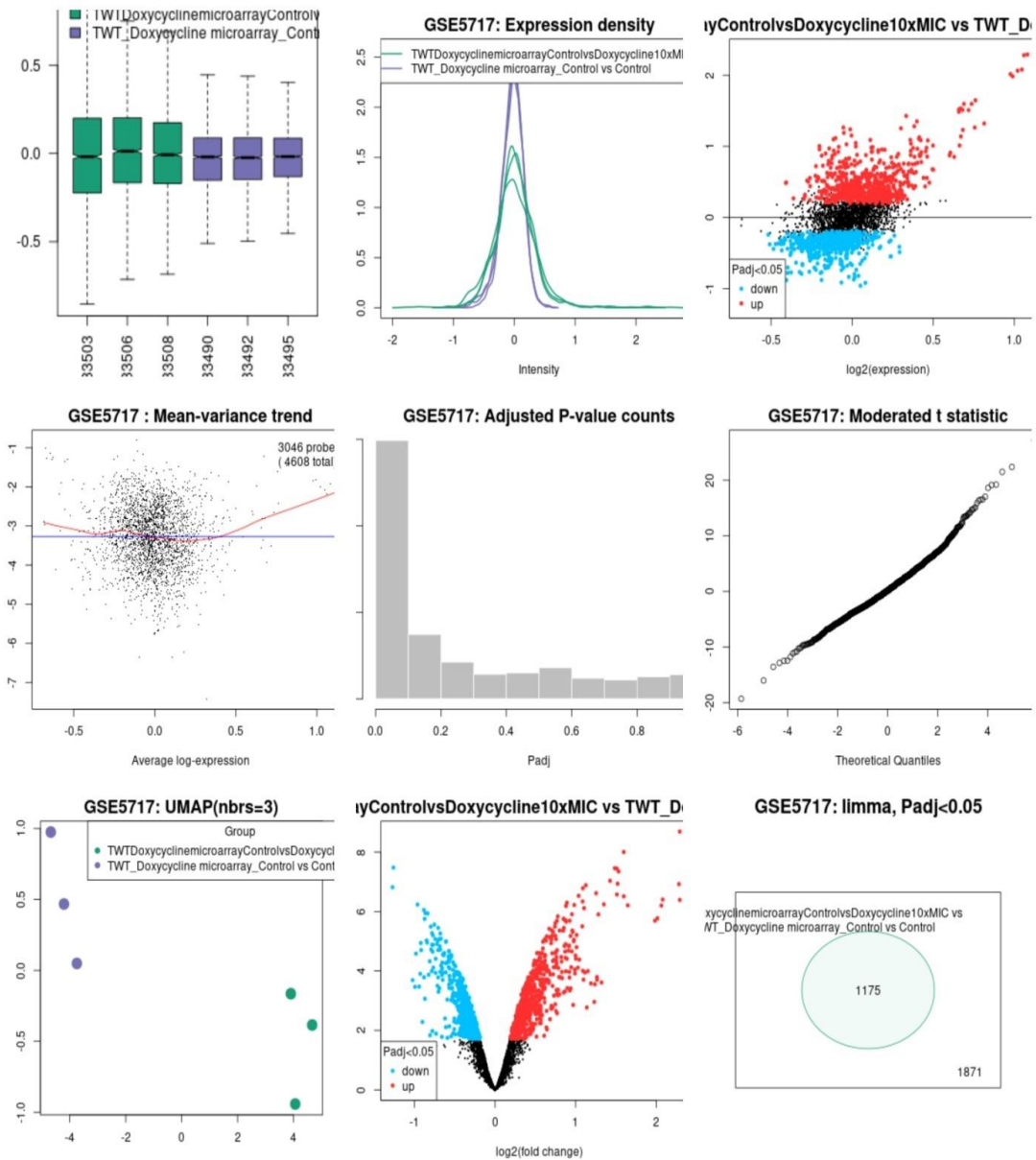


Figure 12. Microarray analysis plots for group Doxycycline @ 5 mg/l vs *Tropheryma whipplei* Twist strain

c. GSE7453

It is the evolutionary data set series where comparative analysis of *Tropheryma whipplei* Twist strain with other 15 strains were checked for differential expression assessment.

Through microarray-based comparative genomic hybridization, the genetic diversity of 15 clinical isolates of *Tropheryma whipplei* strains was compared to the *T. whipplei* Twist strain. The results showed that there was only a little amount of genetic diversity amongst these *T. whipplei* strains, with just 2.24 percent of the markers showing differential hybridization against the Twist strain. The WiSP family molecules were shown to have the most diversity, supporting the idea that these protein molecules are significant players in immune escape. A 19.2 kbpair deletion was also discovered in the *T. whipplei* DIG15 strain. This loss occurs in similar locus as the previously described massive chromosomal reorganization involving Twist and TW08/27, and hence can be regarded a key site for intra-specific *T. whipplei* divergence. So, there were total 15 groups were considered for microarray data analysis, which was as follows:

1. ART1 Vs Twist

22 genes were found to be down regulated commonly cell division protein *ftsE*, while 42 genes were found to be up regulated *ftsZ* cell division protein, pseudouridylate synthase I (*truA* gene), DNA directed RNA polymerase alpha unit (*rpoA* gene). **Figure13** shows microarray analysis plots for this group.

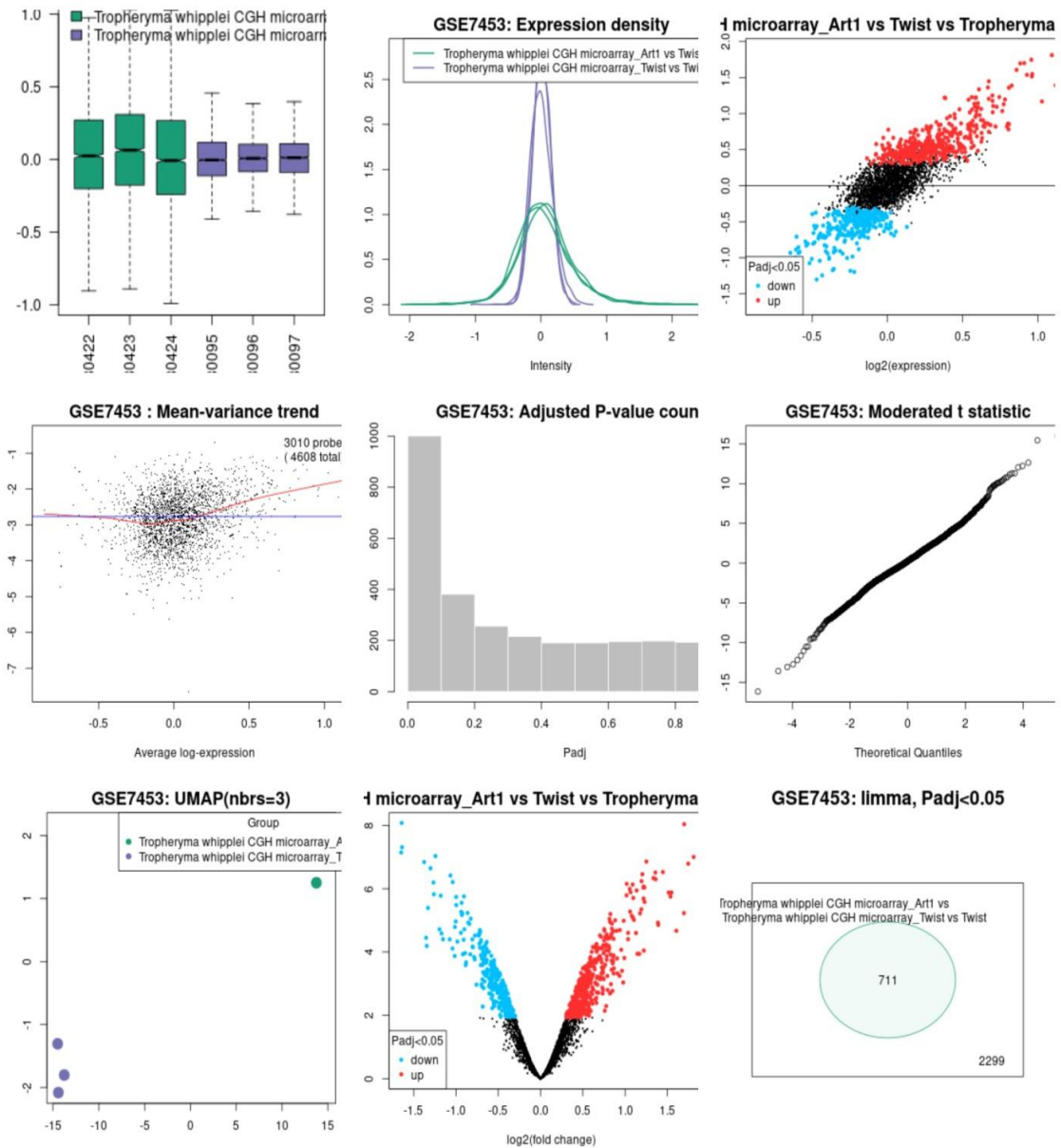


Figure 13. Microarray analysis plots for group ART1 Vs *Tropheryma whipplei* Twist strain

2. DigNeuro14 Vs Twist

3 genes were found to be up regulated and 2 genes were found to be down regulated. The up regulated genes commonly include chromosome partitioning protein (parA1 gene),

while down regulated gene include acetolactate synthase small subunit (ilvH gene). **Figure 14** shows microarray analysis plots for this group.

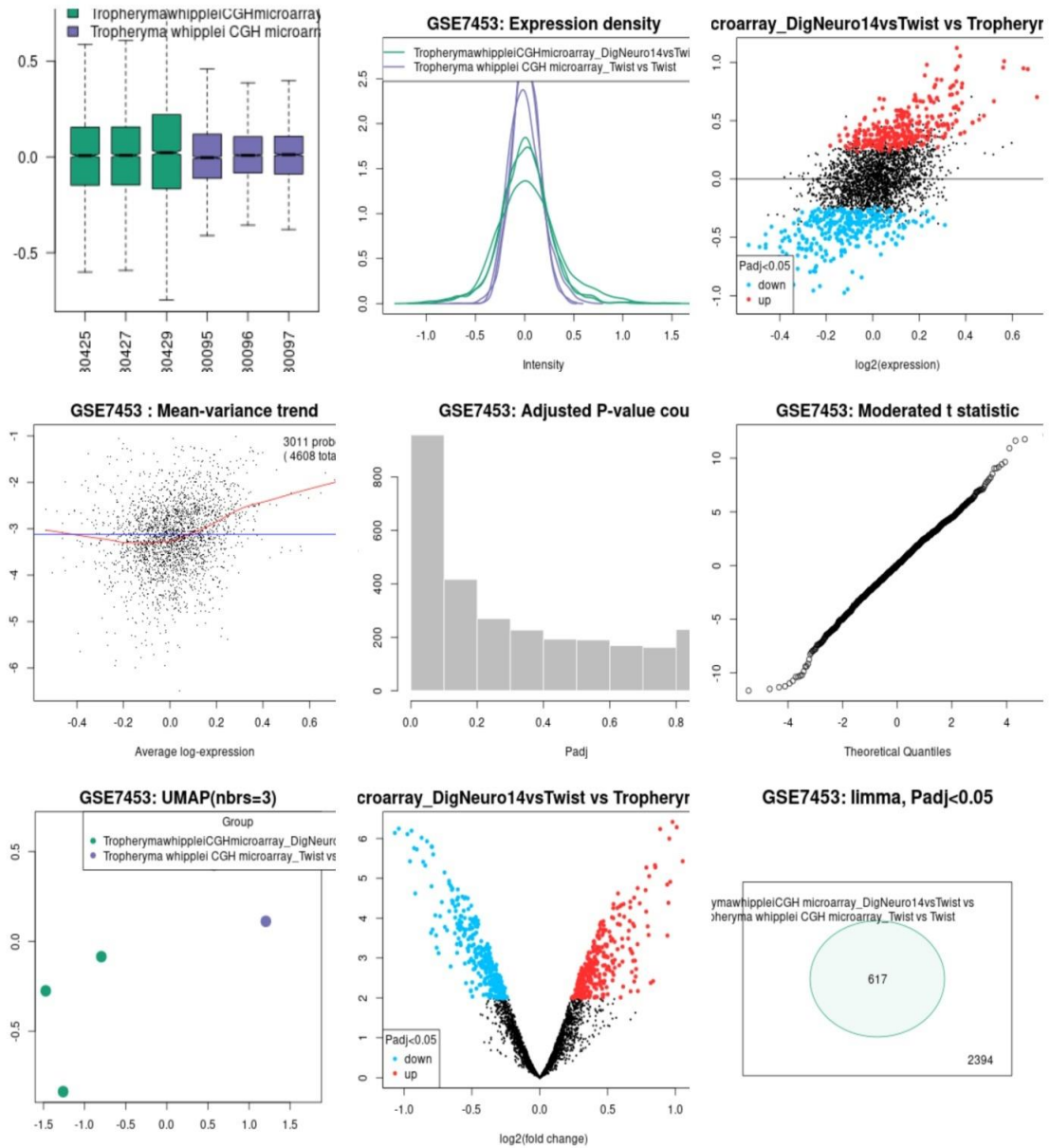


Figure 14. Microarray analysis plots for group DigNeuro14 Vs *Tropheryma whipplei* Twist strain

3. Dig7 Vs Twist

7 genes were found to be down regulated while no significant up regulated genes were found. Common down-regulated gene was found to be ftsE cell division protein.

Figure15 shows microarray analysis plots for this group.

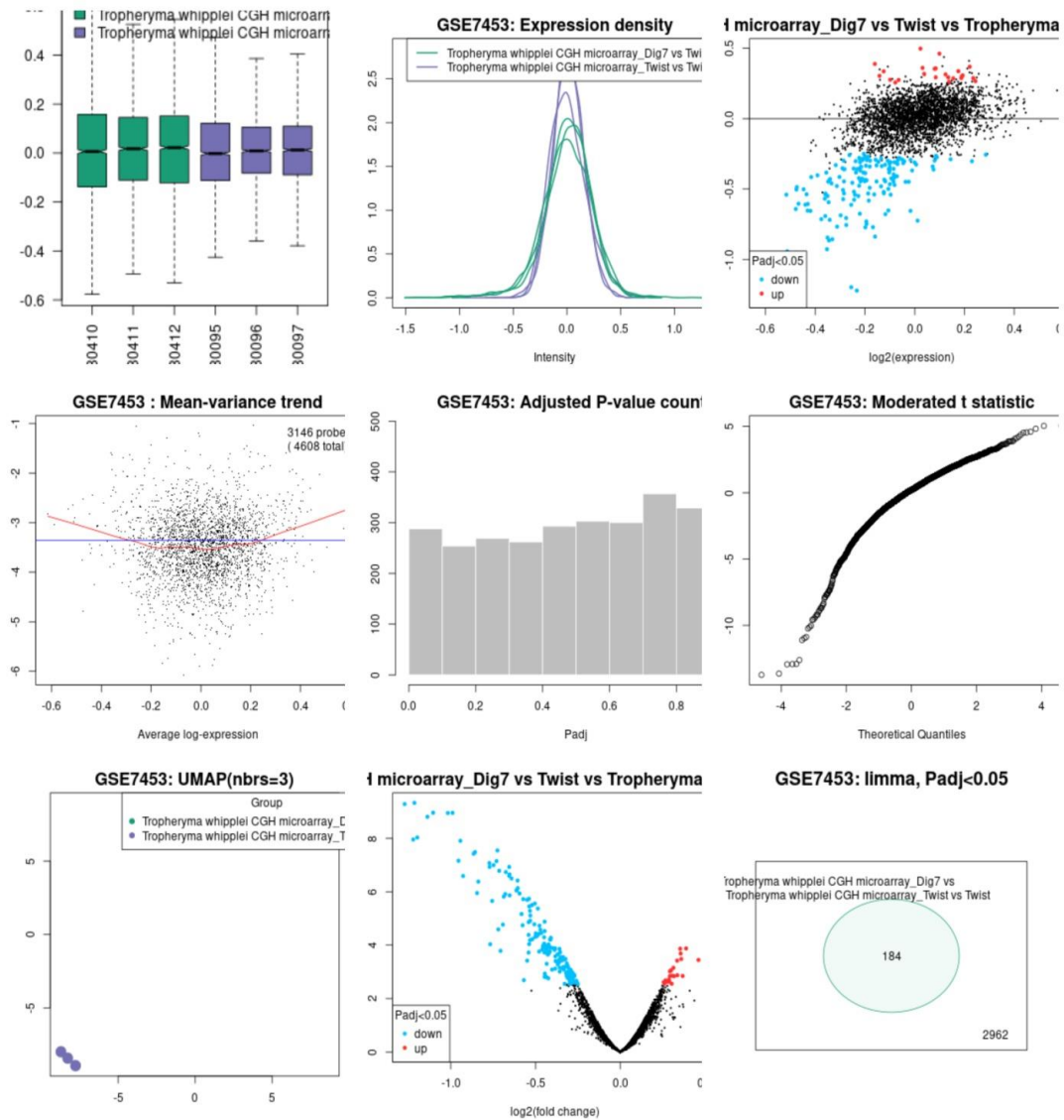


Figure 15. Microarray analysis plots for group Dig7 Vs *Tropheryma whipplei* Twist strain

4. Dig9 Vs Twist

7 genes were found to be down regulated while no significant up regulated genes were found. And most of the proteins were hypothetical proteins. **Figure16** shows microarray analysis plots for this group.

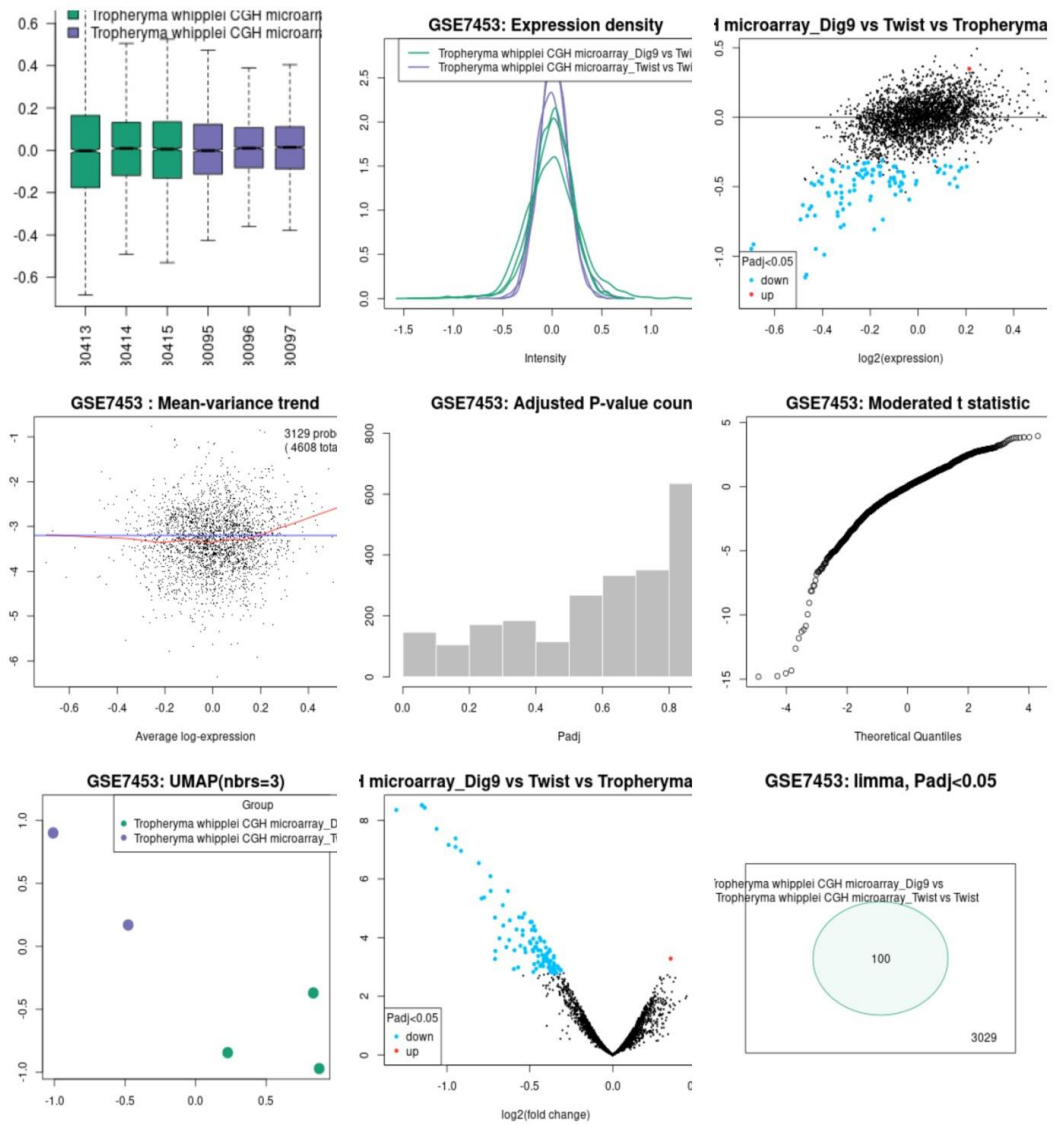


Figure 16. Microarray analysis plots for group Dig9 Vs *Tropheryma whipplei* Twist strain

5. Dig10 Vs Twist

4 genes for hypothetical proteins (commonly marked as TWT151 gene) were found to be up regulated and 2 genes for hypothetical proteins (TWT594 and TWT099) were found to be down regulated. **Figure17** shows microarray analysis plots for this group.

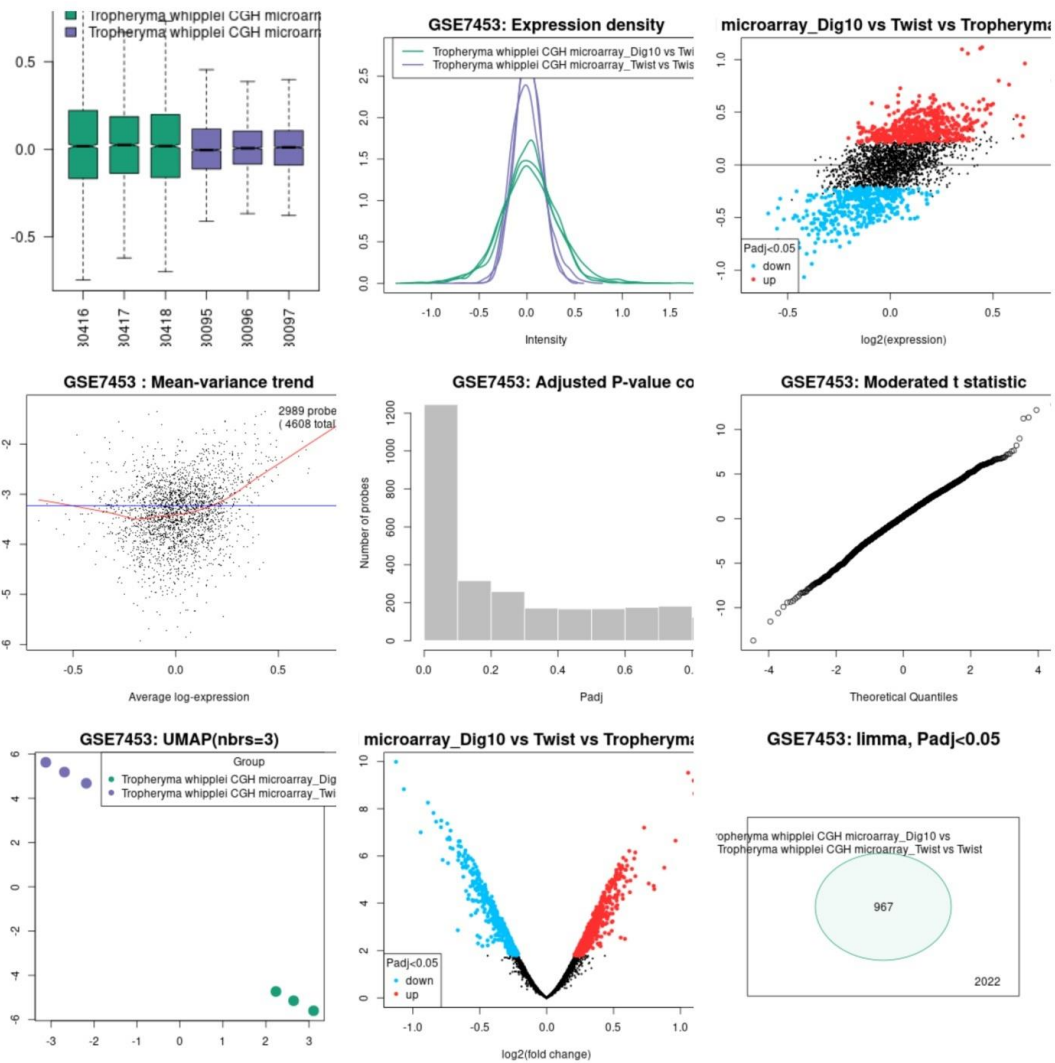


Figure 17. Microarray analysis plots for group Dig10 Vs *Tropheryma whipplei* Twist strain

6. Dig15 Vs Twist

36 genes were down regulated but no significant up regulation genes was observed. Commonly *ftsK* cell division protein, dimethyladenosine transferase, and many hypothetical proteins were found to be under expressed. **Figure18** shows microarray analysis plots for this group.

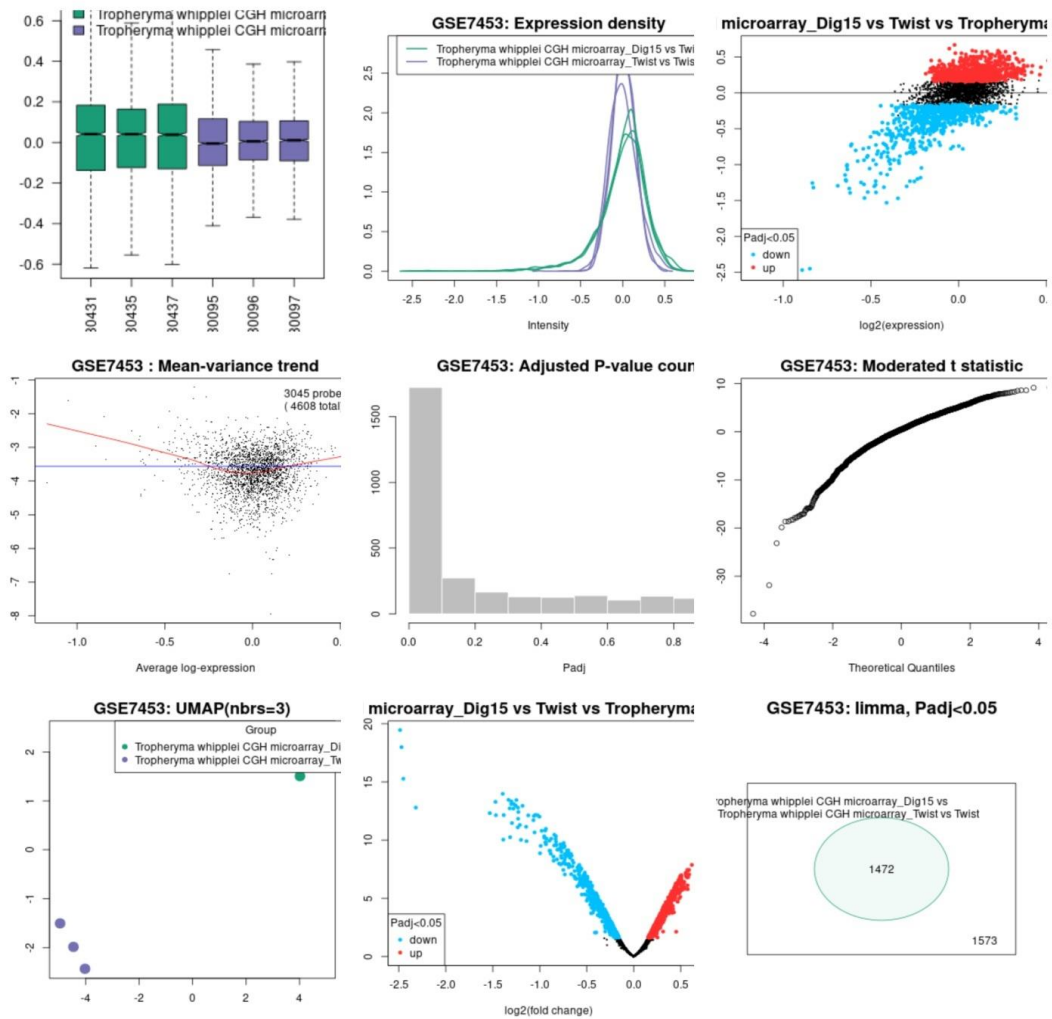


Figure 18. Microarray analysis plots for group Dig15 Vs *Tropheryma whipplei* Twist strain

7. DigADP11 Vs Twist

9 genes were found to be up regulated but no significant down regulated genes were found during analysis, also all up regulated genes were hypothetical proteins. **Figure19** shows microarray analysis plots for this group.

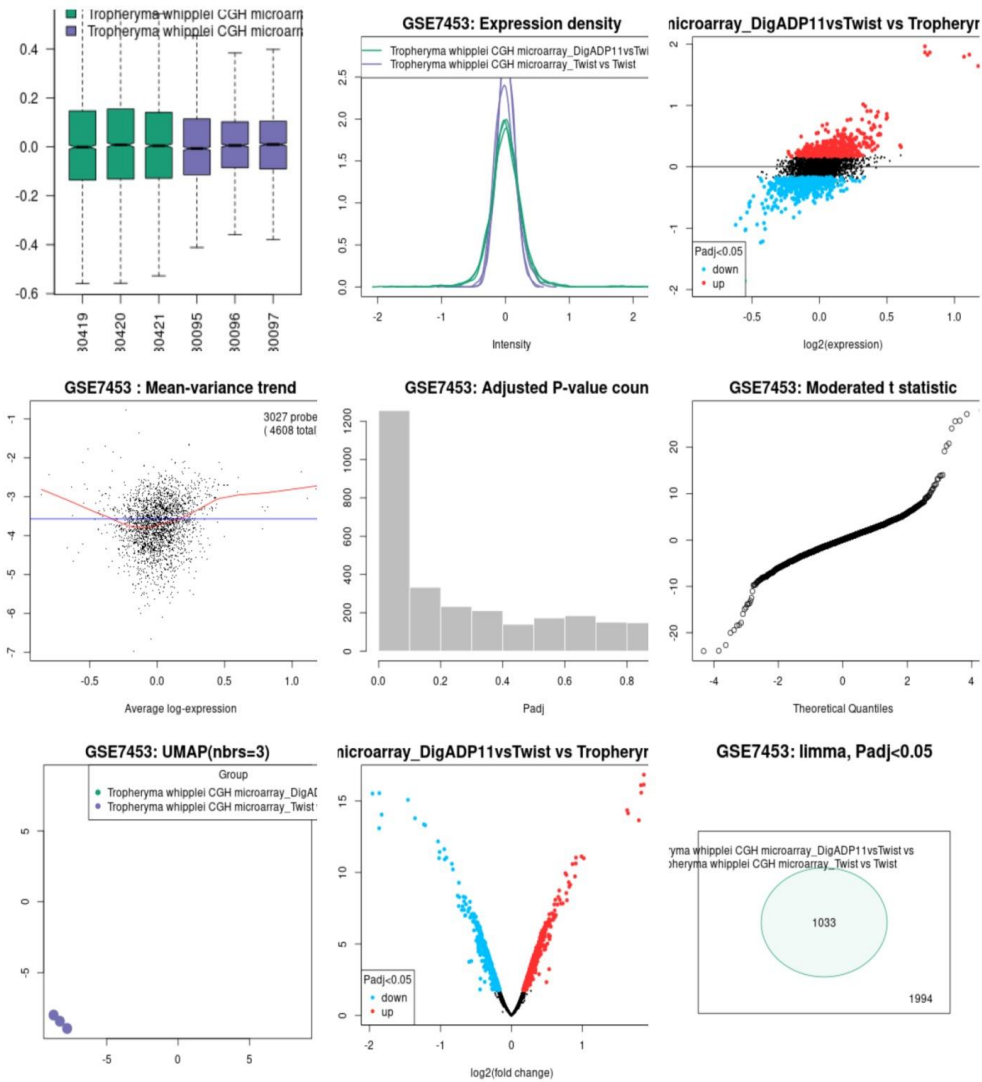


Figure 19. Microarray analysis plots for group DigADP11 Vs *Tropheryma whipplei* Twist strain

8. DigMusc17 Vs Twist

5 genes were found to be down regulated but no significant up regulated genes were found, particularly *ftsE* gene was found to be down regulated. **Figure20** shows microarray analysis plots for this group.

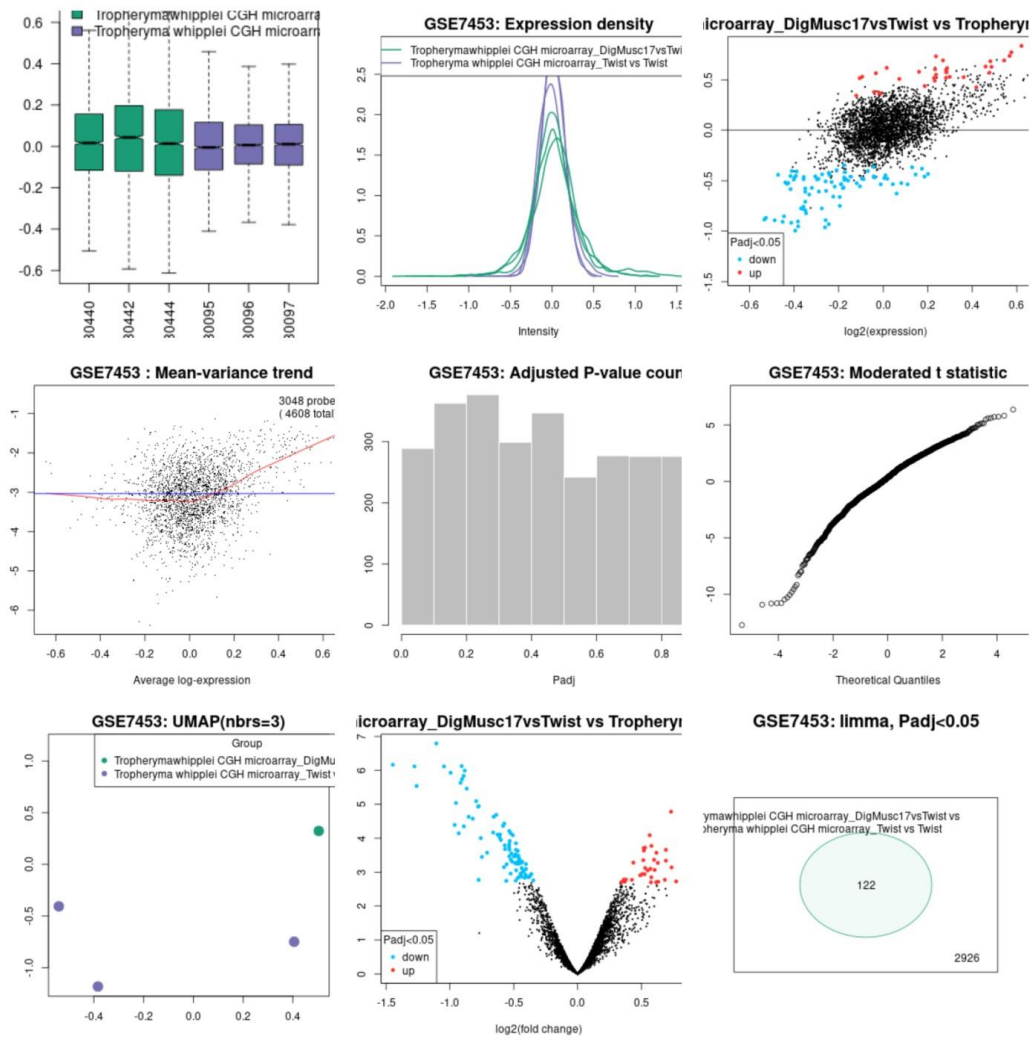


Figure 20. Microarray analysis plots for group DigMusc17 Vs *Tropheryma whipplei* Twist strain

9. DigNeuro18 Vs Twist

10 genes were found to be down regulated but no significant up regulated genes were found during analysis. Particularly, *ftsE* gene was found to be down-regulated. **Figure 21** shows microarray analysis plots for this group.

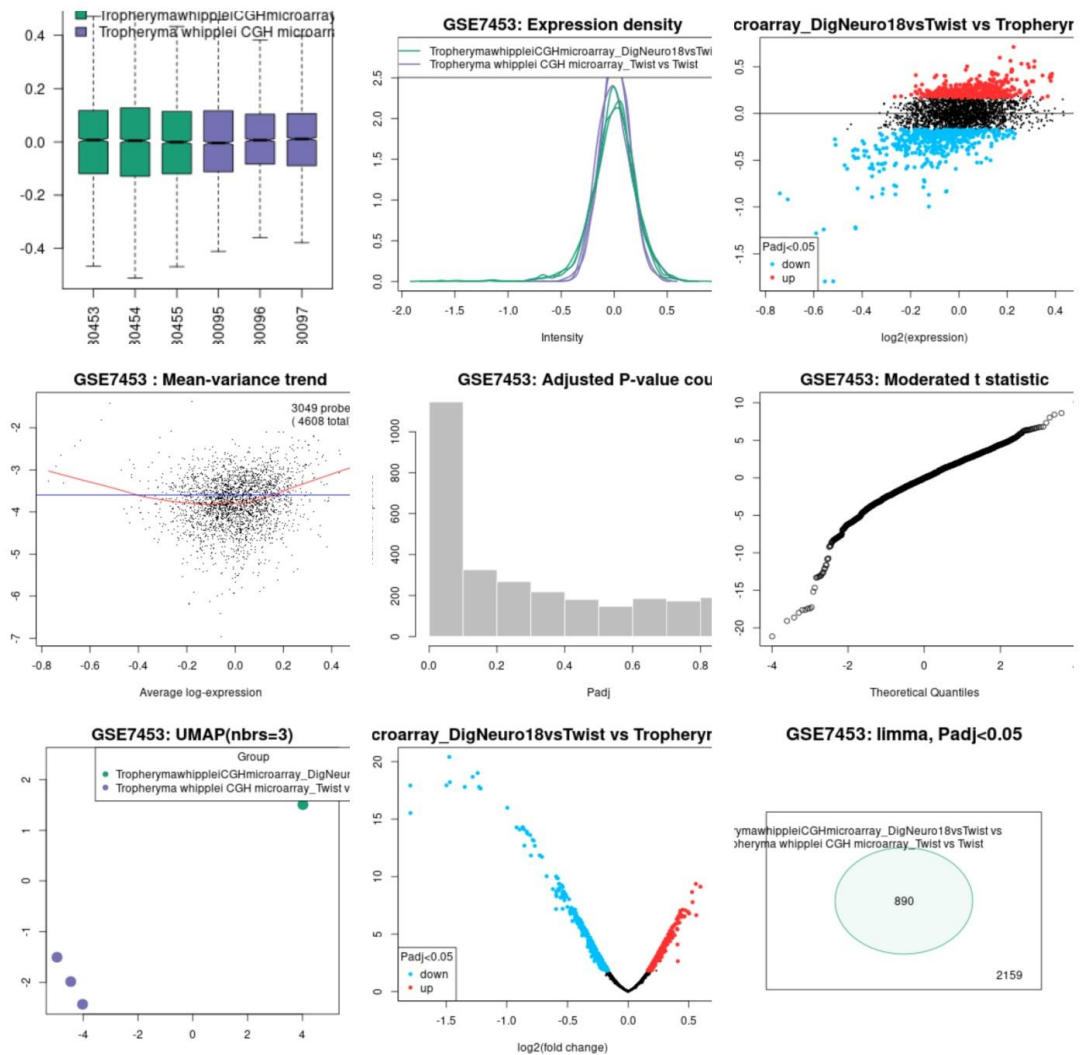


Figure 21. Microarray analysis plots for group DigNeuro18 Vs *Tropheryma whipplei* Twist strain

10. Endo5 Vs Twist

8 genes were found to be down regulated and no significant up regulated genes were found during analysis. Also, all the down regulated genes were found to be hypothetical proteins. **Figure22** shows microarray analysis plots for this group.

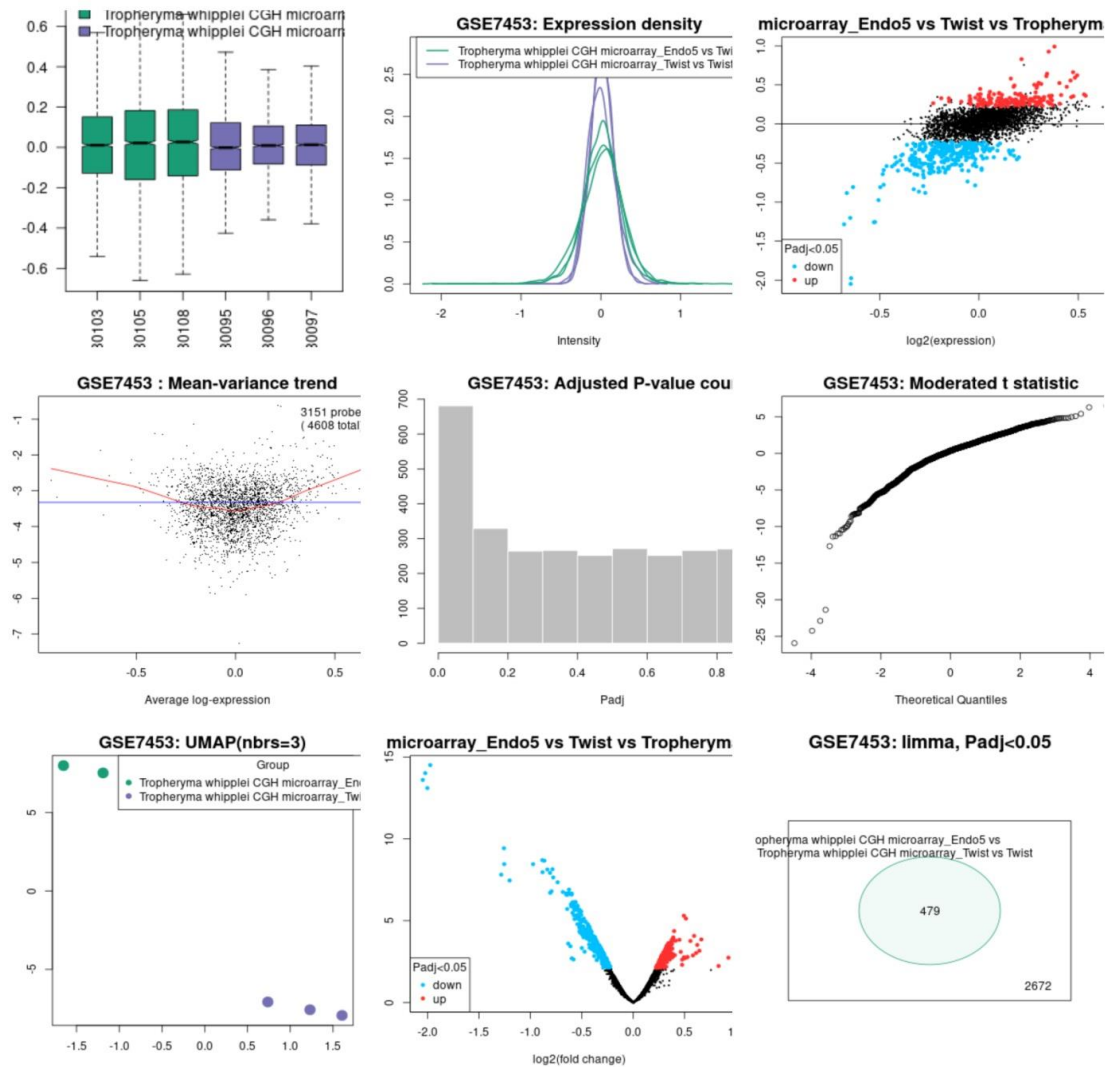


Figure 22. Microarray analysis plots for group Endo5 Vs *Tropheryma whipplei* Twist strain

11. Endo7 Vs Twist

2 genes for hypothetical proteins were found to be down regulated and no significant up regulated genes were found. **Figure23** shows microarray analysis plots for this group.

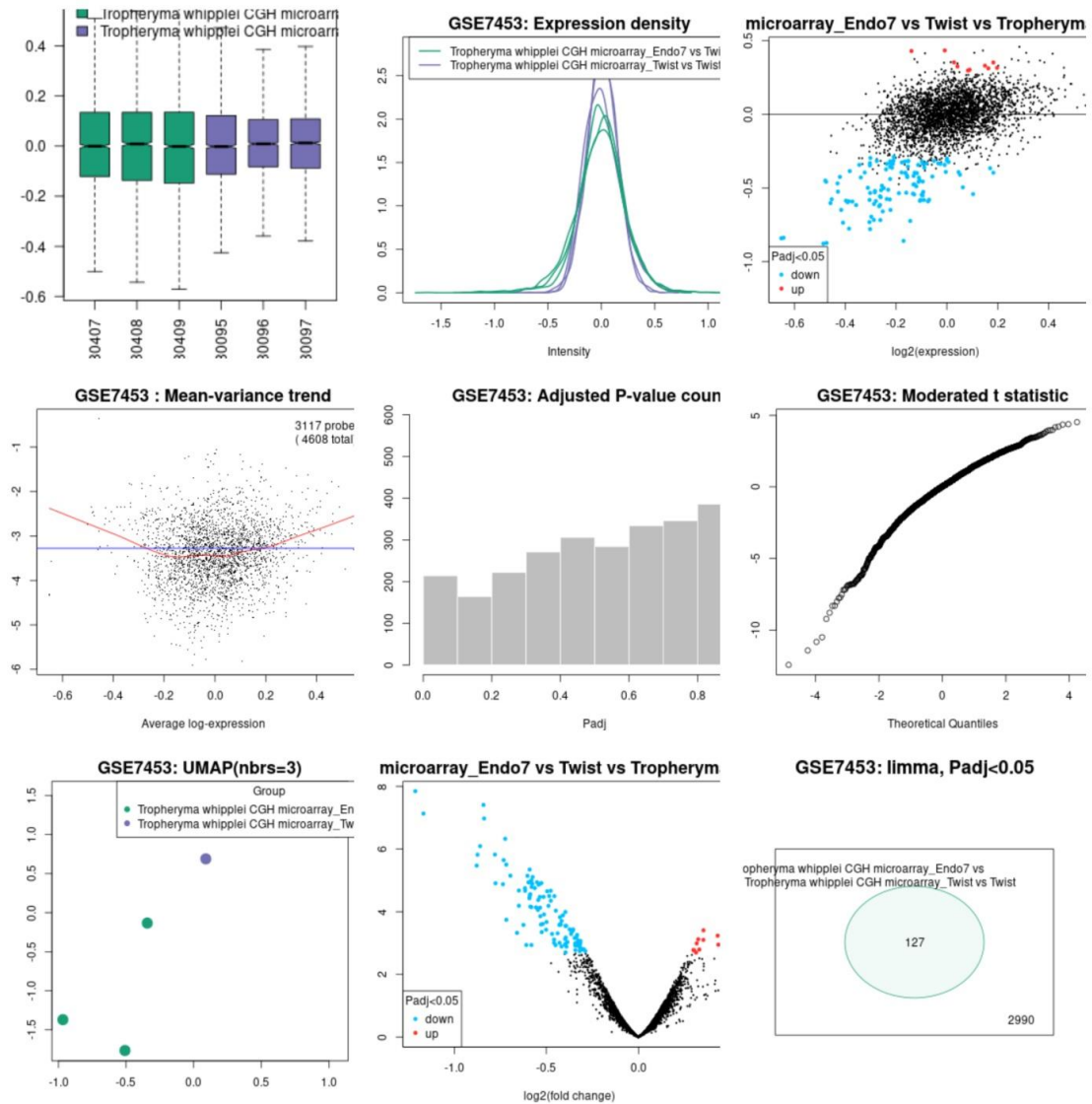


Figure 23. Microarray analysis plots for group Endo7 Vs *Tropheryma whipplei* Twist strain

12. Neuro1 Vs Twist

4 genes for hypothetical proteins were found to be down regulated and no significant up regulated genes were found. **Figure24** shows microarray analysis plots for this group.

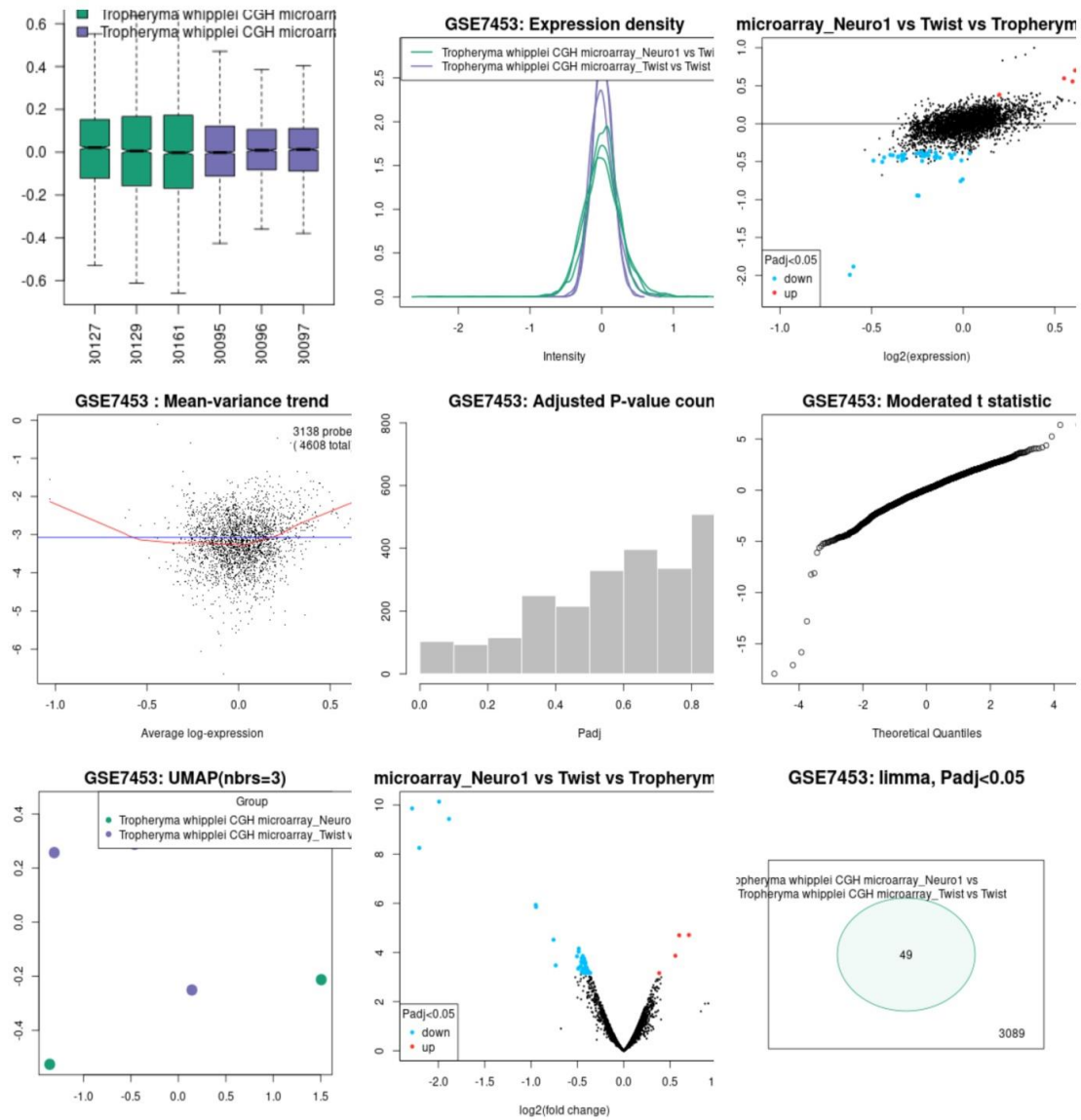


Figure 24. Microarray analysis plots for group Neuro1 Vs *Tropheryma whipplei* Twist strain

13. Neuro2 Vs Twist

No significant down regulated and up regulated genes were found during analysis. **Figure25** shows microarray analysis plots for this group.

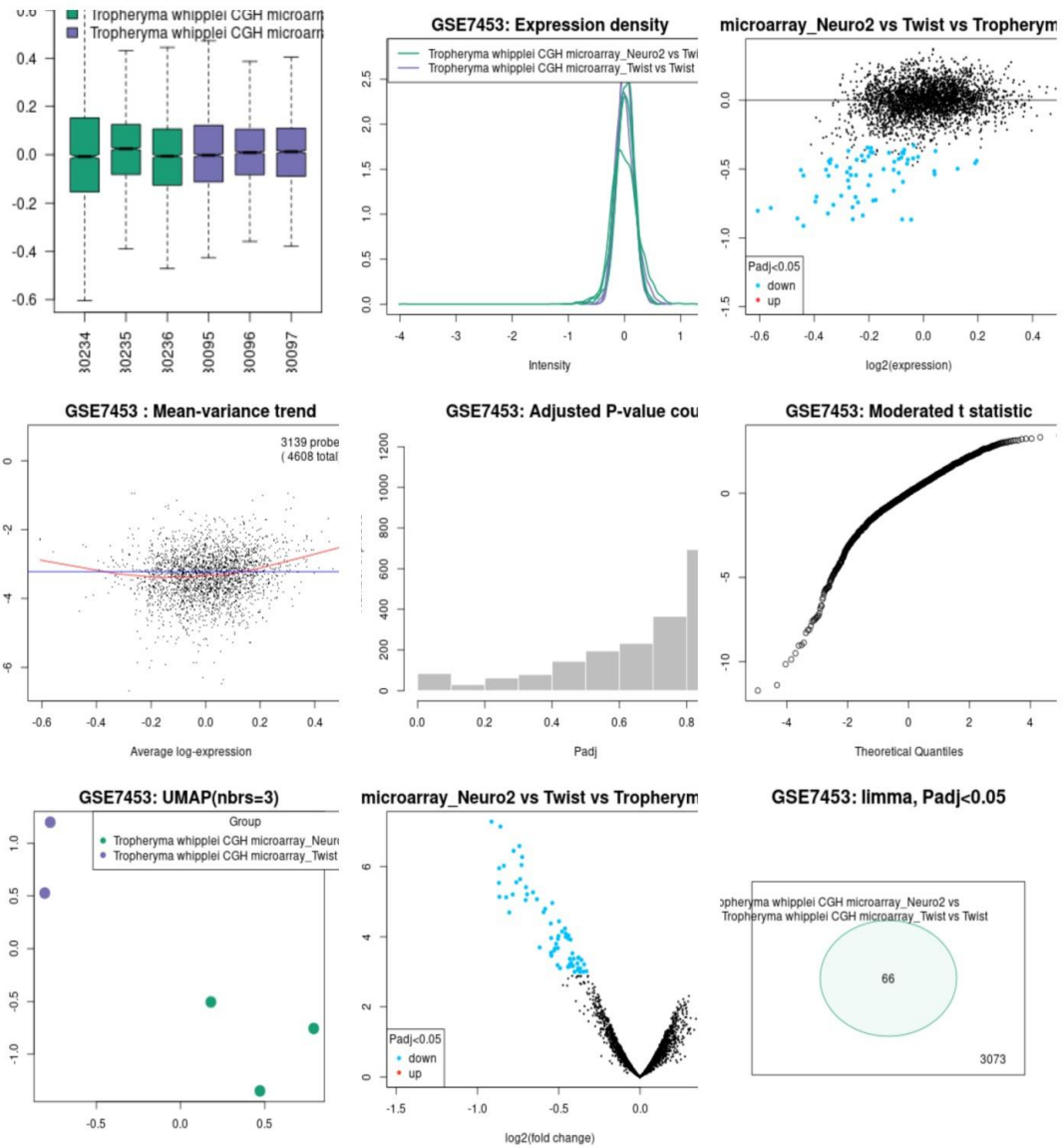


Figure 25. Microarray analysis plots for group Neuro2 Vs *Tropheryma whipplei* Twist strain

1. Slow1B Vs Twist

No significant down regulated and up regulated genes were found during analysis. **Figure26** shows microarray analysis plots for this group.

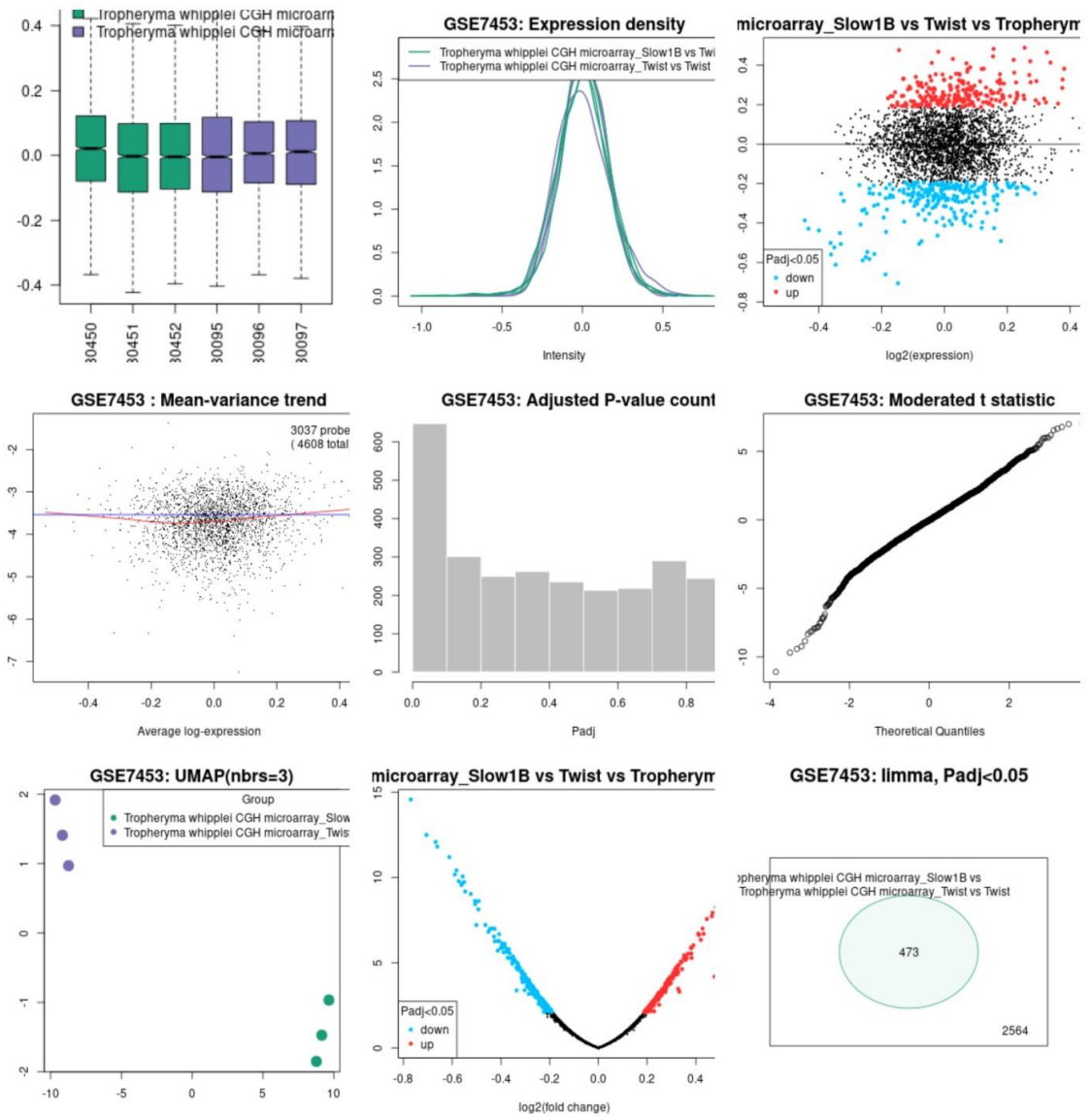


Figure 26. Microarray analysis plots for group Slow1B Vs *Tropheryma whipplei* Twist strain

2. Slow2 Vs Twist

7 genes were found to be down regulated while no significant up regulated genes were found. And most of the proteins were hypothetical proteins.

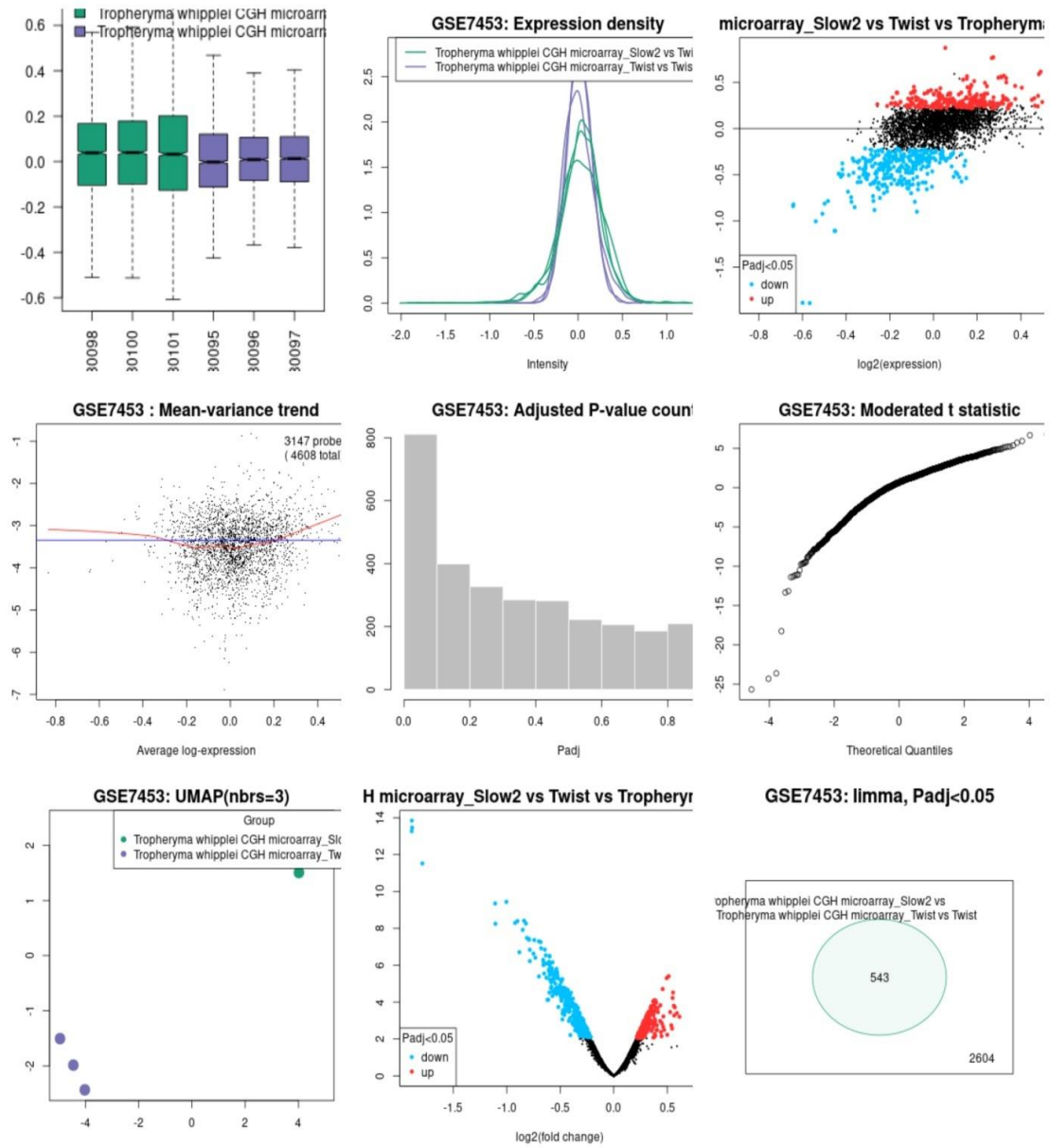


Figure 27. Microarray analysis plots for group Slow2 Vs *Tropheryma whipplei* Twist strain

d. GSE16180:

In this data series microarray data of bone marrow derived macrophages (BMDM) from mice were retrieved from *Tropheryma whipplei* infected individual and from normal individual was retrieved and analysed. The single core group was constructed and analysed by using GEO2R tool.

1. BMDM_Tw Vs BMDM_Control

Here we obtained 114 down regulated genes, and 206 up regulated genes. Up regulated genes mainly include chemokines, TNF receptor associated factor1 (traf1), fas receptor, and phosphatidylinositol 3-kinase. **Figure28** shows microarray analysis plots for this group.

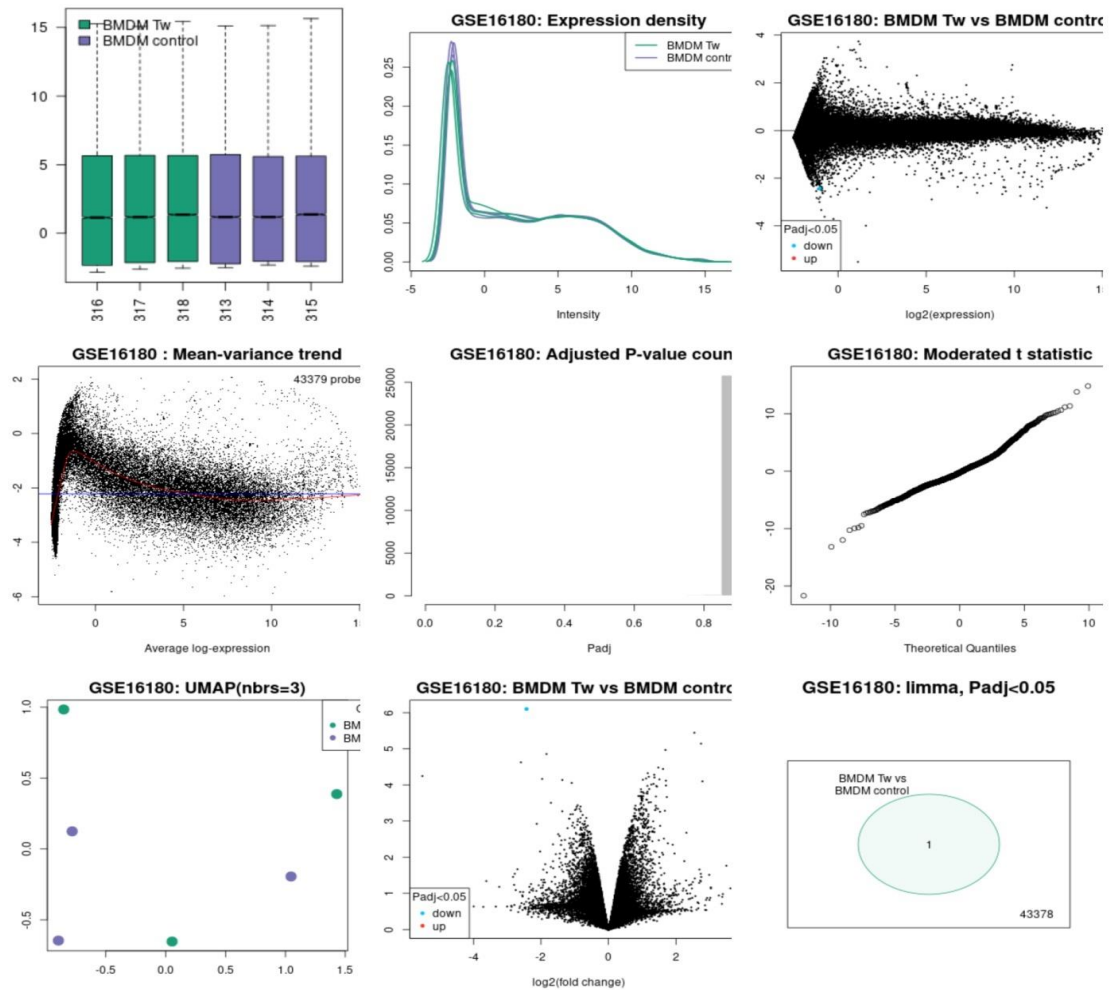


Figure 28. Microarray analysis plots for group BMDM_*Tropheryma whipplei* Twist strain Vs BMDM_Control

e. GSE20209:

In this series IL-16 Knock out Vs BMDM_Tw from mice model core group was constructed and analysed by using GEO2R tool. The data was collected from the NCBI series demonstrating entire genome microarray expression profile to detect differentially expressed genes after *T. whipplei* infection of interleukin-16-knock-out bone marrow derived macrophage. *T. whipplei* (multiplicity of infection50:1) was inoculated to

macrophages for 6 hours, and a pattern was discovered between infected and control subjects.

1. IL-16 Knock out Vs BMDM_Tw

783 genes were up regulated while 79 genes were down regulated. Common down regulated genes were Interferon beta 1, cdc7, cdkn1b (Cyclin dependent kinase Inhibitor 1b), collagen, and retinol binding protein, while common up regulated genes were found to be chemokine receptor, tubulin gamma complex protein, ubiquitin, fibronectin, and interleukin-2. **Figure 29** shows microarray analysis plots for this group.

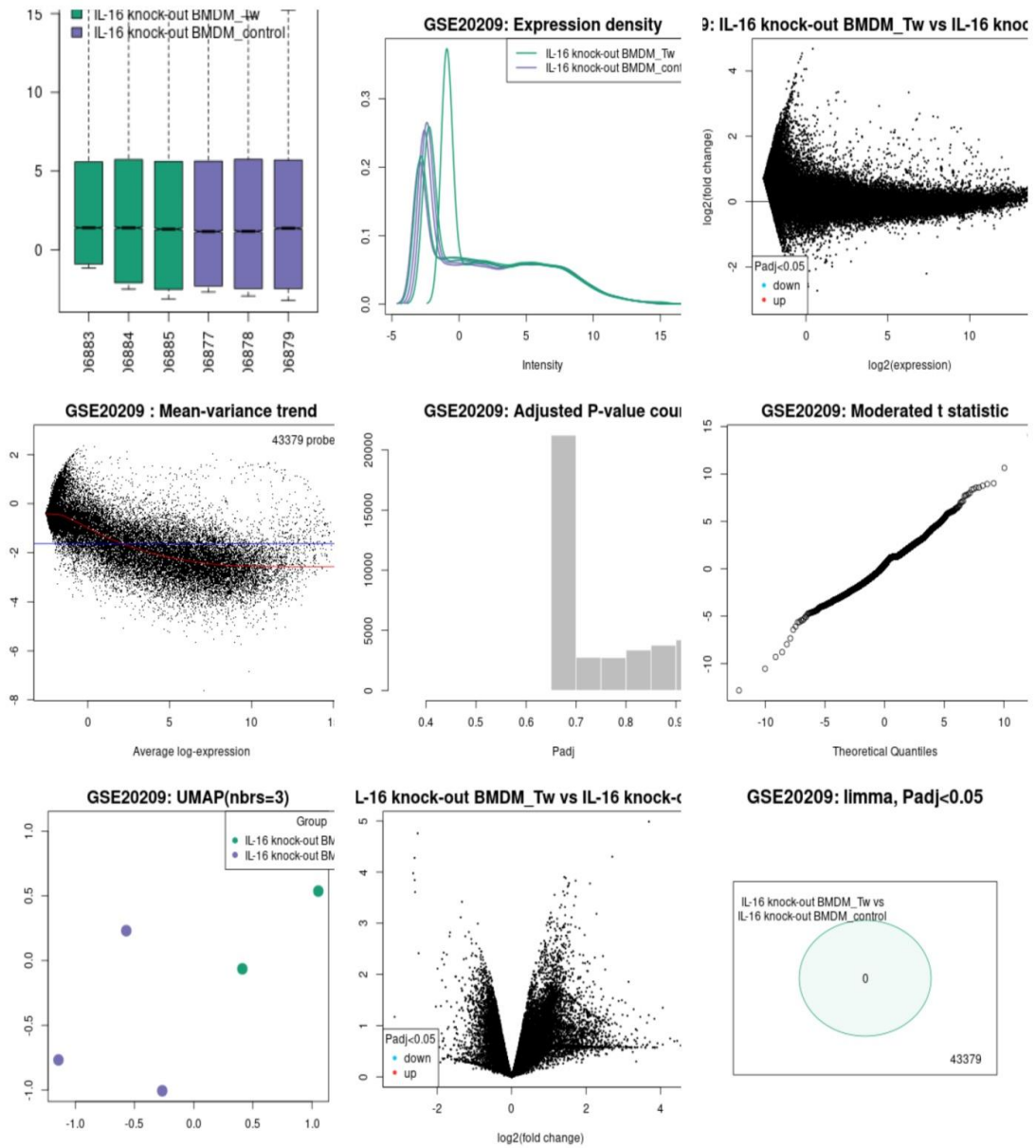


Figure 29. Microarray analysis plots for group IL-16 Knock out Vs BMDM_ *Tropheryma whipplei* Twist strain

F. GSE20210:

In this series four groups were made for analysis plan from the data extended from prior two series. As series d, e, f were from same super series. Following groups were analysed for data analysis:

1. BMDM_Tw Vs BMDM_Control

Here we obtained 114 down regulated genes, and 206 up regulated genes. Up regulated genes mainly include chemokines, TNF receptor associated factor1 (traf1), fas receptor, and phosphatidylinositol 3-kinase.

2. BMDM_LPS Vs BMDM_Control

Here we obtained 3626 gene up regulated and 10188 genes down regulated. Most of the up regulated genes were found to be metabolic. **Figure 30** shows microarray analysis plots for this group.

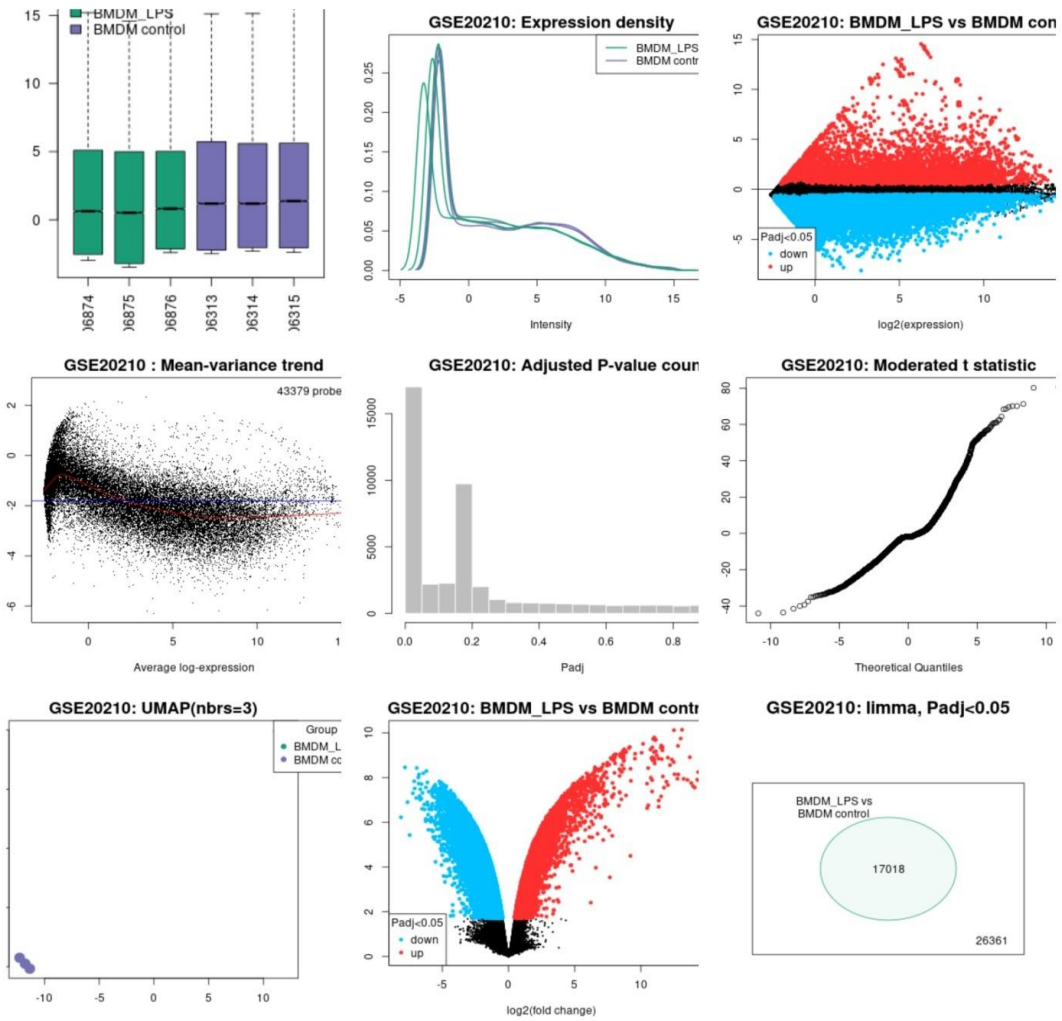


Figure 30. Microarray analysis plots for group BMDM_LPS Vs BMDM_Control

3. IL-16 Knockout BMDM_LPS Vs IL-16 Knockout BMDM_Control

Here we obtained 9834 down regulated and 3675 up regulated genes. Most of the up regulated genes were found to be metabolic. **Figure 31** shows microarray analysis plots for this group.

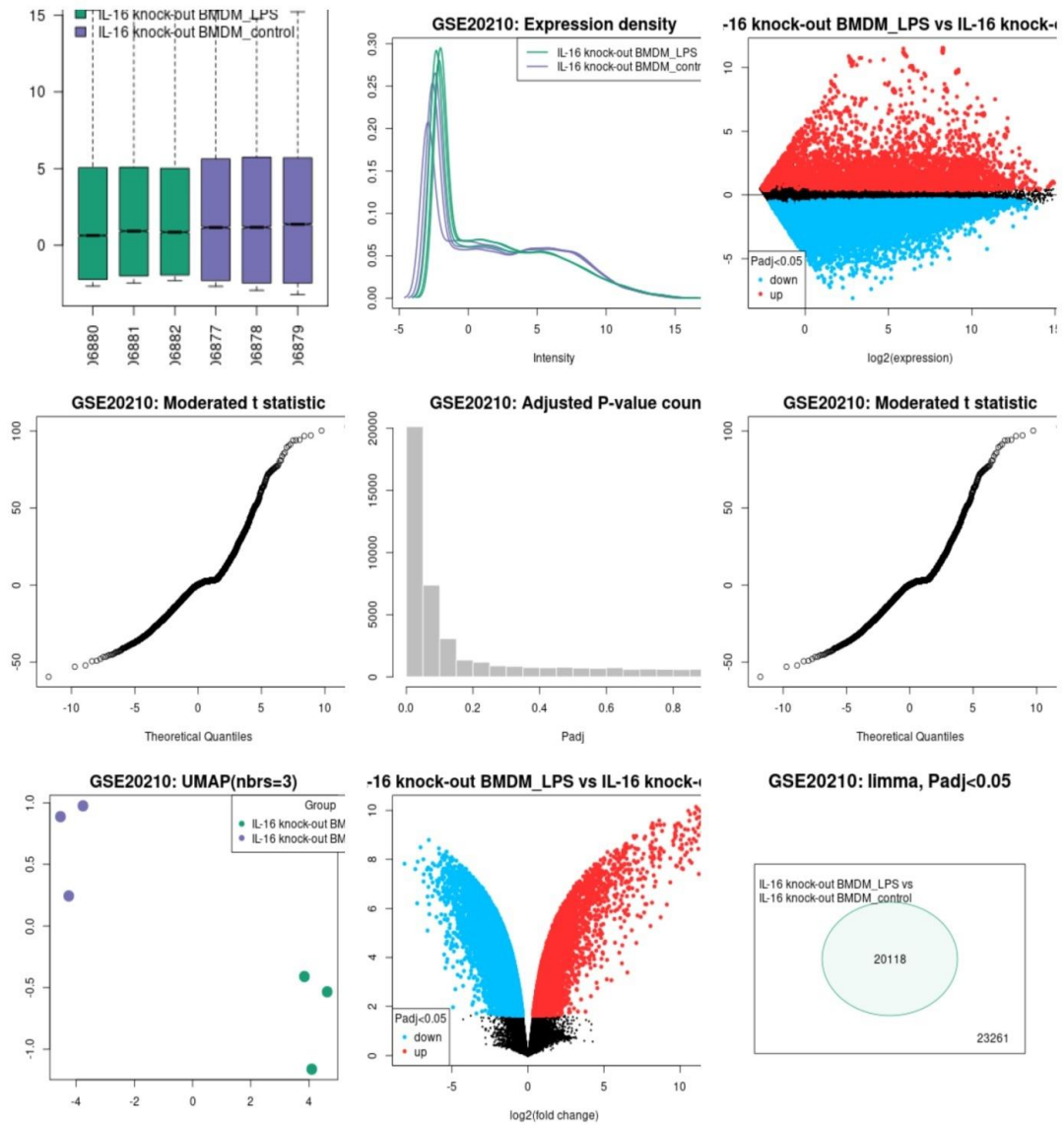


Figure 31. Microarray analysis plots for group IL-16 Knockout BMDM_LPS Vs IL-16 Knockout BMDM_Control

4. IL-16 Knockout BMDM_Tw Vs IL-16 Knockout BMDM_Control

Here we obtained 783 genes were up regulated while 79 genes were down regulated. Common down regulated genes were Interferon beta 1, cdc7, cdkn1b (Cyclin dependent kinase Inhibitor 1b), collagen, and retinol binding protein, while common up regulated

genes were found to be chemokine receptor, tubulin gamma complex protein, ubiquitin, fibronectin, and interleukin-2.

G. GSE49016:

Here particularly one specific group was considered where *Tropheryma whipplei* infected dendritic cells and unstimulated dendritic cells microarray data was retrieved and analysed. When human monocyte-derived dendritic cells were infected with *Tropheryma whipplei*, we looked at how they responded. To characterise common and distinct transcriptional responses to bacterial pathogens, whole genome microarrays were used. Despite the fact that *Coxiella burnetii*, *Orientia tsutsugamushi*, and *Brucella abortus* all triggered dendritic cell maturation as measured by reduced endocytosis, triggering lymphocyte proliferation, surface expression of HLA class II molecules, and phenotypic alterations, *Tropheryma whipplei* did not. Here we primarily focused on group *T. whipplei* infected dendritic cells Vs Unstimulated dendritic cells.

1. *T. whipplei* infected DC's Vs Unstimulated DC's

Here we obtained 138 down regulated genes and 207 up regulated genes. Here fasL (Fas ligand) gene was found to be up regulated, this gene shows gene ontology of cytokine activity, death receptor binding functions, also activates IL-6 (interleukin 6) which allows cytokine activity and T-helper 17 cell lineage commitment; similarly, wnt membrane protein expression also up regulated here which is responsible for wnt signalling and also direct cell response to transcription. Down regulation of IL2RA gene was expressed that was responsible for Ras guanyl-nucleotide exchange factor activity and drug binding, also under expression of DRD5 was observed and it was found to be responsible for dopamine neurotransmitter receptor activity on gene ontology. **Figure 32** shows microarray analysis plots for this group.

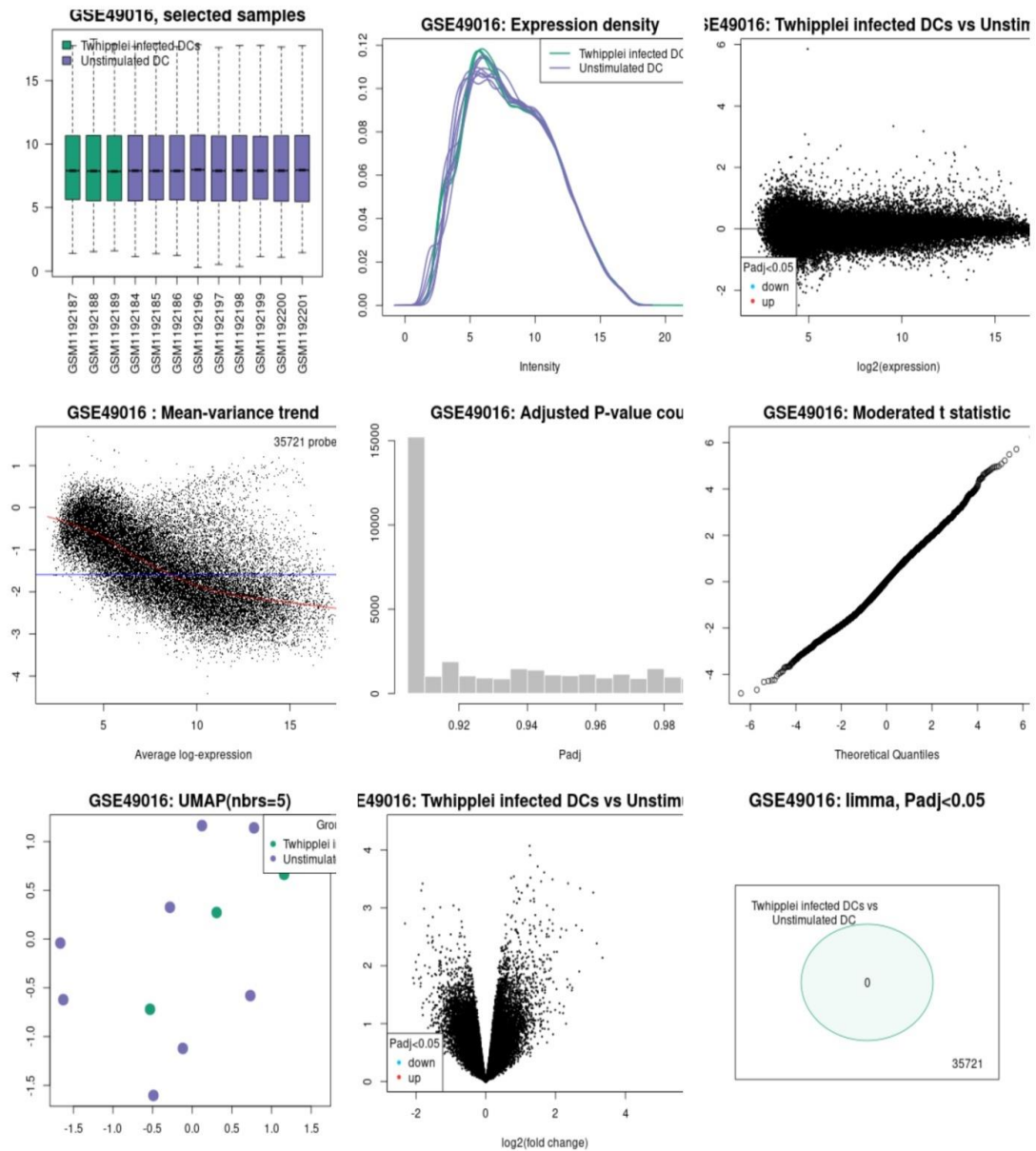


Figure 32. Microarray analysis plots for group *T. whipplei* infected DC's Vs Unstimulated DC's

H. GSE102862:

In this series a very recent study on IRF-4 haploinsufficiency in the PBMC (Peripheral blood mononuclear cells) cells derived from French patients and healthy wild type IRF-4

control individuals was compared. Here we designed core 6 groups within each group 3 subgroups were designed for microarray analysis by deploying Geo2R tool. As, in this experimental study PBMC from 6 IRF-4 haplo insufficient individuals was considered which includes three patients (P1 to P3) and three healthy individuals (HET1 to HET3), also in this study six IRF-4 wild type homozygous individuals were considered which includes 4 healthy wild types WT1-WT4 along with 2 unrelated controls (Control1 and Control 2). This study suggests IRF-4 deficiency or loss of function like R98W mutation in IRF-4 can cause Whipple's disease with age dependent incomplete penetrance. The up regulated and down regulated for such data was provided here.

1. Group1 (Patient 1,2,3 Vs Control1)

1a. BCG Treated:

In this analysis we found 1404 down regulated genes while 398 up regulated genes. **Figure 33** shows microarray analysis plots for this group.

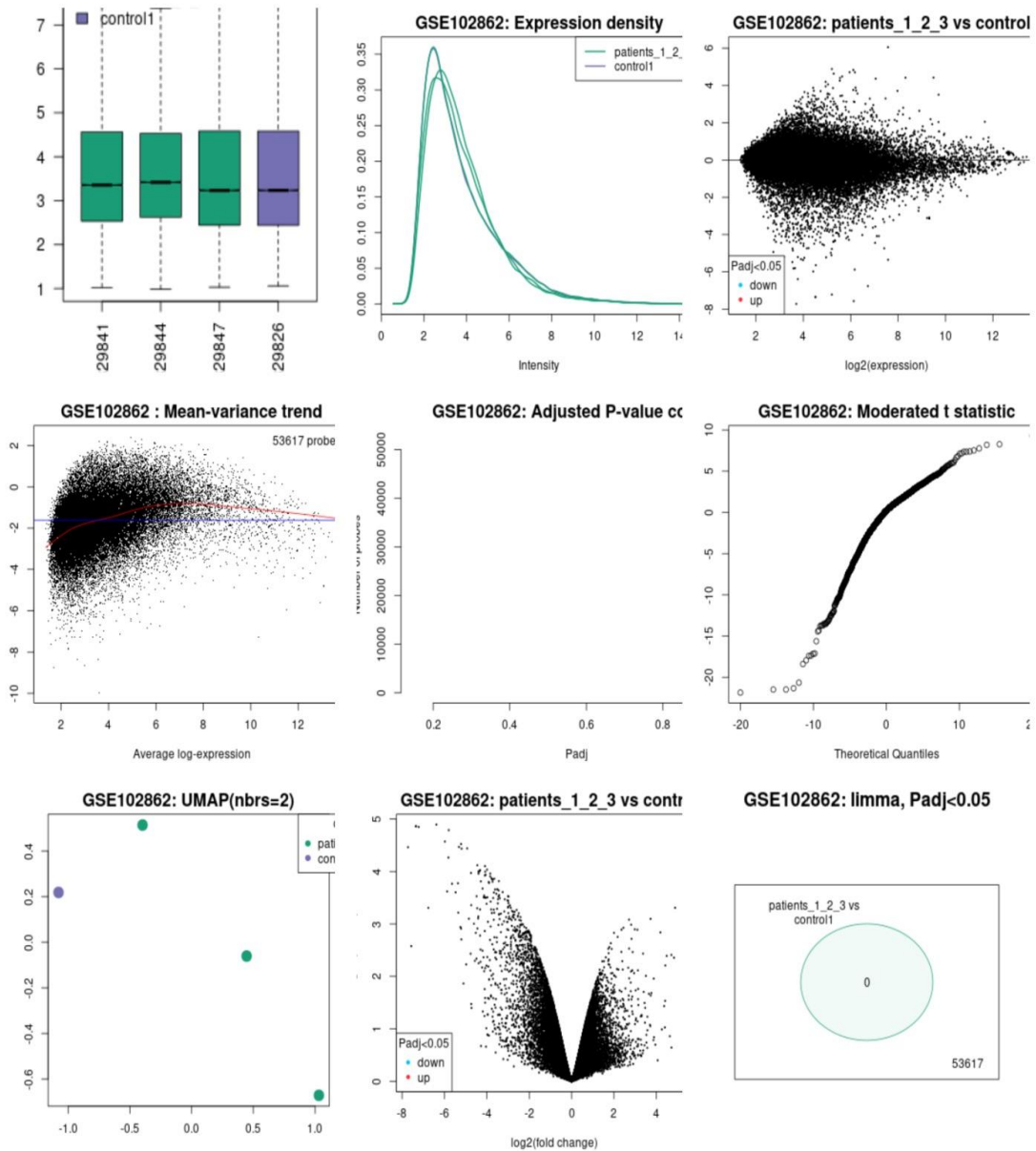


Figure 33. Microarray analysis plots for BCG Treated group (Patient 1,2,3 Vs Control1)

1b. Non-stimulated:

In this analysis we found 1039 down regulated genes while 244 up regulated genes.

Figure 34 shows microarray analysis plots for this group.

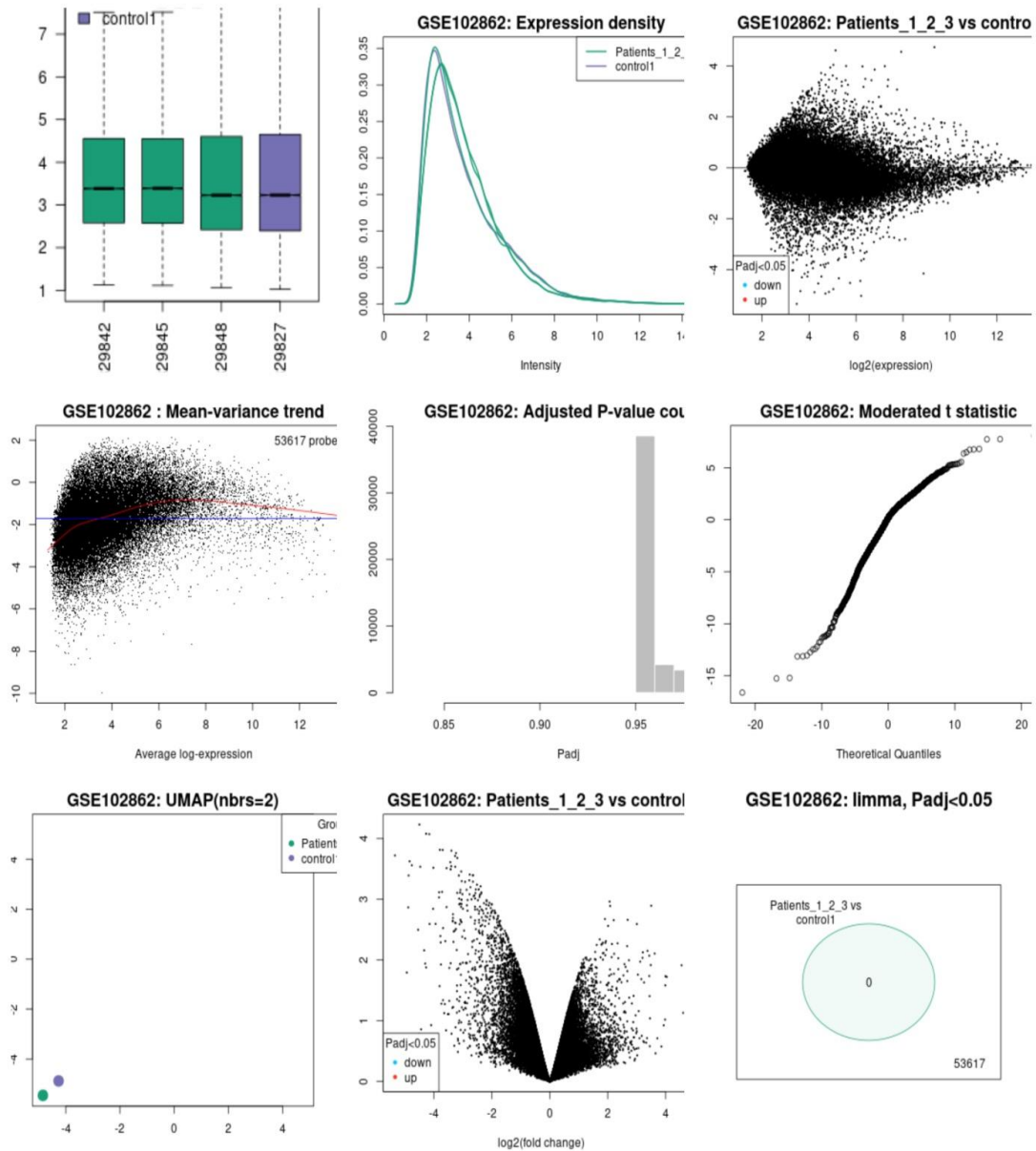


Figure 34. Microarray analysis plots for Non-stimulated group (Patient 1,2,3 Vs Control1)

1c. *Tropheryma whipplei* infected

In this analysis we found 965 down regulated genes while 334 up regulated genes.

Figure 35 shows microarray analysis plots for this group.

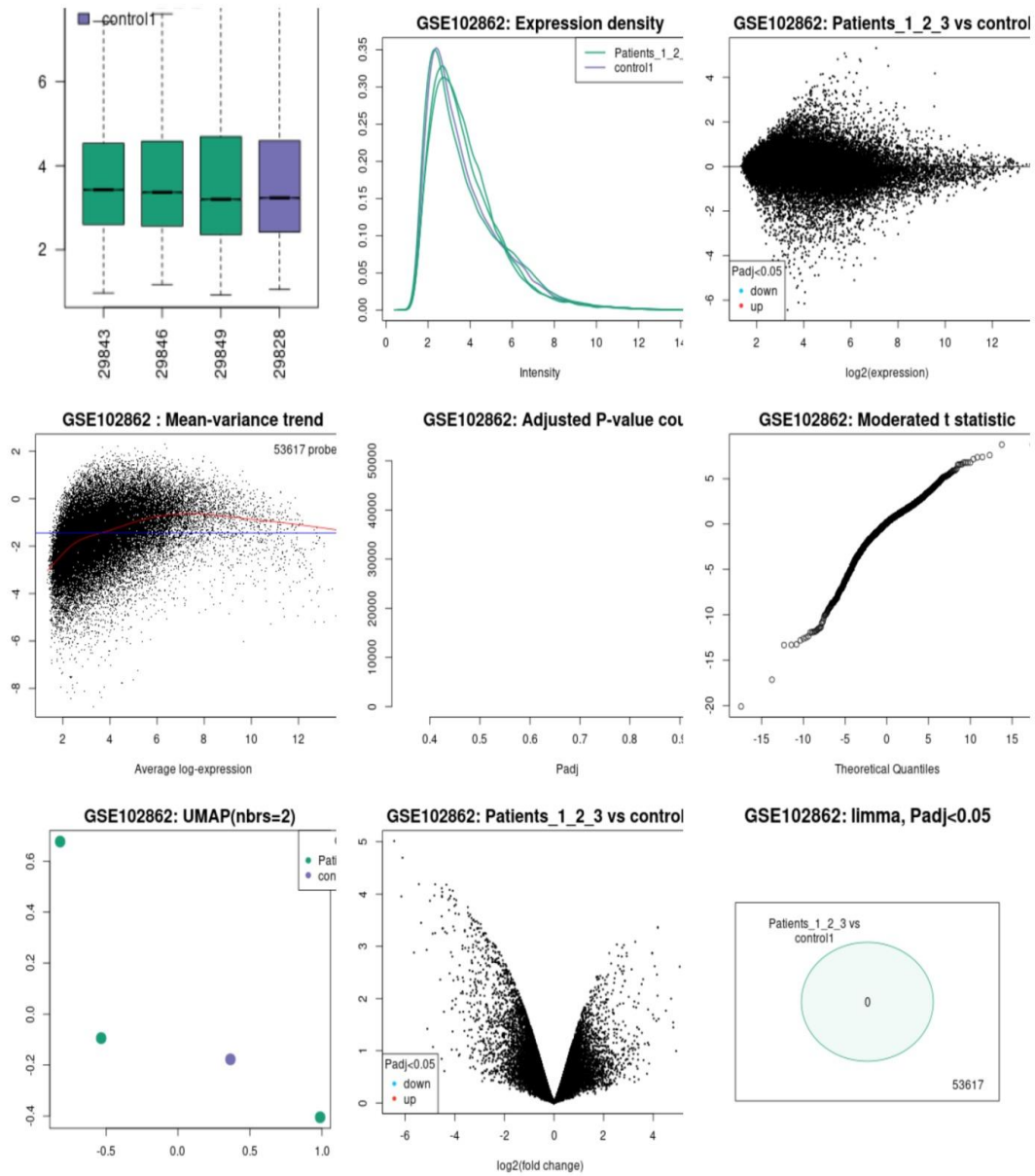


Figure 35. Microarray analysis plots for *Tropheryma whipplei* infected group (Patient 1,2,3 Vs Control1)

2. Group2 (HET1,2,3 Vs Control1)

2a. BCG Treated:

In this analysis we found 1177 down regulated genes while 590 up regulated genes. **Figure 36** shows microarray analysis plots for this group.

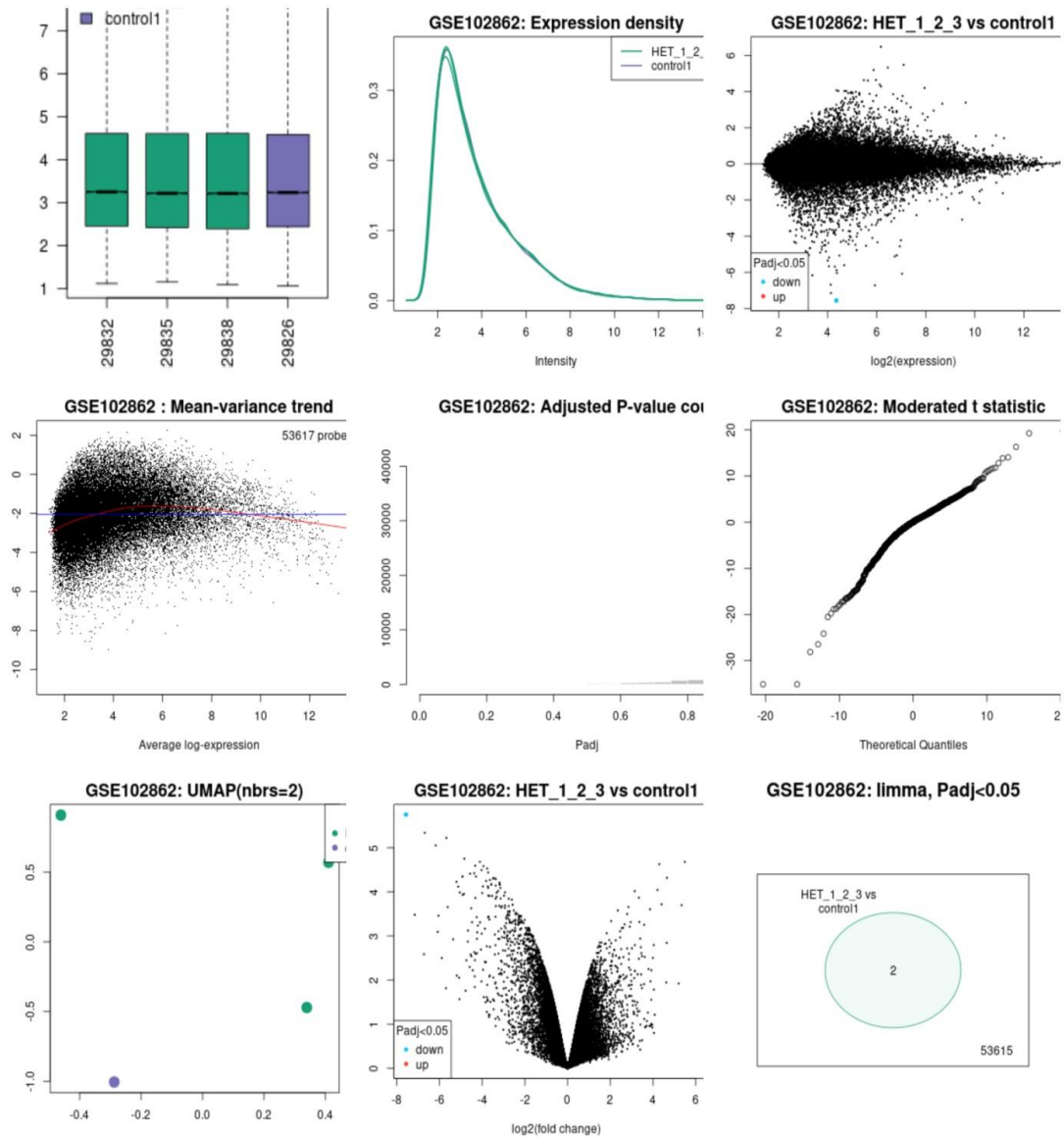


Figure 36. Microarray analysis plots for BCG Treated group (HET1,2,3 Vs Control1)

2b. Non-stimulated:

In this analysis we found 837 down regulated genes while 234 up regulated genes. **Figure 37** shows microarray analysis plots for this group.

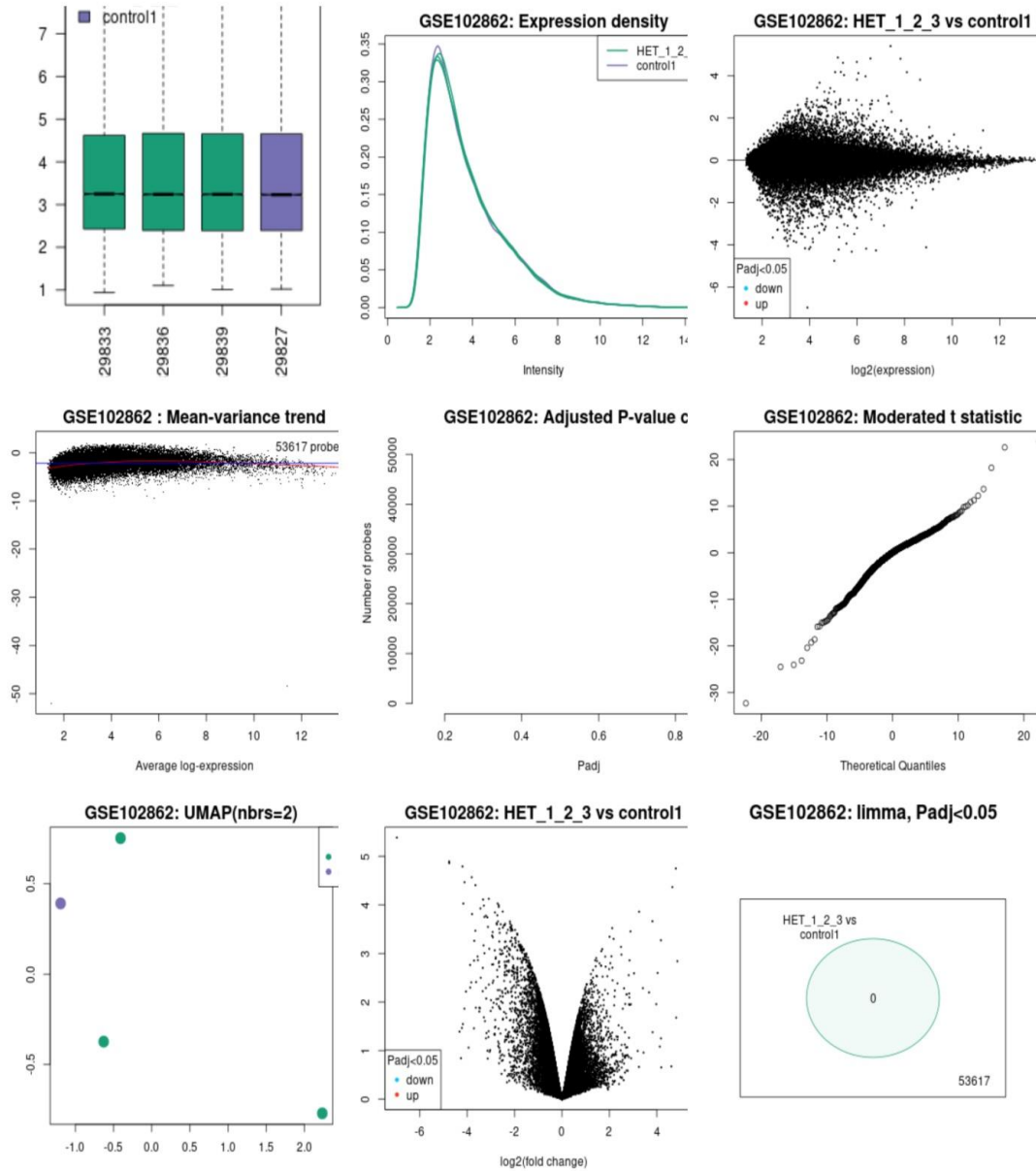


Figure 37. Microarray analysis plots for Non-stimulated group (HET1,2,3 Vs Control1)

2c. *Tropheryma whipplei* infected

In this analysis we found 1156 down regulated genes while 700 up regulated genes. **Figure 38** shows microarray analysis plots for this group.

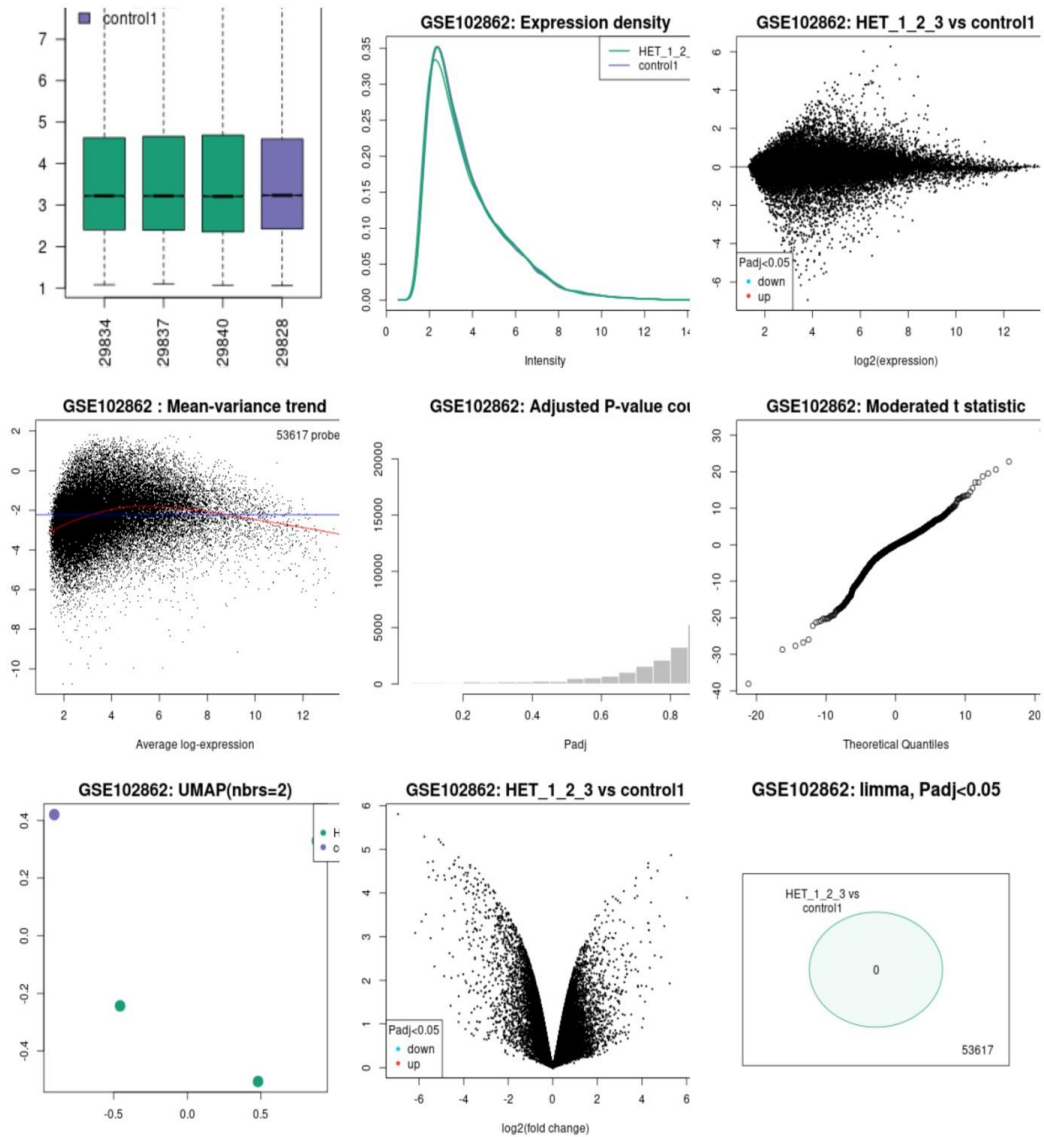


Figure 38. Microarray analysis plots for *Tropheryma whipplei* infected group (HET1,2,3 Vs Control1)

3. Group3 (WT1,2,3,4 Vs Control1)

3a. BCG Treated

In this analysis we found 961 down regulated genes while 184 up regulated genes.

Figure 39 shows microarray analysis plots for this group.

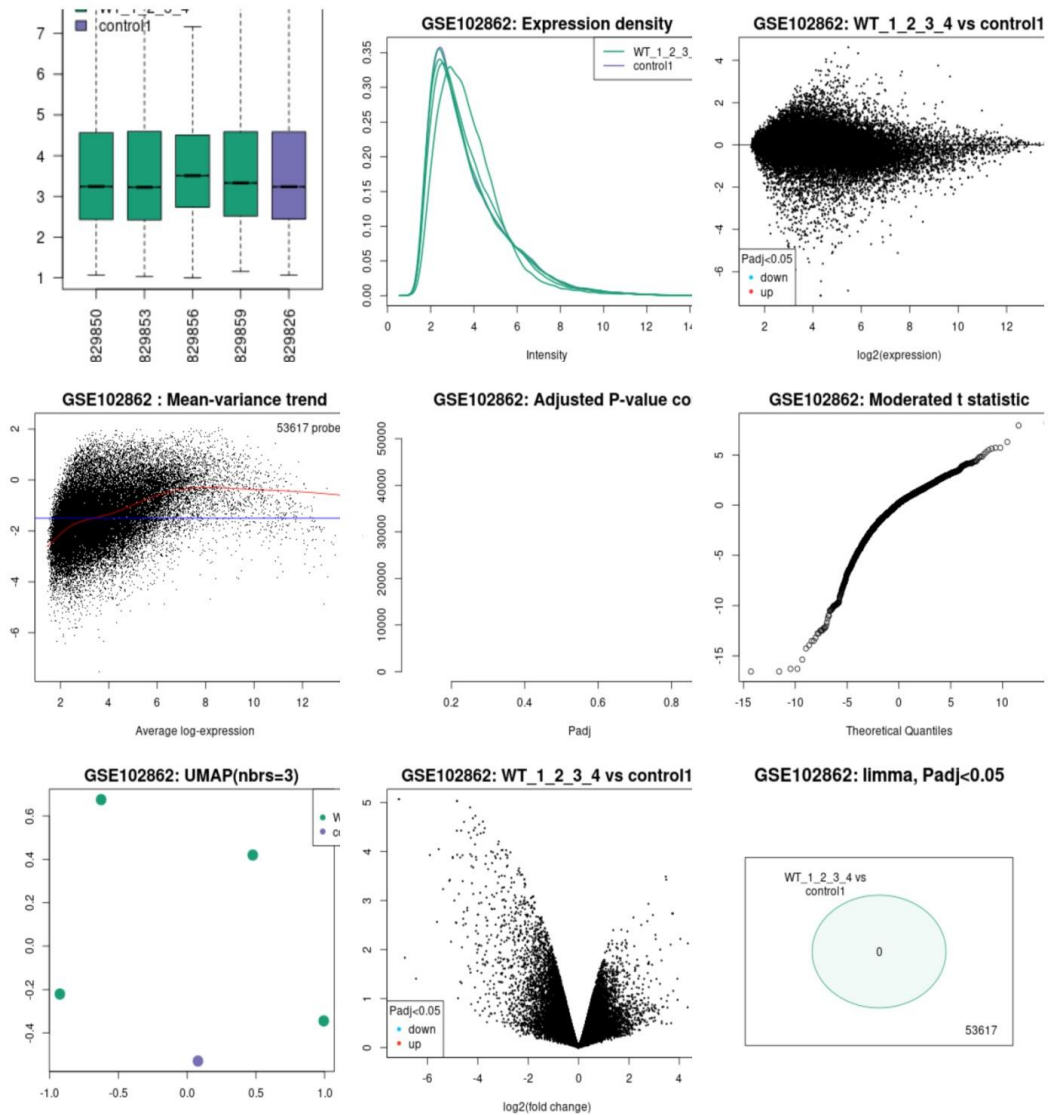


Figure 39. Microarray analysis plots for BCG Treated group (WT1,2,3,4 Vs Control1)

3b. Non-stimulated:

In this analysis we found 928 down regulated genes while 156 up regulated genes. **Figure 40** shows microarray analysis plots for this group.

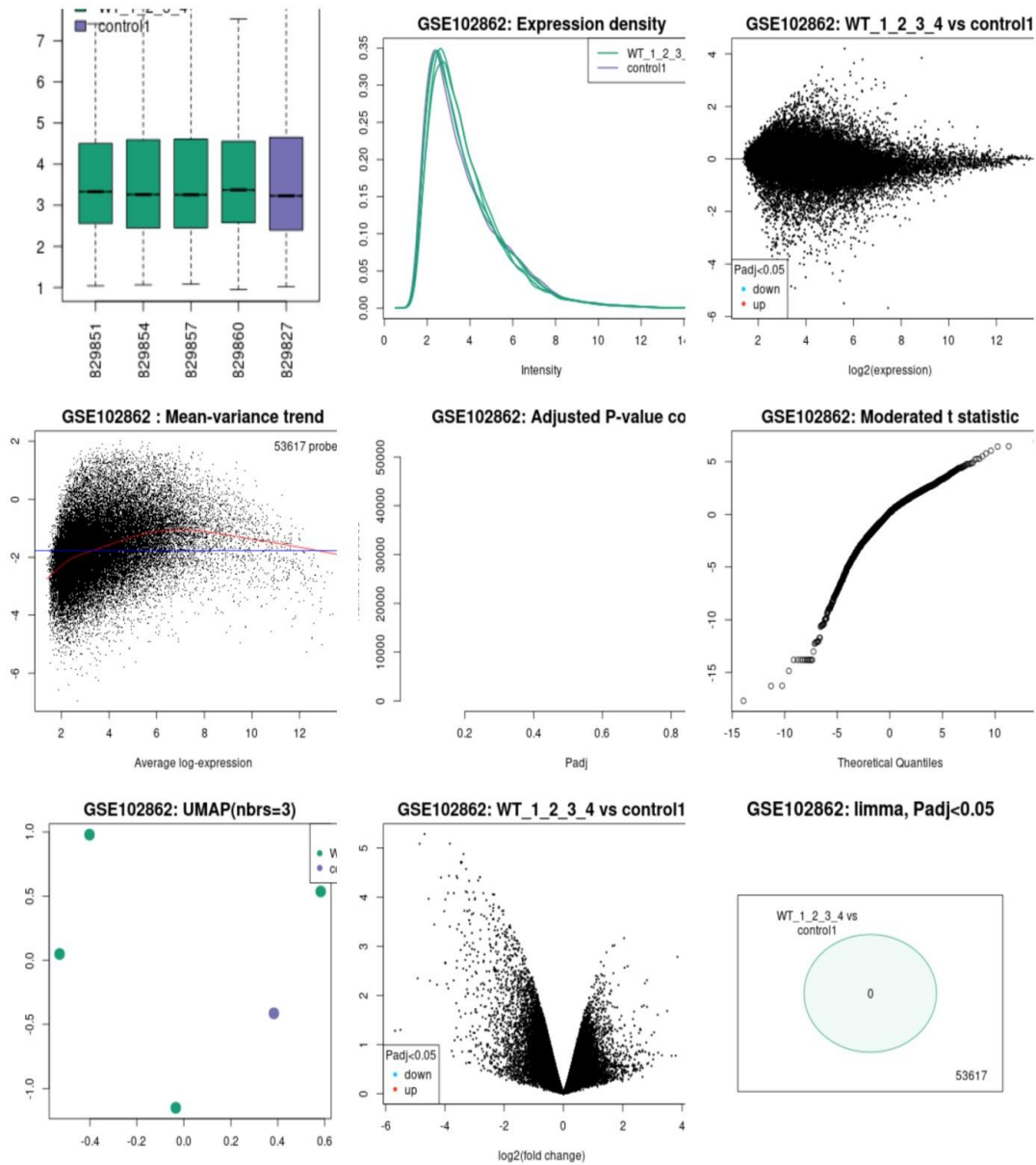


Figure 40. Microarray analysis plots for Non-stimulated group (WT1,2,3,4 Vs Control1)

3c. *Tropheryma whipplei* infected

In this analysis we found 975 down regulated genes while 240 up regulated genes. **Figure 41** shows microarray analysis plots for this group.

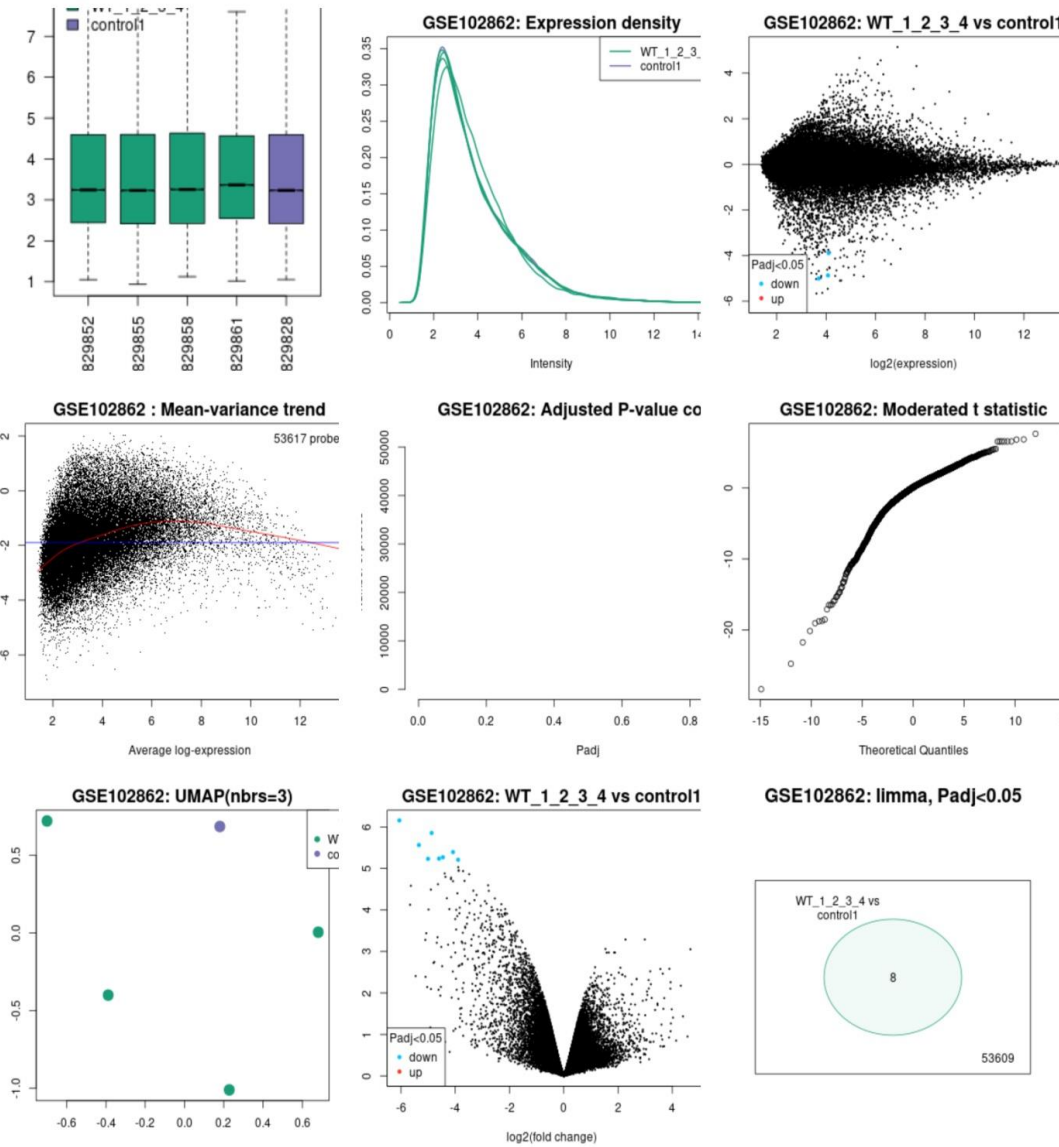


Figure 41. Microarray analysis plots for *Tropheryma whipplei* infected group (WT1,2,3,4 Vs Control1)

4. Group4 (Patient 1,2,3 Vs Control2)

4a. BCG Treated:

In this analysis we found 1459 down regulated genes while 409 up regulated genes.

Figure 42 shows microarray analysis plots for this group.

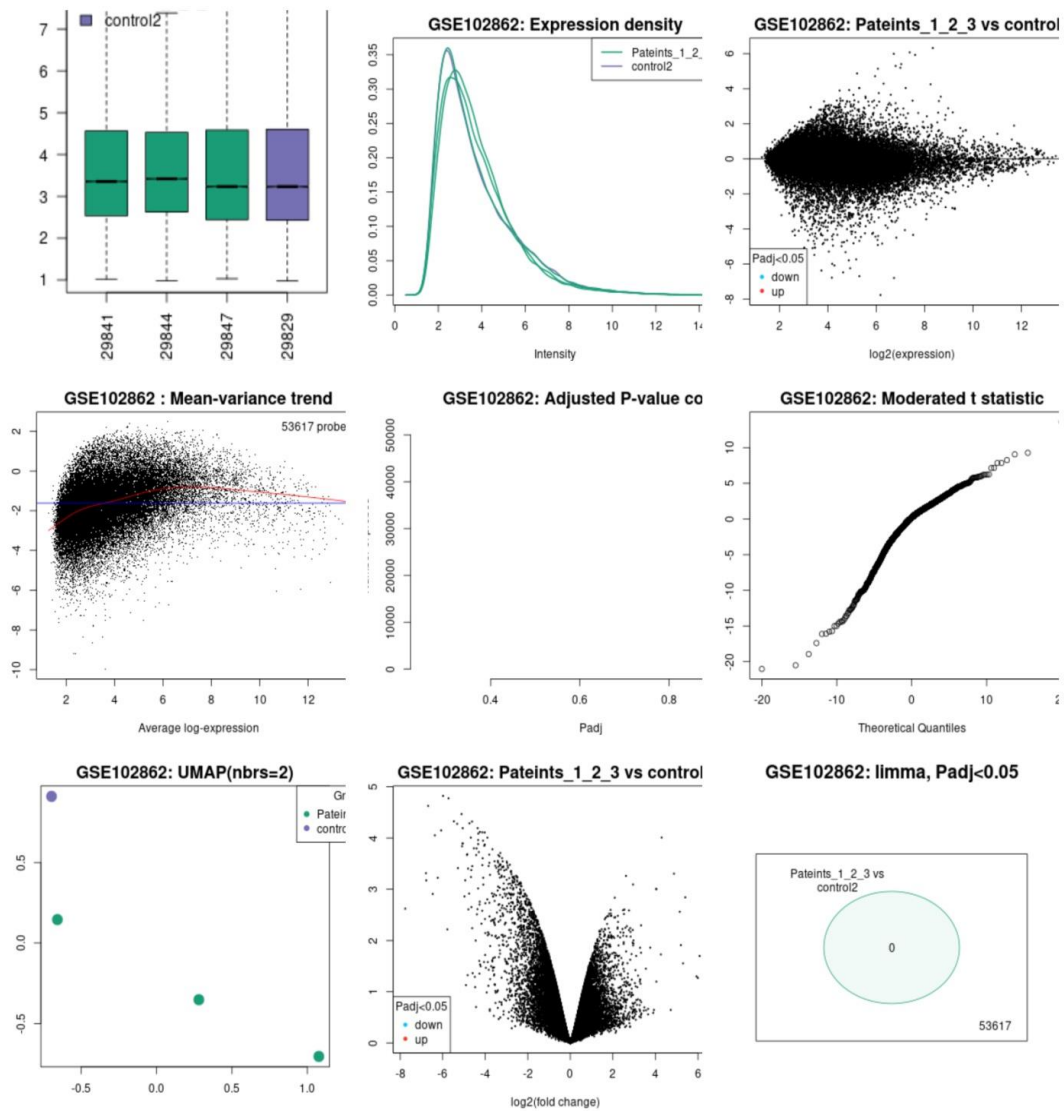


Figure 42. Microarray analysis plots for BCG treated group (Patient 1,2,3 Vs Control2)

4b. Non-stimulated:

In this analysis we found 1342 down regulated genes while 260 up regulated genes. **Figure 43** shows microarray analysis plots for this group.

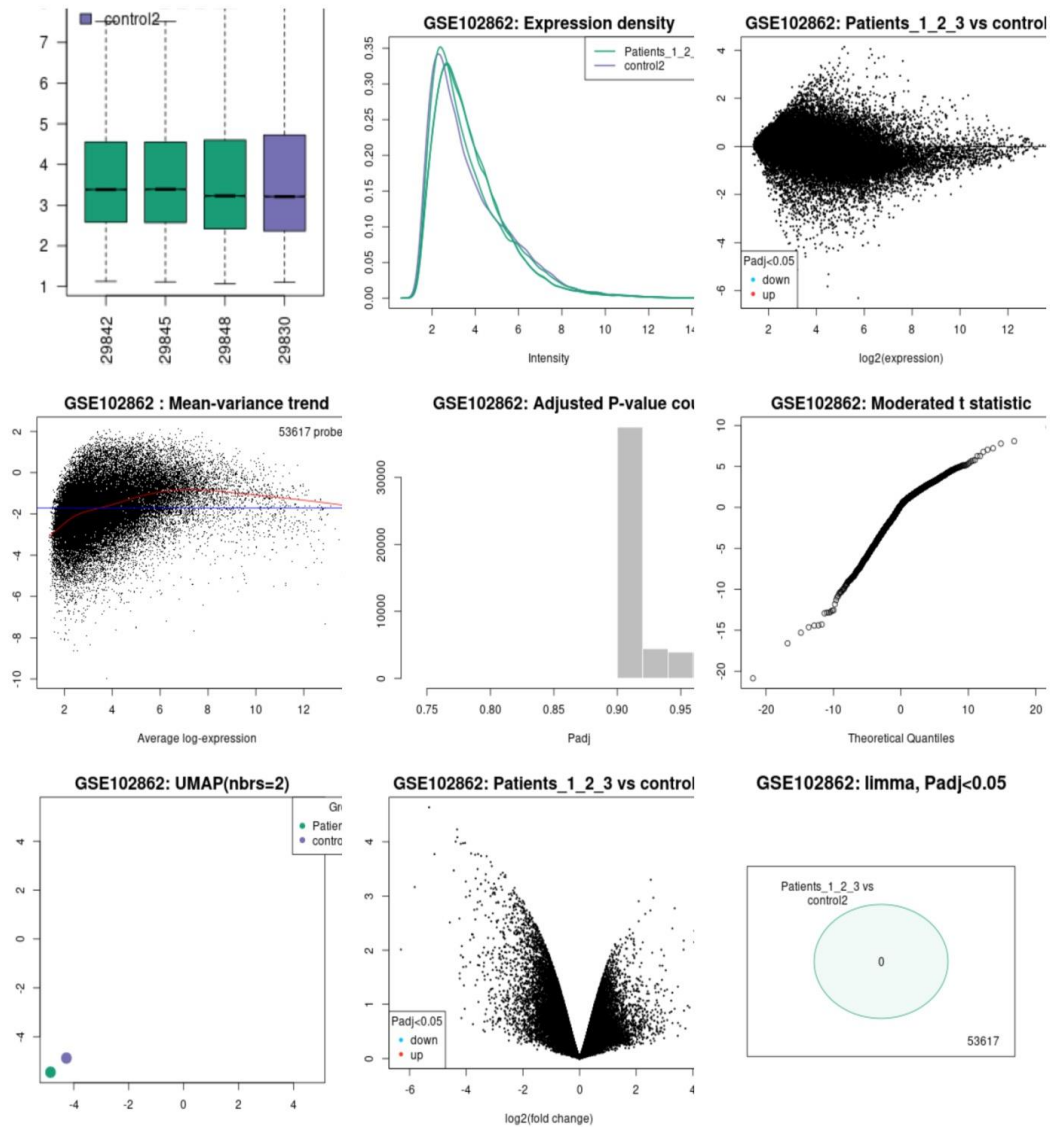


Figure 43. Microarray analysis plots for Non-stimulated group (Patient 1,2,3 Vs Control2)

4c. *Tropheryma whipplei* infected

In this analysis we found 1139 down regulated genes while 409 up regulated genes. **Figure 44** shows microarray analysis plots for this group.

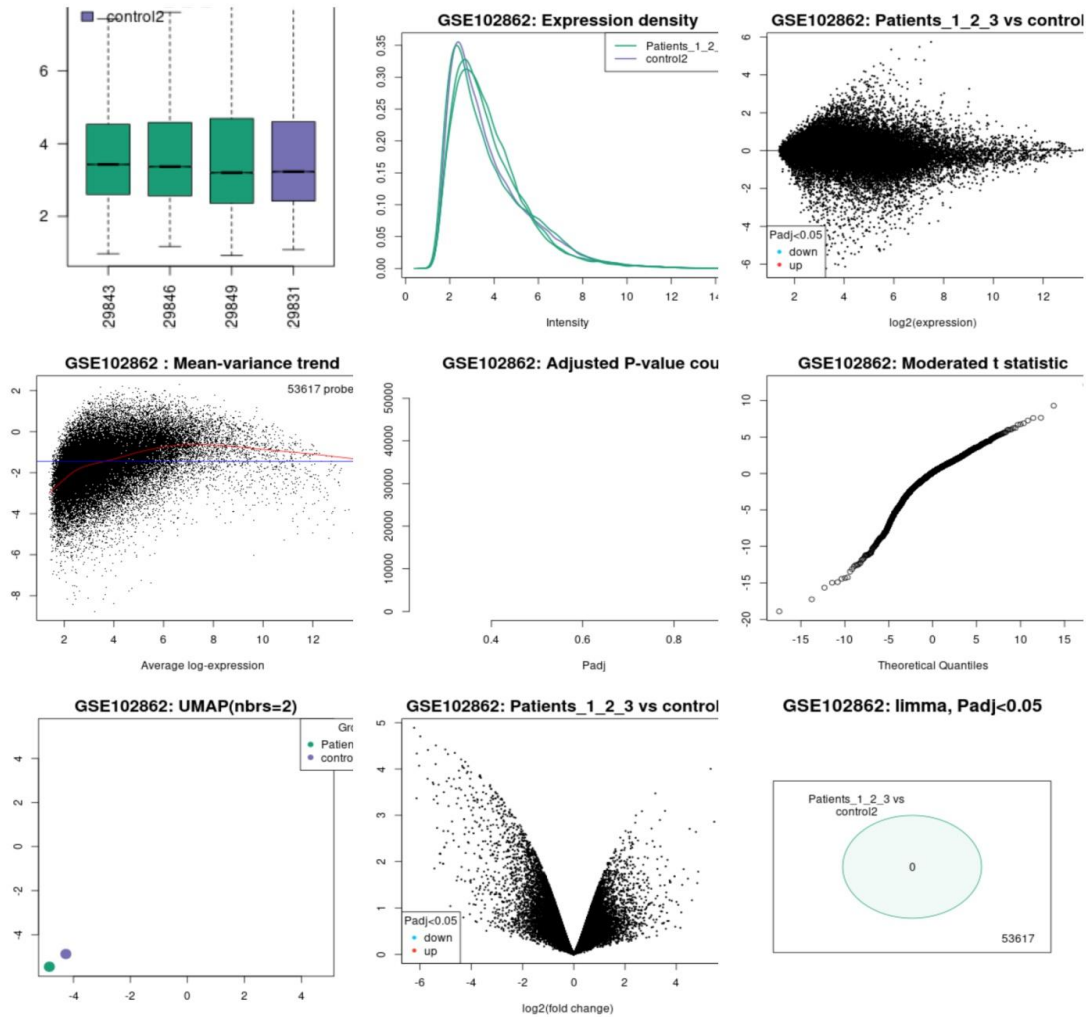


Figure 44. Microarray analysis plots for *Tropheryma whipplei* infected group (Patient 1,2,3 Vs Control2)

5. Group5 (HET1,2,3 Vs Control2)

5a. BCG Treated:

In this analysis we found 1199 down regulated genes while 519 up regulated genes. **Figure 45** shows microarray analysis plots for this group.

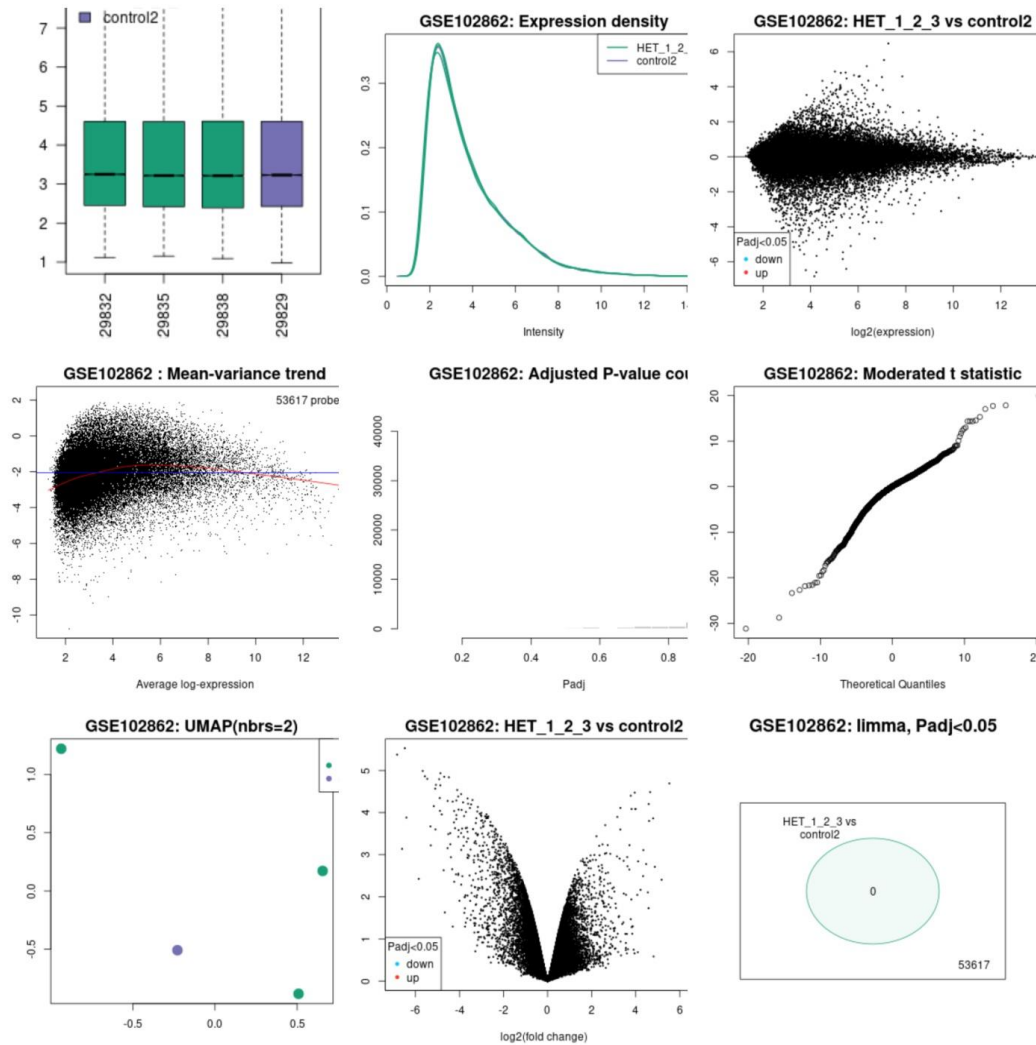


Figure 45. Microarray analysis plots for BCG Treated group (HET1,2,3 Vs Control2)

5b. Non-stimulated:

In this analysis we found 883 down regulated genes while 203 up regulated genes. **Figure 46** shows microarray analysis plots for this group.

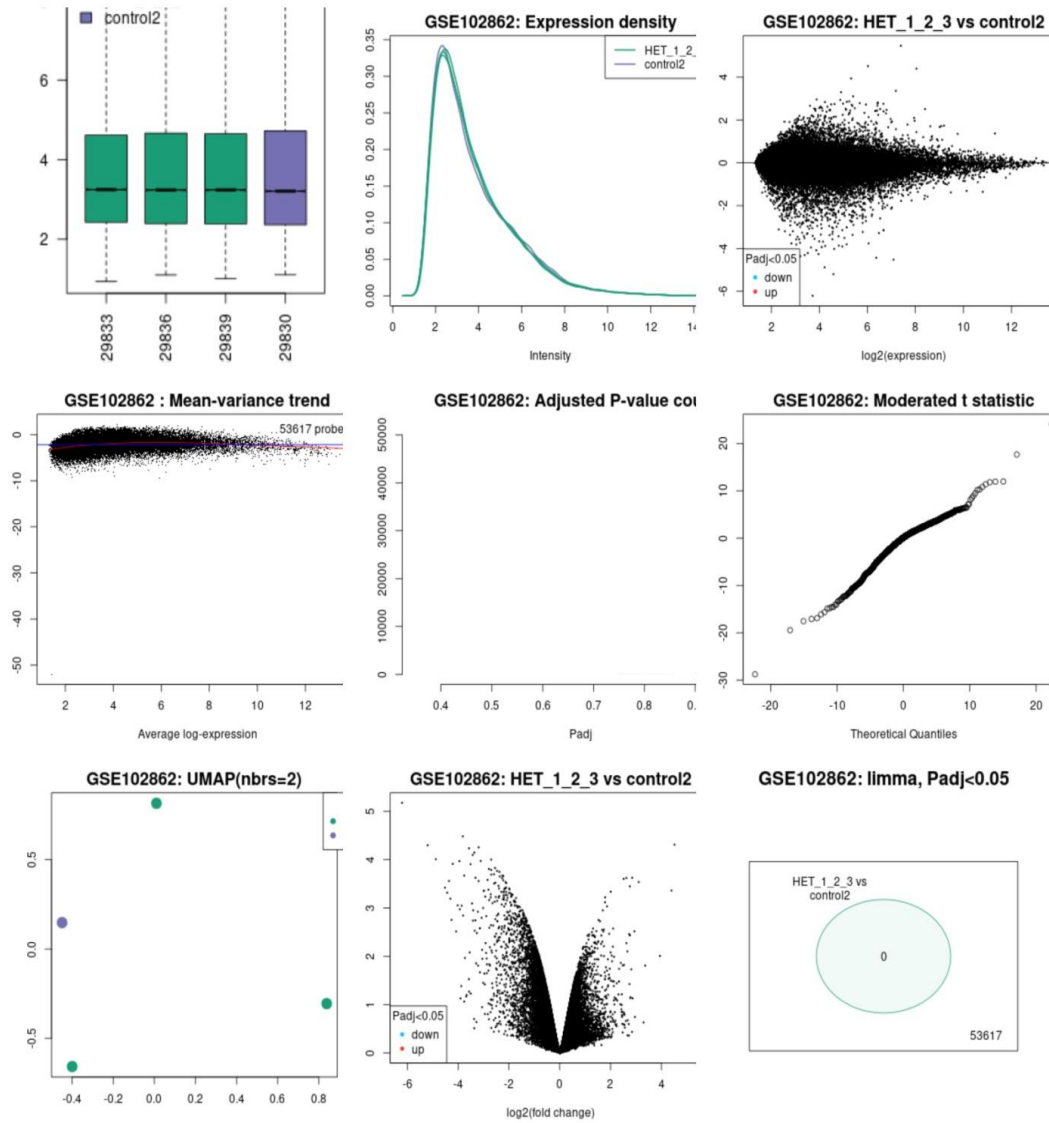


Figure 46. Microarray analysis plots for Non-stimulated group (HET1,2,3 Vs Control2)

5c. *Tropheryma whipplei* infected

In this analysis we found 1208 down regulated genes while 672 up regulated genes. **Figure 47** shows microarray analysis plots for this group.

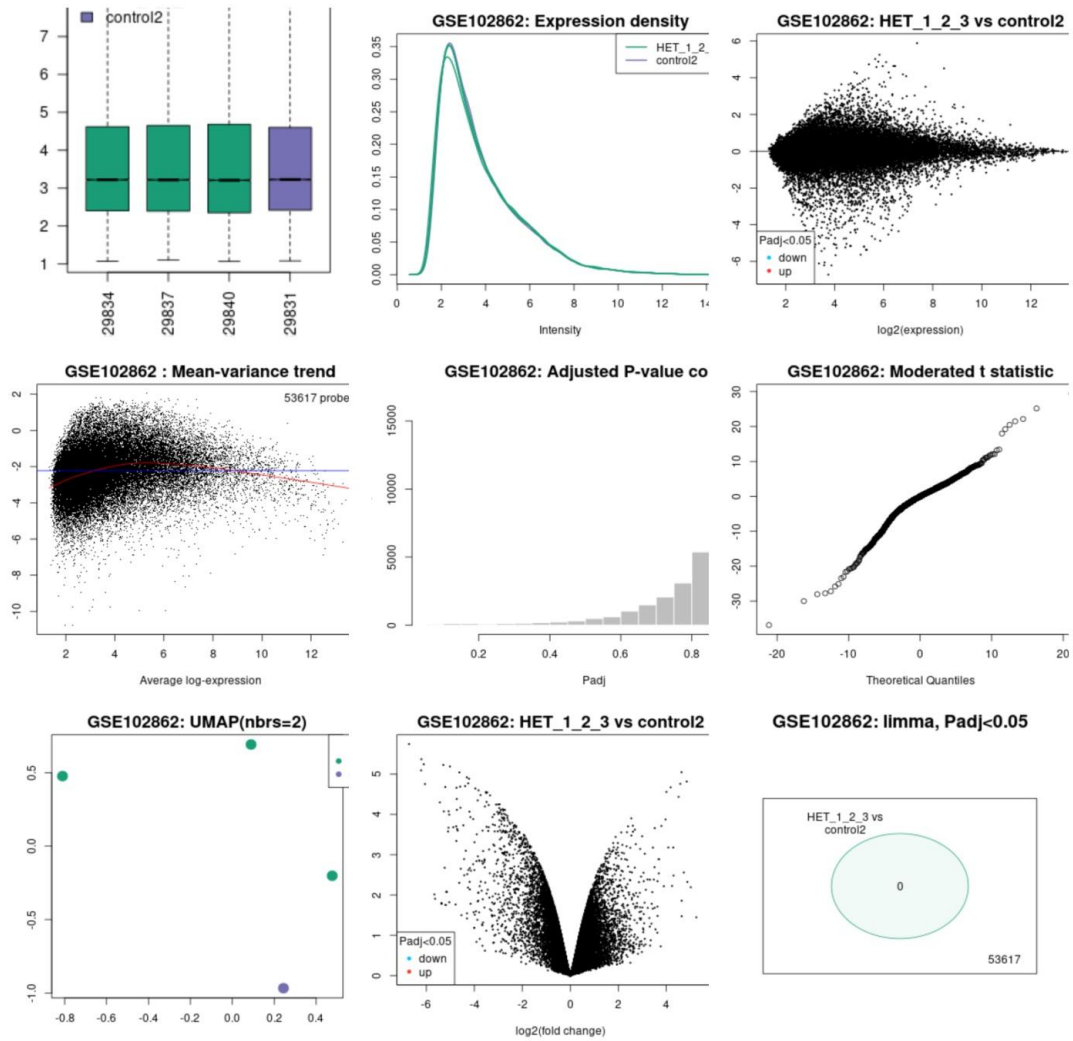


Figure 47. Microarray analysis plots for *Tropheryma whipplei* infected group (HET1,2,3 Vs Control2)

6. Group6 (WT1,2,3,4 Vs Control2)

6a. BCG Treated

In this analysis we found 998 down regulated genes while 180 up regulated genes. **Figure 48** shows microarray analysis plots for this group.

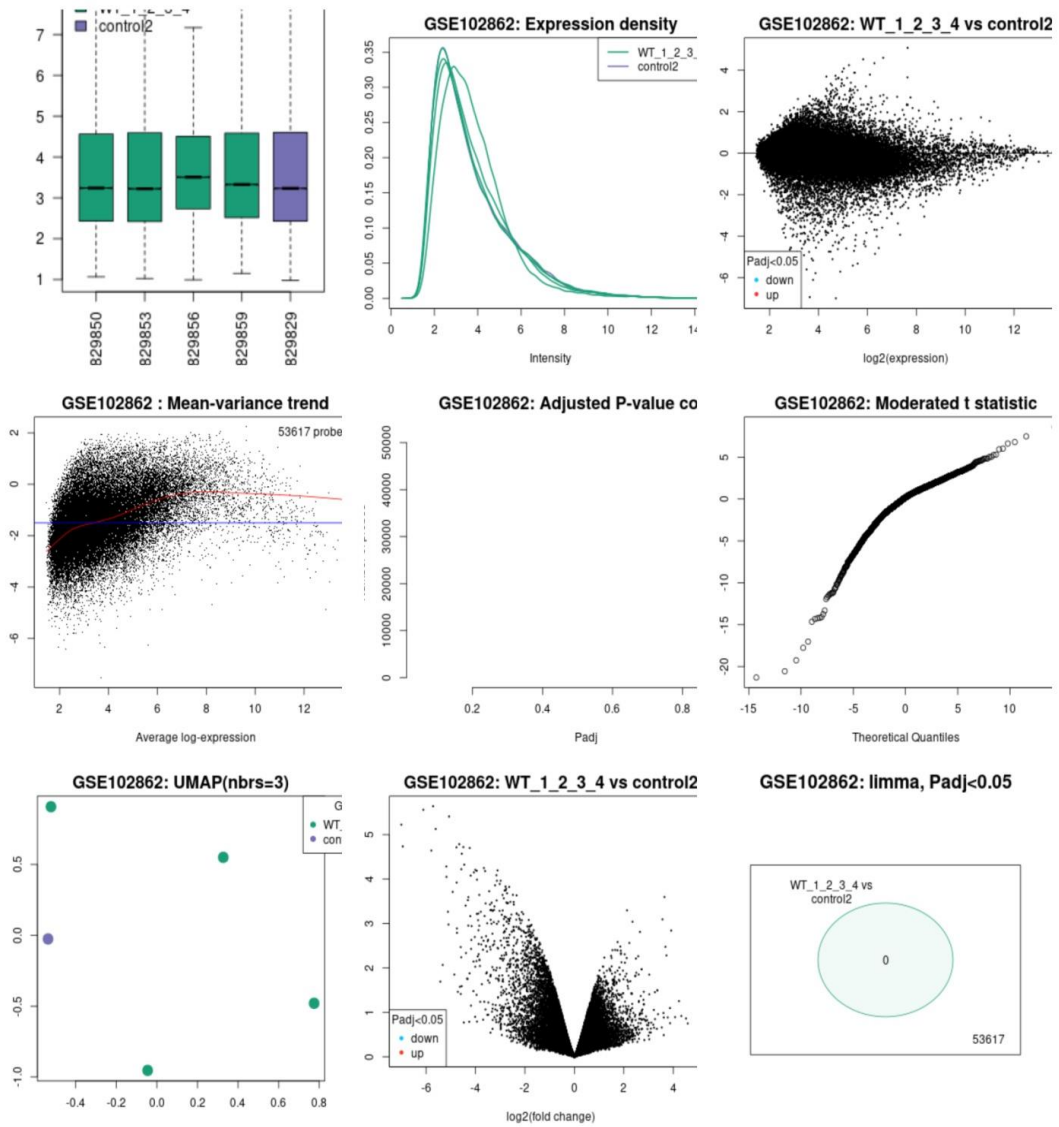


Figure 48. Microarray analysis plots for BCG Treated group (WT1,2,3,4 Vs Control2)

6b. Non-stimulated:

In this analysis we found 1144 down regulated genes while 184 up regulated genes. **Figure 49** shows microarray analysis plots for this group.

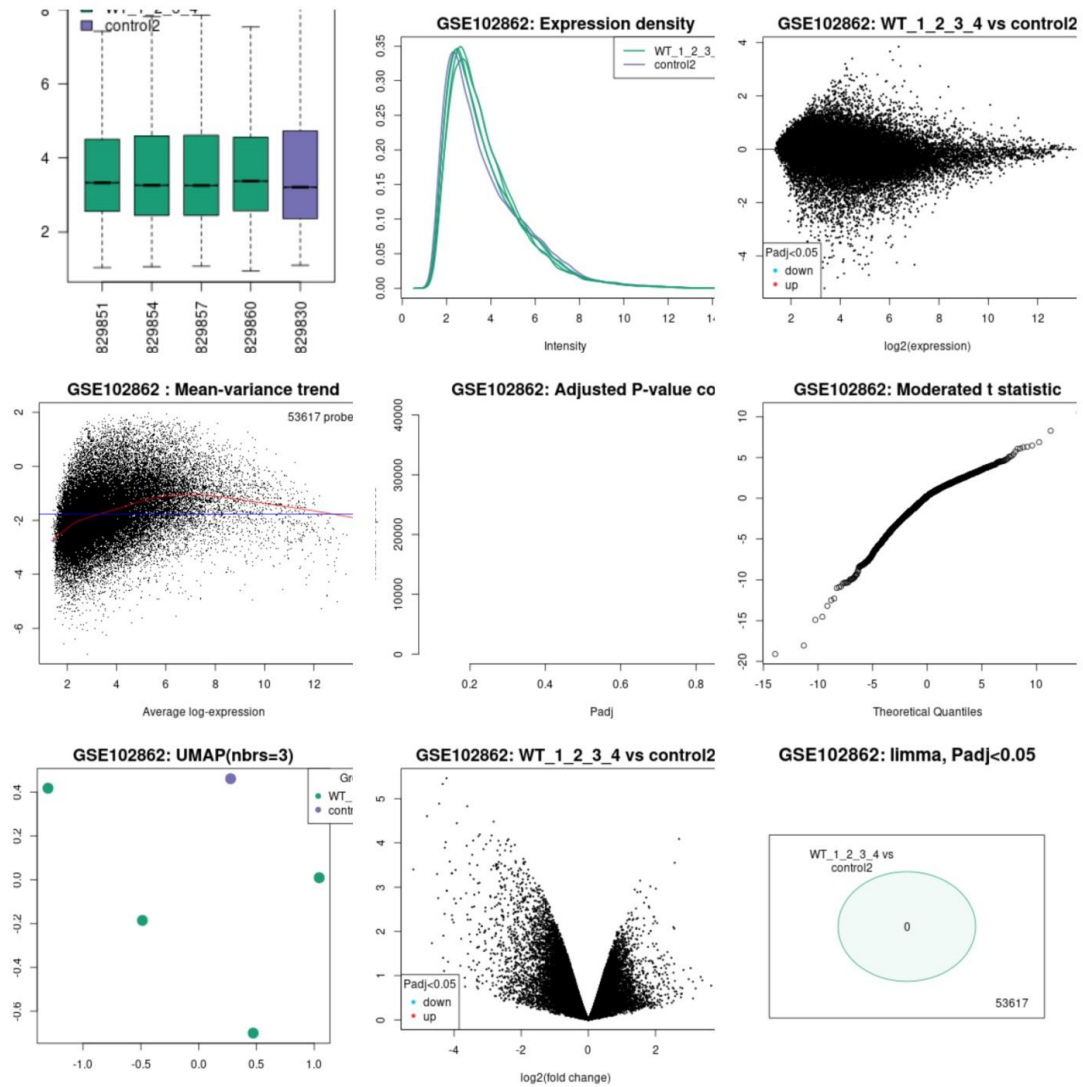


Figure 49. Microarray analysis plots for Non-stimulated group (WT1,2,3,4 Vs Control2)

6c. *Tropheryma whipplei* infected

In this analysis we found 1176 down regulated genes while 238 up regulated genes. **Figure 50** shows microarray analysis plots for this group.

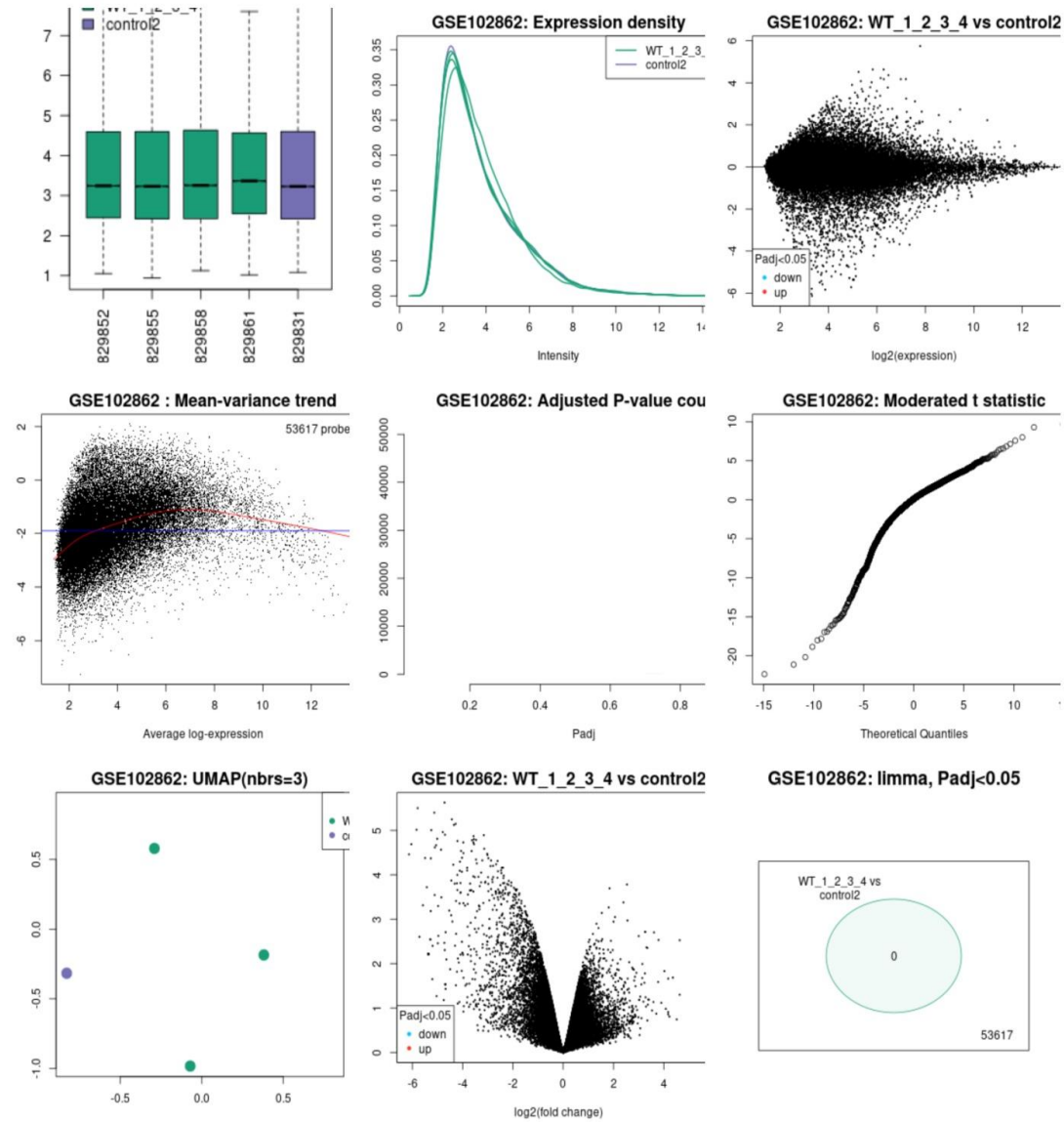


Figure 50. Microarray analysis plots for *Tropheryma whipplei* infected group (WT1,2,3,4 Vs Control2)

5.2 Codon usage and Amino acid usage patterns in *Tropheryma whipplei*

The sequences of various strains of *T. whipplei* are retrieved from the NCBI website and then the BLAST-N tool is deployed to analyze the similarities between available genome sets for various bacterial strains. This genomic blast produces a dendrogram (figure 51). This depicts the phylogenetic relationship between all the considered strains of *T. whipplei*.

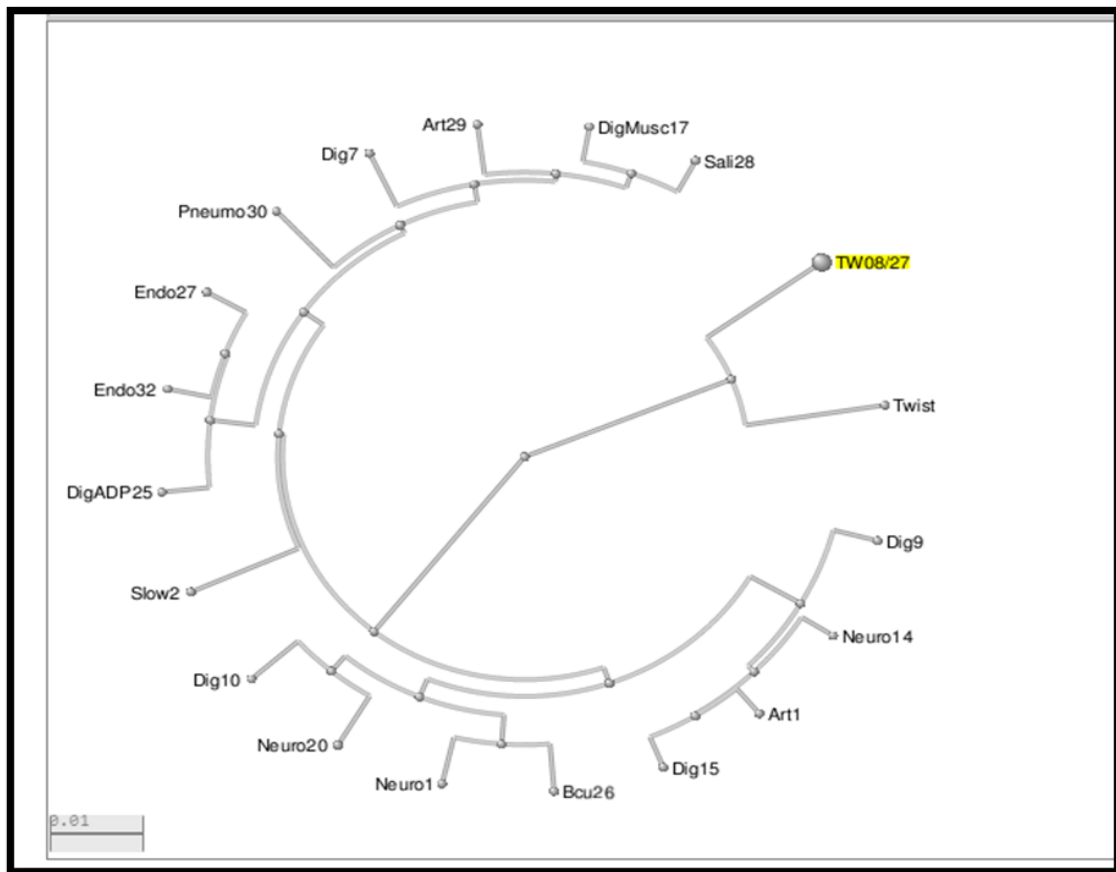


Figure 51. Dendrogram based on genomic blast

The above Dendrogram shows a relevant relationship between different strains of *T. Whipple* bacterium (<https://www.ncbi.nlm.nih.gov/genome/?term=tropheryma+whipplei>). It is based on genomic comparisons to analyze phylogenetic relationships between various strains of the selected bacterium.

In **Figure 52** BLAST (Basic local search alignment tool) of NCBI is utilized for comparing nucleotide sequences of *T. whipplei* Twist and TW 08/27 strains as their complete genomes are available. It is useful in predicting that Twist strain and TW08/27; both strains have greater level of similarity from 408-555 K base pair regions in their genome sets. Score per cent identity in **Figure 53**, that indicates 99.66 per cent similarity in genomic analysis both considered organisms.

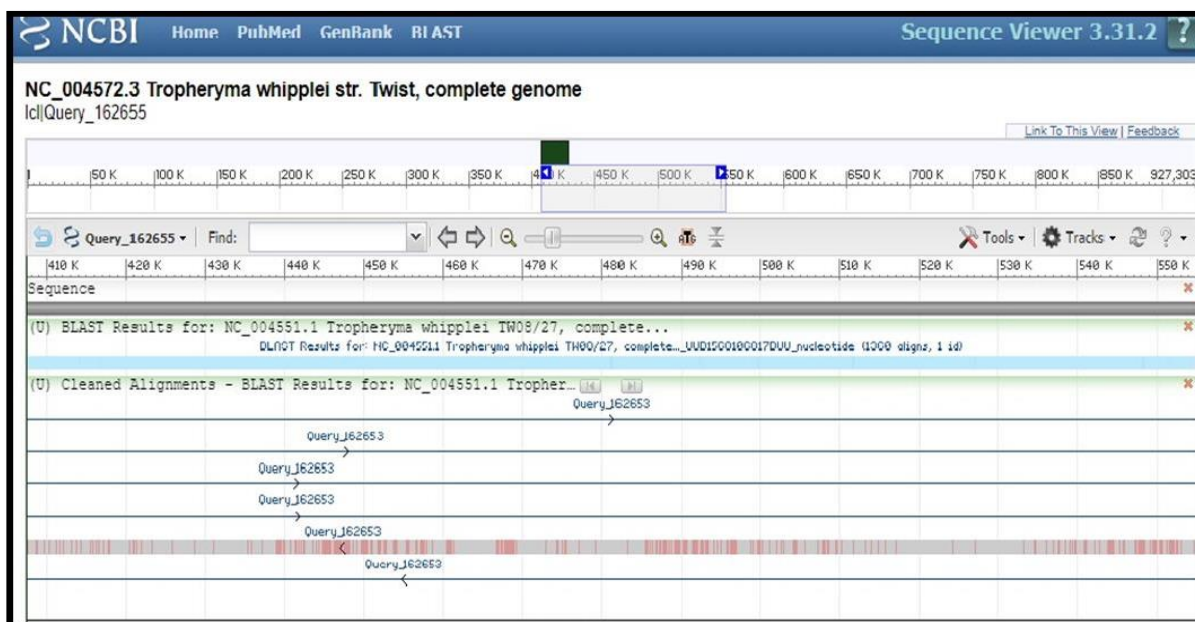


Figure 52. BLAST results for Twist and TW 08/27 Strains

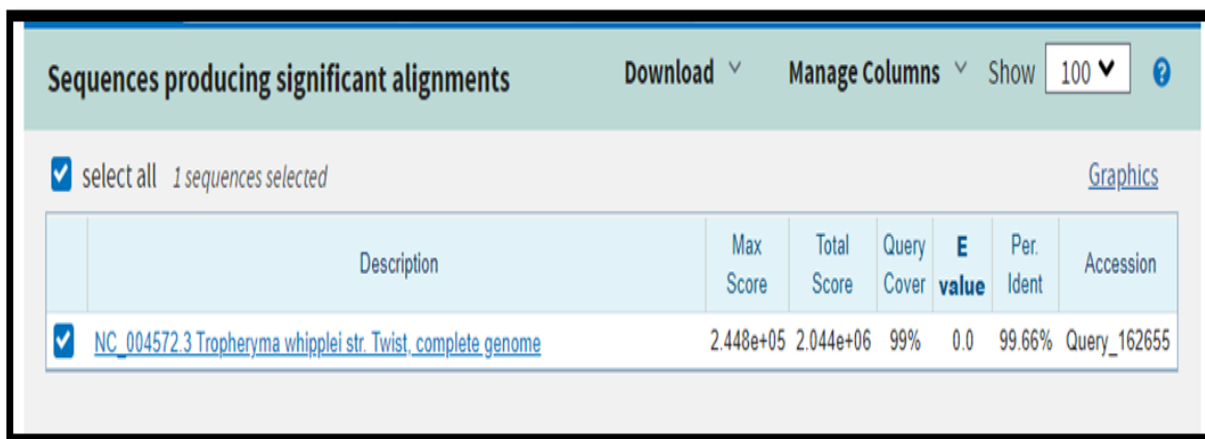


Figure 53. BLAST score percent identity for Twist and TW08/27 strains

This also indicates similarity in proteomic sets that can be deployed in achieving the crucial task of vaccine designing by reverse vaccinology approach. In simple words, one can easily depict that the current study will circumscribe the welfare domain of human health.

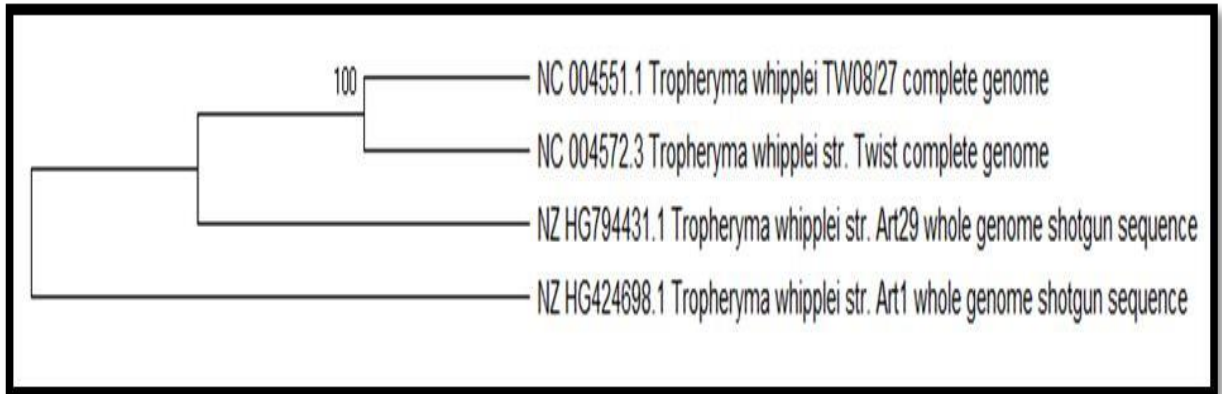


Figure 54. Results of Phylogenetic analysis among 4 strains of *Tropheryma whipplei*

In **Figure 54** the evolutionary antiquity was analyzed by deploying the UPGMA approach. The optimal tree with the sum of branch length = 1.4633 is presented. The transformative gaps were determined by deploying the p-distance approach and it is in the number of amino acids unit variability persite. This investigation ramified 4 amino acid stretches. The respective stances with site indemnity of less than 95 percent were waived out, i.e., aligned or calibrated gaps of less than 5 percent, omitted data, and dubious bases were acquiesced at random position. Concluding datasets contain 925938 positions out of total locales under consideration. Evolution based transformative investigation were operated in MEGA X program. In **figure 55** TW08/27 and Twist coding sequences on Ist and IInd major axes were procreated by correspondence analysis on RSCU parametrical findings. The primary or first axis accounts for 9.21% and 9.18%, whereas secondary axis accounts for 5.13% and 5.08% of total variable alterations in TW08/27 and Twist genomic patterns respectively. Above two plots show open quadrilaterals that represent genes transcribed in the leading strand while closed or dark quadrilaterals indicate gene transcribed in lagging strand. While in the bottom two plots

show dark or filled circles that represent transcribed genes in leading strand while open circles or unfilled ones indicate highly expressed genes in the leading strand. Such a scattered plot depicts the mutational biases which are strand-specific. Predicted sense strand sequences were found to be greater in leading strand (75.3% for TW 08/27 and 77% for TWIST strains genome) as compared to lagging strand. All such observations lead to conclude replication, as well as transcriptional selection together with asymmetric mutational bias, is the major cause for intragenomic variations in genomes of considered strains. This same figure also indicates most of the genes are clustered at the extreme end of the first axis in lower pictures, here strong correlation is predicted with leading strand sensed or coding regions. Also, low divergence at synonymous regions of highly expressed portions is inferred ($dS = 0.0057$), this helps in concluding synonymous codon selection occurs at the translational level also.

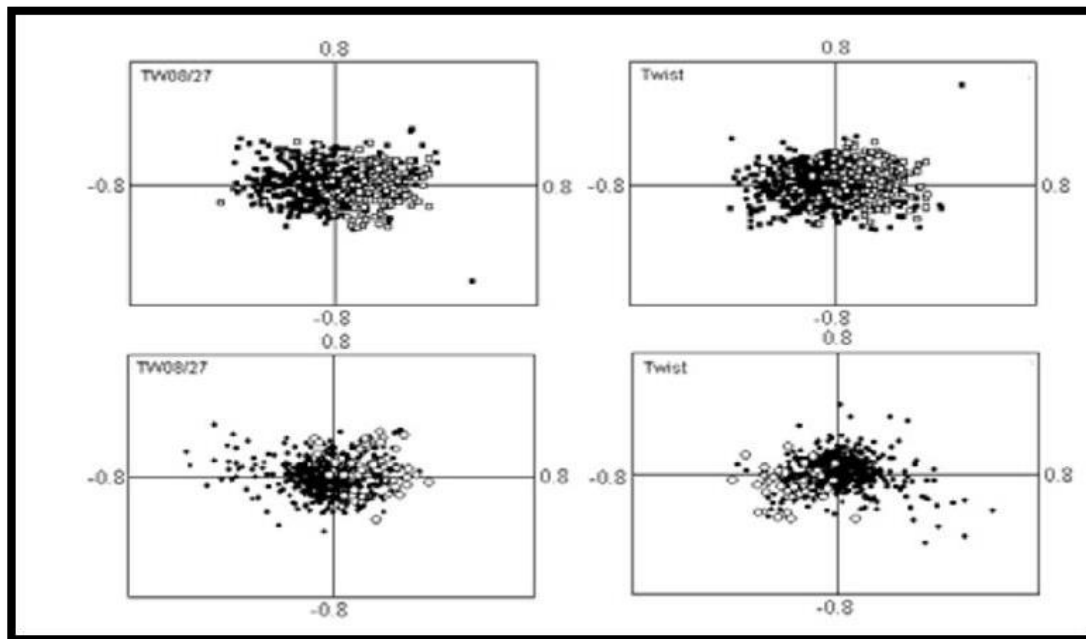


Figure 55. Correspondence study on synonymous transcriptome codon usage (RSCU values)

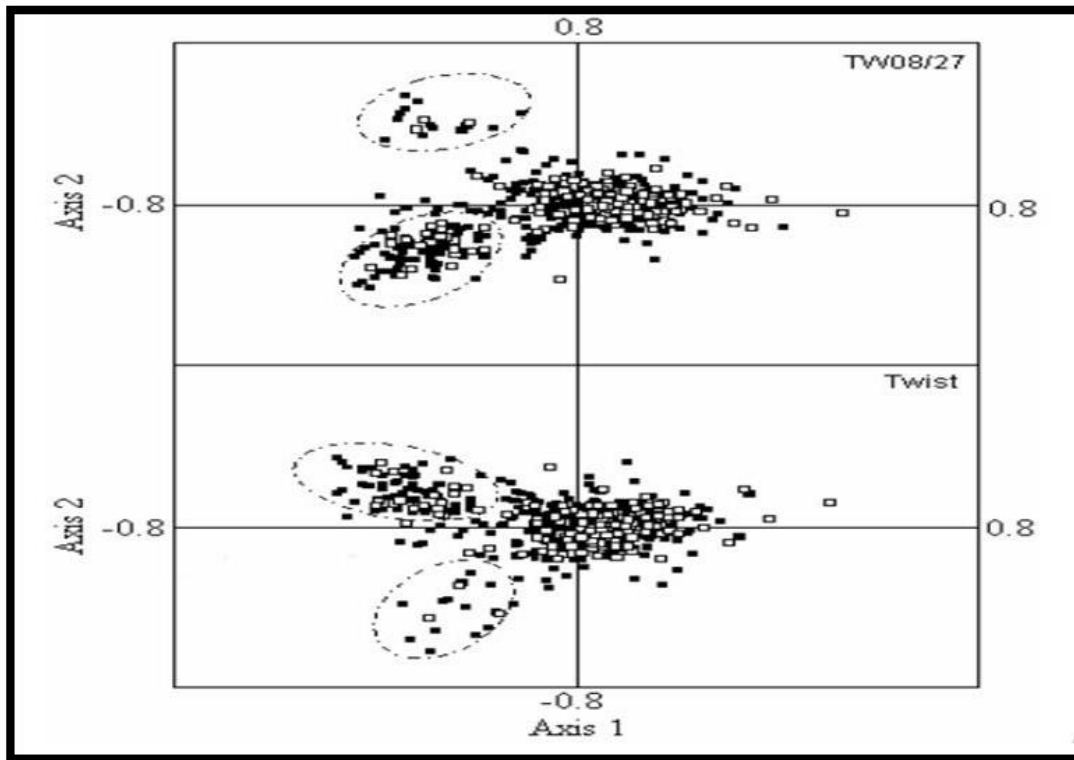


Figure 56. Correspondence study on Relative amino acid usage (RAAU)

In **figure 56**, plot show filled squares that exhibit genes transcribed in leading strand while empty squares show genes transcribed in lagging strand of replication sites of DNA. Correspondence study on Amino acid Usage of encoded genomic outcomes or peptides depicts nullified segregation of proteins coded by leading and lagging strand nucleotide stretches. So separate analysis for each strand was deployed, and it was found Phe and Val are undoubtedly exhibited in G+U abundant codons of leading strand while Lys, Thr, and Asn are found to be abundant in A+T rich codons of lagging strand. In **Table 2** we clearly state that there is a good correlation in terms of hydrophobicity and aromaticity of expressed out proteins, as both are considered the crucial factors for amino acid variability in *T. whipplei* strains. This investigation leads to represent those bunch of genes depicting membrane-linked peptides, in conjunction with WiSP/WND group members and some debatable hypothetical proteinous structures. The cistronic groups of

large clusters were found to hold the key for some crucial proteins for bacterial survival like cytochrome C subunits, integral membranous proteins and transporters of various nutrients. But one can conclude that most of them are associated with membrane proteins.

Table 2. Multiple Parameters for encoded proteins

	TW08/27 strain			TWIST strain		
	Variability (%)	Source of variation	Correlation Coefficient (r)	Variability (%)	Source of variation	Correlation Coefficient (r)
1 st AXIS	18.8	Gravy Score	-0.82	18.5	Gravy Score	-0.87
		Aromaticity	-0.67		Aromaticity	-0.68
2 nd AXIS	14.6	Gravy Score	-0.58	14.3	Gravy Score	-0.47
		Aromaticity	-0.46		Aromaticity	-0.38
3 rd AXIS	8.9	Size/Complexity	-0.79	8.9	Size/Complexity	-0.74
		GC12	0.65		GC12	0.66

5.3 T-Cell Epitope based vaccine prediction

5.3.1 Proteins sequence retrieval and non-Allergen determination –

Proteomes were retrieved in fasta format from NCBI-GenBank and UniProtKB databases. Five proteins of different functionality were selected with following accession no's: WP_042507409.1 DNA-directed RNA polymerase subunit beta (RPO-B), WP_033800049.1 co-chaperone GroES, WP_038104819.1 TerC/Alx family metal homeostasis membrane protein, WP_042505650.1 membrane protein insertase YidC, WP_042505746.1 murein biosynthesis integral membrane protein. This selection depicts the variability to include crucial proteins of pathogenic domain (**Table3**).

Table 3. Allergenicity results for analyzed proteins of *T. whipplei*

S. no.	UniProtKB Accession No.	GenBank Accession No.	Allergen FP Score (Tanimoto similarity index)
1.	Q76L83	WP_042507409.1	0.85 NON-ALLERGEN
2.	Q96P47	WP_033800049.1	0.81 NON-ALLERGEN
3.	P00846	WP_038104819.1	0.91 NON ALLERGEN
4.	O14569	WP_042505650.1	0.83 NON ALLERGEN
5.	Q7XBS0	WP_011096746.1	0.89 NON ALLERGEN

5.3.2 T-Cell Epitope prediction - Net MHCII PAN 3.2 server was deployed to identify promiscuous epitopes and probable HLA alleles of MHC Class II that interacts together by analyzing their 1-log50k values and binding affinities, then the VaxiJen scores were used with threshold of ≥ 0.7 with all informative details are presented in **Table4**.

Table 4: Predicted epitopes based on NetMHCII 3.2 server and VaxiJen score. Threshold value of 0.7 and above was elected.

PROTEIN ID	ALLELE (HLA)	POS	PEPTIDE	1-log50k(aff)	VaxiJen	Antigen/Non-Antigen
Q76L83	DRB1_0101	328	IRYLAALHL	0.581	0.9461	Antigen
	DRB1_0103	328	IRYLAALHL	0.311	0.9461	Antigen
	DRB1_0301	801	LSAEERLLR	0.271	0.2947	Non-Antigen
	DRB1_0401	225	FLRAIGMTD	0.292	-0.3463	Non-Antigen
	DRB1_0404	328	IRYLAALHL	0.375	0.9461	Antigen
	DRB1_0405	328	IRYLAALHL	0.336	0.9461	Antigen
	DRB1_0406	328	IRYLAALHL	0.255	0.9461	Antigen
	DRB1_0701	328	IRYLAALHL	0.496	0.9461	Antigen
	DRB1_0802	324	IIATIRYLA	0.284	-0.5053	Non-Antigen
	DRB1_1101	1010	YMYVLKLHH	0.451	1.2693	Antigen
	DRB1_1302	154	FVINGTERV	0.445	-0.7514	Non-Antigen

Q96P47	DRB1_0101	84	YILASRDV	0.410	0.2401	Non Antigen
	DRB1_0406	86	ILASRDVLA	0.200	-0.1277	Non Antigen
P00846	DRB1_0101	222	FFSLTGLRQ	0.478	-0.0649	Non Antigen
	DRB1_0103	244	YMKFGVAAL	0.194	0.3077	Non Antigen
	DRB1_0301	152	GLLDKVMIR	0.249	0.6173	Non Antigen
	DRB1_0401	198	MFALDSIPA	0.343	0.4086	Non Antigen
	DRB1_0404	295	IIALSVALS	0.339	0.8054	Antigen
	DRB1_0405	222	FFSLTGLRQ	0.350	-0.0649	Non Antigen
	DRB1_0406	295	IIALSVALS	0.237	0.8054	Antigen
	DRB1_0701	87	FRFAVPEIF	0.420	1.2093	Antigen
	DRB1_0802	18	MLVTVRRPA	0.314	-0.5928	Non Antigen
	DRB1_0901	87	FRFAVPEIF	0.407	1.2093	Antigen
	DRB1_0901	244	YMKFGVAAL	0.386	0.3077	Non Antigen
	DRB1_1001	222	FFSLTGLRQ	0.477	-0.0649	Non Antigen
	DRB1_1101	18	MLVTVRRPA	0.402	-0.5928	Non Antigen
	DRB1_1302	159	IRMNVSKNY	0.420	0.7822	Antigen
	DRB1_1602	222	FFSLTGLRQ	0.335	-0.0649	Non Antigen
O14569	DRB1_0101	175	FYALQAGQA	0.539	0.6085	Non Antigen
	DRB1_0301	278	LAFELRRKR	0.220	0.7910	Antigen
	DRB1_0401	175	FYALQAGQA	0.284	0.6085	Non Antigen
	DRB1_0404	8	FLQNILLPI	0.299	0.1551	Non Antigen
	DRB1_0405	8	FLQNILLPI	0.341	0.1551	Non Antigen
	DRB1_1001	175	FYALQAGQA	0.467	0.6085	Non Antigen
	DRB1_1101	63	FLKQIRAQR	0.438	-0.0522	Non Antigen
	DRB1_1302	8	FLQNILLPI	0.478	0.1551	Non Antigen
Q7XBS0	DRB1_0101	374	YILQKAFYA	0.556	0.3071	Non Antigen
	DRB1_0103	374	YILQKAFYA	0.274	0.3071	Non Antigen
	DRB1_0401	374	YILQKAFYA	0.323	0.3071	Non Antigen
	DRB1_0404	334	VLMVSAPPL	0.350	1.2114	Antigen
	DRB1_0405	374	YILQKAFYA	0.312	0.3071	Non Antigen
	DRB1_0406	334	VLMVSAPPL	0.250	1.2114	Antigen
	DRB1_0701	334	VLMVSAPPL	0.539	1.2114	Antigen
	DRB1_0802	435	FLAIRVKLG	0.274	1.1141	Antigen

DRB1_0901	334	VLMVSAPPL	0.436	1.2114	Antigen
DRB1_1001	374	YILQKAFYA	0.473	0.3071	Non Antigen
DRB1_1101	503	YFLVITRCR	0.416	-0.2632	Non Antigen
DRB1_1302	334	VLMVSAPPL	0.412	1.2114	Antigen
DRB1_1602	374	VLMVSAPPL	0.366	0.3071	Non Antigen

5.3.3 Molecular 3D modeling of epitopes and HLA alleles - 3D structural models of selected epitopes were designed by using PEP-FOLD 3 web server and then most common HLA DRB1 proteins structural models were obtained from RCSB-PDB database. In **Table5** PDB Id along with HLA alleles is exhibited. Molprobrity results of Ramachandran plot analysis results shows satisfactory structural prediction (>85% residues in favorable region) of epitopes that were finalized at last in **Figure64**.

Table 5. HLA Template model based on Pdb Id derived for MHC Class II alleles Structure from RCSB-PDB

ALLELE NAME	TEMPLATE STRUCTURE (PDB ID)
HLA-DRB1* 01_01	4AH2
HLA-DRB1* 01_03	3PDO
HLA-DRB1* 03_01	1A6A
HLA-DRB1* 04_01	5LAX
HLA-DRB1* 04_04	4IS6
HLA-DRB1* 04_05	4IS6
HLA-DRB1* 04_06	4IS6
HLA-DRB1* 07_01	3C5J
HLA-DRB1* 08_02	3PDO
HLA-DRB1* 09_01	1BX2
HLA-DRB1* 10_01	3PDO
HLA-DRB1*11_01	6CPL
HLA-DRB1*13_02	1FV1
HLA-DRB1*16_02	6CPO

5.3.4 Molecular Docking of epitopes and HLA alleles - PatchDock tool was deployed for interaction between selected structures of epitopes and HLA DRB1 proteins. Then interaction data produced by docked molecules include ACE (Atomic contact energy)

and best model score that leads to the final selection in the way of prediction for each pair. In **Table 6** the selected models and rejected models both were included to enhance the comparative analysis. The two selected epitopes were VLMVSAFPL and IRYLAALHL interacting with 4 and 6 HLA DRB1 alleles respectively. VLMVSAFPL epitope is a part of DNA-directed RNA polymerase subunit beta and IRYLAALHL epitope is a part of membranous protein insertase YidC of *Tropheryma whipplei* and are major identifiers of this bacterium. **Figure 57** and **58** clearly depicts the good interaction between epitopes and HLA Alleles in docked results. In **figure 57** docked result of IRYLAALHL with HLA-DRB1* 01:01 exhibits perfect hydrogen bond due to presence of tyrosine residue in epitope at 3rd position, while most of the other non-polar amino acids of this epitope have van der Waals interactions with in the HLA model. In **figure 58** docked result of VLMVSAFPL with HLA-DRB1* 04:04 exhibits perfect hydrogen bond due to presence of serine residue in epitope at 5th position, while most of the other non-polar amino acids of this epitope have van der Waals interactions with in the HLA model. The effectiveness of epitope-based vaccines for treatment of endocarditis, has already been claimed (**Priyadarshini et al. 2014**). **Figure59** represents the selected epitopes and HLA alleles of MHC II on the basis of ACE values.

Table 6. Molecular docking outcomes of epitopes with HLA alleles using patch dock

Allele	Epitope	Binding Score	ACE	Epitope Selection
DRB1 0101 (4AH2)	IRYLAALHL	7704	-367.28	SELECTED
DRB1 0103 (3PDO)	IRYLAALHL	7452	-321.11	SELECTED
DRB1 0404 (4IS6)	IRYLAALHL	8064	-353.51	SELECTED
DRB1 0405 (4IS6)	IRYLAALHL	8064	-353.51	SELECTED
DRB1 0406 (4IS6)	IRYLAALHL	8064	-353.51	SELECTED
DRB1 0701 (3C5J)	IRYLAALHL	7544	-349.18	SELECTED
DRB1 1101 (6CPL)	YMYVLKLHH	7880	-288.29	REJECTED
DRB1 0404 (4IS6)	IIALSVALS	6562	-136.21	REJECTED
DRB1 0406 (4IS6)	IIALSVALS	6562	-136.21	REJECTED
DRB1 0701 (3C5J)	FRFAVPEIF	8260	-223.61	REJECTED
DRB1 0901 (1BX2)	FRFAVPEIF	9042	-127.68	REJECTED
DRB1 1302 (1FV1)	IRMNVSKNY	8564	344.23	REJECTED
DRB1 0301 (1A6A)	LAFELRRKR	7746	-93.04	REJECTED

DRB1 0404 (4IS6)	VLMVSAFPL	6788	-443.31	SELECTED
DRB1 0406 (4IS6)	VLMVSAFPL	6788	-443.31	SELECTED
DRB1 0701 (3C5J)	VLMVSAFPL	6868	-226.35	SELECTED
DRB1 0802 (3PDO)	FLAIRVKLG	7812	-298.60	REJECTED
DRB1 0901 (1BX2)	VLMVSAFPL	7960	-297.99	SELECTED
DRB1 1302 (1FV1)	VLMVSAFPL	8500	104.94	REJECTED



Figure 57. Molecular docking between epitope IRYLAALHL and HLA: Docked result of IRYLAALHL with HLA-DRB1* 01:01 exhibits perfect hydrogen bond due to presence of tyrosine residue in epitope at 3rd position, while most of the other non-polar amino acids of this epitope are depicting Vander Waals interactions with In the HLA model.

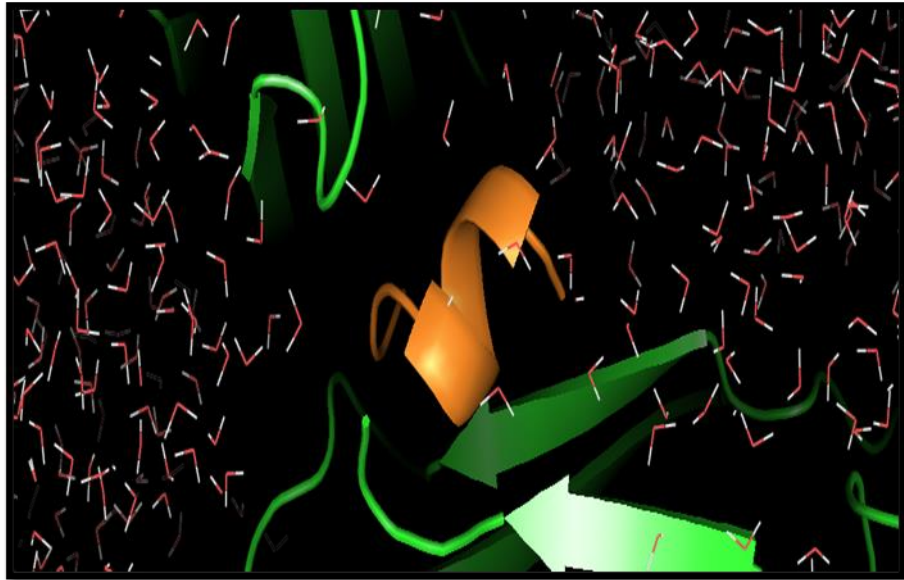


Figure 58. Molecular docking of epitope VLMVSAFPL and HLA: Docked result of VLMVSAFPL with HLA-DRB1* 04:04 exhibits perfect hydrogen bond due to presence of serine residue in epitope at 5th position, while most of the other non-polar amino acids of this epitope are depicting van der Waals interactions with in the HLA model.

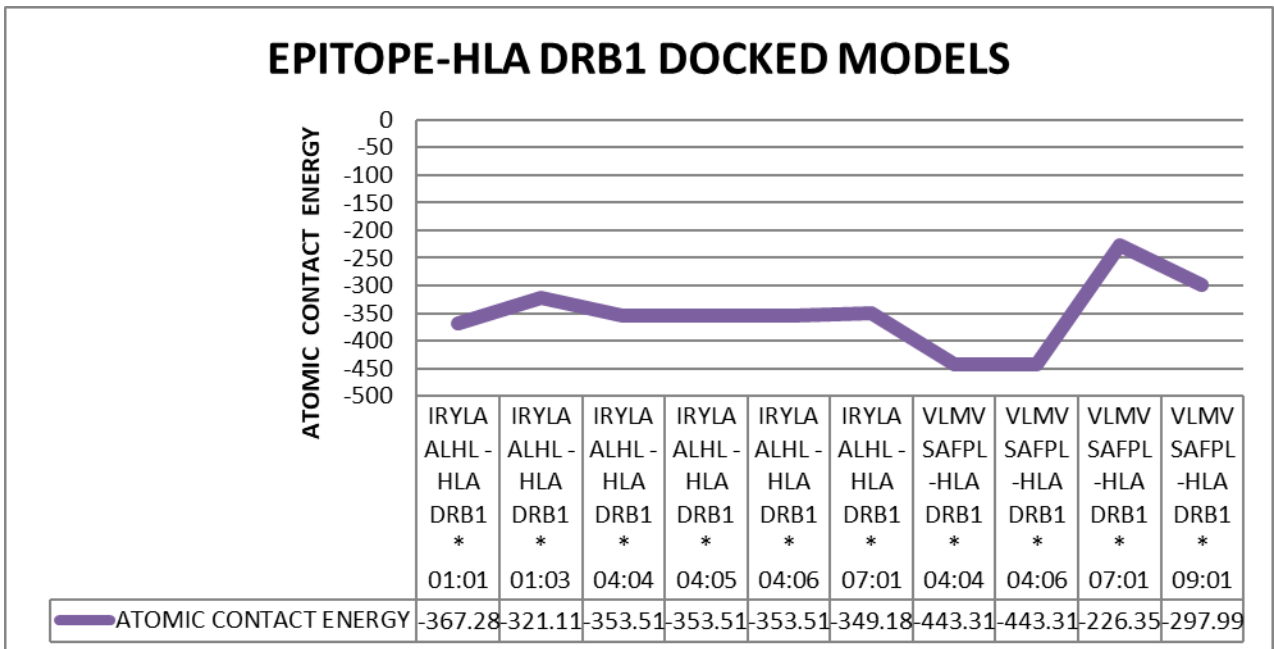


Figure 59. Graphical depiction of docked and selected Epitopes- HLA Alleles corresponding to their atomic contact energies.

5.3.5 Toxicity prognostication of putative vaccine targets

Predicted epitopes VLMVSAFPL and IRYLAALHL have VaxiJen scores 0.9461 and 1.2114 respectively, they are also of non-toxic nature as per the study of Toxin Pred tool and its toxicity scores (SVM score) are provided in **Table 7**. In **Table 8** quantitative estimation of best interaction of epitope with HLADRB1 alleles were achieved with upright IC₅₀ values by using MHCpred tool, this allows confidence of prediction.

Table 7- Selected multi-target epitopes and toxicity score based on TOXIN-PRED tool

EPITOPE	NO. OF HLA BINDERS	TOXICITY SCORE	TOXICITY
VLMVSAFPL	4	-1.19	NON TOXIN
IRYLAALHL	6	-0.61	NON TOXIN

Table 8- MHCpred results for selected epitopes with their desired HLA DRB1 allele's binders (only three alleles were present in this database for MHC II).

Amino acid groups	HLA Allele used in test	Predicted - logIC ₅₀ (M)	Predicted IC ₅₀ Value (nM)	Confidence of prediction (Max = 1)
IRYLAALHL	DRB0101	8.598	2.52	0.89
IRYLAALHL	DRB0701	5.935	1161.45	1.00
VLMVSAFPL	DRB0701	5.954	1111.73	1.00

5.3.6 Population coverage analysis of epitopes

VLMVSAFPL and IRYLAALHL manifest 28.82% and 37.06% respectively elicitation of immune responsiveness by world population by availing IEDB tool. The epitopes VLMVSAFPL and IRYLAALHL shows greater effect in European population by 29.63% and 42.68% respectively, and correspondingly similar results with North American population coverage analysis. This indicates its greater relevance in treatment

of Whipple’s disease as it is mostly seen in Caucasoid population. In **figure 60** and **61** it is clearly represented in a graphical representation.

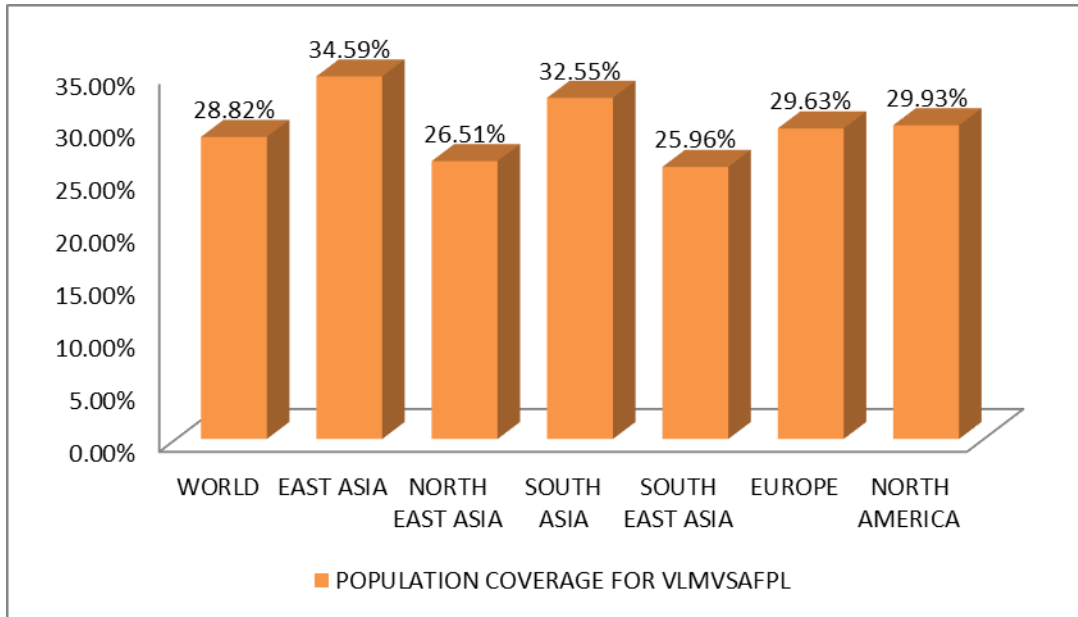


Figure 60. Graphical representation along of IEDB Population coverage for VLMVSAFPL

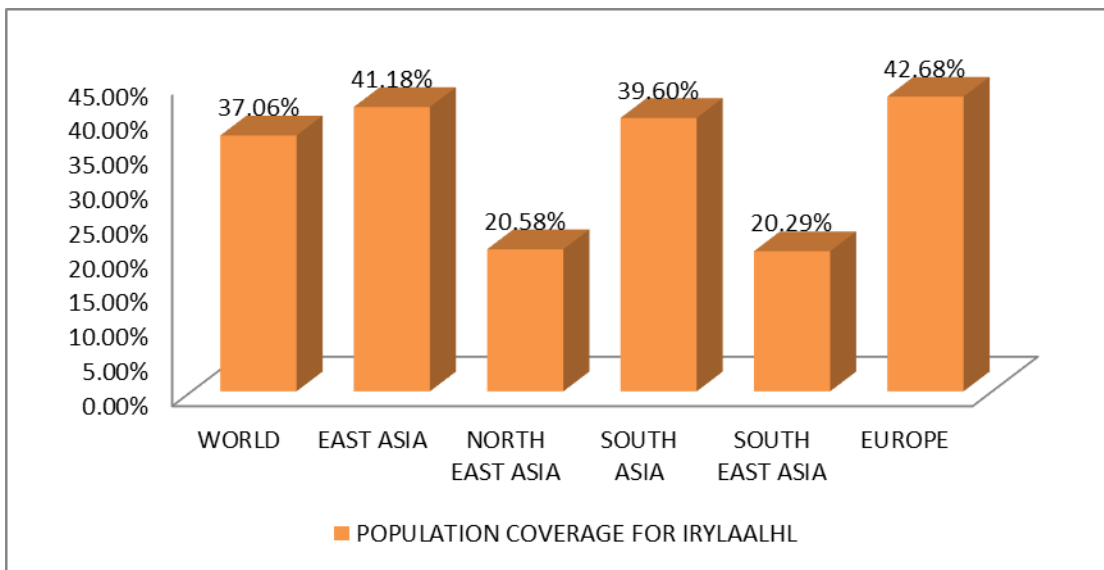


Figure 61. Graphical representation of IEDB Population coverage for IRYLAALHL

5.3.7 Molecular Dynamic and Simulation Studies

NAMD was deployed for simulation studies on docked Epitope – HLA allele sets to obtain RMSD values. Maximum value of RMSD for VLMVSAFPL and IRYLAALHL epitopes were analyzed. This gives more confidentiality in selection of Epitope based putative vaccine against *Tropheryma whipplei*. **Figure 62** and **63** shows RMSD plots vs time that indicates clear picture of selection of these two epitopes.

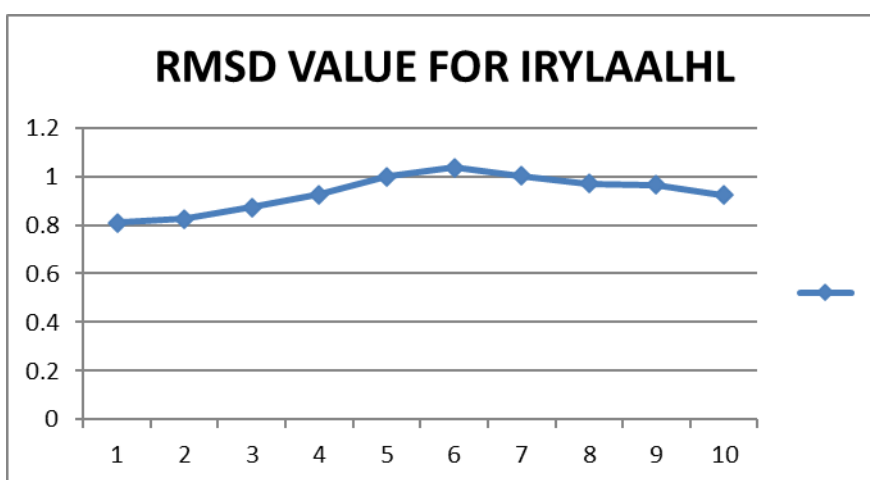


Figure 62. RMSD Value Vs Time (ps) for Epitope IRYLAALHL

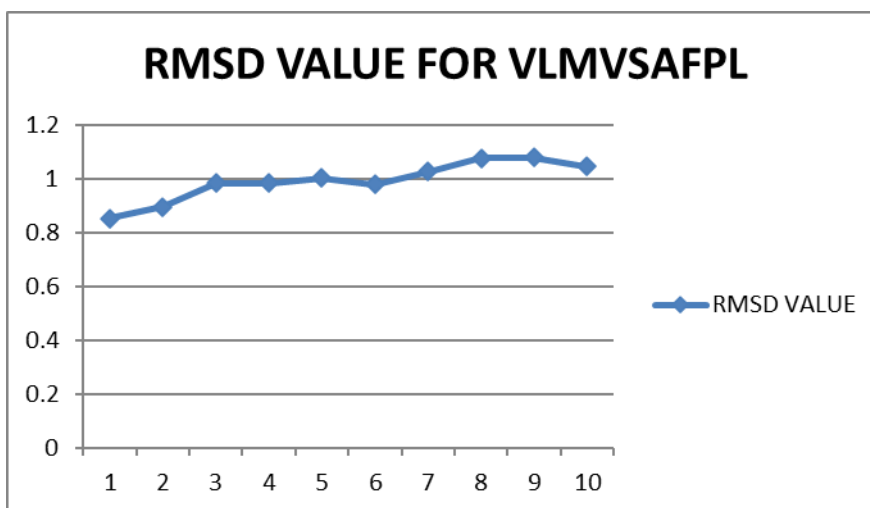


Figure 63. RMSD Value vs Time (ps) plot for epitope VLMVSAFPL

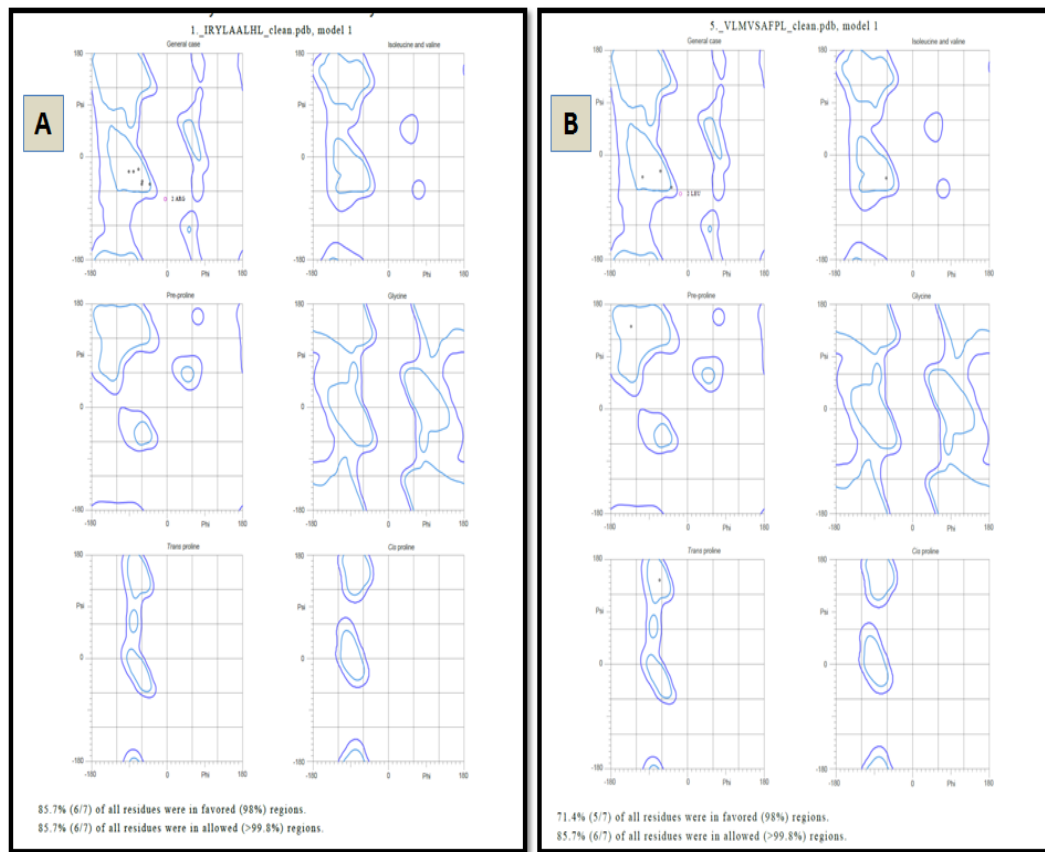


Figure 64. Molprobity Ramachandran Plot analysis results for Epitopes **A.** IRYLAALHL **B.** VLMVSAPFL

Immuno-informatics is the suitable approach as well as novel science method that use the proteomic data with the utilization of computer systems for predicting epitopes without culturing bacterium species (Tang et al. 2012). It allows the choice in hands of human interface for selecting antigens from pathogenic set of DNA and most antigenic areas could be used to synthesize potential immunization to initiate defensive responses against harmful pathogenic species (Kazi et al. 2018). In-silico approach was earlier successful in case of *Staphylococcus aureus* (Delfani et al. 2015), *Mycobacterium tuberculosis* (Mustafa 2013) and numerous bacterial species, but *Tropheryma whipplei* is still not fully explored in this domain. Current regimens include hydroxychloroquine and doxycycline for treatment of Whipple’s disease for 12 to 18 months, but life time follow

up is required (**Lagier et al. 2014**), so it is time consuming treatment process and only few handful trials were conducted in earlier studies (**Feurle et al. 2013**).

In present study we identified two possible epitopes that can interact with MHC-II alleles to elicit immune response on individuals namely VLMVSAFPL epitope (part of DNA-directed RNA polymerase subunit beta), and IRYLAALHL epitope (part of murein biosynthesis integral membrane protein). These Epitopes exhibit better interaction with HLA DRB1 alleles, as confirmed by deploying Molecular-Docking and Molecular-Simulation studies (**Adhikari et al. 2018**). Population coverage analysis was found to be satisfactory and in earlier studies it was used in strengthening vaccine prediction aspects (**Misra et al. 2011**). Very similar studies were also conducted successfully for related bacterium *Mycobacterium avium* and found to be successful in predicting epitopes (**Gurung et al. 2012**). The epitope VLMVSAFPL was found to interact with 4 HLA alleles of HLA-DRB1 domain(04:04, 04:06, 07:01, 09:01) with satisfactory ACE values (-443.3,-443.3, -226.3, -297.9 respectively); and the epitope IRYLAALHL found to interact with 6 alleles of HLA-DRB1 domain (01:01, 01:03, 04:04, 04:05, 04:06, 07:01) with satisfactory ACE values (-367.2, -321.1, -353.5, -353.5, -353.5, -349.1 respectively) in docking results similar type of methodology was seen in recent studies in screening epitopes for SARS-Cov-2 (**Joshi et al. 2020**). Both selected epitopes exhibit structural integrity as possess less than 35% instability index score, and half-life greater than 20 hours for mammalian reticulocytes, this makes the screening criteria more reliable. Also, more than 85% residues of selected epitopes come under favorable region in Ramachandran plot analysis. Still, no one has used vaccine-based treatments for Whipple's disease, as it is thought to be rare and possess reduced genome but considered one of the harmful pathogens of human (**Raoult et al. 2003, La Scola et al. 2001, Marth et al. 2016**). But in our study, we found the short peptides that can easily be synthesized and deployed in developing immunity in Caucasian populations against Whipple's disease. We have used T-cell epitopes not B-cell epitopes because T-cell epitope prediction remains an integral part of T-cell epitope mapping approaches. In contrast, B-cell epitope prediction utility is currently much more limited. There are several reasons to

that. First of all, prediction of B-cell epitopes is still unreliable for both linear and conformational B-cell epitopes. Secondly, linear B-cell epitopes do usually elicit antibodies that do not cross-react with native antigens. Third, the great majority of B-cell epitopes are conformational and yet predicting conformational epitopes have few applications, as they cannot be isolated from their protein context (**Sanchez-Trincado et al., 2017**).

5.4 In-silico drug prediction against *Tropheryma whipplei*

The selected proteins are listed on **Table 9** indicating their KEGG annotation and NCBI-GenBank accession number along with their known functionality (based on DEG and KEGG server) in *Tropheryma whipplei* twist strain. Phyre2 that is aHMM algorithm-based server was deployed to determine protein structure of DNA Ligase and Chorismate synthase enzymes. DNA ligase is crucial enzymatic assembly that is used by bacterium for repair and copying DNA sequences, and inhibited by 2- amino-7-fluoro-5-oxo-5H-chromeno[2,3-b] pyridine-3- carboxamide (2APC) and Nicotinamide mononucleotide (NMN). While Chorismate synthase perform dephosphorylation of 5-O-(1-carboxyvinyl)-3- phosphoshikimate to chorismate. This enzyme not exists in *Homo sapiens*. Chorismate is a precursor for aromatic-ring containing amino acids. This enzyme interacts with riboflavin monophosphate (as per Drug-Bank database). Riboflavin monophosphate (RFMP) is a strong oxidizing agent and has been used as an additive (coloring agent) in the food industry. In earlier studies, Riboflavin was used in combination with antibiotics and shown to control *Staphylococcus aureus* infection efficiently. Similar studies have been deployed to eradicate *T. whipplei* infection and promote its novel treatment strategy. PubChem database was deployed to retrieve structure of drugs interacting with selected proteins. PubChem CID and name of drugs are mentioned in **Table 11** with Swiss-ADME characteristics. In **Figure 65** structure of drugs is represented and for better visualization of pharmacophore analysis PyMOL software is used. The structure of enzymatic complexes was retrieved from phyre2 server, On the basis of detailed homology report of modeled structure of DNA ligase and Chorismate synthase enzymatic assembly of

Tropheryma whipplei. Best model result was retrieved to obtain PDB file of their structure. Out of 120 best structures, one was finalized for DNA ligase while out of 99 models one was finalized for Chorismate synthase. Docking studies reveals binding energies for docked complexes and perfect binding energies for all docked complexes represented in **table 10**. The perfectly docked complex of inhibitory drugs and enzymatic assembly are represented in **Figure 66T (A, B, C)**.

These results satisfy the perfect interaction of complexes suggests that 2-amino-7-fluoro-5-oxo-5H-chromeno [2,3- b] pyridine-3-carboxamide as well as Nicotinamide mononucleotide interacts with DNA Ligase while Riboflavin monophosphate can interact with Chorismate synthase. And these selected chemicals can be used as putative drug candidates. Drug 2D interaction pattern with enzymes based on LigPlotv2.2 software represented in **Figure66(P, Q, S)**. Mostly all the selected drugs not only show better binding scores but 3D and 2D interaction pattern reveals hydrogen bonds interaction with considered enzymes in their binding pocket. In **table 11** drug physiochemical parameters were represented based on SwissADME server (www.swissadme.ch). Lipinski rule (**Lipinski., 2004**) was also considered during drug analysis, and found to show zero violations for 2APC. This clearly indicates that all drugs have good inhibitory properties against selected enzymes. All of the drugs don't show blood brain barrier permeability also no inhibition for CYP3A4 inhibition, these results suggest effective drug clearance in body after effective action. 2APC, NMN and RFMP show logKp(cm/s) values -6.97, -8.26and 10.9 respectively. TPSA values for 2APC, NMN, and RFMP was found 112.21, 165.61, and 217.9 respectively.2APC also show high GI absorption. MD simulation for 40 nanoseconds (ns) was conducted by deploying GROMACSsoftwarever.2019. The equilibration steps were set with constant pressure and temperature (NPT) ensemble The MD simulations were carried out at standard temperature of 300 K and pressure level of 1.013 bar. The MD trajectories were estimated by analyzing the root mean square deviation (RMSD) and the root mean square fluctuation (RMSF) of the complexes for a timescale of 40 ns. It was found the best docked complexes were exhibiting perfect characteristics on the basis of RMSD and RMSF plots (**figure 67**).

Simulation study revealed that drug-enzyme complexes did not face any alterations in their binding patterns.

Table 9. Selected proteins information for *Tropheryma whipplei*, that are found to be drug targets by screening NCBI- GenBank database and identification by KAAS (KEGG Automatic annotation server) for pathway analysis of crucial genes.

NCBI-GenBank Accession no.	Identified Protein	KEGG Orthology Number	Functionality
AAO44511	DNA Ligase	K01972	DNA replication and repair
WP_011096348	Chorismate synthase	K01736	Biosynthesis of amino acids

Table 10. AutoDock-vina docked results: Binding energies of best docked complexes.

Best Docked Model	Binding Energy (Kcal/mol)
DNA Ligase and 2APC	-8.3
DNA Ligase and NMN	-8.2
Chorismate Synthase and RFMP	-7.3

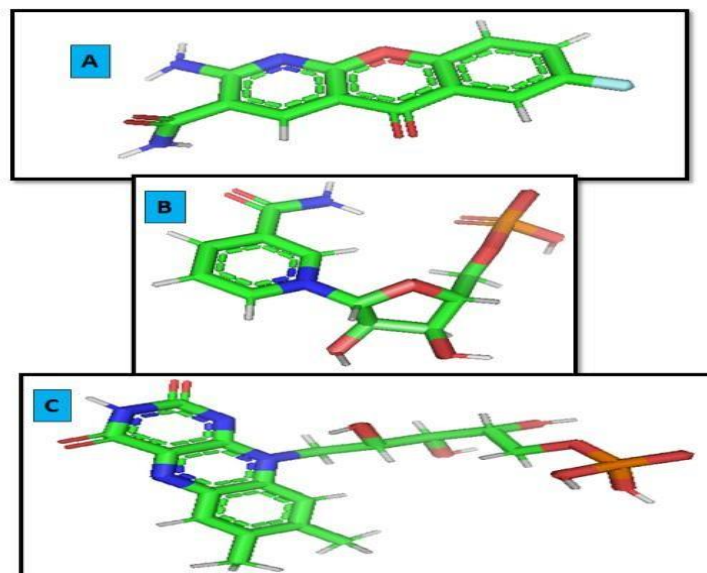


Figure 65. Chemical structure of drugs obtained from PubChem database and analyzed in Pymol **A.** 2-amino-7-fluoro-5-oxo- 5H-chromeno[2,3-b] pyridine-3-carboxamide (CID-10038928) **B.** Nicotinamide mononucleotide (CID- 14180) **C.** Riboflavin monophosphate (CID- 643976)

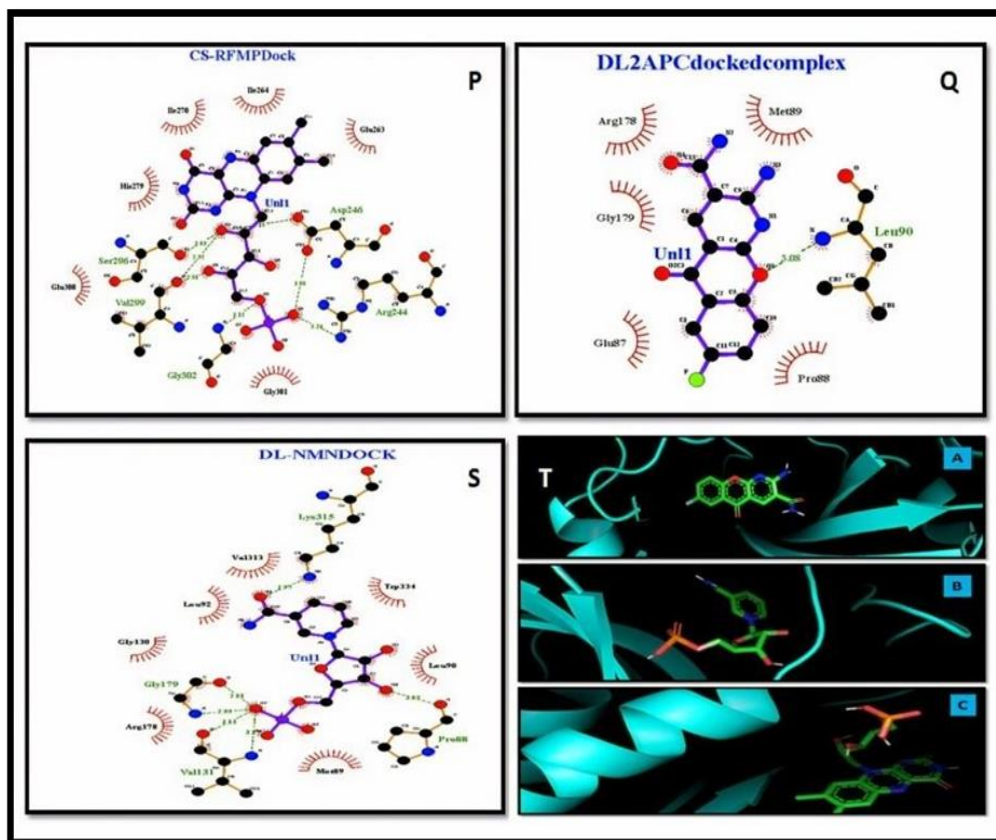


Figure 66. Molecular interactions between drugs and receptor: **P)** Chorismate synthase interaction with RFMP: Val299, Gly 302, Ser 296, Asp 246, Arg 244 interacts with RFMP drug via hydrogen bond (2.00 to 3.50 Å). **Q)** DNA Ligase interaction with 2APC drug: Leu at 90 position interacts with 2APC drug via hydrogen bond of strength 3.08Å. **S)** DNA Ligase interaction with NMN drug: Lys at 315 position interacts with NMN drug via hydrogen bond of strength 2.97Å, also Gly 179 & Val 131 show hydrogen bonding with the NMN drug. **T)** AutoDock vina docking results of drugs interacting with proteins- **A.** DNA Ligase with 2-amino-7-fluoro-5-oxo-5H- chromeno[2,3-b] pyridine-3-carboxamide **B.** DNA Ligase with Nicotinamide mononucleotide **C.** Chorismate synthase with Riboflavin monophosphate.

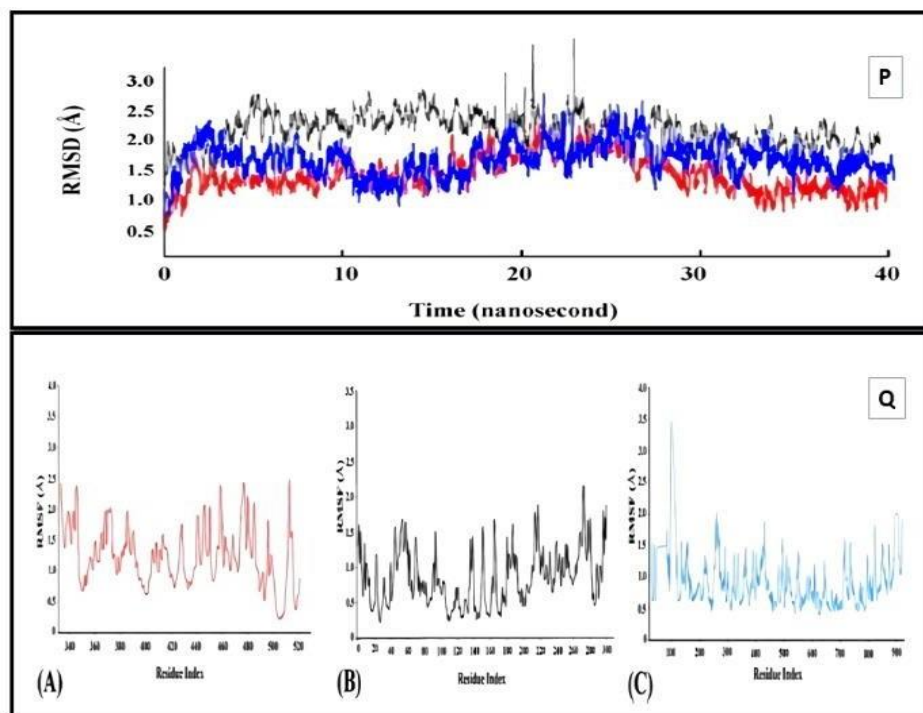


Figure 67. Molecular simulation analysis of drugs complexed with enzymes of *T. whipplei*: **P)** RMSD plot: Black color-(RFMP-Chorismate synthase complex), Red Color-(2APC-DNA Ligase complex), Blue color- (NMN-DNA Ligase complex); **Q)** RMSF plots: **A.** RFMP interaction with Chorismate synthase. **2APC** interacting DNA Ligase **C.**NMN drug interacting DNA Ligase

Table 11-Drug characteristics analyzed by PubChem database and SwissADME

Molecule	Formula	MW	Lipinski #violations
2-amino-7-fluoro-5-oxo-5H-chromeno [2,3-b] pyridine-3-carboxamide (CID- 10038928)	C ₁₃ H ₈ FN ₃ O ₃	273.22	0
Nicotinamide mononucleotide (CID- 14180)	C ₂₇ H ₂₉ N ₉ O ₉	511.52	1
Riboflavin monophosphate (CID- 643976)	C ₁₇ H ₂₁ N ₄ O ₉ P	456.34	2

5.5. Codon usage studies and epitope-based peptide vaccine prediction

To measure the codon usage bias retrieved codon usage tables were retrieved from Codon and Codon Pair Usage Tables (CoCoPUTs) database. This database showed the relative frequency that different codons are used in genes in *T. whipple* RefSeq data. Similarly, Codon-Pair Usage Tables displayed the counts of each codon-pair in the CDSs of *T. Whipple* genomic data (RefSeq) and calculated codon-pair usage bias. The complete nucleotide sequences of *T. whipplei* strains. The selected FASTA sequences of Twist 16S ribosomal RNA and 23S ribosomal RNA were retrieved from the NCBI RefSeq database (<https://www.ncbi.nlm.nih.gov/nucleotide>). The codon usage dataset was retrieved from the Codon Usage Database (<http://www.kazusa.or.jp/codon/>).

All codons in the original sequence of *T. whipplei* strains are replaced with the corresponding redundant codon having the highest codon usage frequency. ATGme tool (Daniel et al., 2015) was used to identify rare codons and accordingly optimize genomic sequences (<http://www.atgme.org/>). Genomic sequences in FASTA format were pasted in the search box and codon usage table was pasted in the respective interface and processed the data for analysis of rare codons and sequence optimization was processed.

From the identified genomic sequences of ribosomal RNA nucleotide composition were computed. The G+C composition of 1st, 2nd, 3rd positions and GC1s, GC2s, GC3s in the codons were discovered for the frequency and mean frequency identification. The frequency of synonymous third position codon and percentage i.e., A3, T3, G3, and C3 and (%A3s, %C3s, %T3s, and %G3s respectively calculated. To measure the bias of synonymous codons and the effective number of codons (ENC) were identified. Additional codon usage, codon usage per thousand, and Relative Synonymous Codon Usage (RSCU) were also calculated using 'CAIcal' tool available from <https://ppuigbo.me/programs/CAIcal/>.

The codon pair usage table and dinucleotide usage data were identified from the CoCoPUTs database (Alexaki et al., 2019; Athey et al., 2017). The *T. Whipple* taxonomy ID or taxid (2039) was verified by NCBI's taxonomy tool and the taxonomy

was illustrated in **Figure 68**. The log-transformed codon pair frequency heat map was discovered from the data analysis as illustrated in **Figure: 69**. The degree of ENC values ranges from 20 to 61(**Wright, 1990**). If the value is 20 then one codon coding for each amino acid and value ranged to 61 means all the synonymous codon was used for each amino acid. The ENC value computed in our analysis was 56.138, which means more than one codon was used for each amino acid. The ENC value should be ≤ 35 for significant codon bias (**Butt et al., 2016**). So, the higher ENC value indicates the low codon usage bias in *T. whipplei*. The ENC value details demonstrated in **Table 12**.

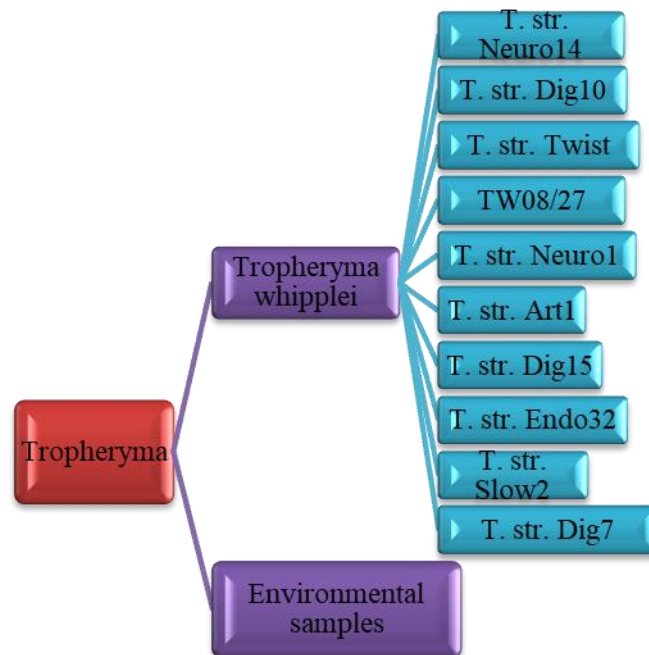


Figure 68. Taxonomy and strains of *Tropheryma whipplei*

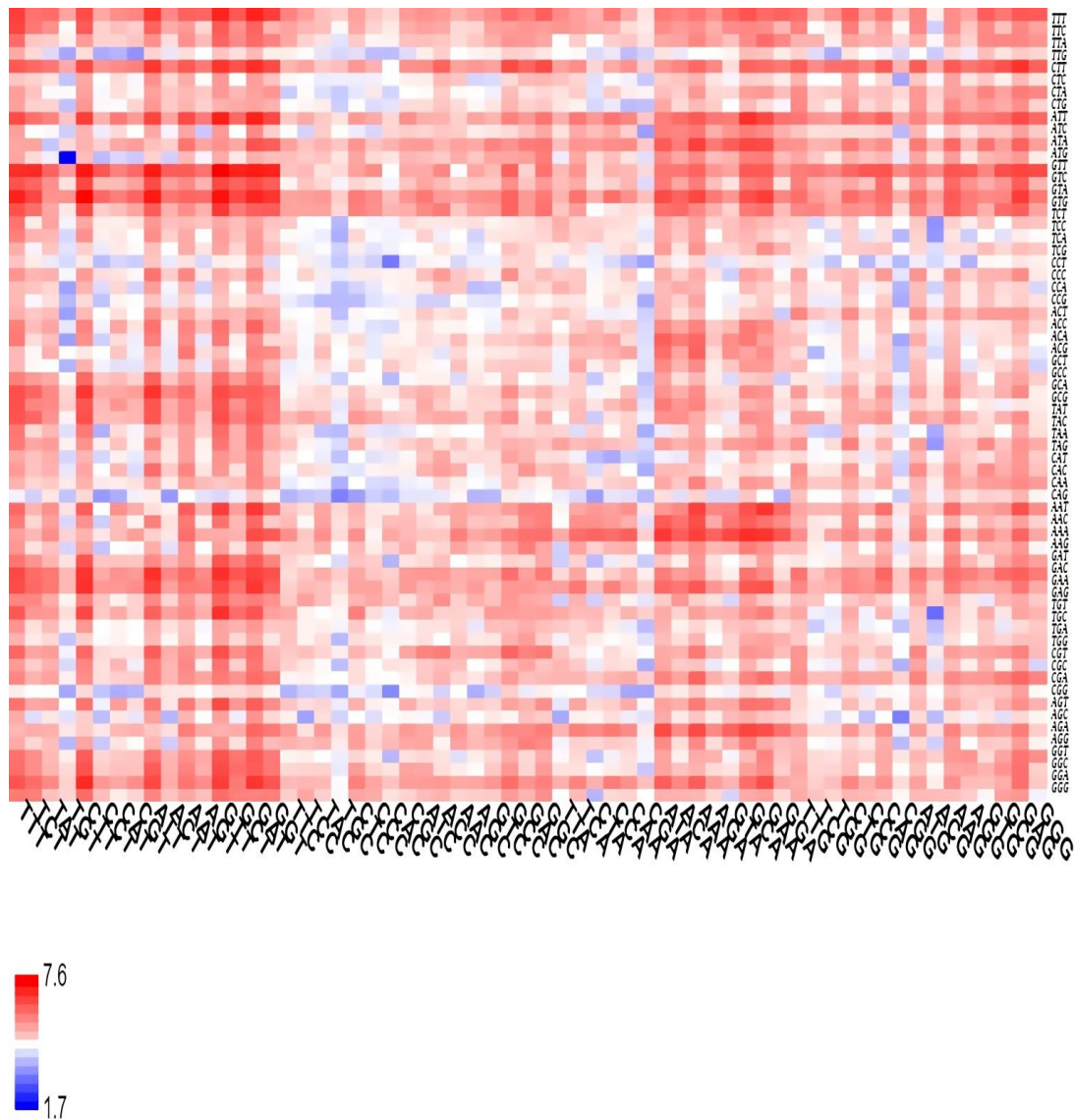


Figure 69. Heatmap of log-transformed codon usage.

Table 12- Effective Number of Codon Pairs for each *T. Whipplei*

ENc	ENcp	ENc (GC corrected)	ENcp (GC corrected)	Genetic Code
56.138	54.026	57.212	54.910	Standard code

The Codon usage details are summarized in the **table 12** and codon usage frequency per 1000 codons illustrated in **Figure 70**. The RefSeq (n = 859) of *T. Whipplei* had 88597 CDSs and 28006357 codons. **Table: 13** illustrated the CDS and its codon pair. The

codons GTT (37.06), GAT (37.03), CTT (32.53), and TTT (30.88) were identified as the highest usage frequency (frequency value showed in bracket). Dinucleotide frequencies per 1000 dinucleotide is demonstrated in **Figure71**.

Table13-*Tropheryma whipplei* RefSeq codon table contains 88597 CDSs (28006357 codons)

Codon	Usage frequency	No. codons	Codon	Usage frequency	No. codons	Codon	Usage frequency	No. codons	Codon	Usage frequency	No. codons
TTT	30.88	(864933)	TCT	19.41	(543618)	TAT	16.82	(470999)	TGT	7.23	(202580)
TTC	11.07	(309980)	TCC	9.94	(278337)	TAC	10.24	(286912)	TGC	6.09	(170524)
TTA	10.79	(302319)	TCA	15.06	(421837)	TAA	1.00	(27960)	TGA	1.11	(31013)
TTG	18.49	(517778)	TCG	10.24	(286768)	TAG	1.14	(31981)	TGG	10.09	(282675)
CTT	32.53	(910922)	CCT	9.92	(277793)	CAT	12.79	(358300)	CGT	11.66	(326630)
CTC	13.10	(366773)	CCC	9.27	(259711)	CAC	7.92	(221772)	CGC	11.85	(331835)
CTA	10.11	(283081)	CCA	13.02	(364568)	CAA	12.40	(347268)	CGA	6.07	(169892)
CTG	18.23	(510468)	CCG	12.02	(336672)	CAG	17.96	(503036)	CGG	6.98	(195455)
ATT	31.40	(879451)	ACT	12.53	(350913)	AAT	23.82	(667060)	AGT	13.41	(375656)
ATC	13.10	(366940)	ACC	13.40	(375396)	AAC	11.87	(332502)	AGC	10.61	(297274)
ATA	23.41	(655721)	ACA	17.74	(496749)	AAA	26.46	(741084)	AGA	13.59	(380649)
ATG	18.25	(511118)	ACG	8.30	(232504)	AAG	21.39	(598923)	AGG	13.55	(379598)
GTT	37.06	(1037894)	GCT	21.35	(597805)	GAT	37.03	(1037065)	GGT	24.77	(693723)
GTC	12.15	(340402)	GCC	19.26	(539448)	GAC	16.30	(456523)	GGC	18.74	(524935)
GTA	14.34	(401675)	GCA	26.13	(731733)	GAA	25.97	(727374)	GGA	14.66	(410657)
GTG	16.39	(459063)	GCG	16.86	(472115)	GAG	25.54	(715420)	GGG	15.16	(424597)

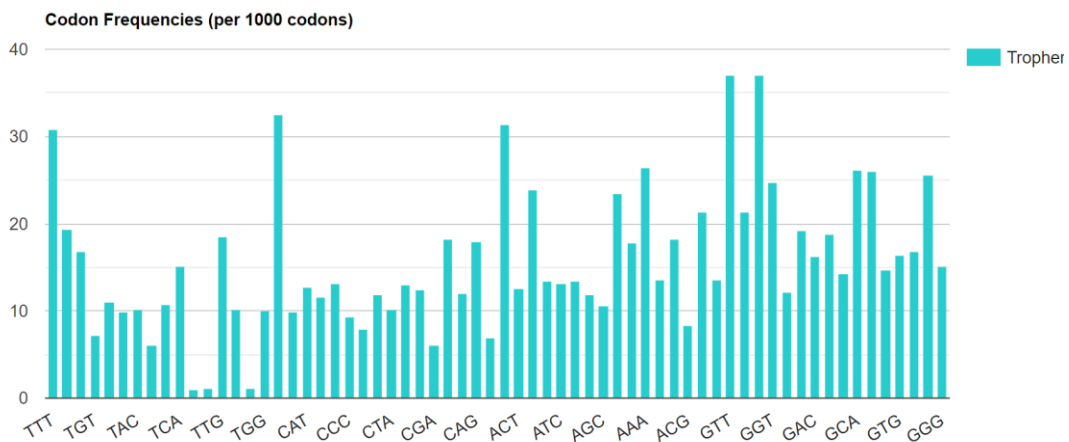


Figure 70. Codon frequencies of *Tropheryma whipplei*

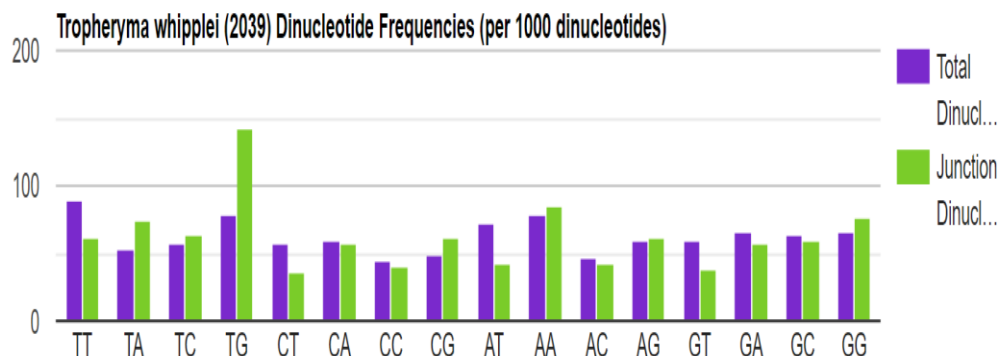


Figure 71. Dinucleotide frequencies of *Tropheryma whipplei*

5.5.1. *Tropheryma whipplei* str. Twist codon usage

Tropheryma whipplei strain Twist complete sequence of 23S & 16 S ribosomal RNA genes was composed of 3102 base pairs and 1521 base pairs respectively. *Tropheryma whipplei* Twist strain’s CDS, codons, frequency per thousand, and the number of codons details are summarized in **Table 14 and 15**. These codon usage tables were used for the identification of rare codons and sequence optimization.

Table14-*Tropheryma whipplei* str. Twist 808 CDS' (266294 codons) Codons, frequency per thousand and in bracket number of codons.

Codon	Frequency (No. of codon)	Codon	Frequency (No. of codon)	Codon	Frequency (No. of codon)	Codon	Frequency (No. of codon)
UUU	30.5(8121)	UCU	19.7(5246)	UAU	17.2(4590)	UGU	7.3(1938)
UUC	11.5(3066)	UCC	10.1(2690)	UAC	10.5(2790)	UGC	6.1(1626)
UUA	10.9(2906)	UCA	15.4(4100)	UAA	0.9(250)	UGA	1.1(281)
UUG	18.4(4894)	UCG	9.9(2643)	UAG	1.0(277)	UGG	10.2(2710)
CUU	31.8(8461)	CCU	10.6(2826)	CAU	12.8(3409)	CGU	11.6(3079)
CUC	13.1(3492)	CCC	9.8(2620)	CAC	7.9(2111)	CGC	11.6(3085)
CUA	10.6(2832)	CCA	13.5(3588)	CAA	12.5(3316)	CGA	6.0(1585)
CUG	18.3(4871)	CCG	11.6(3095)	CAG	18.4(4889)	CGG	6.9(1832)
AUU	30.6(8157)	ACU	12.7(3392)	AAU	23.7(6313)	AGU	13.1(3497)
AUC	13.2(3503)	ACC	14.2(3776)	AAC	11.9(3179)	AGC	10.7(2855)
AUA	23.3(6209)	ACA	19.6(5223)	AAA	26.2(6970)	AGA	13.6(3613)
AUG	18.0(4784)	ACG	8.2(2176)	AAG	21.2(5642)	AGG	13.2(3516)
GUU	36.7(9774)	GCU	21.3(5660)	GAU	36.3(9679)	GGU	24.9(6640)
GUC	12.2(3247)	GCC	19.4(5172)	GAC	16.1(4283)	GGC	18.8(5007)
GUA	14.7(3916)	GCA	26.1(6939)	GAA	25.2(6702)	GGA	14.8(3952)

GUG	16.6(4431)	GCG	16.3(4340)	GAG	24.7(6586)	GGG	14.8(3942)
GC Percent Information				Coding GC	1 st letter GC	2nd letter GC	3rd letter GC
				46.46 %	54.59%	42.30%	42.48%

Table15-*Tropheryma whipplei* TW08/27 783 CDSs and 261028 codons, frequency per thousand and in bracket number of codons.

Codon	Frequency (No. of codon)	Codon	Frequency (No. of codon)	Codon	Frequency (No. of codon)	Codon	Frequency (No. of codon)
UUU	30.4(7947)	UCU	19.8(5158)	UAU	17.4(4531)	UGU	6.9(1813)
UUC	11.4(2984)	UCC	10.3(2683)	UAC	10.5(2743)	UGC	5.7(1496)
UUA	10.7(2802)	UCA	15.6(4063)	UAA	1.0 (251)	UGA	1.0 (265)
UUG	17.7(4611)	UCG	9.8 (2567)	UAG	1.0(267)	UGG	10.0(2603)
CUU	31.9(8314)	CCU	10.6(2762)	CAU	12.8(3343)	CGU	11.5(2996)
CUC	13.4(3509)	CCC	9.8(2560)	CAC	7.8(2034)	CGC	11.5(3008)
CUA	10.8(2829)	CCA	13.8(3610)	CAA	12.6(3276)	CGA	5.8(1513)
CUG	18.2(4741)	CCG	11.5(3014)	CAG	18.4(4793)	CGG	6.7(1747)
AUU	30.7(8013)	ACU	12.8(3352)	AAU	23.7(6193)	AGU	13.1(3413)
AUC	12.9(3377)	ACC	14.6(3803)	AAC	12.1(3149)	AGC	10.7(2781)
AUA	23.6(6166)	ACA	20.1(5243)	AAA	26.2(6829)	AGA	13.6(3546)
AUG	17.9(4662)	ACG	8.1(2108)	AAG	21.2(5533)	AGG	13.1(3409)
GUU	36.9(9638)	GCU	21.3(5567)	GAU	36.7(9577)	GGU	25.0(6521)
GUC	12.2(3193)	GCC	19.6(5111)	GAC	16.2(4218)	GGC	18.7(4884)
GUA	14.7(3835)	GCA	26.1(6821)	GAA	25.2(6578)	GGA	14.9(3879)
GUG	16.3(4256)	GCG	16.2(4239)	GAG	24.9(6488)	GGG	14.6(3813)
GC Percent Information				Coding GC	1 st letter GC	2nd letter GC	3rd letter GC
				46.41 %	54.66%	42.27%	42.29%

5.5.2. Rare and very rare codons

The analysis resulted from usage data, original sequence, and optimized sequence. *Tropheryma whipplei* strain Twist 23S ribosomal RNA gene sequence analyzed usage data predicted GTT and GAT (36.7%& 36.3 %) had the high frequency in codon usage. TAA, TAG, and TGA code as ‘STOP’ had the lowest usage frequency percentage ((0.9 %, 1.0 %& 1.1 %) and found these are the very rare codons. The rare codons are CGA, TGC, CGG, TGT, CAC, ACG, CCC, and TCG. The stop codons are terminating the

protein translation process (Seligmann et al., 2019). The details of rare codons and very rare codons (code as, count, and percentage of usage frequency) of 23 s & 16 S rRNA were summarized in **Tables 16 and 17**.

Table16- *Tropheryma whipplei* strain Twist 23S ribosomal RNA gene

Codon	Codes as	Usage frequency ‰	Count
TAA	STOP	0.9	14
TAG	STOP	1	26
TGA	STOP	1.1	14
CGA	Arg	6	31
TGC	Cys	6.1	12
CGG	Arg	6.9	22
TGT	Cys	7.3	19
CAC	His	7.9	8
ACG	Thr	8.2	15
CCC	Pro	9.8	21
TCG	Ser	9.9	16

Table17- *Tropheryma whipplei* str. Twist 16S ribosomal RNA

Codon	Codes as	Usage frequency ‰	Count
TAA	STOP	0.9	8
TAG	STOP	1	3
TGA	STOP	1.1	5
CGA	Arg	6	5
TGC	Cys	6.1	8
CGG	Arg	6.9	15
TGT	Cys	7.3	3
CAC	His	7.9	6
ACG	Thr	8.2	7
CCC	Pro	9.8	6
TCG	Ser	9.9	10

5.5.3. Codon usage measurements

The calculated compositional properties for the coding sequences of the *Tropheryma whipplei* Twist strain overall frequency of nucleotides A% (25.11 & 23.54), C% (22.76 & 24.0), T% (20.76 & 19.4), and G% (31.37 and 33.07) in 23s and 16 s ribosomal RNA

gene respectively. The synonymous codons had the base content in 3rd position were calculated as A3S% (24.47 & 22.88), C3S% (20.99 & 22.88), T3S% (21.47 & 19.53), and G3S% (33.08 & 34.71) for 23s and 16 s rRNA respectively. GC3S% (52.85 & 57.85) the third synonymous codon position in GC content of 23s and 16 s rRNA respectively. **Figure72 and 73** shows rRNA characteristic features like length, nucleotide composition. In **Figure 74** rRNA synonymous codons percentage is given, while in **Figure 75** codon measurements were indicated.

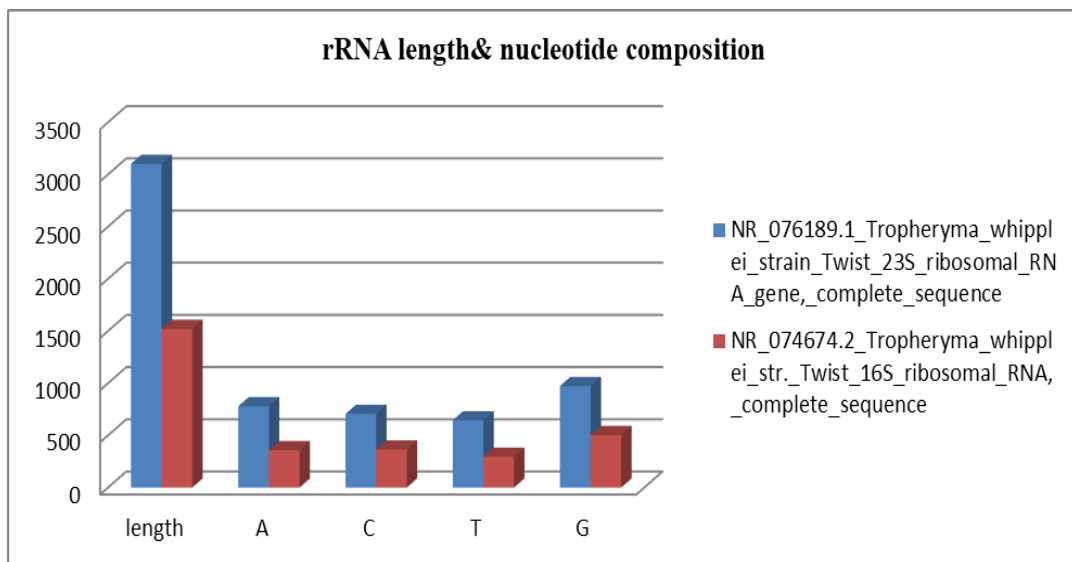


Figure72. rRNA length & nucleotide composition

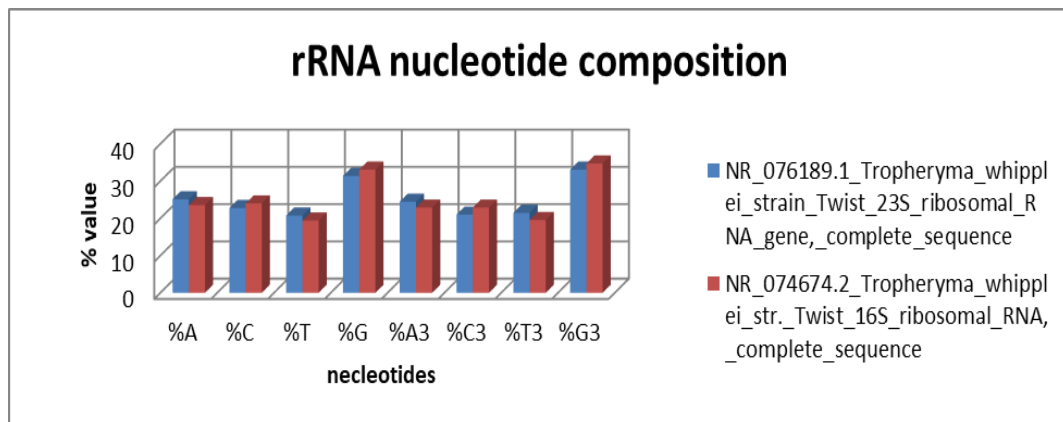


Figure73. Percentage of rRNA nucleotide composition

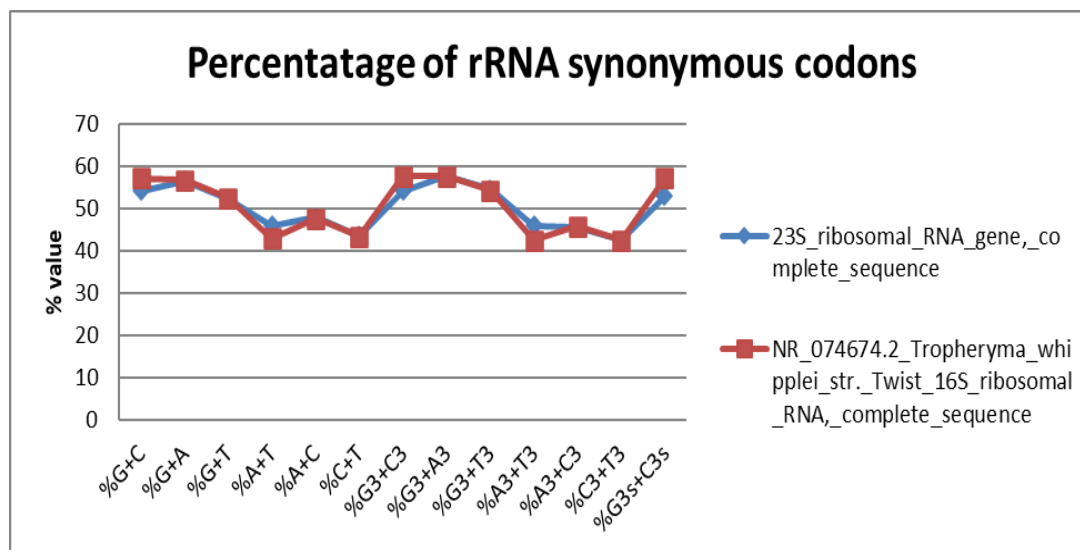


Figure 74. Percentage of rRNA synonymous codons

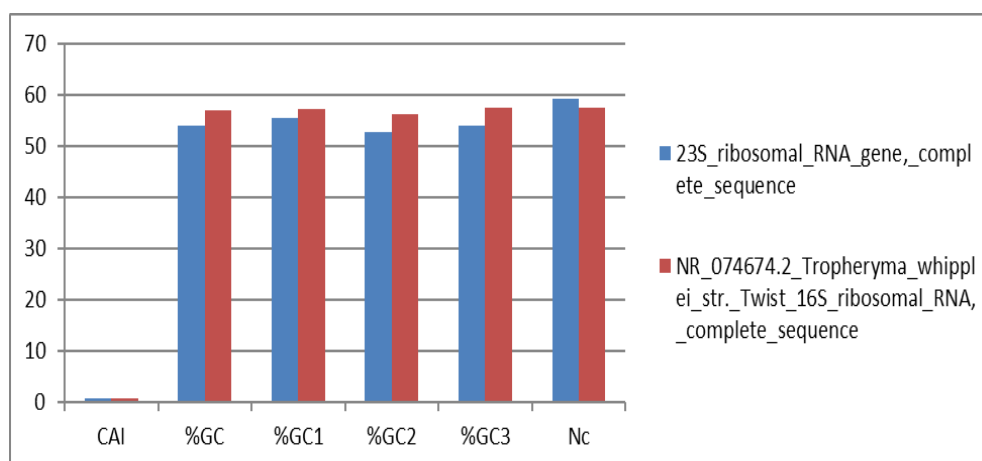


Figure 75. Codon measurement values plot

Table18-Allergenfp score and proteins considered for *Tropheryma whipplei*

Proteins/ No. of Amino acid residues	GenBank-Accession No.	Function	Allergen FP score	Inference
Pro-lipoprotein diacylglyceryl transferase [<i>Tropheryma whipplei</i>] 272 aa protein	WP_042506957.1	Catalyzes the transition of the diacylglyceryl group from phosphatidylglycerol to the sulfhydryl	0.87	Non-allergen

		group of the N-terminal cysteine of a prolipoprotein, the first step in the development of mature lipoproteins.		
excinuclease ABC subunit UvrC [<i>Tropheryma whipplei</i>] 607 aa protein	WP_042506954.1	DNA Excision repair	0.82	Non-allergen
Holliday junction resolvase RuvX [<i>Tropheryma whipplei</i>] 145 aa protein	WP_042506082.1	Nuclease activity, rRNA Processing	0.82	Non-allergen
exodeoxyribonuclease VII large subunit [<i>Tropheryma whipplei</i>] 404 aa protein	WP_042506175.1	Degrades single-stranded DNA bidirectionally, first into massive acid-insoluble oligonucleotides, then into small acid-soluble oligonucleotides.	0.82	Non-allergen
isoprenyl transferase [<i>Tropheryma whipplei</i>] 249 aa protein	WP_042506056.1	Isopentenyl diphosphate (IPP) condensation with allylic pyrophosphates is catalysed, resulting in a number of terpenoids.	0.80	Non-allergen
3-oxoacyl-ACP reductase FabG [<i>Tropheryma whipplei</i>] 238 aa protein	WP_011096407.1	Catalyzes the NADPH-dependent reduction of beta-ketoacyl-ACP substrates to beta-hydroxyacyl-ACP products, the first reductive step in the elongation cycle of	0.82	Non-allergen

		fatty acid biosynthesis.		
ABC transporter permease subunit [<i>Tropheryma whipplei</i>] 332 aa protein	WP_206536426.1	Transmembrane transportation of molecules	0.90	Non-allergen

Table 19- Peptides showing interaction to HLA-DRB0101, NETMHCII PAN 4.0 server results and VaxiJen score

Pos	Peptide	ID	Score	Rank	VaxiJen Score	Inference
39	NRRFIVLTGNREFTA	WP_042506957.1	0.958934	0.16	-0.4516	Non-Antigenic
316	KPSYLSALSAHLNDK	WP_042506954.1	0.978324	0.06	0.7208	Antigenic
384	LQKYLNLNSLPVRIE	WP_042506954.1	0.968518	0.11	1.1646	Antigenic
580	IEDISALPGFGVKTA	WP_042506954.1	0.960251	0.15	0.7039	Antigenic
227	RDKIQAQTVLSRSA	WP_042506954.1	0.805061	0.85	0.1459	Antigenic
77	EFSRFLVSSGVQVRF	WP_042506082.1	0.651559	1.60	0.4449	Antigenic
235	KTPLISAIGHEADRP	WP_042506175.1	0.966542	0.12	-0.0952	Non-Antigenic
231	DDFWAALRAYSGRSR	WP_042506056.1	0.960550	0.15	0.2368	Antigenic
24	FKSFNYNVAIGVRQP	WP_011096407.1	0.916978	0.35	0.7126	Antigenic
3	PARFFFVSPLSCVKP	WP_206536426.1	0.691033	1.40	0.6685	Antigenic

Table 20-ProtParam results: Biochemical properties of epitopes

Peptides	Molecular mass	pI	Gravy score	Aliphatic index	Instability index	Half life Mammalian reticulocytes
KPSYLSALSAHLNDK	1643.86	8.51	-0.553	91.33	5.83	1.3 hours
LQKYLNLNSLPVRIE	1800.13	8.59	-0.147	149.33	86.04	5.5 hours
IEDISALPGFGVKTA	1517.74	4.37	0.573	110.67	62.39	20 hours
FKSFNYNVAIGVRQP	1739.99	9.99	-0.180	71.33	24.99	1.1 hours
PARFFFVSPLSCVKP	1695.06	9.57	0.673	71.33	61.23	>20 hours

Table21- ACE Value, Global energy, and Binding energy for selected docked complexes (epitopes to HLA DRB0101).

Epitope	ACE value (Kcal/Mol)	Global Energy (Kcal/Mol)	Binding energy (Kcal/Mol)
KPSYLSALSAHLNDK	-6.59	-36.93	-2.80
FKSFNYNVAIGVRQP	-3.79	-1.19	-3.40

5.5.4. Epitope based vaccine prediction: Application of Codon usage studies

The *in-silico* analysis reveals two epitopes of 15 amino acid residues (i.e., KPSYLSALSAHLNDK and FKSFNYNVAIGVRQP) that holds perfect interaction with HLA-DRB-0101 (MHC Class II allelic determinant). In **Table18** retrieved sequences were shown with accession numbers, and allergenicity was also presented by deploying Allergen FP tool (this tool generates Tanimoto similarity index). Epitopes were determined by using NETMHCII Pan 4.0 Server, that gathers core information from IEDB database and uses Artificial neural networks (ANN) to access interaction of peptide stretches to HLA allelic determinants. Amino acids like Valine, Aspartate, Leucine, and Phenylalanine holds high codon usage frequency, and also found to be present in these screened epitopes from excinuclease ABC subunit UvrC and 3-oxoacyl-ACP reductase FabG. In **Table 19**, all 10 peptides holding good VaxiJen score, and NETMHCII Pan 4.0 scores are provided, but there were total of 2151 epitopes discovered. VaxiJen score indicates antigenicity for peptides. ProtParam results reveals only two finalized epitopes to be stable (**Table20**). Epitope's structure was predicted by using PepFold 3.5(**Thévenet et al., 2012**), and HLA Allelic determinant HLA DRB1_0101 (PDB-ID: 1AQD) was retrieved from rcsb-pdb database to perform molecular docking analysis. Molecular docking of selected epitopes with HLA-DRB0101 shows possible lead (**Table21**). **Figure 76** indicates docked complexes of selected epitopes with HLA-DRB-0101 visualized in PyMOL software.

Molecular Docking Results

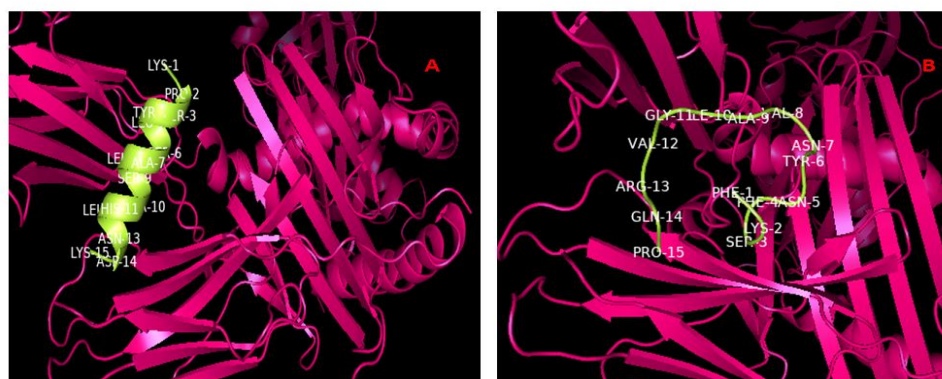


Figure 76. Molecular Docking results of epitopes with HLA-DRB-0101: **A)** KPSYLSALS AHLNDK from protein excinuclease ABC subunit UvrC and **B)** FKSFN YNVAIGVRQP from protein 3-oxoacyl-ACP reductase FabG.

Considerable biases in codon usage and amino acid usage indicate clearly that *T. whipplei* has a low codon bias. The synonymous codons had the base content in 3rd position were calculated as A3S% (24.47 & 22.88), C3S% (20.99 & 22.88), T3S% (21.47 & 19.53), and G3S% (33.08 & 34.71) for 23s and 16 s rRNA respectively. Also Codon-usage patterns clearly indicates that there will be less chances of variational or evolutionary alterations in *T. Whipple* genomic sets. The analysis could be targeted for disease evolution prediction, developing drugs, or vaccine candidates. We also found KPSYLSALS AHLNDK and FKSFN YNVAIGVRQP, two epitopes can possibly act as vaccine candidates against *T. whipplei*. A future development requires wet-lab validations for these epitopes that are highly expressed in this bacterium and have therapeutic peptide formation capability.

CHAPTER 6

SUMMARY AND CONCLUSION

The recent reverse vaccinology strategies against *Tropheryma whipplei* can resolve the multisystemic deadly malady Whipple's disease. This disease not only causes gastroenteritis, lipodystrophy but also causes severe neuronal damages, endocarditis, and impaired immune-system in small children's, workers (sewage, agriculture) among different human populations (severely affecting Caucasians) along with other mammals. Current treatments include hydroxychloroquine and doxycycline medications but these strategies require a yearly planned medications and very long-life time follow-up. Reverse vaccinology strategy has been found to be fast and effective to treat different diseases associated with bacterium like *Meningococcus* species. In this strategy complete proteomic and genomic data assessments are performed to obtain few peptides that can elicit immune response among patients. This would assist in developing novel therapeutic pathways against not only bacterial diseases but also against viral diseases. Till now various *in-silico* research epitopes against SARS-CoV2, Dengue virus, Nipah virus, Zika virus and also against *Tropheryma whipplei* have been obtained.

So, to reduce such extreme treatment pressure from patients and medics side, it is necessary to move towards some good available alternatives. Availability of genome due to efforts of various scientist groups working on *Tropheryma whipplei* in GenBank-NCBI database makes it easier to access for all. Many bioinformatics servers and tools were used to achieve goals of the current study. A) To identify genomic sets based on codon usage analysis: Codon-w software, python scripts, DNA star, and CUSP of EMBOSS tools were used to analyze RSCU (Relative synonymous codon usage) values that ultimately assist in determining which amino acids are mostly expressed in proteomic sets of *Tropheryma whipplei*. B) To retrieve proteomic data sets: Available protein sequences for the bacterium was obtained from primary databases like, NCBI, EMBL, and DDBJ etc. C) To screen proteins of interest: Allergenicity and Toxicity parameters were checked by deploying free web servers like AllergenFP, ToxinPred etc. D) To screen peptides or epitopes of interest from selected non-allergenic proteins: The epitopes was screened from protein sequences by using modern Artificial neural network filter screens like NetMHC server, to predict epitope binding with different HLA (Human

Leukocyte Antigen) allelic protein determinants. Antigenicity of epitopes was determined by VaxiJen, SVMTriP, PREDITOP tools. E) 3D Structure determination: Structures for HLA was obtained from RCSB-PDB (Protein Data Bank), while peptides (epitopes) structure was obtained either by homology modeling tools (Phyre) or from De-novo modeling tools (PepFold3.5). Biochemical properties for each selected peptide were determined by using ExPASy tools like ProtParam, Ramachandran plot was obtained by deploying Molprobit server. F) To study Molecular interactions between epitopes and HLA protein: Molecular docking between HLA and epitopes was found useful for identifying major binding pockets, and tools like DINC server, PatchDock, Auto-Dock vina made it handier for everyone, such docking tools are also useful in determining Binding energies for interacting protein-protein species. G) Molecular dynamics and Simulation provide first line validation: MD-simulation tools like Gromacs, Desmond, and MD-Web server are available to simulate the cellular environment by setting force-fields, and fundamental pressure, temperature conditions where epitopes can interact with HLA allelic protein determinants for set time span (usually 10 nanoseconds to 1 microseconds), depending on system quality (better in supercomputing facility). This assists primarily in generating RMSD (Root mean square deviation) plot, and RMSF (Root mean square fluctuation) plot, that identify stability among molecules for selected time span. H) Wet-lab validations: After finalizing epitopes, these epitopes can be synthesized and joined together by using linker peptides along with adjuvants. Then animal cell line testings or model organism testing will be the second line validations.

Currently we identified VLMVSAFPL and IRYLAALHL interact with four and six HLA DRB1 MHC Class II alleles, respectively. The VLMVSAFPL epitope is found in DNA-directed RNA polymerase subunit beta, and the IRYLAALHL epitope is found in this bacterium's membranous protein insertase YidC. These novel methods are not only financially efficient but also save time from long hit and trials. IL-16 up-regulation in macrophages results in under-expression of defensive genes (Interferon-gamma, ICAM, and Complement C3etc). The method applied for comparative analysis of microarray data was found to be easy and fast approach to reveal transcriptomic insights of molecular

pathophysiology and dual-relationship between host and pathogen interaction. This *in-silico* approach holds the future key to reveal the deep mysteries within host and pathogen genomes. In this study we obtained VLMVSAFPL and IRYLAALHL as predicted epitopes for vaccine crafting.

This novel approach in crafting Vaccine based treatment of *Tropheryma whipplei* will open new doors in research for creating regimens to treat such harmful bacterium by developing adaptive immune response and eradicating it globally before any future escalations takes place. The predicted epitopes can be deployed in crafting vaccines against *Tropheryma whipplei* bacterium after Molecular-wet lab corroboration. *In-silico* investigation discovers that 2-APC, NMN, and RFMP as possible medications to treat Whipple's disease and can be used for animal testing and clinical trials. All the pharmacokinetics clearly depicts that these medications would perfectly interacts with bacterial biocatalysts to hamper their activity. This approach was found to be rapid for prediction of possible lead or drug candidates and even effective against harmful organism like *Tropheryma whipplei* having reduced genome.

Considerable biases in codon usage and amino acid usage indicate clearly that *T. whipplei* has a low codon bias. The synonymous codons had the base content in 3rd position were calculated as A3S% (24.47 & 22.88), C3S% (20.99 & 22.88), T3S% (21.47 & 19.53), and G3S% (33.08 & 34.71) for 23s and 16 s rRNA respectively. Also, Codon-usage patterns clearly indicates that there will be less chances of variational or evolutionary alterations in *T. Whipple* genomic sets. The analysis could be targeted for disease evolution prediction, developing drugs, or vaccine candidates. We also found KPSYLSALSAHLNDK and FKSFNYNVAIGVRQP, two epitopes can possibly act as vaccine candidates against *T. whipplei*. A future development requires wet-lab validations for these epitopes that are highly expressed in this bacterium and have therapeutic peptide formation capability.

In our findings of codon-usage analysis we found amino acids Phe and Val are undoubtedly exhibited in G+U abundant codons of leading strand while Lys, Thr, and

Asn are found to be abundant in A+T rich codons of lagging strand in membrane proteins. Also, after prediction of epitope-based vaccine peptides VLMVSAFPL and IRYLAALHL interact with four and six HLA DRB1 MHC Class II alleles, respectively (The VLMVSAFPL epitope is found in DNA-directed RNA polymerase subunit beta, and the IRYLAALHL epitope is found in this bacterium's membranous protein insertase YidC), it was observed they holds these amino acids which were found to be abundant in codon usage analysis. In the microarray data analysis of Whipple's disease patients from Geo Database of NCBI it was found that volcano plots exhibit log (fold change) value for IL-16 to be greater than zero in a positive scale, that shows up-regulation. Which clearly correlate with the study that shows the production of IL-16, a cytokine recognized because of its chemoattractant but also proinflammatory qualities. IL-16 upregulation is promoting multiplication of *Tropheryma whipplei* inside macrophages. The study on the series GSE5717 it is quite evident that higher doses of Doxycycline drug treatments over the bacterium *Tropheryma whipplei* have not much impact as bacterium survival strategies also enhances both directly and indirectly. Therefore the search for new drug or novel vaccine candidate's research by *in-silico* mode is required for rapid development of treatment. While analyzing the series GSE3693, it was found that thermic stress on bacterium can activate bacterium survival genetic makeup. While analysing series GSE49016 it was found that *Tropheryma whipplei* does not promote much impact on human dendritic cells, so by producing vaccine candidates we can induce adaptive immunity among patients. Similarly with mice BMDM cells study in three series GSE20210, GSE20209, and GSE16180; it was found that IL-16 activity within BMDM cells promote *Tropheryma whipplei* replication. In series GSE7453 analysis all 14 strains have much resemblance to twist strain of *Tropheryma whipplei* only a small variation was observed within the WiSP membrane protein regions, that's why in vaccine candidate's search we targeted variety of membrane proteins for epitope predictions. In the series GSE102862 IRF-4 deficiency or loss of function type mutation can cause *T. whipplei* rapid growth, as these genes are responsible for growth of variety of immune system cells (NK cells, T-cells, B-cells). Also, Molecular Docking and MD-simulation

structural biology tools show greater indication of 2-APC, NMN, and RFMP as possible medications to treat Whipple's disease, as these possible lead drugs target bacterial enzymes like Chorismate-Synthase involved in shikimate metabolic pathways is responsible for synthesis of aromatic amino acids like phenylalanine, tryptophan and tyrosine. The 2-APC, NMN, and RFMP also target DNA ligase which plays important role in multiple functions like DNA replication and DNA-repair mechanisms.

CHAPTER 7

FUTURE ASPECTS

In current research work, all the methods and results were as per considered objectives where we primarily focused on predicting epitope or peptide dependent vaccine candidates for developing new treatment strategies against *Tropheryma whipplei*. As a future scope we can use different methods for wet lab validations:

A. Model organism dependent vaccine validations

In one of the recent studies, it was found that *Toxoplasma gondii* ROP proteins were used to find out T-cell and B-cell epitopes, then these multiple epitopes used to form vaccine that was injected to infected mice models, which in turn increased IgG antibodies, and IFN-gamma (**Foroutan et al., 2020**). In another study murine models were used for in-vivo testing against *Mycobacterium tuberculosis* bacterium where scientific group injected *palmitoyl* linked immunogenic peptides and noted increase of IFN-gamma and T-cell proliferation (**Horváti et al., 2019**).

B. Cell-lines associated vaccine validations

In recent literature it was observed that *Lactococcus* species causing hemolysis of RBCs. In HeLa cell cultures when (**Tanhaieian et al., 2018**) group performs MTT viability tests they found no toxicity and less hemolysis even at low concentrations of peptide chimeric vaccine candidates (**Tanhaieian et al., 2018**). Similar extensions can be performed for this research work.

C. Nanoparticles-linked peptide vaccine test

Another recent approach was found where antibacterial peptides linked to nano particles assist in target delivery of peptides from HIV proteins, and this group used not only MTT cell viability assays but RT-PCR for quantification of cellular physiological responses and inflammatory reactions. Surprisingly, they found increased TNF-alpha and Interleukin 1beta via RT-PCR studies and greater cellular viability via MTT assays. Amalgamation of novel strategies will not only evolve the delivery mechanisms of peptide vaccines but also improve the specific testing and validations of epitope-based vaccine candidates that were predicted by Immunoinformatics approach (**Becker et al., 2004**).

D. Genome wide associated scanning and biochemical testing

In recent pandemic era, many scientists have performed full genome wide analysis after DNA sample collection from patients. The expression level of HLA alleles and haplotypes, was measured by using tools like pLink, arlequin for sequencing data analysis. Also, Immunoinformatics based peptide selection and considered HLA allelic determinants molecular docking and simulation validations provided finalized selection. After adding adjuvants, linker peptides and designing multi-epitope vaccine assist in biochemical testing in model organisms by deploying RT-PCR methods (**Sylvester-Hvid et al.,2004, Singh et al., 2020**). Future scope exists here for conducting microarray study for human macrophage cell cultures were we can use SNP Chips for data generation related to different populations.

BIBLIOGRAPHY

1. Adhikari UK, Tayebi M, Rahman MM. Immunoinformatics approach for epitope-based peptide vaccine design and active site prediction against polyprotein of emerging oropouche virus. *Journal of immunology research*. 2018 Oct 8;2018.
2. Ahmed RK, Maeurer MJ. T-cell epitope mapping. In *Epitope Mapping Protocols* 2009 (pp. 427-438). Humana Press.
3. Akhtar N, Joshi A, Singh B, Kaushik V. Immuno-informatics quest against COVID-19/SARS-COV-2: determining putative T-cell epitopes for vaccine prediction. *Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders)*. 2021 Jun 1;21(4):541-52.
4. Akhtar N, Joshi A, Singh J, Kaushik V. Design of a novel and potent multivalent epitope based human cytomegalovirus peptide vaccine: an immunoinformatics approach. *Journal of Molecular Liquids*. 2021 Aug 1;335:116586.
5. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PV, Hamasaki-Katagiri N, Osipova E, Simonyan V, Bar H, Komar AA. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *Journal of molecular biology*. 2019 Jun 14;431(13):2434-41.
6. Ashfaq UA, Saleem S, Masoud MS, Ahmad M, Nahid N, Bhatti R, Almatroudi A, Khurshid M. Rational design of multi epitope-based subunit vaccine by exploring MERS-COV proteome: Reverse vaccinology and molecular docking approach. *PLoS One*. 2021 Feb 3;16(2):e0245072.
7. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C. A new and updated resource for codon usage tables. *BMC bioinformatics*. 2017 Dec;18(1):1-0.
8. Azzouz EB, Boumaza A, Mezouar S, Bardou M, Carlini F, Picard C, Raoult D, Mège JL, Desnues B. *Tropheryma whipplei* increases expression of human Leukocyte Antigen-G on monocytes to reduce tumor necrosis factor and promote bacterial replication. *Gastroenterology*. 2018 Nov 1;155(5):1553-63.

9. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *Journal of cheminformatics*. 2015 Dec;7(1):1-3.
10. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*. 2012 Nov 26;41(D1):D991-5.
11. Becker ML, Bailey LO, Wooley KL. Peptide-derivatized shell-cross-linked nanoparticles. 2. Biocompatibility evaluation. *Bioconjugate chemistry*. 2004 Jul 21;15(4):710-7.
12. Bibi S, Ullah I, Zhu B, Adnan M, Liaqat R, Kong WB, Niu S. In silico analysis of epitope-based vaccine candidate against tuberculosis using reverse vaccinology. *Scientific reports*. 2021 Jan 13;11(1):1-6.
13. Bui HH, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC bioinformatics*. 2007 Dec;8(1):1-6.
14. Bureš J, Kopáčová M, Douda T, Bártová J, Tomš J, Rejchrt S, Tachecí I. Whipple's disease: our own experience and review of the literature. *Gastroenterology Research and Practice*. 2013 Jun 17;2013.
15. Butt AM, Nasrullah I, Qamar R, Tong Y. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerging microbes & infections*. 2016 Jan 1;5(1):1-4.
16. Chandra SR, Raj P, Pai AR, Reddy N. A case of Whipple's disease: A very rare cause for rapidly progressive dementia. *Indian journal of psychological medicine*. 2018 May;40(3):280-3.
17. Cohen AS, Schimmel EM, Holt PR, Isselbacher KJ. Ultrastructural abnormalities in Whipple's disease. *Proceedings of the Society for Experimental Biology and Medicine*. 1960 Nov;105(2):411-4.

18. Compain C, Sacre K, Puéchal X, Klein I, Vital-Durand D, Houeto JL, De Broucker T, Raoult D, Papo T. Central nervous system involvement in Whipple disease: clinical study of 18 patients and long-term follow-up. *Medicine*. 2013 Nov;92(6):324.
19. Crapoulet N, Barbry P, Raoult D, Renesto P. Global transcriptome analysis of *Tropheryma whippelii* in response to temperature stresses. *Journal of bacteriology*. 2006 Jul 15;188(14):5228-39.
20. Craven M, Mansfield CS, Simpson KW. Granulomatous colitis of boxer dogs. *Veterinary Clinics: Small Animal Practice*. 2011 Mar 1;41(2):433-45.
21. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*. 2017 Mar 3;7(1):1-3.
22. Daniel E, Onwukwe GU, Wierenga RK, Quaggin SE, Vainio SJ, Krause M. ATGme: Open-source web application for rare codon identification and custom DNA sequence optimization. *BMC bioinformatics*. 2015 Dec;16(1):1-6.
23. Del Tordello E, Rappuoli R, Delany I. Reverse vaccinology: exploiting genomes for vaccine design. In *Human vaccines 2017* Jan 1 (pp. 65-86). Academic Press.
24. Delfani S, Fooladi AA, Mobarez AM, Emaneini M, Amani J, Sedighian H. In silico analysis for identifying potential vaccine candidates against *Staphylococcus aureus*. *Clinical and experimental vaccine research*. 2015 Jan;4(1):99.
25. Desnues B, Raoult D, Mege JL. IL-16 is critical for *Tropheryma whippelii* replication in Whipple's disease. *The Journal of Immunology*. 2005 Oct 1;175(7):4575-82.
26. Dimitrov I, Naneva L, Doytchinova I, Bangov I. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics*. 2014 Mar 15;30(6):846-51.
27. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics*. 2007 Dec;8(1):1-7.

28. Elchert JA, Mansoor E, Abou-Saleh M, Cooper GS. Epidemiology of Whipple's disease in the USA between 2012 and 2017: a population-based national study. *Digestive diseases and sciences*. 2019 May;64(5):1305-11.
29. Evans CF, Davtyan H, Petrushina I, Hovakimyan A, Davtyan A, Hannaman D, Cribbs DH, Agadjanyan MG, Ghochikyan A. Epitope-based DNA vaccine for Alzheimer's disease: translational study in macaques. *Alzheimer's & Dementia*. 2014 May 1;10(3):284-95.
30. Feurle GE, Moos V, Bläker H, Loddenkemper C, Moter A, Stroux A, Marth T, Schneider T. Intravenous ceftriaxone, followed by 12 or three months of oral treatment with trimethoprim-sulfamethoxazole in Whipple's disease. *Journal of Infection*. 2013 Mar 1;66(3):263-70.
31. Foroutan M, Ghaffarifar F, Sharifi Z, Dalimi A. Vaccination with a novel multi-epitope ROP8 DNA vaccine against acute *Toxoplasma gondii* infection induces strong B and T cell responses in mice. *Comparative immunology, microbiology and infectious diseases*. 2020 Apr 1;69:101413.
32. Fournier PE, Thuny F, Richet H, Lepidi H, Casalta JP, Arzouni JP, Maurin M, Célard M, Mainardi JL, Caus T, Collart F. Comprehensive diagnostic strategy for blood culture-negative endocarditis: a prospective study of 819 new cases. *Clinical Infectious Diseases*. 2010 Jul 15;51(2):131-40.
33. Frank AC, Lobry JR. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics*. 2000 Jun 1;16(6):560-1.
34. Gerard A, Sarrot-Reynauld F, Liozon E, Cathebras P, Besson G, Robin C, Vighetto A, Mosnier JF, Durieu I, Rousset H. Neurologic presentation of Whipple disease: report of 12 cases and review of the literature. *Medicine*. 2002 Nov 1;81(6):443-57.
35. Ghigo E, Barry AO, Pretat L, Al Moussawi K, Desnues B, Capo C, Kornfeld H, Mege JL. IL-16 promotes *T. whipplei* replication by inhibiting phagosome

- conversion and modulating macrophage activation. *PloS one*. 2010 Oct 21;5(10):e13561.
36. Gorvel L, Textoris J, Banchereau R, Ben Amara A, Tantibhedhyangkul W, von Bargen K, Ka MB, Capo C, Ghigo E, Gorvel JP, Mege JL. Intracellular bacteria interfere with dendritic cell functions: role of the type I interferon pathway. *PloS one*. 2014 Jun 10;9(6):e99420.
37. Guan P, Doytchinova IA, Zygouri C, Flower DR. MHCpred: bringing a quantitative dimension to the online prediction of MHC binding. *Applied bioinformatics*. 2003 Jan 1;2:63-6.
38. Guérin A, Kerner G, Marr N, Markle JG, Fenollar F, Wong N, Boughorbel S, Avery DT, Ma CS, Bougarn S, Bouaziz M. IRF4 haploinsufficiency in a family with Whipple's disease. *Elife*. 2018 Mar 14;7:e32340.
39. Gupta N, Kumar A. Designing an efficient multi-epitope vaccine against *Campylobacter jejuni* using immunoinformatics and reverse vaccinology approach. *Microbial Pathogenesis*. 2020 Oct 1;147:104398.
40. Gurung RB, Purdie AC, Begg DJ, Whittington RJ. In silico identification of epitopes in *Mycobacterium avium* subsp. *paratuberculosis* proteins that were upregulated under stress conditions. *Clinical and Vaccine Immunology*. 2012 Jun 1;19(6):855-64.
41. Hopp TP, Woods KR. A computer program for predicting protein antigenic determinants. *Molecular immunology*. 1983 Apr 1;20(4):483-9.
42. Horváti K, Pályi B, Henczkó J, Balka G, Szabó E, Farkas V, Biri-Kovács B, Szeder B, Fodor K. A convenient synthetic method to improve immunogenicity of mycobacterium tuberculosis related T-cell epitope peptides. *Vaccines*. 2019 Sep;7(3):101.
43. Humphrey, W., Dalke, A. and Schulten, K., 1996. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1), pp.33-38.

44. Jain P, Joshi A, Akhtar N, Krishnan S, Kaushik V. An immunoinformatics study: designing multivalent T-cell epitope vaccine against canine circovirus. *Journal of Genetic Engineering and Biotechnology*. 2021 Dec;19(1):1-1.
45. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, Nielsen M. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. 2018 Jul;154(3):394-406.
46. Joshi A, Joshi BC, Mannan MA, Kaushik V. Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. *Informatics in Medicine Unlocked*. 2020 Jan 1;19:100338.
47. Joshi A, Krishnan S, Kaushik V. Codon usage studies and epitope-based peptide vaccine prediction against *Tropheryma whipplei*. *Journal of Genetic Engineering and Biotechnology*. 2022 Dec;20(1):1-2.
48. Kanampalliwar AM, Rajkumar S, Girdhar A, Archana T. Reverse Vaccinology: Basics and Applications 2013. *J Vaccines Vaccin* 4: 194. doi: 10.4172/2157-7560.1000 194 Page 2 of 5 Volume 4• Issue 6• 1000194 *J Vaccines Vaccin* ISSN: 2157-7560
49. Kazi A, Chuah C, Majeed AB, Leow CH, Lim BH, Leow CY. Current progress of immunoinformatics approach harnessed for cellular-and antibody-dependent vaccine design. *Pathogens and global health*. 2018 Apr 3;112(3):123-31.
50. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*. 2015 Jun;10(6):845-58.
51. Krishnan S, Joshi A, Akhtar N, Kaushik V. Immunoinformatics designed T cell multi epitope dengue peptide vaccine derived from non structural proteome. *Microbial Pathogenesis*. 2021 Jan 1;150:104728.
52. La MV, Crapoulet N, Barbry P, Raoult D, Renesto P. Comparative genomic analysis of *Tropheryma whipplei* strains reveals that diversity among clinical isolates is mainly related to the WiSP proteins. *BMC genomics*. 2007 Dec;8(1):1-1.

53. La Scola B, Fenollar F, Fournier PE, Altwegg M, Mallet MN, Raoult D. Description of *Tropheryma whippelii* gen. nov., sp. nov., the Whipple's disease bacillus. *International Journal of Systematic and Evolutionary Microbiology*. 2001 Jul 1;51(4):1471-9.
54. Lagier JC, Fenollar F, Lepidi H, Giorgi R, Million M, Raoult D. Treatment of classic Whipple's disease: from in vitro results to clinical outcome. *Journal of Antimicrobial Chemotherapy*. 2014 Jan 1;69(1):219-27.
55. Lagier JC, Fenollar F, Raoult D. Whipple's disease and *Tropheryma whippelii* infections in internal medicine. When to think about it? How to treat? *La Revue de medecine interne*. 2014 Jun 2;35(12):801-7.
56. Lefébure, T., & Stanhope, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome biology*, 8(5), 1-17.
57. Lipinski CA. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies*. 2004 Dec 1;1(4):337-41.
58. Lloyd-Williams P, Albericio F, Giralt E. Chemical approaches to the synthesis of peptides and proteins. CRC Press; 2020 Aug 18.
59. Mahdavi M, Keyhanfar M, Jafarian A, Mohabatkar H, Rabbani M. Immunization with a novel chimeric peptide representing B and T cell epitopes from HER2 extracellular domain (HER2 ECD) for breast cancer. *Tumor Biology*. 2014 Dec;35(12):12049-57.
60. MAIZEL H, RUFFIN JM, DOBBINS III WO. Whipple's disease: a review of 19 patients from one hospital and a review of the literature since 1950. *Medicine*. 1970 May 1;49(3):175-206.
61. Marth T, Moos V, Müller C, Biagi F, Schneider T. *Tropheryma whippelii* infection and Whipple's disease. *The Lancet Infectious Diseases*. 2016 Mar 1;16(3):e13-22.

62. Matthews BR, Jones LK, Saad DA, Aksamit AJ, Josephs KA. Cerebellar ataxia and central nervous system Whipple disease. *Archives of neurology*. 2005 Apr 1;62(4):618-20.
63. Misra N, Panda PK, Shah K, Sukla LB, Chaubey P. Population coverage analysis of T-Cell epitopes of *Neisseria meningitidis* serogroup B from Iron acquisition proteins for vaccine design. *Bioinformation*. 2011;6(7):255.
64. Moisa AA, Kolesanova EF. Synthetic peptide vaccines. *Insight and Control of Infectious Disease in Global Scenario*. 2012 Mar 21:201-28.
65. Mondal SI, Ferdous S, Jewel NA, Akter A, Mahmud Z, Islam MM, Afrin T, Karim N. Identification of potential drug targets by subtractive genome analysis of *Escherichia coli* O157: H7: an in-silico approach. *Advances and applications in bioinformatics and chemistry: AABC*. 2015;8:49.
66. Moos V, Kunkel D, Marth T, Feurle GE, LaScola B, Ignatius R, Zeitz M, Schneider T. Reduced peripheral and mucosal *Tropheryma whippelii*-specific Th1 response in patients with Whipple's disease. *The Journal of Immunology*. 2006 Aug 1;177(3):2015-22.
67. Moos V, Schmidt C, Geelhaar A, Kunkel D, Allers K, Schinnerling K, Loddenkemper C, Fenollar F, Moter A, Raoult D, Ignatius R. Impaired immune functions of monocytes and macrophages in Whipple's disease. *Gastroenterology*. 2010 Jan 1;138(1):210-20.
68. Mustafa AS. In silico analysis and experimental validation of *Mycobacterium tuberculosis*-specific proteins and peptides of *Mycobacterium tuberculosis* for immunological diagnosis and vaccine development. *Medical Principles and Practice*. 2013;22(Suppl. 1):43-51.
69. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*. 1986 Sep 1;3(5):418-26.

70. Özcan ÖÖ, Karahan M, Kumar PV, Tan SL, Tee YN. New generation peptide-based vaccine prototype. In *Current and Future Aspects of Nanomedicine 2019* Oct 7. IntechOpen.
71. Paulley JW. A case of Whipple's disease (intestinal lipodystrophy). *Gastroenterology*. 1952 Sep 1;22(1):128-33.
72. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*. 2005 Dec;26(16):1781-802.
73. Pouriayevali MH, Bamdad T, Parsania M, Sari RD. Full length antigen priming enhances the CTL epitope-based DNA vaccine efficacy. *Cellular immunology*. 2011 Jan 1;268(1):4-8.
74. Puéchal X. Whipple disease and arthritis. *Current opinion in rheumatology*. 2001 Jan 1;13(1):74-9.
75. Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM. *Tropheryma whippelii* Twist: a human pathogenic Actinobacteria with a reduced genome. *Genome Research*. 2003 Aug 1;13(8):1800-9.
76. Relman DA, Schmidt TM, MacDermott RP, Falkow S. Identification of the uncultured bacillus of Whipple's disease. *New England Journal of Medicine*. 1992 Jul 30;327(5):293-301.
77. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T-and B-cell epitope prediction. *Journal of immunology research*. 2017 Oct;2017.
78. Schneider T, Moos V, Loddenkemper C, Marth T, Fenollar F, Raoult D. Whipple's disease: new aspects of pathogenesis and treatment. *The Lancet infectious diseases*. 2008 Mar 1;8(3):179-90.
79. Seligmann H. Localized context-dependent effects of the "ambush" hypothesis: more off-frame stop codons downstream of shifty codons. *DNA and Cell Biology*. 2019 Aug 1;38(8):786-95.

80. Serruto D, Rappuoli R. Post-genomic vaccine development. *FEBS letters*. 2006 May 22;580(12):2985-92.
81. Sharma P, Sharma P, Ahmad S, Kumar A. Chikungunya Virus Vaccine Development: Through Computational Proteome Exploration for Finding of HLA and cTAP Binding Novel Epitopes as Vaccine Candidates. *International Journal of Peptide Research and Therapeutics*. 2022 Mar;28(2):1-5.
82. Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*. 1987 Feb 11;15(3):1281-95.
83. Shi Y, Chang D, Li W, Zhao F, Ren X, Hou B. Identification of core genes and clinical outcomes in tumors originated from endoderm (gastric cancer and lung carcinoma) via bioinformatics analysis. *Medicine*. 2021 Mar 26;100(12):e25154.
84. Singh KK, Chaubey G, Chen JY, Suravajhala P. Decoding SARS-CoV-2 hijacking of host mitochondria in COVID-19 pathogenesis. *American Journal of Physiology-Cell Physiology*. 2020 Aug 1;319(2):C258-67.
85. Skwarczynski M, Toth I. Peptide-based synthetic vaccines. *Chemical science*. 2016;7(2):842-54.
86. Sylvester-Hvid C, Nielsen M, Lamberth K, Røder G, Justesen S, Lundegaard C, Worning P, Thomadsen H, Lund O, Brunak S, Buus S. SARS CTL vaccine candidates—HLA supertype, genome-wide scanning and biochemical validation. *Scandinavian Journal of Immunology*. 2004 Apr;59(6):632-632.
87. Tang H, Liu X, Fang Y, Pan L, Zhang Z, Zhou P, Lv J, Jiang S, Hu W, Zhang P, Wang Y. The epitopes of foot and mouth disease. *Asian Journal of Animal and Veterinary Advances*. 2012;7(12):1261-5.
88. Tanhaieian A, Sekhavati MH, Ahmadi FS, Mamarabadi M. Heterologous expression of a broad-spectrum chimeric antimicrobial peptide in *Lactococcus lactis*: its safety and molecular modeling evaluation. *Microbial pathogenesis*. 2018 Dec 1;125:51-9.

89. Tettelin H. The bacterial pan-genome and reverse vaccinology. In *Microbial Pathogenomics 2009* (Vol. 6, pp. 35-47). Karger Publishers.
90. Thévenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tuffery P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic acids research*. 2012 May 11;40(W1):W288-93.
91. Tolfvenstam T, Lindblom A, Schreiber MJ, Ling L, Chow A, Ooi EE, Hibberd ML. Characterization of early host responses in adults with dengue disease. *BMC infectious diseases*. 2011 Dec;11(1):1-7.
92. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*. 2010 Jan 30;31(2):455-61.
93. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. *Journal of computational chemistry*. 2005 Dec;26(16):1701-18.
94. Van La M, Barbry P, Raoult D, Renesto P. Molecular basis of *Tropheryma whipplei* doxycycline susceptibility examined by transcriptional profiling. *Journal of antimicrobial chemotherapy*. 2007 Mar 1;59(3):370-7.
95. Whipple GH. A hitherto undescribed disease characterized anatomically by deposits of fat and fatty acids in the intestinal and mesenteric lymphatic tissues. *Bull Johns Hopkins Hosp*. 1907; 18:382-91.
96. Wilson KH, Frothingham R, Wilson JA, Blichington R. Phylogeny of the Whipple's-disease-associated bacterium. *The Lancet*. 1991 Aug 24;338(8765):474-5.
97. Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990 Mar 1;87(1):23-9.
98. YARDLEY J, HENDRIX T. Combined Electron and Light Microscopy in Whipple's Disease. Demonstration of " Bacillary Bodies" in the Intestine. *Bulletin of the Johns Hopkins Hospital*. 1961 Aug; 109:80-98.

99. Zhang Y, Center DM, David MH, Cruikshank WW, Yuan J, Andrews DW, Kornfeld H. Processing and activation of pro-interleukin-16 by caspase-3. *Journal of Biological Chemistry*. 1998 Jan 9;273(2):1144-9.



In-Silico Proteomic Exploratory Quest: Crafting T-Cell Epitope Vaccine Against Whipple's Disease

Amit Joshi¹ · Vikas Kaushik¹

Accepted: 12 May 2020
© Springer Nature B.V. 2020

Abstract

Whipple's disease is one of the rare maladies in terms of spread but very fatal one as it is linked with many disorders (like Gastroenteritis, Endocarditis etc.). Also, current regimens include less effective drugs which require long duration follows up. This exploratory study was conducted to commence the investigation for crafting multi target epitope vaccine against its bacterial pathogen *Tropheryma whipplei*. The modern bioinformatics tools like VaxiJen, NETMHCII PAN 3.2, ALLERGEN-FP, PATCH-DOCK, TOXIC-PRED, MHC-PRED and IEDB were deployed, which makes the study more intensive in analyzing proteome of *T. whipplei* as these methods are based on robust result generating statistical algorithms ANN, HMM, and ML. This Immuno-Informatics approach leads us in the prediction of two epitopes: VLMVSAFPL and IRYLAALHL interacting with 4 and 6 HLA DRB1 alleles of MHC Class II respectively. VLMVSAFPL epitope is a part of DNA-directed RNA polymerase subunit beta, and IRYLAALHL epitope is a part of membranous protein insertase YidC of this bacterium. Molecular-Docking and Molecular-Simulation analysis yields the perfect interaction based on Atomic contact energy, binding scores along with RMSD values (0 to 1.5 Å) in selection zone. The IEDB (Immune epitope database) population coverage analysis exhibits satisfactory relevance with respect to world population.

Keywords Epitopes · *Tropheryma whipplei* · Population coverage · Alleles · Simulation · Vaccine

Introduction

George Hoyt Whipple in 1907 explains Whipple's disease, as a multisystemic chronic infectious disease. He identified silver stained rod shaped bacterium in vacuoles associated with macrophages of patients, he initially did not think of them as the cause for the disease rather he think that intestinal lipodystrophy (Whipple's disease) was caused due to some novel disturbances in fat metabolic schemes (Whipple 1907). When the first successful treatment started by using antibiotics in 1952, determined that this bacterium might be the major causative agent of this disease (Paulley 1952). An electron microscopic study in 1960's provided additional support for this hypothesis (Cohen et al. 1960;

Yardley and Hendrix 1961). Whipple's disease occurs uncommonly, as a multisystemic disorder (inexact annual frequency less than 1 per 1,000,000 populace) that specially affects middle-aged Caucasian men (Fenollar et al. 2007; Ramharter et al. 2014, Dobbins et al. 1981). This bacterium was found to mostly affect small children (Keita et al. 2015) and sewage workers (Schöniger-Hekele et al. 2007). Since, its first portrayal by Whipple in very beginning of first decade in twentieth century (Whipple 1907), a limited progresses with in pathogenesis, prognosis, and treatment of the malady have been made. The bacterium gets internalized in to lamina propria of intestine and then make its way to mucosal macrophages, as this bacterium induces the decreased expression of CD11b in such macrophages (CD11b on macrophages frequently mediates the intracellular degradation of bacteria) causes flip in the scenario (inappropriate antigen presentation by such macrophages and dendritic cells). This specially reasons the boom in IL-10, TGF- β and CCL-18 expression and decrease in IFN- γ , which in turn causes destroy in maturation of phagosomes and decrease in thioredoxin expression, lead them unable to kill bacterium and antigen presentation (Moss et al. 2006, 2010).

✉ Vikas Kaushik
vikas.14664@lpu.co.in

Amit Joshi
amit34655@gmail.com

¹ Domain of Bioinformatics, School of Bio-Engineering and Bio-Sciences, Lovely Professional University, Punjab, India

An unseemly development of proficient antigen-presenting cells caused by the presence of interleukin 10 and interleukin 16, and the non appearance of interferon γ and interleukin 12 might lead to inadequate antigen-presentation and hinder the incitement of antigen-specific T-helper 1 cells enhancing growth and systemic spread of *Tropheryma whipplei*. The nearby generation of provocative cytokines through macrophages and endothelial cells within the fringe might actuate lymphocyte invasion through a defective endothelial obstruction taken after by central aggravation, indeed in immunologically ensured tissues such as joints or the neuronal domain (Schneider et al. 2008). Currently hydroxychloroquine (600 mg/day) and doxycycline (200 mg/day) used for treatment of whipple's disease for 12–18 months, but life time follow up is required (Lagier et al. 2014), so it is time consuming treatment process and only few handful trials were conducted in earlier studies (Feurle et al. 2013). Nowadays epitope based vaccines provide better options in search of good treatment strategy for such type of harmful and rare malady, even if the individuals are genetically predisposed as in case of classical Whipple's disease (Trotta et al. 2017). This modern approach of putative vaccine determination which involves the use of proteomic databases is very handy and easy to use method not only for rare bacterial pathogens, but also very effective in case of harmful viruses like Nipah (Kaushik 2019). *Tropheryma whipplei* was found to be associated with major ailments like gastroenteritis and endocarditis (Fenollar et al. 2013).

In this research work, five proteins from proteomic data of *T. whipplei* were analyzed for allergenicity. Non-allergenic proteins were deployed for predicting epitopes. Predicted epitopes were subjected for immunogenic properties, structural modeling and the docking with corresponding MHC II alleles to investigate the strong binding affinity. Method is more economic, time efficient, and harmless when compared to the vaccine designing and testing in wet lab and animal

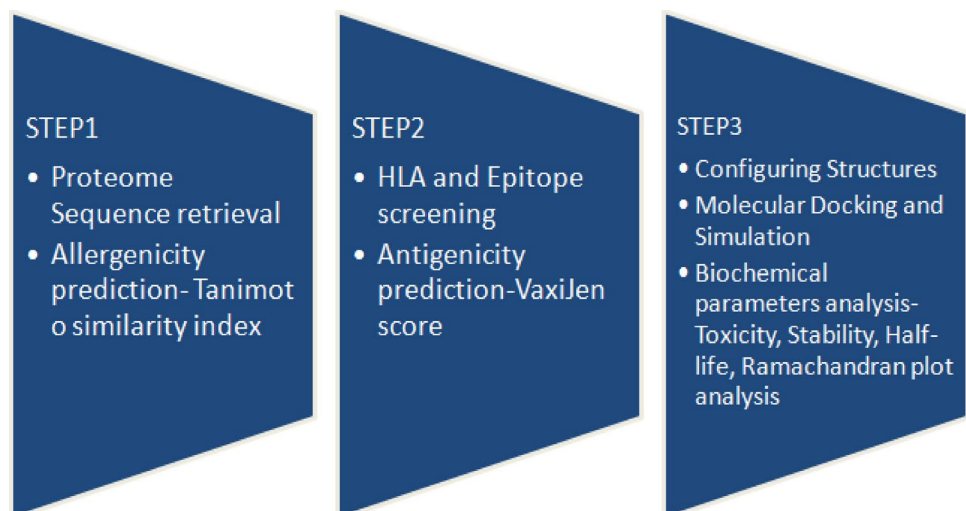
testing strategies (Kumar et al. 2015). Reverse vaccinology is the suitable approach as well as novel science method that use the genomic data with the utilization of computer for the arrangement of antibodies without culturing bacterium species (Kanampalliwar et al. 2013; Tang et al. 2012). It allow the choice in hands of human interface for selecting antigens from pathogenic set of DNA and most antigenic areas could be used to synthesize potential immunization to initiate defensive responses against such pathogenic species (Ada et al. 2018). Epitopes based antibodies selection and production is explicitly less time consuming, economical and considered safest approach in vaccine designing. Earlier computational methods were found to be successful in analyzing genome and prediction of putative drugs for *T. whipplei* (Palanisamy 2018), such studies provide motivation to craft vaccine targets by deploying in-silico approach. T-cell epitopes were screened out in this study may effectively elicit immune responses against this bacterium, and also similar type of recent study was found to be successful in determining epitope based vaccine agents for SARS-Cov2 (Joshi et al. 2020). Brief flow chart of the study used to determine putative epitope based vaccine candidates against *T. whipplei* is presented in Fig. 1.

Methodology

Retrieval of Proteins for *T. whipplei*

Proteomes were retrieved in fasta format from NCBI-Genbank and UniProtKB databases. Five proteins of different functionality were selected with following accession no's: WP_042507409.1 DNA-directed RNA polymerase subunit beta (RPO-B), WP_033800049.1 co-chaperone GroES, WP_038104819.1 TerC/Alx family metal homeostasis membrane protein, WP_042505650.1 membrane protein

Fig. 1 Flow chart of the study used to determine putative epitope



insertase YidC, WP_042505746.1 murein biosynthesis integral membrane protein. This selection depicts the variability to include crucial proteins of pathogenic domain (Table 1).

Allergenicity Prediction for Proteins

The protein sequences were then deployed for further analysis based on Allergen FP V 1.0 for predicting allergenicity (Dimitrov et al. 2014).

T-Cell Epitope Prediction

Net MHCII PAN 3.2 server is used to find and screen out HLA alleles which have good interaction with selected non-allergens of pathogenic origin (Jensen et al. 2018). To bring higher confidence in selecting epitope, VaxiJen server is deployed to determine antigenicity with threshold ≥ 0.7 for selected rare bacterium (Doytchinova et al. 2007). By subjecting proteomic sequences to Net MHCII PAN 3.2 server we obtained 1147 epitopes for WP_042507409.1, 90 epitopes for WP_033800049.1, 309 epitopes for WP_038104819.1, 302 epitopes for WP_042505650.1, and 510 epitopes for WP_011096746.1, this server was used because of its neural networking algorithm based approaches for fine predictions. $1-\log_{50k}$ (affinity score) ≤ 0.6 is used to screen out possible epitopes presented in Table 2. These epitopes were further subjected for antigenicity analysis based on VaxiJen scores.

Prediction of Epitope's Toxic Behavior

Toxicity for putative peptides was designated by using SVM scores from Toxin Pred web server (Gupta et al. 2013). Non toxic peptides were finalized for further analysis.

Table 1 Allergenicity results for analyzed proteins of *T. whipplei*

S. no	UniProtKB Accession No.	GenBank Accession No.	Allergen FP Score (Tanimoto similarity index)
1	Q76L83	WP_042507409.1	0.85 NON ALLER-GEN
2	Q96P47	WP_033800049.1	0.81 NON ALLER-GEN
3	P00846	WP_038104819.1	0.91 NON ALLER-GEN
4	O14569	WP_042505650.1	0.83 NON ALLER-GEN
5	Q7XBS0	WP_011096746.1	0.89 NON ALLER-GEN

Molecular Structural Modeling for Epitopes and Probable HLA Alleles

The tertiary structure or 3D structure for epitope is determined by using PEP-FOLD 3 web server (Lamiabile et al. 2016; Shen et al. 2014; Thévenet et al. 2012). And predicted Human leukocyte antigen alleles 3D structure was obtained from RCSB PDB database (Berman et al. 2000). Also Ramachandran plot analysis was conducted for verification of results by using Molprobit server (Williams et al. 2018).

Epitopes 3D Interaction or Molecular Docking with HLA Alleles

The docking experiments was conducted by using Patch-Dock tool (Schneidman-Duhovny et al. 2005), The predicted docked models of putative epitope and HLA alleles was selected on the basis of score, which relies on highest geometric shape complementarities and Atomic contact energy (Zhang et al. 1997). This allows the best selection of epitope and HLA allele interaction. This tool is easy to deploy for all life science domains.

Analysis of Population Coverage

Immune Epitope Database (IEDB) analysis Resource tool of population coverage was used to predict population coverage of the putative epitopes that are exhibiting interaction to HLA alleles and based on MHC-II restriction data (Bui et al. 2006). MHCpred tool was deployed for quantitative prediction of selected epitopes interacting to major Histocompatibility complexes (Guan et al. 2003).

Molecular Docking Simulations

Epitope-HLA allele docked sets were then used for simulation and dynamics analysis by deploying NAMD (Phillips et al. 2005) associated with VMD (Visual Molecular Dynamics) tool (Humphrey et al. 1996).

Results

Non-allergen Determination

Total 5 protein sequences were analyzed for allergenicity and depicted as non-allergen in Table 1 by using AllergenFP tool.

T-Cell Epitope Prediction

Net MHCII PAN 3.2 server is deployed to identify promiscuous epitopes and probable HLA alleles of MHC Class II

Table 2 List of predicted epitopes based on NetMHCII 3.2 server and VaxiJen score (threshold value of 0.7 and above was selected)

Protein ID	Allele (HLA)	POS	Peptide	1-log50k(aff)	VaxiJen	Antigen/non antigen	
Q76L83	DRB1_0101	328	IRYLAALHL	0.581	0.9461	Antigen	
	DRB1_0103	328	IRYLAALHL	0.311	0.9461	Antigen	
	DRB1_0301	801	LSAEERLLR	0.271	0.2947	Non antigen	
	DRB1_0401	225	FLRAIGMTD	0.292	-0.3463	Non antigen	
	DRB1_0404	328	IRYLAALHL	0.375	0.9461	Antigen	
	DRB1_0405	328	IRYLAALHL	0.336	0.9461	Antigen	
	DRB1_0406	328	IRYLAALHL	0.255	0.9461	Antigen	
	DRB1_0701	328	IRYLAALHL	0.496	0.9461	Antigen	
	DRB1_0802	324	IIATIRYLA	0.284	-0.5053	Non antigen	
	DRB1_1101	1010	YMYVLKLHH	0.451	1.2693	Antigen	
	DRB1_1302	154	FVINGTERV	0.445	-0.7514	Non antigen	
	Q96P47	DRB1_0101	84	YTIASRDV	0.410	0.2401	Non antigen
		DRB1_0406	86	ILASRDVLA	0.200	-0.1277	Non antigen
P00846	DRB1_0101	222	FFSLTGLRQ	0.478	-0.0649	Non antigen	
	DRB1_0103	244	YMKFGVAAL	0.194	0.3077	Non antigen	
	DRB1_0301	152	GLLDKVMIR	0.249	0.6173	Non antigen	
	DRB1_0401	198	MFALDSIPA	0.343	0.4086	Non antigen	
	DRB1_0404	295	IIALSVALS	0.339	0.8054	Antigen	
	DRB1_0405	222	FFSLTGLRQ	0.350	-0.0649	Non antigen	
	DRB1_0406	295	IIALSVALS	0.237	0.8054	Antigen	
	DRB1_0701	87	FRFAVPEIF	0.420	1.2093	Antigen	
	DRB1_0802	18	MLVTVRRPA	0.314	-0.5928	Non antigen	
	DRB1_0901	87	FRFAVPEIF	0.407	1.2093	Antigen	
	DRB1_0901	244	YMKFGVAAL	0.386	0.3077	Non antigen	
	DRB1_1001	222	FFSLTGLRQ	0.477	-0.0649	Non antigen	
	DRB1_1101	18	MLVTVRRPA	0.402	-0.5928	Non antigen	
	DRB1_1302	159	IRMNVSKNY	0.420	0.7822	Antigen	
	DRB1_1602	222	FFSLTGLRQ	0.335	-0.0649	Non antigen	
O14569	DRB1_0101	175	FYALQAGQA	0.539	0.6085	Non antigen	
	DRB1_0301	278	LAFELRRKR	0.220	0.7910	Antigen	
	DRB1_0401	175	FYALQAGQA	0.284	0.6085	Non antigen	
	DRB1_0404	8	FLQNILLPI	0.299	0.1551	Non antigen	
	DRB1_0405	8	FLQNILLPI	0.341	0.1551	Non antigen	
	DRB1_1001	175	FYALQAGQA	0.467	0.6085	Non antigen	
	DRB1_1101	63	FLKQIRAQR	0.438	-0.0522	Non antigen	
	DRB1_1302	8	FLQNILLPI	0.478	0.1551	Non antigen	
Q7XBS0	DRB1_0101	374	YILQKAFYA	0.556	0.3071	Non antigen	
	DRB1_0103	374	YILQKAFYA	0.274	0.3071	Non antigen	
	DRB1_0401	374	YILQKAFYA	0.323	0.3071	Non antigen	
	DRB1_0404	334	VLMVSAFPL	0.350	1.2114	Antigen	
	DRB1_0405	374	YILQKAFYA	0.312	0.3071	Non antigen	
	DRB1_0406	334	VLMVSAFPL	0.250	1.2114	Antigen	
	DRB1_0701	334	VLMVSAFPL	0.539	1.2114	Antigen	
	DRB1_0802	435	FLAIRVCLG	0.274	1.1141	Antigen	
	DRB1_0901	334	VLMVSAFPL	0.436	1.2114	Antigen	
	DRB1_1001	374	YILQKAFYA	0.473	0.3071	Non antigen	
	DRB1_1101	503	YFLVITRCR	0.416	-0.2632	Non antigen	
	DRB1_1302	334	VLMVSAFPL	0.412	1.2114	Antigen	
DRB1_1602	374	VLMVSAFPL	0.366	0.3071	Non antigen		

that interacts together by analyzing their 1-log50k values and binding affinities, then the VaxiJen scores were used with threshold of ≥ 0.7 with all informative details are presented in Table 2.

Molecular 3D Modeling of Epitopes and HLA Alleles

3D structural models of selected epitopes were designed by using PEP-FOLD 3 web server and than most common HLA

Table 3 HLA template model based on Pdb Id derived for MHC Class II alleles structure from RCSB-PDB

Allele name	Template structure (PDB ID)
HLA-DRB1*01_01	4AH2
HLA-DRB1*01_03	3PDO
HLA-DRB1*03_01	1A6A
HLA-DRB1*04_01	5LAX
HLA-DRB1*04_04	4IS6
HLA-DRB1*04_05	4IS6
HLA-DRB1*04_06	4IS6
HLA-DRB1*07_01	3C5J
HLA-DRB1*08_02	3PDO
HLA-DRB1*09_01	1BX2
HLA-DRB1*10_01	3PDO
HLA-DRB1*11_01	6CPL
HLA-DRB1*13_02	1FV1
HLA-DRB1*16_02	6CPO

Table 4 Molecular docking results of screened epitopes with HLA alleles

Allele	Epitope	Binding score	Ace	Epitope selection
DRB1_0101 (4AH2)	IRYLAALHL	7704	- 367.28	Selected
DRB1_0103 (3PDO)	IRYLAALHL	7452	- 321.11	Selected
DRB1_0404 (4IS6)	IRYLAALHL	8064	- 353.51	Selected
DRB1_0405 (4IS6)	IRYLAALHL	8064	- 353.51	Selected
DRB1_0406 (4IS6)	IRYLAALHL	8064	- 353.51	Selected
DRB1_0701 (3C5J)	IRYLAALHL	7544	- 349.18	Selected
DRB1_1101 (6CPL)	YMYVLKHH	7880	- 288.29	Rejected
DRB1_0404 (4IS6)	IIALSVALS	6562	- 136.21	Rejected
DRB1_0406 (4IS6)	IIALSVALS	6562	- 136.21	Rejected
DRB1_0701 (3C5J)	FRFAVPEIF	8260	- 223.61	Rejected
DRB1_0901 (1BX2)	FRFAVPEIF	9042	- 127.68	Rejected
DRB1_1302 (1FV1)	IRMVSKNY	8564	344.23	Rejected
DRB1_0301 (1A6A)	LAFELRRKR	7746	- 93.04	Rejected
DRB1_0404 (4IS6)	VLMVSAFPL	6788	- 443.31	Selected
DRB1_0406 (4IS6)	VLMVSAFPL	6788	- 443.31	Selected
DRB1_0701 (3C5J)	VLMVSAFPL	6868	- 226.35	Selected
DRB1_0802 (3PDO)	FLAIRVKLG	7812	- 298.60	Rejected
DRB1_0901 (1BX2)	VLMVSAFPL	7960	- 297.99	Selected
DRB1_1302 (1FV1)	VLMVSAFPL	8500	104.94	Rejected

DRB1 proteins structural models were derived by using RCSB-PDB database. In Table 3 PDB Id along with HLA alleles is exhibited. Molprobit results of Ramachandran plot analysis results shows satisfactory structural prediction ($> 85\%$ residues in favorable region) of epitopes that were finalized at last in Fig. 8.

Molecular Docking of Epitopes and HLA Alleles

PatchDock tool was deployed for interaction between selected structures of epitopes and HLA DRB1 proteins. Then interaction data produced by docked molecules include ACE (Atomic contact energy) and best model score that leads to the final selection in the way of prediction for each pair. In Table 4 the selected models and rejected models both were included to enhance the comparative analysis. The two selected epitopes were VLMVSAFPL and IRYLAALHL interacting with 4 and 6 HLA DRB1 alleles respectively. VLMVSAFPL epitope is a part of DNA-directed RNA polymerase subunit beta and IRYLAALHL epitope is a part of murein biosynthesis integral membrane protein of *T. whipplei* and are major identifiers of this bacterium. Figure 2 clearly depicts the good interaction between epitopes and HLA Alleles in docked results. In Fig. 2a Docked result of IRYLAALHL with HLA-DRB1* 01:01 exhibits perfect hydrogen bond due to presence of tyrosine residue in epitope at 3rd position, while most of the other non polar amino acids of this epitope are depicts vander waals interactions with in the HLA model and in Fig. 2c Docked result of VLMVSAFPL with HLA-DRB1* 04:04 exhibits perfect

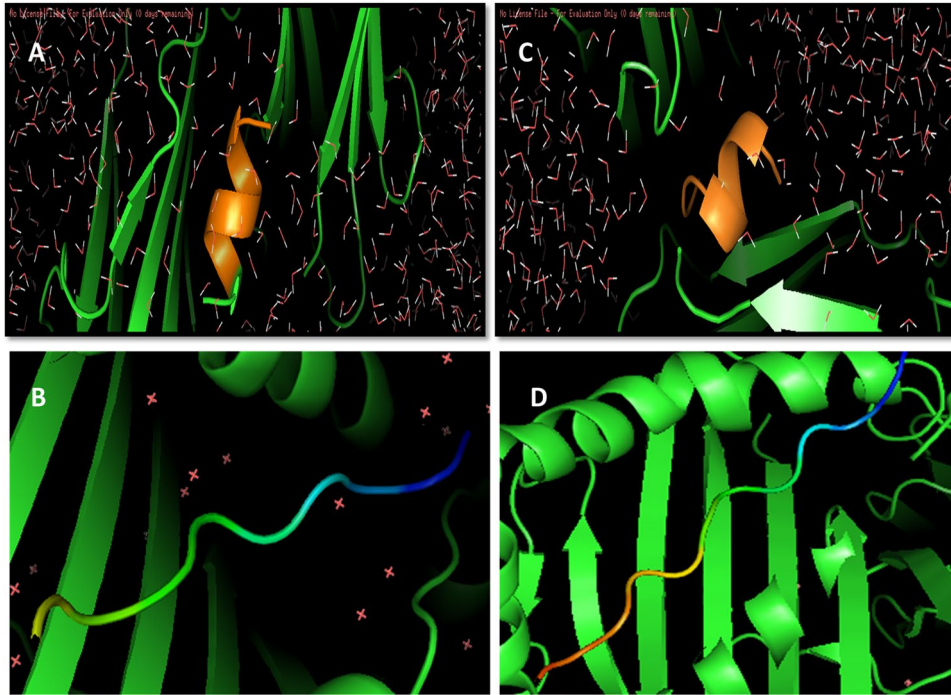


Fig. 2 **a** Docked result of IRYLAALHL with HLA-DRB1* 01:01 exhibits perfect hydrogen bond due to presence of tyrosine residue in epitope at 3rd position, while most of the other non polar amino acids of this epitope are depicting Vander waals interactions with in the HLA model. **b** MKMRMATPLLMQAL interacting with HLA-DRB1*01:01 (4AH2) structure considered as reference. **c** Docked

result of VLMVSAFPL with HLA-DRB1* 04:04 exhibits perfect hydrogen bond due to presence of serine residue in epitope at 5th position, while most of the other non polar amino acids of this epitope are depicting Vander waals interactions with in the HLA model. **d** RQLYPEWTEAQR epitope interacting with HLA-DRB* 04:04 (4IS6) structure considered as reference

hydrogen bond due to presence of serine residue in epitope at 5th position, while most of the other non polar amino acids of this epitope are depicts Vander waals interactions with in the HLA model. In Fig. 2b, d both RCSB-PDB structures for selected HLA DRB1 alleles (i.e. 4AH2 and 4IS6) with predicted peptide considered as reference structure (Schlundt et al. 2012; Chen et al. 2013). The reference docked peptides have great difference in amino acid sequence in comparison to our screened epitopes but exhibits some resemblance alike of our epitopes in interaction towards antigen binding pocket. Figure 3a, b represents the free undocked HLA-DRB1 receptors (4AH2, 4IS6 respectively), while Fig. 3c, d represents free unbound putative epitopes (IRYLAALHL, VLMVSAFPL respectively) and their side chains. Figure 4 graphically represents the selected epitopes and HLA alleles of MHC II on the basis of ACE values.

Toxicity, Half Life, and Stability Analysis of Putative Epitopes

Predicted epitopes VLMVSAFPL and IRYLAALHL have VaxiJen scores 0.9461 and 1.2114 respectively, they are also of non toxic nature as per the study of Toxin Pred tool and its toxicity scores (SVM score) represented in Table 5. In

Table 6 quantitative estimation of best interaction of epitope with HLADRB1 alleles were achieved with upright IC_{50} values by using MHCpred tool, this allows confidence of prediction. Table 7 shows Half-life and instability index for putative epitopes by deploying ProtParam expasy tool.

Population Coverage Analysis of Epitopes

VLMVSAFPL and IRYLAALHL manifest 28.82% and 37.06% elicitation of immune responsiveness by world population by availing IEDB tool. The epitopes VLMVSAFPL and IRYLAALHL shows greater effect in European population by 29.63% and 42.68% respectively, and correspondingly similar results with North American population coverage analysis. This indicates its greater relevance in treatment of Whipple's disease as it is mostly seen in Caucasoid population. In Figs. 5 and 6 it is clearly represented in a graphical representation.

Molecular Dynamic Simulation Studies

NAMD was deployed for simulation studies on docked Epitope—HLA allele sets to obtain RMSD values. Maximum value of RMSD for VLMVSAFPL and IRYLAALHL

Fig. 3 Free HLA-DRB1 receptors and putative epitopes structure: **a** 4AH2 **b** 4IS6 **c** IRYLAALHL **d** VLMVSAFPL

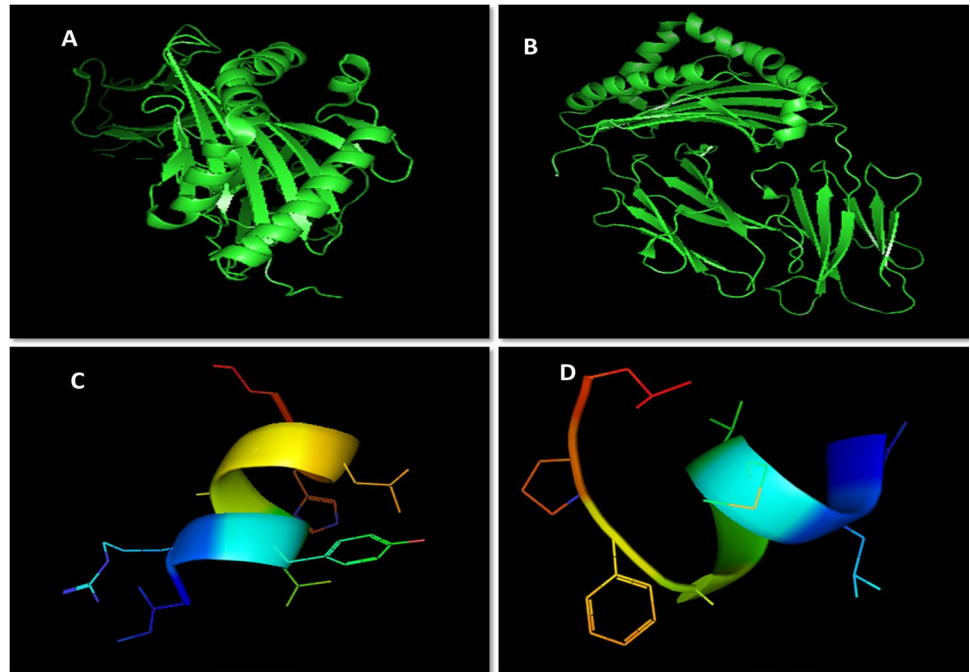


Fig. 4 Graphical representation of atomic contact energy (ACE) for docked complexes of putative epitopes and HLA DRB1 alleles

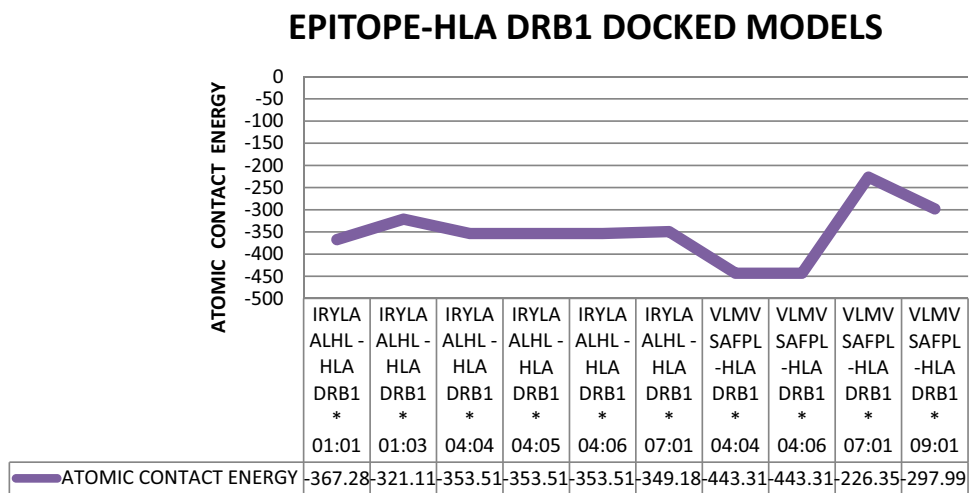


Table 5 Toxicity scores for putative epitopes

Epitope	No. of HLA binders	Toxicity score	Toxicity
VLMVSAFPL	4	- 1.19	Non toxin
IRYLAALHL	6	- 0.61	Non toxin

epitopes were analyzed, this gives more confidentiality in selection of vaccine candidate against *Tropheryma whipplei*. Figures 7 and 8 shows RMSD plots that indicates clear picture of selection of these two epitopes.

Discussion

Immuno-informatics is the suitable approach as well as novel science method that use the proteomic data with the utilization of computer systems for predicting epitopes without culturing bacterium species (Kanampalliwar et al. 2013; Tang et al. 2012). It allow the choice in hands of human interface for selecting antigens from pathogenic set of DNA and most antigenic areas could be used to synthesize potential immunization to initiate defensive responses against harmful pathogenic species (Ada et al. 2018). In-silico approach was earlier successful in case of *Staphylococcus aureus* (Delfani et al. 2015), *Mycobacterium*

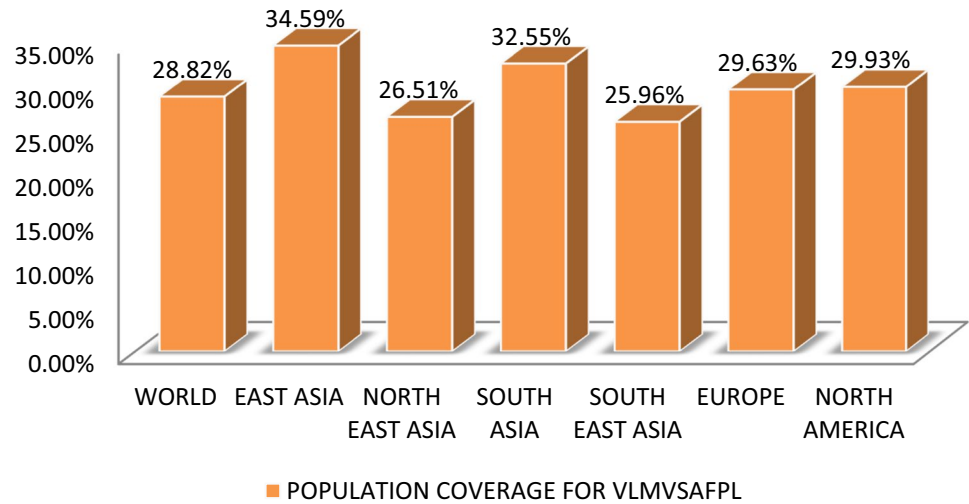
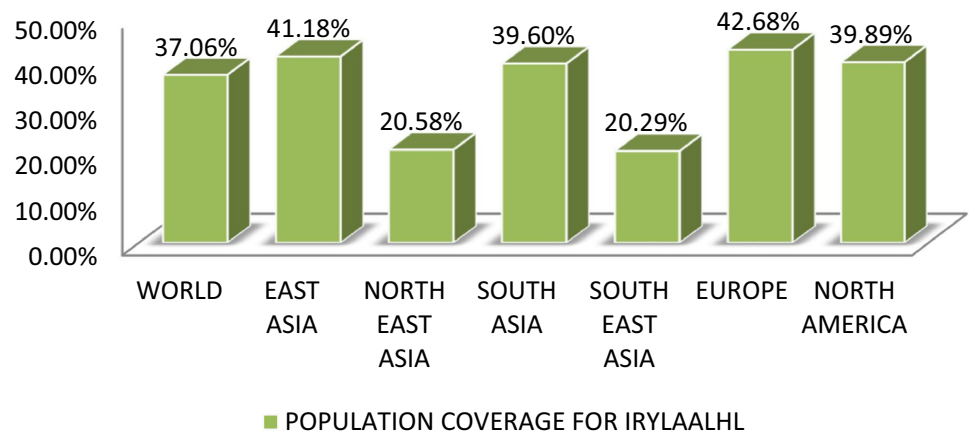
Table 6 MHC-PRED results for interacting epitopes and HLA DRB1 alleles

Epitope	HLA allele used in test	Predicted $-\log IC_{50}$ (M)	Predicted IC_{50} value (nM)	Confidence of prediction (Max = 1)
IRYLAALHL	DRB0101	8.598	2.52	0.89
IRYLAALHL	DRB0701	5.935	1161.45	1.00
VLMVSAFPL	DRB0701	5.954	1111.73	1.00

Only two alleles were present in this database for MHC II i.e. DRB0101 and DRB0701

Table 7 ProtParam tool used for predicting biochemical parameters (GRAVY, half-life, and instability index)

Epitope	Theoretical PI	Half life (for mammalian reticulocytes)	Instability index (percentage)	GRAVY
IRYLAALHL	8.75	20 h	0.51 (stable)	1.167
VLMVSAFPL	5.49	100 h	30.29 (stable)	2.233

Fig. 5 Graphical representation of Population coverage for VLMVSAFPL**Fig. 6** Graphical representation of population coverage for IRYLAALHL

tuberculei (Mustafa 2013) and numerous bacterial species, but *T. whipplei* is still not fully explored in this domain. Current regimens include hydroxychloroquine

and doxycycline for treatment of Whipple's disease for 12–18 months, but life time follow up is required (Lagier et al. 2014), so it is time consuming treatment process and

RMSD VALUE FOR IRYLAALHL

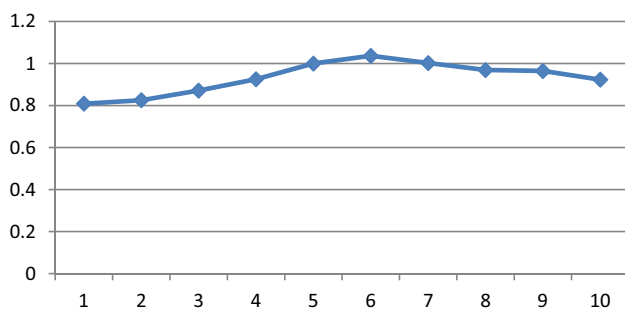


Fig. 7 Graphical representation of RMSD values for epitope IRYLAALHL

RMSD VALUE FOR VLMVSAPFL

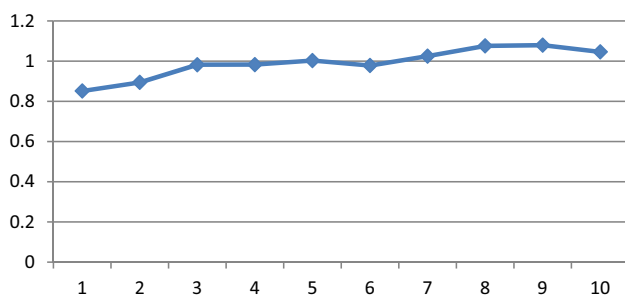


Fig. 8 Graphical representation of RMSD values for epitope VLMVSAPFL

only few handful trials were conducted in earlier studies (Feurle et al. 2013).

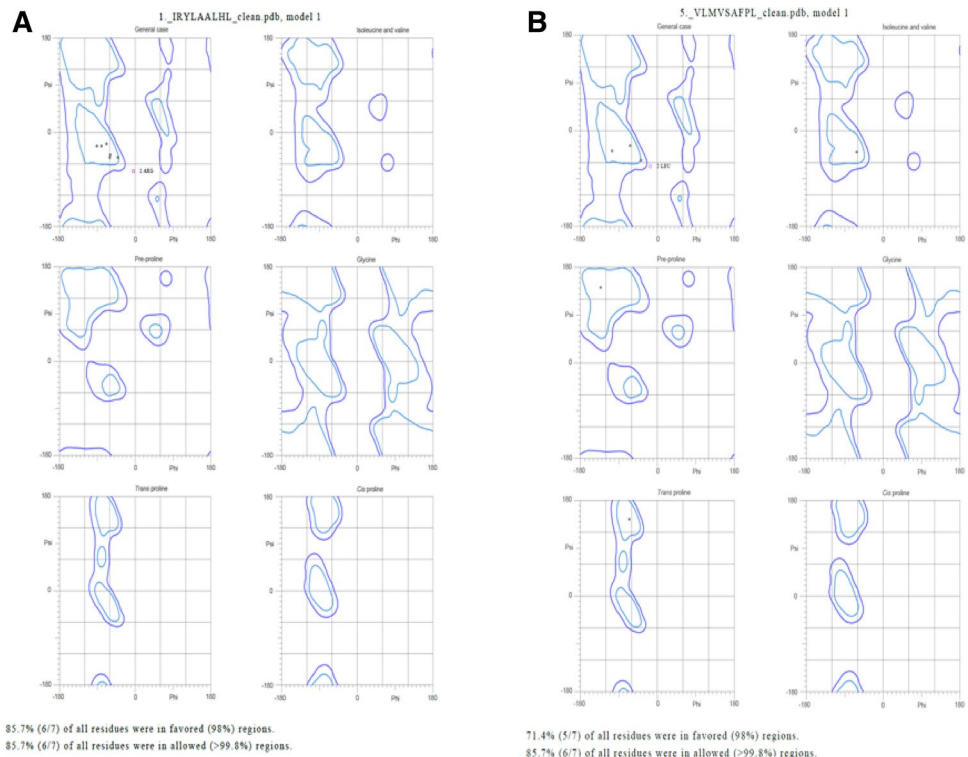
In present study we identified two possible epitopes that can interact with MHC-II alleles to elicit immune response on individuals namely VLMVSAPFL epitope (part of DNA-directed RNA polymerase subunit beta), and IRYLAALHL epitope (part of murein biosynthesis integral membrane protein). These epitopes exhibit better interaction with HLA DRB1 alleles, as confirmed by deploying Molecular-Docking and Molecular-Simulation studies (Adhikari et al. 2018). Population coverage analysis was found to be satisfactory and in earlier studies it was used in strengthening vaccine prediction aspects (Misra et al. 2011). Very similar

studies were also conducted successfully for related bacterium *Mycobacterium avium* and found to be successful in predicting epitopes (Gurung et al. 2012). The epitope VLMVSAPFL was found to interact with 4 HLA alleles of HLA-DRB1 domain (04:04, 04:06, 07:01, 09:01) with satisfactory ACE values (− 443.3, − 443.3, − 226.3, − 297.9 respectively); and the epitope IRYLAALHL found to interact with 6 alleles of HLA-DRB1 domain (01:01, 01:03, 04:04, 04:05, 04:06, 07:01) with satisfactory ACE values (− 367.2, − 321.1, − 353.5, − 353.5, − 353.5, − 349.1 respectively) in docking results similar type of methodology was seen in recent studies in screening epitopes for SARS-Cov-2 (Joshi et al. 2020). Both selected epitopes exhibit structural integrity as possess less than 35% instability index score, and half life greater than 20 h for mammalian reticulocytes, this makes the screening criteria more reliable. Also, more than 85% residues of selected epitopes come under favorable region in Ramachandran plot analysis (Fig. 9). Still no one has used vaccine based treatments for Whipple's disease, as it is thought to be rare and possess reduced genome but considered one of the harmful pathogen of human (Raoult et al. 2003; La Scola et al. 2001; Marth et al. 2016). The effectiveness of epitope based vaccines for treatment of endocarditis has already been claimed (Priyadarshini et al. 2014). But in our study we found the short peptides that can easily be synthesized and deployed in developing immunity in Caucasian populations against Whipple's disease.

Conclusion

In this study we obtained VLMVSAPFL and IRYLAALHL as predicted epitopes for vaccine crafting. This novel approach in crafting vaccine based treatment of *T. whipplei* will open new doors in research for creating regimens to treat such harmful bacterium by developing adaptive immune response and eradicating it globally before any future escalations takes place. The predicted epitopes can be deployed in crafting vaccines against *T. whipplei* bacterium after Molecular-wet lab corroboration.

Fig. 9 Ramachandran plot for putative epitopes **a** IRY-LAALHL **b** VLMVSAFPL



References

- Ada K, Candy C, Abu B, Abdul M, Chuan CH, Boon HL, Chuan YL (2018) Current progress of immunoinformatics approach harnessed for cellular and antibody-dependent vaccine design. *Pathog Glob Health* 112:3
- Adhikari UK, Tayebi M, Rahman MM (2018) Immunoinformatics approach for epitope-based peptide vaccine design and active site prediction against polyprotein of emerging *Oropouche virus*. *J Immunol Res*. <https://doi.org/10.1155/2018/6718083>
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
- Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A (2006) Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinform* 7(1):153
- Chen S, Li Y, Depontieu FR, McMiller TL, English AM, Shabanowitz J, Kos F, Sidney J, Sette A, Rosenberg SA, Hunt DF, Mariuzza RA, Topalian SL (2013) Structure-based design of altered MHC class II-restricted peptide ligands with heterogeneous immunogenicity. *J Immunol* 191:5097–5106. <https://doi.org/10.4049/jimmunol.1300467>
- Cohen AS, Schimmel EM, Holt PR, Isselbacher KJ (1960) Ultrastructural abnormalities in Whipple's disease. *Proc Soc Exp Biol Med* 105:411–414
- Delfani S, Imani Fooladi AA, Mobarez AM, Emameini M, Amani J, Sedighian H (2015) In silico analysis for identifying potential vaccine candidates against *Staphylococcus aureus*. *Clin Exp Vaccin Res* 4(1):99–106. <https://doi.org/10.7774/cevr.2015.4.1.99>
- Dimitrov I, Naneva L, Doytchinova I, Bangov I (2014) AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* 30:846–851. <https://doi.org/10.1093/bioinformatics/btt619>
- Dobbins WO (1981) (1981) Is there an immune deficit in Whipple's disease? *Dig Dis Sci* 26:247–252
- Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinform* 8:4
- Fenollar F, Célard M, Lagier JC, Lepidi H, Fournier PE, Raoult D (2013) *Tropheryma whipplei* endocarditis. *Emerg Infect Dis* 19:1721–1730. <https://doi.org/10.3201/eid1911.121356>
- Fenollar F, Puechal X, Raoult D (2007) Whipple's disease. *N Engl J Med* 356:55–66
- Feurle GE, Moos V, Blaker H, Lodenkemper C, Morter A, Stroux A, Marth T, Schneider T (2013) Intravenous seftriaxone, followed by 12 or three months of oral treatment with trimethoprim-sulfamethoxazole in Whipple's disease. *J Infect* 66:263–270
- Guan P, Doytchinova IA, Zygouri C, Flower DR (2003) MHCpred: a server for quantitative prediction of peptide—MHC binding. *Nucleic Acids Res* 31:3621–3624
- Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R (2013) In silico approach for predicting toxicity of peptides and proteins. *PLoS ONE* 8:e73957
- Gurung RB, Purdie AC, Begg DJ, Whittington RJ (2012) In silico identification of epitopes in *Mycobacterium avium* subsp. *paratuberculosis* proteins that were upregulated under stress conditions. *Clin Vaccine Immunol* 19:855–864. <https://doi.org/10.1128/CVI.00114-12>
- Humphrey W, Dalke A, Schulten K (1996) VMD—visual molecular dynamics. *J Mol Graph* 14:33–38
- Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, Nielsen M (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. <https://doi.org/10.1111/imm.12889>
- Joshi A, Joshi BC, Amin-ul Mannan M, Kaushik V (2020) Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. *Inform Med Unlocked*. <https://doi.org/10.1016/j.imu.2020.100338>

- Kanampalliwar AM, Soni R, Gridhar A, Tiwari A (2013) Reverse vaccinology: basics and applications. *J Vaccines Vaccin* 4(6):1–5
- Kaushik V (2019) In Silico identification of epitope based peptide vaccine for Nipah virus. *Int J Pept Res Ther*. <https://doi.org/10.1007/s10989-019-09917-0>
- Keita AK, Dubot-Pérès A, Phommason K, Sibounheuang B, Vongsouvath M, Mayxay M, Raoult D, Newton PN, Fenollar F (2015) High prevalence of *Tropheryma whippelii* in Lao kindergarten children. *PLoS Negl Trop Dis*. <https://doi.org/10.1371/journal.pntd.0003538>
- Kumar A, Hays M, Lim F, Foster LJ, Zhou M, Zhu G, Miesner T (2015) Protective enterotoxigenic *Escherichia coli* antigens in a murine intranasal challenge model. *PLoS Negl Trop Dis*. <https://doi.org/10.1371/journal.pntd.0003924>
- La Scola B, Fenollar F, Fournier PE, Altwegg M, Mallet MN, Raoult D (2001) Description of *Tropheryma whippelii* gen. nov., sp. nov., the Whipple's disease bacillus. *Int J Syst Evol Microbiol* 51:1471–1479. <https://doi.org/10.1099/00207713-51-4-1471>
- Lagier JC, Fenollar F, Lepidi H, Glorgi R, Million M, Raoult D (2014) Treatment of classical Whipple's disease: from in vitro results to clinical outcome. *J Antimicrob Chemother* 69:219–227
- Lamiable A, Thévenet P, Rey J, Vavrusa M, Derreumaux P, Tufféry P (2016) PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res* 44(W1):W449–W454
- Marth T, Moos V, Müller C, Biagi F, Schneider T (2016) *Tropheryma whippelii* infection and Whipple's disease. *Lancet Infect Dis* 16:e13–e22. [https://doi.org/10.1016/S1473-3099\(15\)00537-X](https://doi.org/10.1016/S1473-3099(15)00537-X)
- Misra N, Panda PK, Shah K, Sukla LB, Chaubey P (2011) Population coverage analysis of T-Cell epitopes of *Neisseria meningitidis* serogroup B from iron acquisition proteins for vaccine design. *Bioinformation* 6(7):255–261. <https://doi.org/10.6026/97320630006255>
- Moos V, Kunkel D, Marth T, Feurle GE, La Scola B, Ignatius R, Zeitz M, Schneider T (2006) Reduced peripheral and mucosal *Tropheryma whippelii* specific Th1 response in patients with Whipple's disease. *J Immunol* 177:2015–2022
- Moos V, Schmidt C, Geelhaar A, Kunkel D, Allers K, Schinnerling K, Loddenkemper C, Fenollar F, Morter A, Raoult D, Ignatius R, Schneider T (2010) Impaired immune functions of monocytes and macrophages in Whipple's disease. *Gastroenterology* 138:210–220
- Mustafa AS (2013) In silico analysis and experimental validation of *Mycobacterium tuberculosis*-specific proteins and peptides of *Mycobacterium tuberculosis* for immunological diagnosis and vaccine development. *Med Princ Pract*. <https://doi.org/10.1159/000354206>
- Palanisamy N (2018) Identification of putative drug targets and annotation of unknown proteins in *Tropheryma whippelii*. *Comput Biol Chem*. <https://doi.org/10.1016/j.compbiolchem.2018.05.024>
- Paulley JW (1952) A case of Whipple's disease (intestinal lipodystrophy). *Gastroenterology* 22:128–133
- Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Christophe C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–1802
- Priyadarshini V, Pradhan D, Munikumar M, Swargam S, Umamaheswari A, Rajasekhar D (2014) Genome-based approaches to develop epitope-driven subunit vaccines against pathogens of infective endocarditis. *J Biomol Struct Dyn* 32(6):876–889
- Ramharther M, Harrison N, Bühler T, Herold B, Lagler H, Lötsch F, Mombo-Ngoma G, Müller C, Adegnika AA, Kremsner PG, Makristathis A (2014) Prevalence and risk factor assessment of *Tropheryma whippelii* in a rural community in Gabon: a community based cross-sectional study. *Clin Microbiol Infect* 20:1189–1194. <https://doi.org/10.1111/1469-0691.12724>
- Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM (2003) *Tropheryma whippelii* Twist: a human pathogenic Actinobacteria with a reduced genome. *Genome Res* 13:1800–1809. <https://doi.org/10.1101/gr.1474603>
- Schlundt A, Günther S, Sticht J, Wieczorek M, Roske Y, Heinemann U, Freund C (2012) Peptide Linkage to the α -subunit of MHCII creates a stably inverted antigen presentation complex. *J Mol Biol* 423(3):294–302. <https://doi.org/10.1016/j.jmb.2012.07.008>
- Schneider T, Moos V, Loddenkemper C, Marth T, Fenollar F, Raoult D (2008) Whipple's disease: new aspects of pathogenesis and treatment. *Lancet Infect Dis* 8:179–190
- Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33:W363–W367. <https://doi.org/10.1093/nar/gki481>
- Schöniger-Hekele M, Petermann D, Weber B, Müller C (2007) *Tropheryma whippelii* in the environment: survey of sewage plant influges and sewage plant workers. *Appl Environ Microbiol* 73:2033–2035. <https://doi.org/10.1128/AEM.02335-06>
- Shen Y, Maupetit J, Derreumaux P, Tufféry P (2014) Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *J Chem Theor Comput* 10:4745–4758
- Tang H, Liu XS, Fang YZ, Pan L, Zhang ZW, Zhou P, Lv JL, Jiang ST, Hu WF, Zhang P, Wang YL, Zhang YG (2012) The epitopes of foot and mouth disease. *Asian J Anim Vet Adv* 7:1261–1265
- Thévenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tufféry P (2012) PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Res* 40:W288–293
- Trotta L, Weigt K, Schinnerling K, Geelhaar-Karsch A, Oelkers G, Biagi F, Corazza GR, Allers K, Schneider T, Erben U, Moos V (2017) Peripheral T-cell reactivity to heat shock protein 70 and its cofactor GrpE from *Tropheryma whippelii* is reduced in patients with classical Whipple's disease. *Infect Immun* 85:e00363–e417
- Whipple GH (1907) A hitherto undescribed disease characterized anatomically by deposits of fat and fatty acids in the intestinal and mesenteric lymphatic tissues. *Bull Johns Hopkins Hosp* 18:382–393
- Williams CJ et al (2018) MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* 27:293–315
- Yardley JH, Hendrix TR (1961) Combined electron and light microscopy in Whipple's disease. Demonstration of "bacillary bodies" in the intestine. *Bull Johns Hopkins Hosp* 109:80–98
- Zhang C, Vasmatzis G, Cornette JL, DeLisi C (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol* 267(3):707–726

Application of Computational Methods for Identification of Drugs against *Tropheryma Whipplei*

Amit Joshi¹ and Vikas Kaushik^{2,*}

¹ Department of Biochemistry, Lovely Professional University, Punjab, India

² Department of Bioinformatics, Lovely Professional University, Punjab, India

Abstract: *Tropheryma whipplei* causes severe malady termed as Whipple's disease, a multisystemic lethal problem and we still require modified best regimens. To treat it successfully, 3 medications were distinguished in this investigation by using in-silico methods. 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3b]pyridine-3-carboxamide (2APC), Nicotinamide mononucleotide (NMN), and Riboflavin Monophosphate (RFMP) were seen as putative medications. 2APC and NMN restrain DNA Ligase catalytic activity for *Tropheryma whipplei* and compelling in impeding genomic copying and repairing mechanisms, RFMP shows the inhibitory impact on Chorismate synthase that drives hindrance in metabolic biosynthesis of amino acids. Our investigation used modern advanced in-silico assemblies. BLAST, CDART, CD-HIT were utilized to choose target catalytic biomolecules of a bacterium. Phyre2, dependent on HMM calculation, was applied to discover the best auxiliary models of chosen biocatalysts. AutoDock-Vina assembly was utilized for molecular docking and scoring restricting energies of these medications with catalytic proteins of the bacterium. 2APC and NMN hindering DNA Ligase show - 8.3 and - 8.2 kcal/mol individually while RFMP represses Chorismate synthase - 7.3 kcal/mol binding energy. Sub-atomic re-enactment or simulative mechanistic analysis gives further approval to concluding 2APC as impeccable inhibitory medication having remedial activity against *T. whipplei*. This escalated and novel examination is simple, quick, and valuable in anticipating drugs by incorporating computational insights in medicinal sciences.

Keywords: Biomolecules, Computational approach, Drug Discovery, Molecular dynamics Simulation, *Tropheryma whipplei*.

* Corresponding author Vikas Kaushik: Department of Bioinformatics, Lovely Professional University, Punjab, India; E-mail: vikas31bt@gmail.com

INTRODUCTION

Whipple's malady is an uncommon multisystem disorder affecting the gastrointestinal and central nervous systems. *Tropheryma whipplei*, Gram +ve actinobacterium about which little is known. The total DNA arrangement of its genome consists of 925 938, not complete genes set with an absence for significant biosynthesis-pathways and a diminished limit concerning vitality in metabolic mechanisms [1]. *T. whipplei* normally contaminates sewage laborers in Caucasian populations and among little children under 7 years in poor unclean surroundings [2]. Still, no data about the transmission of this microscopic organism distinguishing that this bacterium is associated with Homo sapiens and transferred from other animals. *T. whipplei* make duplicates inside macrophages, living in mucous and monocytes of blood vasculature in patients. Like Mycobacterium tuberculosis, it gradually develops under culture. Currently, hydroxychloroquine, doxycycline, and trimethoprim/sulfamethoxazole drugs, and also an injection of ceftriaxone given to patients with *T. whipplei* but the problem with this medication approach was very long and time taking up to 2 to 3 years with lifetime follow up [3]. Modern computational approaches can be useful in predicting drugs for the rare bacterium that causes severe health effects like *T. whipplei* [4]. In this study, we used genomic and proteomic analysis along with molecular docking and simulation analysis to find out possible drugs. Evaluation of pharmacokinetic characteristics in computer-aided drug designing is integral for hit-to-lead improvements. Exceptional unpredictability of the present research design for medication search, scientists strongly sought molecular docking and simulation mechanistic models to characterize patterns in ADMET information to develop practical insights [5]. This investigation deployed ADMET analysis along with 2D and 3D interaction of drugs to biocatalysts and it was found to be very successful in drug predictions.

MATERIAL AND METHODS

Enzyme Selection Bias

Proteomic sequences of *Tropheryma whipplei* were retrieved from NCBI-Genbank for two significant compounds to be specific, DNA Ligase (AAO44511) and Chorismate synthase (WP_011096348), and these enzymatic edifices are engaged with DNA replication, biosynthesis of amino acids individually (Table 1). The choice of these fundamental proteins depends on DEG (database of basic qualities) server investigation [6].

Table1. Selected proteins information for *Tropheryma whipplei*, that are found to be drug targets by screening NCBI- Genbank database and identification by KAAS (KEGGAutomatic annotation server) for pathway analysis of crucial genes.

NCBI-Genbank Accession no.	Identified Protein	KEGG Orthology Number	Functionality
AAO44511	DNA Ligase	K01972	Replicative and reparative Mechanisms of DeoxyriboNucleicAcids
WP_011096348	Chorismate synthase	K01736	Peptide and amino acid biosynthetic mechanisms

Protein Drug Analysis

After this CD-HIT server [7] was utilized to distinguish paralogs, it depends on a fast heuristic examination approach and accommodating in deciding likeness investigation between peptide stretches. Basic local search alignment was utilized for deciding homology [8] between considered proteins of *Tropheryma whipplei* and proteomic spaces of *Homo sapiens*. This gives more approval to the determination from escalated proteomic sets of the bacterium. To examine pharmacogenetics or medication capacity of considered proteins, a drug bank web-server (<http://www.drugbank.ca/>) was applied. To recognize the space homogeneity of protein groupings, 2 WebServers conserved-domain-architecture-retrieval-tool (CDART) & Pfam was utilized. KEGG automatic annotation server (KAAS) assisted in distinguish metabolic pathways for selected biocatalysts and here *T. whipplei* Twist and *T. whipplei* TW08/27 strains were browsed from the NCBIgenbank organisms list at the time of the investigation.

Structural Analysis: Docking & Simulation

The selected proteins, after intensive investigation and KEGG annotation, was exposed to homology displaying using Phyre2, it is a hidden Markov model-based server for structural predictions of catalytic enzymes. A quick overview of medications acting on chosen proteins and their 3D structure was acquired by utilizing Pubchem web server and RCSB- PDB databank. Sub-atomic docking was led using Autodock-vina assembly to examine the interaction energies of ligand-protein docked structures. SwissADME tool was deployed to examine biochemical properties like pharmo-kinetics, drug-likeness, and inhibitory action on cytochrome P450 isoforms, structural properties, bioavailability, and synthetic accessibility. Molecular simulation studies were performed for 40ns by using GROMACS ver.2019 simulation suite.

RESULT AND DISCUSSION

The selected proteins listed in Table 1 indicate KEGG annotation and NCBI-Genbank accession no. along with their known functionality (based on DEG and KEGG server) in *Tropheryma whipplei* twist strain. Phyre2 itself is an HMM algorithm-based server deployed to determine the protein structure of DNA Ligase and Chorismate synthase enzymes. DNA ligase is a crucial enzymatic assembly that is used by bacterium for repair and copying DNA sequences and inhibited by 2-amino-7-fluoro-5-oxo-5H-chromeno [2,3b]-pyridine-3-carboxamide (2APC) and Nicotinamide mononucleotide (NMN). While Chorismate Synthase performs dephosphorylation of 5O(1-carboxyvinyl)-3-phosphoshikimate to chorismate. This enzyme does not exist in *Homo sapiens*. Chorismate is a precursor for aromatic-ring containing amino acids. This enzyme interacts with riboflavin monophosphate (as per the Drug-Bank database). Riboflavin monophosphate (RFMP) is a potent oxidizing agent which has been used in the food processing industry as a colorant. In earlier studies, Riboflavin has been seen to effectively regulate *Staphylococcus aureus* colonization when used in conjunction with antibiotics. Similar studies deployed to eradicate *T. whipplei* infection and promote its novel treatment strategy. Pubchem database was deployed to retrieve the structure of drugs interacting with selected proteins. Pubchem CID and the name of drugs were mentioned in Table 3 with Swiss-ADME characteristics. In Fig. (1) structure of drugs is represented and for better visualization of pharmacophore analysis, pymol software is used. The structure of enzymatic complexes was retrieved from phyre2 server, based on a detailed homology report of modeled structure of DNA ligase and Chorismate synthase enzymatic assembly of *Tropheryma whipplei*. The best model result was retrieved to obtain the PDB file of their structure. Out of 120 best structures, one was finalized for DNA ligase, while out of 99 models, one was finalized for Chorismate synthase. Docking studies reveal binding energies for docked complexes and perfect binding energies for all docked complexes represented in Table 2. The perfectly docked complex of inhibitory drugs and enzymatic assembly was represented in Fig. 2(A,B and C). These results satisfy the perfect interaction of complexes suggests that 2-amino-7-fluoro-5-oxo-5H-chromeno [2,3b]-pyridine-3-carboxamide as well as Nicotinamide mononucleotide interacts with DNA Ligase while Riboflavin monophosphate can interact with Chorismate synthase. And these selected chemicals can be used as putative drug candidates. Drug 2D interaction pattern with enzymes based on ligPlot v2.2 software represented in Fig. 2(P,Q and T). Mostly all the selected drugs not only show better binding scores but 3D and 2D interaction pattern reveals hydrogen bonds interaction with considered enzymes in their binding pocket. In Table 3 drug physicochemical parameters were represented based on the SwissADME server (www.swissadme.ch). Lipinski rule was also considered during drug analysis and

found to show zero violations for 2APC. This indicates that all drugs have good inhibitory properties against selected enzymes. All of the drugs don't show blood-brain barrier permeability also any inhibition for CYP3A4 inhibition; these results suggest effective drug clearance in the body after effective action. 2APC, NMN and RFMP show $\log K_p(\text{cm/s})$ values -6.97, -8.26 and 10.9 respectively. TPSA values for 2APC, NMN, and RFMP were found 112.21, 165.61, and 217.9, respectively. 2APC also shows high GI absorption. MD simulation for 40 nanoseconds (ns) was conducted by deploying GROMACS software ver.2019. Constant pressure and temperature (NPT) ensembles were used to set the equilibration measures. At a normal temperature of 300 K and a pressure level of 1.013 bar, MD simulations were run. The MD trajectories were calculated using the perfectly docked-complexes to determine the root mean square deviation (RMSD) and root mean square fluctuation (RMSF) for a timescale of 40-ns. It was found the best-docked complexes were exhibiting perfect characteristics based on RMSD and RMSF plots (Fig. 3). The simulation study revealed that drug-enzyme complexes did not face any alterations in their binding patterns. RMSD values for docked complexes were found in the range of 1- 3 Angstrom, while RMSF values for docked complexes 1 – 4 Angstrom. These values indicate perfect interaction between considered drugs and enzymes of *T.whipplei*.

Table 2. AutoDock-vina docked results: Binding energies of best-docked complexes.

Best Docked Model	Binding-Energy (Kcal/mol)
DNA Ligase and 2APC	-8.3
DNA Ligase and NMN	-8.2
Chorismate Synthase and RFMP	-7.3

Table 3. -Drug characteristics analysed by Pubchem database and SwissADME.

Molecule	Formula	MW	Lipinski #violations
2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide (CID- 10038928)	C13H8FN3O3	273.22	0
Nicotinamide mononucleotide (CID- 14180)	C27H29NO9	511.52	1
Riboflavin monophosphate (CID- 643976)	C17H21N4O9P	456.34	2

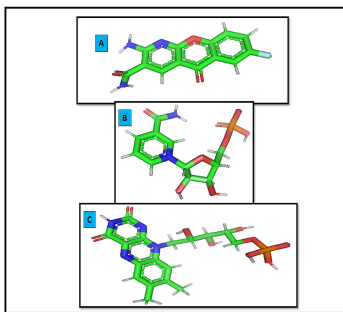


Fig. (1). Chemical structure of drugs obtained from Pubchem database and analyzed in Pymol **A.** 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide (CID-10038928), **B.** Nicotinamide mononucleotide (CID- 14180), **C.** Riboflavin monophosphate (CID- 643976).

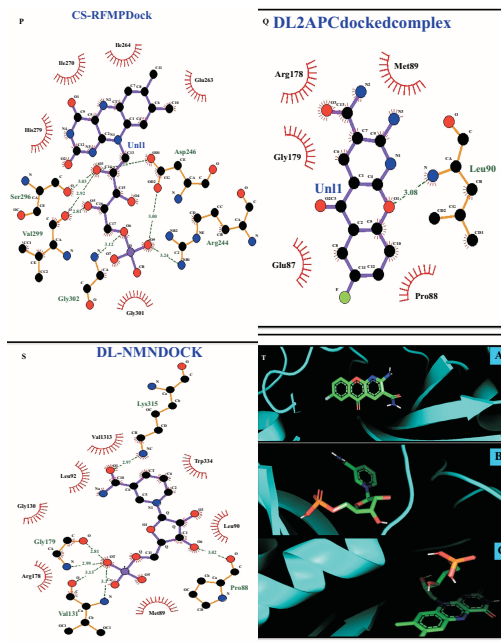


Fig. (2). Molecular interactions between drugs and receptor: **P)** Chorismate synthase interaction with RFMP: Val299, Gly 302, Ser 296, Asp 246, Arg 244 interacts with RFMP drug *via* hydrogen bond (2.00 to 3.50 Å). **Q)** DNA Ligase interaction with 2APC drug: Leu at 90 position interacts with 2APC drug *via* hydrogen bond of strength 3.08Å. **S)** DNA Ligase interaction with NMN drug: Lys at 315 position interacts with NMN drug *via* hydrogen bond of strength 2.97Å, also Gly 179 & Val 131 show hydrogen bonding with the NMN drug. **T)** AutoDock vina docking results of drugs interacting with proteins- **A.** DNA Ligase with 2-amino-7-fluoro-5-oxo-5H-chromeno [2,3b]pyridine-3-carboxamide **B.** DNA Ligase with Nicotinamide mononucleotide **C.** Chorismate synthase with Riboflavin monophosphate.

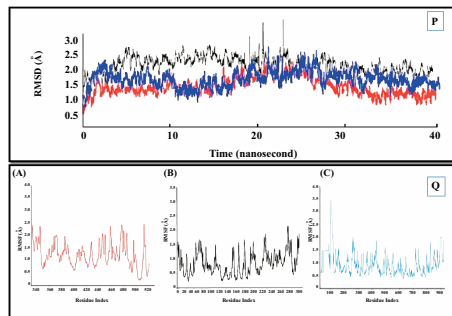


Fig. (3). Molecular simulation analysis of drugs complexed with enzymes of *T. whipplei*: **P**) RMSD plot: Black color-(RFMP-Chorismate synthase complex), Red Color- (2APC-DNA Ligase complex), Blue color-(NMN-DNA Ligase complex); **Q**) RMSF plots: **A.** RFMP interaction with Chorismate synthase, **B.** 2APC interacting DNA Ligase **C.** NMN drug interacting DNA Ligase.

CONCLUSION

This in-silico investigation discovers that 2-APC, NMN, and RFMP as possible medications to treat Whipple's disease and can be used for animal testing and clinical trials. All the pharmaco-kinetics depict that these medications would perfectly interact with bacterial biocatalysts to hamper their activity. This approach was found to be rapid for predicting drug candidates and even effective against harmful organisms like *Tropheryma whipplei* having a reduced genome.

CONSENT FOR PUBLICATION

Not Applicable.

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

All the authors A.J. and V.K. conducted this research study and also manuscript preparation under the Department of bioinformatics, LPU, Punjab India.

REFERENCES

- [1] S.D. Bentley, M. Maiwald, L.D. Murphy, M.J. Pallen, C.A. Yeats, L.G. Dover, H.T. Norbertczak, G.S. Besra, M.A. Quail, D.E. Harris, A. von Herbay, A. Goble, S. Rutter, R. Squares, S. Squares, B.G. Barrell, J. Parkhill, and D.A. Relman, "Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*", *Lancet*, vol. 361, no. 9358, pp. 637-644, 2003. [[http://dx.doi.org/10.1016/S0140-6736\(03\)12597-4](http://dx.doi.org/10.1016/S0140-6736(03)12597-4)] [PMID: 12606174]
- [2] F. Fenollar, M. Célard, J-C. Lagier, H. Lepidi, P-E. Fournier, and D. Raoult, "Tropheryma whipplei endocarditis", *Emerg. Infect. Dis.*, vol. 19, no. 11, pp. 1721-1730, 2013. [<http://dx.doi.org/10.3201/eid1911.121356>] [PMID: 24207100]
- [3] A. Joshi, and V. Kaushik, "In-Silico Proteomic Exploratory Quest: Crafting T-Cell Epitope Vaccine

- Against Whipple's Disease", *Int. J. Pept. Res. Ther.*, vol. 27, no. 1, pp. 1-11, 2020.
[<http://dx.doi.org/10.1007/s10989-020-10077-9>] [PMID: 32427224]
- [4] N. Palanisamy, "Identification of putative drug targets and annotation of unknown proteins in *Tropheryma whipplei*", *Comput. Biol. Chem.*, vol. 76, pp. 130-138, 2018.
[<http://dx.doi.org/10.1016/j.compbiolchem.2018.05.024>] [PMID: 30005292]
- [5] L.L.G. Ferreira, and A.D. Andricopulo, "ADMET modeling approaches in drug discovery", *Drug Discov. Today*, vol. 24, no. 5, pp. 1157-1165, 2019.
[<http://dx.doi.org/10.1016/j.drudis.2019.03.015>] [PMID: 30890362]
- [6] R. Zhang, H.Y. Ou, and C.T. Zhang, "DEG: a database of essential genes", *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D271-D272, 2004.
[<http://dx.doi.org/10.1093/nar/gkh024>] [PMID: 14681410]
- [7] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences", *Bioinformatics*, vol. 26, no. 5, pp. 680-682, 2010.
[<http://dx.doi.org/10.1093/bioinformatics/btq003>] [PMID: 20053844]
- [8] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezuk, S. McGinnis, and T.L. Madden, "NCBI BLAST: a better web interface," *Nucleic Acids Research*, vol. 36, no. 5, *Web Server*, no. May, pp. W5-W9, 2008.

In-Silico therapeutic approach for identification of drug targets for *Tropheryma whipplei*

Amit Joshi¹ and Vikas Kaushik^{1*}

1- Domain of bioinformatics, School of bioengineering and Biosciences, Lovely Professional University

Abstract

Tropheryma whipplei is main causative agent of whipple ailment as it is multisystemic fatal disorder and there is still need of developing best regimens for its regulation. To control its spread and to treat it effectively three drugs were identified in this study by deploying pharmaco-informatic approach. 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide (2APC), Nicotinamide mononucleotide(NMN), and Riboflavin monophosphate(RFMP) were found to be putative drugs. 2APC and NMN inhibits activity of DNA Ligase enzyme of this bacterium and effective in impairing replication and repair mechanisms, while RFMP exhibits inhibitory effect on Chorismate synthase that leads impairment of amino acid metabolism or biosynthesis. In this study effective alignment tools like BLAST, CDART, CD-HIT were used to select enzymes. Phyre2 based on HMM algorithm is deployed to find best structural models of selected enzymatic proteins. AutoDock-Vina tool is used for docking and scoring binding energies of these drugs with considered enzymatic domains. 2APC and NMN inhibiting DNA Ligase exhibits -8.3 and -8.2 kcal/mol respectively while RFMP inhibiting Chorismate synthase -7.3 kcal/mol. This intensive and novel study is easy, fast, and useful in predicting drugs by *In-silico* approach.

Keywords- Drug, *In-silico*, Binding energy, Alignment, DNA-Ligase, Chorismate synthase, Docking

Introduction

Tropheryma whipplei (*T. whipplei*) is associated with Actinobacteria class, the another bacterium of this class include *Mycobacterium tuberculosis*, *Mycobacterium leprae*, *Corynebacterium diphtheriae* and *Micrococcus tetragenus* (Raoult et al., 2003) that causes other harmful ailments in human. Actinobacteriums have greater GC concentration (de Lima Procópio et al., 2012; Raoult et al., 2003) in their genomic strands. Though *T. whipplei* is classified as Gram-positive (La Scola et al., 2001; Raoult et al., 2003) bacteria. *T. whipplei* usually infects sewage workers in caucasian countries and among toddlers less than 7 years with worse sanitary conditions (Fenollar et al., 2008; Keita et al., 2015; Ramharter et al., 2014; Schöniger-Hekele et al., 2007). presently, no information about zoonosis of this bacteria identifying that this bacterial specie is directly related to humans. This bacteria make copies inside macrophages residing in mucous and mononuclear cells of blood vasculature in *Homo sapiens* (Marth et al., 2016). Similar to *Mycobacterium*, it

also slowly grows under culture (Marth et al., 2016). This study was focused to identify drugs that can exhibit interaction with biochemical constituents of this bacterium and hampers its growth. This study involves use of pharmaco-Informatic approach and computer based drug discovery methods to eradicate this bacterium from human system.

Methodology

Protein sequences of *Tropheryma whipplei* were retrieved from NCBI-Genbank for two major enzymes namely, DNA Ligase (AAO44511) and Chorismate synthase (WP_011096348), and these enzymatic complexes are involved in DNA replication, biosynthesis of amino acids respectively (Table 1). The selection of these essential proteins is based on DEG (database of essential genes) server analysis (Zhang et al., 2004).

After retrieval of sequence CD-HIT server (Huang et al., 2010) was used to identify paralogs. It is based on rapid heuristic analysis approach, and helpful in determining similarity analysis between peptide stretches. BLAST-p tool of NCBI (Johnson et al., 2008) is used for determining homology between selected proteins of *Tropheryma whipplei* and proteomic domains of *Homo sapiens*. This provides more validation or filtrations to the selection from intensive proteomic sets of bacterium. To analyze pharmacogenecity or drug-ability of selected proteins drugbank web-server (<http://www.drugbank.ca/>) was deployed (Wishart et al., 2006). To identify domain homology in unknown protein sequences, two web servers namely conserved domain architecture retrieval tool (CDART) (Geer et al., 2002) and Pfam (Finn et al., 2016) were used. KEGG automatic annotation server (KAAS) was deployed to identify metabolic pathway of selected proteins (Moriya et al., 2007) and here *T. whipplei* Twist and *T. whipplei* TW08/27 strains were chosen from the organism list during analysis.

The selected proteins after intensive analysis and KEGG annotation was subjected to homology modeling via Phyre2, it is a Hidden markov model based server for perfect analytical prediction based structural study of proteins (Kelley et al., 2015). Rapid survey of drugs targeting selected proteins and their 3D structure was obtained by using Pubchem web server and RCSB-PDB databank. Molecular docking was conducted via Autodock-vina tool to analyze binding energies of ligand-protein interactions (Trott and Olson, 2010).

Results and Discussion

Selected proteins listed on Table 1 clearly indicate KEGG annotation and NCBI-Genbank accession no. along with their known functionality (based on DEG and KEGG server) in *Tropheryma whipplei* twist strain. Phyre2 itself a HMM algorithm based server deployed to determine protein structure of DNA Ligase and Chorismate synthase enzymes. DNA ligase is crucial enzymatic assembly that is used by bacterium for repair and copying DNA sequences, and inhibited by 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide (2APC) and Nicotinamide mononucleotide (NMN). While Chorismate synthase perform dephosphorylation of 5-O-(1-carboxyvinyl)-3-phosphoshikimate to chorismate (Pitchandi et al., 2013). This enzyme not exists in *Homo sapiens*. Chorismate is a precursor for aromatic-ring containing amino acids. This enzyme interacts

with riboflavin monophosphate (as per Drug-Bank database). Riboflavin monophosphate (RFMP) is a strong oxidizing agent and has been used as an additive (coloring agent) in the food industry. In earlier studies, Riboflavin was used in combination with antibiotics and shown to control *Staphylococcus aureus* infection efficiently (Dey and Bishayi, 2016). Similar studies deployed to eradicate *T. whipplei* infection and promote its novel treatment strategy.

Table1. Selected proteins information for *Tropheryma whipplei*, that are found to be drug targets by screening NCBI-Genbank database and identification by KAAS (KEGG Automatic annotation server) for pathway analysis of crucial genes.

NCBI-Genbank Accession no.	Identified Protein	KEGG Orthology Number	Functionality
AAO44511	DNA Ligase	K01972	DNA replication and repair
WP_011096348	Chorismate synthase	K01736	Biosynthesis of amino acids

Pubchem database was deployed to retrieve structure of drugs interacting with selected proteins. Pubchem CID and name of drugs were mentioned in **Table 2**. In **Figure 1** structure of drugs is represented and for better visualization of pharmacophore analysis pymol software is used.

The structure of Proteins or enzymatic complexes were retrieved from phyre2 server, **Figure 2 and 3** shows detailed homology report of modeled structure of DNA ligase and Chorismate synthase enzymatic assembly of *Tropheryma whipplei* twist strain. Best model result is retrieved to obtain PDB file of their structure. Out of 120 models 11 were depicted in Figure 2 for DNA ligase while out of 99 models 11 were depicted in Figure 3 for Chorismate synthase.

Table2. Therapeutic drugs identified for analyzing pharmacophore interaction with screened proteins of *Tropheryma whipplei*

S.No.	Protein name	Interacting Potential Drug / Pubchem CID
1.	DNA Ligase	2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide (CID-10038928)
		Nicotinamide mononucleotide (CID- 14180)
2.	Chorismate synthase	Riboflavin monophosphate (CID- 643976)

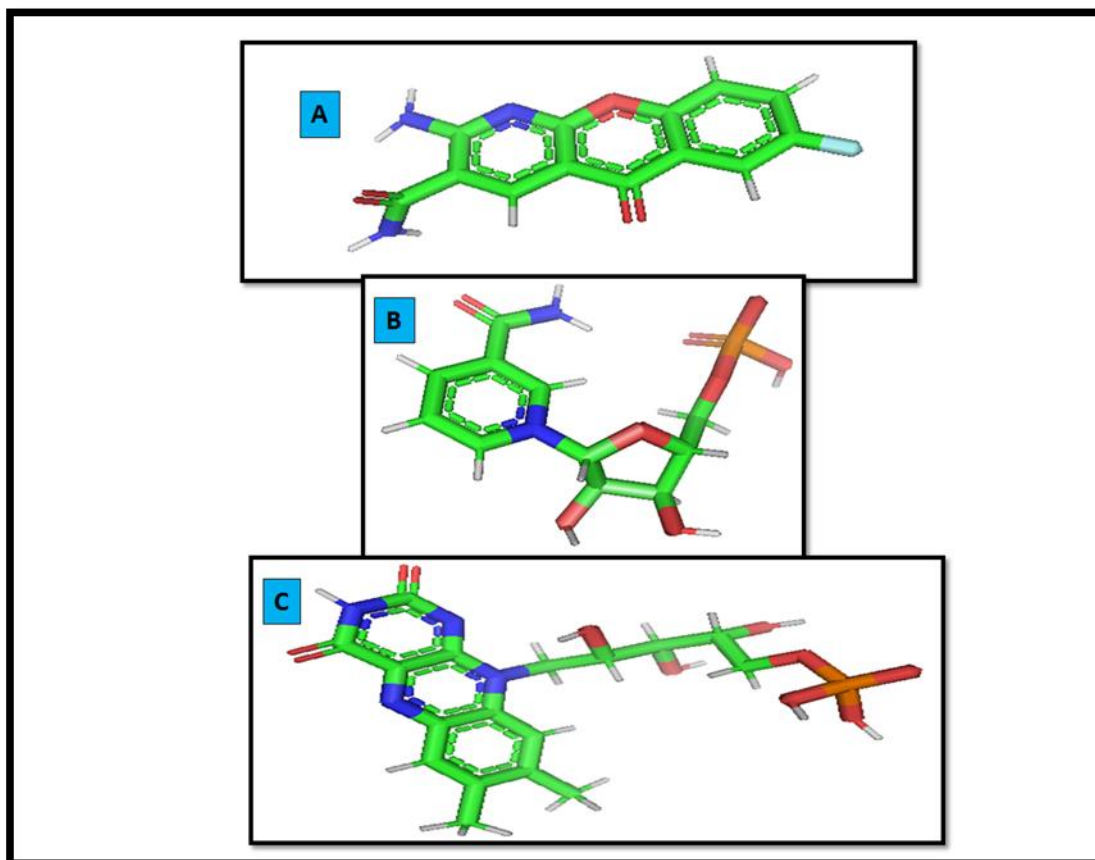


Figure 1. Chemical structure of drugs obtained from Pubchem database and analyzed in Pymol **A.** 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide (CID-10038928) **B.** Nicotinamide mononucleotide (CID- 14180) **C.** Riboflavin monophosphate (CID- 643976)

#	Template	Alignment Coverage	3D Model	Confidence	% I.d.	Template Information
1	c2ow0A	Alignment		100.0	35	PDB header: ligase/dna Chain: A; PDB Molecule: dna ligase; PDBTitle: last stop on the road to repair: structure of e.coli dna ligase bound2 to nicked dna-adenylate
2	c1v9pB	Alignment		100.0	35	PDB header: ligase Chain: B; PDB Molecule: dna ligase; PDBTitle: crystal structure of nad+-dependent dna ligase
3	c1dqsB	Alignment		100.0	35	PDB header: ligase Chain: B; PDB Molecule: dna ligase; PDBTitle: crystal structure of nad+-dependent dna ligase from t.2 filliformis
4	c4qlxA	Alignment		100.0	35	PDB header: ligase/ligase inhibitor/dna Chain: A; PDB Molecule: dna ligase; PDBTitle: dna ligase a in complex with inhibitor
5	c3sglA	Alignment		100.0	42	PDB header: ligase Chain: A; PDB Molecule: dna ligase; PDBTitle: crystal structure of dna ligase a brct domain deleted mutant of2 mycobacterium tuberculosis
6	c3nn1A	Alignment		100.0	37	PDB header: ligase/ligase inhibitor Chain: A; PDB Molecule: dna ligase; PDBTitle: novel bacterial nad+-dependent dna ligase inhibitors with broad2 spectrum potency and antibacterial efficacy in vivo
7	d1ta8a	Alignment		100.0	36	Fold: ATP-grasp Superfamily: DNA ligase/mRNA capping enzyme, catalytic domain Family: Adenylation domain of NAD+-dependent DNA ligase
8	d1b0da	Alignment		100.0	36	Fold: ATP-grasp Superfamily: DNA ligase/mRNA capping enzyme, catalytic domain Family: Adenylation domain of NAD+-dependent DNA ligase
9	c3hlA	Alignment		100.0	36	PDB header: ligase Chain: A; PDB Molecule: dna ligase; PDBTitle: crystal structure of the adenylation domain of nad+-2 dependent dna ligase from staphylococcus aureus
10	c1znuA	Alignment		100.0	37	PDB header: ligase Chain: A; PDB Molecule: dna ligase; PDBTitle: adenylation domain of nad+-dependent dna ligase from m.tuberculosis
11	d1v9na3	Alignment		100.0	33	Fold: ATP-grasp Superfamily: DNA ligase/mRNA capping enzyme, catalytic domain Family: Adenylation domain of NAD+-dependent DNA ligase

Figure 2. Phyre2 homology report for modeled Structure of DNA Ligase *Tropheryma whipplei* twist strain.

#	Template	Alignment Coverage	3D Model	Confidence	% I.d.	Template Information
1	dlqxo_a_	Alignment		100.0	43	Fold: Chorismate synthase, AroC Superfamily: Chorismate synthase, AroC Family: Chorismate synthase, AroC
2	c1zta_A	Alignment		100.0	55	PDB header: ligase Chain: A; PDB Molecule: chorismate synthase; PDBTitle: crystal structure of chorismate synthase from mycobacterium2 tuberculosis
3	dium0a_	Alignment		100.0	34	Fold: Chorismate synthase, AroC Superfamily: Chorismate synthase, AroC Family: Chorismate synthase, AroC
4	dlqla_	Alignment		100.0	48	Fold: Chorismate synthase, AroC Superfamily: Chorismate synthase, AroC Family: Chorismate synthase, AroC
5	c4lj2A_	Alignment		100.0	37	PDB header: lyase Chain: A; PDB Molecule: chorismate synthase; PDBTitle: crystal structure of chorismate synthase from acinetobacter baumannii2 at 3.15a resolution
6	dlsq1a_	Alignment		100.0	37	Fold: Chorismate synthase, AroC Superfamily: Chorismate synthase, AroC Family: Chorismate synthase, AroC
7	dlr53a_	Alignment		100.0	37	Fold: Chorismate synthase, AroC Superfamily: Chorismate synthase, AroC Family: Chorismate synthase, AroC
8	c4ecdB_	Alignment		100.0	57	PDB header: lyase Chain: B; PDB Molecule: chorismate synthase; PDBTitle: 2.5 angstrom resolution crystal structure of bifidobacterium longum2 chorismate synthase
9	c5z9aB_	Alignment		100.0	41	PDB header: lyase Chain: B; PDB Molecule: chorismate synthase; PDBTitle: crystal structure of chorismate synthase from pseudomonas aeruginosa
10	c2k1hA_	Alignment		77.1	13	PDB header: structural genomics, unknown function Chain: A; PDB Molecule: uncharacterized protein ser13; PDBTitle: solution nmr structure of ser13 from staphylococcus epidermidis.2 northeast structural genomics consortium target ser13
11	d2ffma1	Alignment		74.1	13	Fold: Hypothetical protein SAV1430 Superfamily: Hypothetical protein SAV1430 Family: Hypothetical protein SAV1430

Figure 3. Phyre2 homology report for modeled Structure of Chorismate synthase *Tropheryma whippelii* twist strain.

Docking studies reveals binding energies for docked complexes and perfect binding energies for all docked complexes represented in **table 3**. The perfectly docked complex of inhibitory drugs and enzymatic assembly was represented in **Figure 4**.

Table 3. AutoDock-vina docked results: Binding energies of best docked complexes

Best Docked Model	Binding Energy (Kcal/mol)
1. DNA Ligase and 2APC	-8.3
2. DNA Ligase and NMN	-8.2
3. Chorismate Synthase and RFMP	-7.3

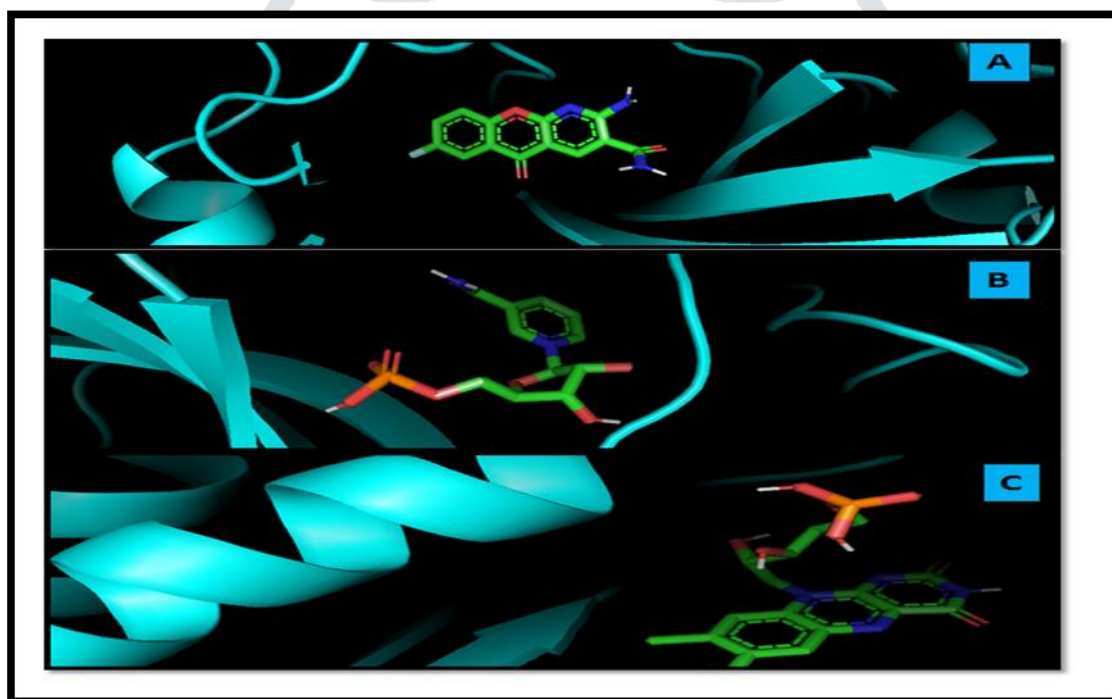


Figure 4. AutoDock vina docking results of drugs interacting with proteins- **A.** DNA Ligase with 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide **B.** DNA Ligase with Nicotinamide mononucleotide **C.** Chorismate synthase with Riboflavin monophosphate.

These results satisfies the perfect interaction of complexes suggests that 2-amino-7-fluoro-5-oxo-5H-chromeno[2,3-b]pyridine-3-carboxamide as well as Nicotinamide mononucleotide interacts with DNA Ligase while Riboflavin monophosphate can interact with Chorismate synthase. And these selected chemicals can be used as putative drug candidates.

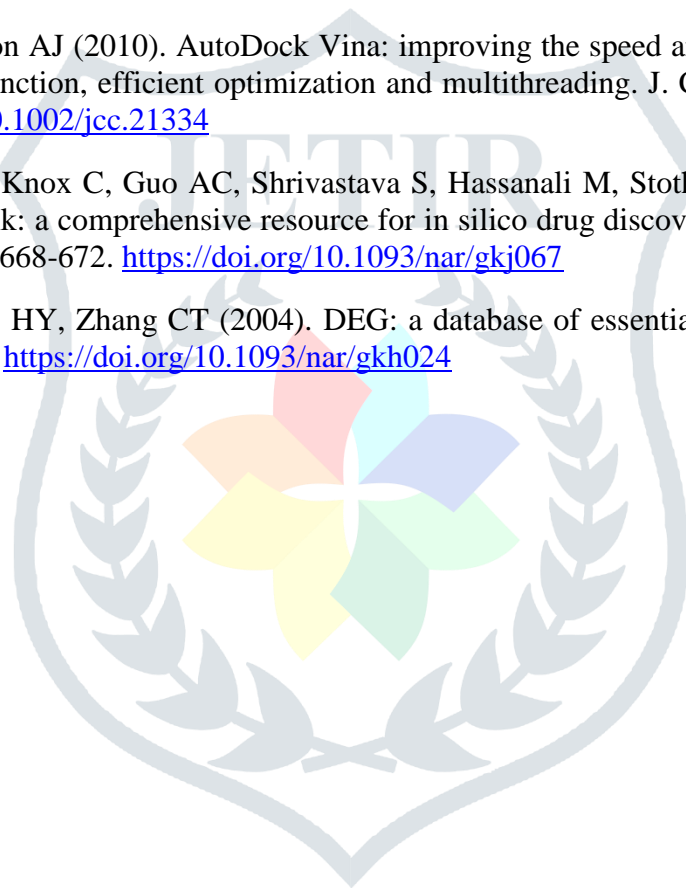
Conclusion

This intensive pharmaco-informatic study reveals that 2-amino-7-fluoro-5-oxo-5H-chromeno [2,3-b]pyridine-3-carboxamide, Nicotinamide mononucleotide, and Riboflavin Monophosphate considered as putative drugs and they can be used for further wet-lab validation. All the Pharmacophore of analyzed drugs were found to be perfectly interacting with enzymatic assemblies to inhibit them. This is easy and fast method to predict drug and even effective against organisms like *Tropheryma whipplei* having reduced genome.

References

- [1] de Lima Procópio RE, da Silva IR, Martins MK, de Azevedo JL, de Araújo JM (2012). Antibiotics produced by *Streptomyces*. *Braz. J. Infect. Dis.* 16, 466–471. <https://doi.org/10.1016/j.bjid.2012.08.014>
- [2] Dey S, Bishayi B (2016). Riboflavin along with antibiotics balances reactive oxygen species and inflammatory cytokines and controls *Staphylococcus aureus* infection by boosting murine macrophage function and regulates inflammation. *J. Inflamm. Lond. Engl.* 13. <https://doi.org/10.1186/s12950-016-0145-0>
- [3] Fenollar F, Célard M, Lagier JC, Lepidi H, Fournier PE, Raoult D (2013). *Tropheryma whipplei* endocarditis. *Emerg. Infect. Dis.* 19, 1721–1730. <https://doi.org/10.3201/eid1911.121356>
- [4] Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. <https://doi.org/10.1093/nar/gkv1344>
- [5] Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002). CDART: protein homology by domain architecture. *Genome Res.* 12, 1619–1623. <https://doi.org/10.1101/gr.278202>
- [6] Huang Y, Niu B, Gao Y, Fu L, Li W (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. <https://doi.org/10.1093/bioinformatics/btq003>
- [7] Johnson M, Zaretskaya I, Raytselis Y, Merezuk Y, McGinnis S, Madden TL (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9. <https://doi.org/10.1093/nar/gkn201>
- [8] Keita AK, Dubot-Pères A, Phommasone K, Sibounheuang B, Vongsouvath M, Mayxay M, Raoult D, Newton PN, Fenollar F (2015). High prevalence of *Tropheryma whipplei* in Lao kindergarten children. *PLoS Negl. Trop. Dis.* 9.
- [9] Kelly LA. et al., (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10, 845-858.
- [10] La Scola B, Fenollar F, Fournier PE, Altwegg M, Mallet MN, Raoult D (2001). Description of *Tropheryma whipplei* gen. nov., sp. nov., the Whipple's disease bacillus. *Int. J. Syst. Evol. Microbiol.* 51, 1471–1479. <https://doi.org/10.1099/00207713-51-4-1471>
- [11] Marth T, Moos V, Müller C, Biagi F, Schneider T (2016). *Tropheryma whipplei* infection and Whipple's disease. *Lancet Infect. Dis.* 16, e13–e22. [https://doi.org/10.1016/S1473-3099\(15\)00537-X](https://doi.org/10.1016/S1473-3099(15)00537-X)

- [12] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182- 185. <https://doi.org/10.1093/nar/gkm321>
- [13] Pitchandi P, Hopper W, Rao R. (2013). Comprehensive database of Chorismate synthase enzyme from shikimate pathway in pathogenic bacteria. *BMC Pharmacol. Toxicol.* 14, 29. <https://doi.org/10.1186/2050-6511-14-29>
- [14] Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM. (2003). *Tropheryma whippelii* Twist: a human pathogenic Actinobacteria with a reduced genome. *Genome Res.* 13, 1800–1809. <https://doi.org/10.1101/gr.1474603>
- [15] Schöniger -Hekele M., Petermann D, Weber B, Müller C (2007). *Tropheryma whippelii* in the environment: survey of sewage plant influxes and sewage plant workers. *Appl. Environ. Microbiol.* 73, 2033–2035. <https://doi.org/10.1128/AEM.02335-06>
- [16] Trott O, Olson AJ (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* 31,455–461. <https://doi.org/10.1002/jcc.21334>
- [17] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668-672. <https://doi.org/10.1093/nar/gkj067>
- [18] Zhang R, Ou HY, Zhang CT (2004). DEG: a database of essential genes. *Nucleic Acids Res.* 32, D271–D272. <https://doi.org/10.1093/nar/gkh024>





G. Sunil Krishnan, Amit Joshi, and Vikas Kaushik

Abstract

Genomics delivers purposeful biological information, as it is a part of life science dealing about the comprehension and planning of genomes. A genome is the complex arrangement of genetic sets present in a cell or a whole living being. It is a useful measure of information when you consider that the human genome has in excess of 3 billion DNA base sets. It is a stunning measure of data that people have experienced difficulty in wielding, despite the fact that nature figured out how to pack everything into each cell in the human body. Customized medication is clinical consideration related to every patient's hereditary cosmetics. It implies mass, mechanical production system like medication reaches a conclusion, and medication intended to convey greatest advantage to the individual turns into the standard. This would kill a great deal of awful side-effects related with standard medicines presently, decrease or dispose of hypersensitive responses, diminish the expense of medical care, and lessen patient sufferings, as enduring more successful therapies. So as to really perform customized medication, every patient's genome should initially be converted into advanced information which is then handled, put away, and recovered varying. Accordingly the triple play of genomics, bioinformatics and customized medication is vital. Everything sounds so basic yet it is so confounded. Numerous medications and preventive therapies are neglected to convey ideal reaction to wide populace. PM is a combinational way to deal with individual specific health care. The patient specific medication development advanced through bioinformatics tools. Bioinformatics devices may furnish better conclusions in genomic level with prior identification of infection and better focused on treatment through productive PM improvement. Variety in

G. Sunil Krishnan · A. Joshi · V. Kaushik (✉)
Lovely Professional University, Phagwara, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_15

303

sex, racial, ethnic, genetic polymorphisms, and other ecological variables influence the in resistant reaction to a specific therapy.

Keywords

Personalized medicine · Bioinformatics · Genome · Health care · Human body

15.1 Introduction

Medication and medicines are changing due to the advancement of technology development. Many people are suffering in the world due to rare diseases and adverse effect of a therapy. This was because of the unavailability of efficacious drug or personalized medicine (PM). PM is an individual targeted tailor made medication to reduce the drug or disease associated risk this can leads to provide more efficacious treatment. PM is also known as precision or stratified medicine. In this specific approach individual patient's genomics profile plays an important role to find safe and efficacious treatment. PM has the prospective tools to manage different incurable disease stages from detection to prevention. Next-Gen Sequencing (NGS) innovation, frequently observed as the establishment of personalized medication, has been effectively applied in oncology diagnostics and immunotherapy. With propels in quality diagnostics and immunotherapy, there might be an opportunity to control the advancement of malignant growths and mitigate the enduring of patients going through chemotherapy. To advance the interpretation of exactness medication from seat to alongside and from utilization of hereditary testing to customized medication, new investigation techniques for NGS and hereditary information should be created. For instance, the NGS board is very unique in relation to entire or whole-genome sequencing (WGS), focus on less gene sets or locales yet requiring more noteworthy exactness and effectiveness. For complex maladies, for example, malignant growths, the driver genetic elements are normally a bunch of qualities in a positive or negative regulatory organization. Chart speculations, for example, briefest way examination and irregular walk calculations, will help dismember entire genomic communications into key modules or ways whose brokenness is related with infectious propagation. The genomics technologies enabling the search and filter genes and their variants from the whole genome and the pharmacogenomics identify the patient specific drug associated variant genes. In the practical approach of personalized medicine patient's genomic and proteomic data processed to digital data. Then the stored data retrieved and analyzed through bioinformatics tool for this tailor made genomics-based drug discovery and therapy. This selects right drug and dose based on individual's genomic data processed, analyzed, clinical, and along with disease environmental data. This field is young and evolving in the healthcare system. The pharmaco-dynamics, genomics, and kinetics involvement has an important role in the succession of PM proceedings. The information and communication technology (ICT) manage patient's personal and medical data (Louca 2012). Bioinformatics and data mining are combined to

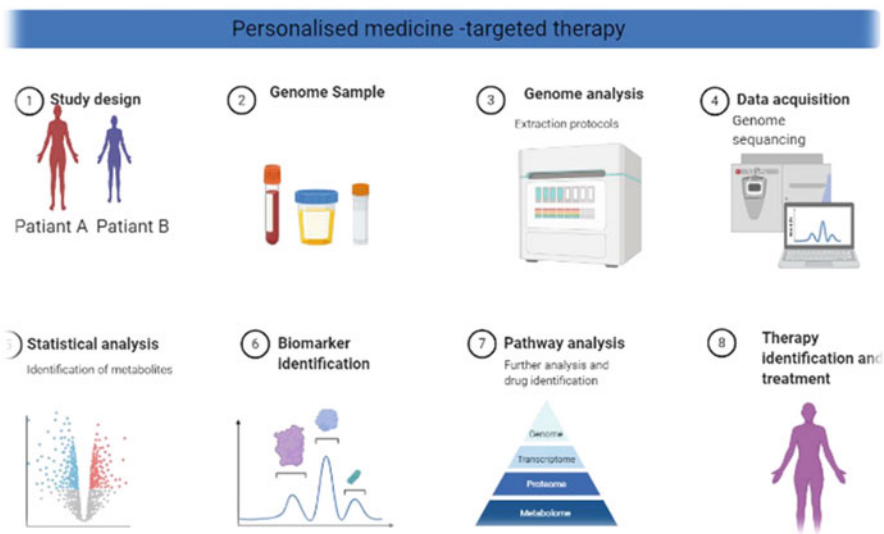


Fig. 15.1 Bioinformatics and genomics protocol in personalized medicine

create tools and procedures for the prediction of emerging, recurrence, progression, response of disease to treatment. Individualizing drug or vaccine therapy with the use of bioinformatics and pharmacogenomics tools have the prospective to transform health care system (Mancinelli et al. 2000). The whole-exome sequencing have proven to be valuable methods for the discovery of the genetic causes of rare and complex diseases (Gonzaga-Jauregui et al. 2012). Bioinformatics and genomics protocol in personalized medicine summarized in Fig. 15.1.

15.2 Significance of Personal Medicine and Bioinformatics

Many drugs and vaccines are failing to deliver optimum response to broad population. PM is a combinational approach to individual health care. This is required for the improvement of early disease diagnosis and treatment at individual level. Each individual's pre- or post-disease clinical, genomic, and environmental information are not unique. Genome-wide association studies helped to identify genes important in serious adverse drug reactions (Daly and Day 2012). In the most recent decade, biochemical science has made numerous advances to personalized medication, including the Human Genome venture, International HapMap task, and genome-wide affiliation contemplates (GWASs). Single nucleotide polymorphisms (SNPs) are currently perceived as the fundamental driver of human hereditary fluctuation and are as of now a significant asset for planning complex hereditary characteristics. A great many DNA variations have been distinguished that are related with ailments and attributes. By joining this hereditary relationship with phenotypes and medication reaction, customized medication will tailor medicines to the patients' particular

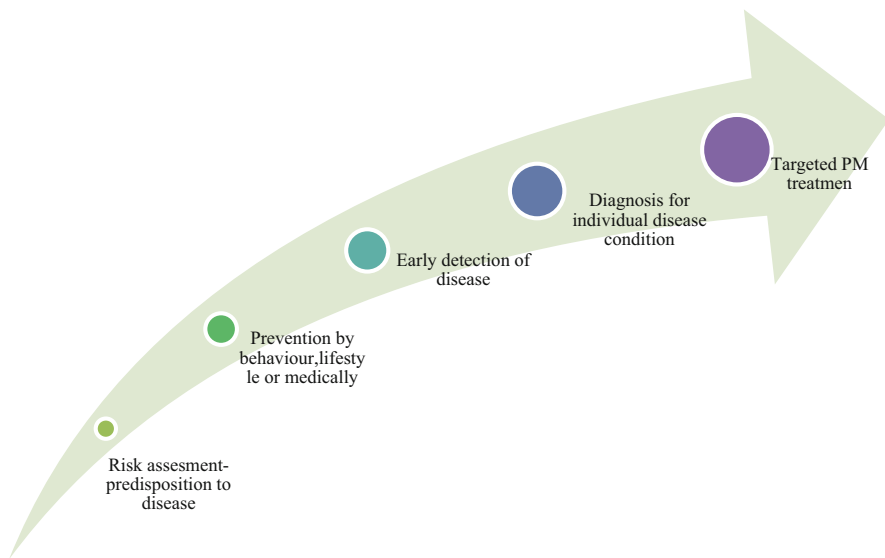


Fig. 15.2 Stages of personal medicine design

genotype. Albeit entire genome groupings are not utilized in ordinary practice today, there are as of now numerous instances of customized medication in current practice. Figures 15.2 and 15.3 explain various stages and steps of personal medicine design, respectively. Chemotherapy prescriptions, for example, trastuzumab and imatinib target explicit diseases, a focused on pharmacogenetic dosing calculation is utilized for warfarin and the frequency of unfavorable occasions is decreased by checking for powerless genotypes for drugs like abacavir, carbamazepine and clozapine.

Customized medication is required to profit by consolidating genomic data with customary checking of physiological states by different high-throughput methods. Over the previous decade, upgrades in instrument affectability, speed, exactness, and throughput, combined with the improvement of innovations, for example, various responses observing. Under the direction of the Human Proteome Organization over 80% of the proteins anticipated by the human genome have now been recognized utilizing either mass spectrometric or immunizer based procedures, and the staying “missing proteins” are as a rule consistently represented. Assets, for example, the Human MRM Atlas, a far reaching asset intended to empower researchers to perform quantitative examination of every human protein, are being created to encourage reproducible exchange of quantitative tests between labs. Such turns of events and activities currently empower both top to bottom disclosure and focused on/quantitative work processes, making the way for the clinical analytic field. Combined with this, the foundation of exhaustive information bases and the improvement of amazing in silico methods is empowering viable information mining. Specifically this has empowered interactome examines permitting the recognizable proof of key flagging pathways prompting potential new medication

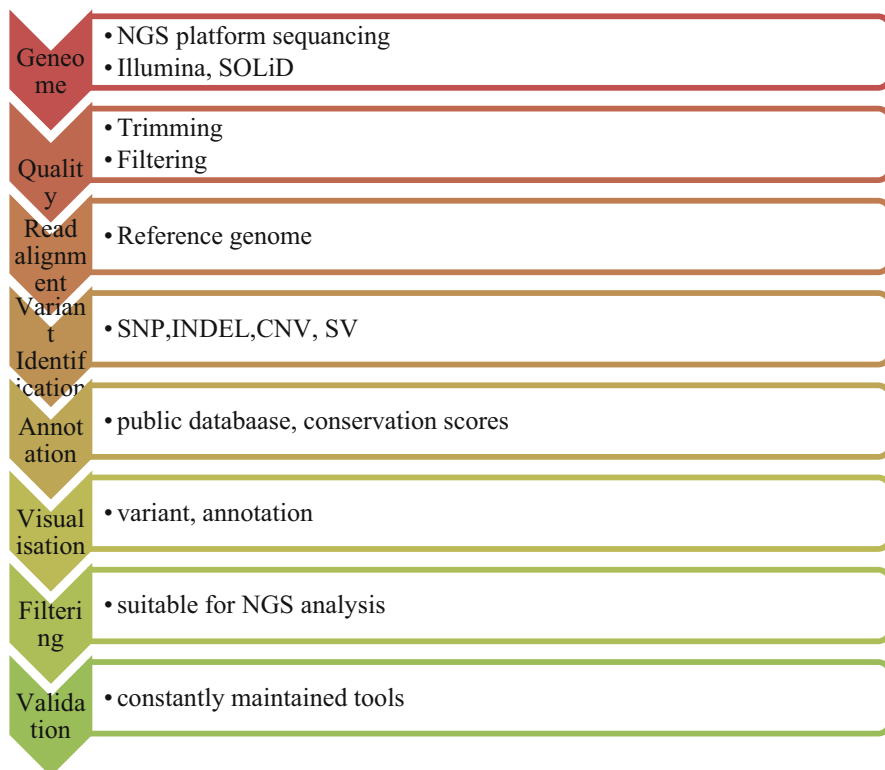


Fig. 15.3 Bioinformatics steps in genomic data processing for PM

targets, despite the fact that to date it has been assessed that under 20% of the protein communications in people, not including dynamic, tissue-or infection explicit associations, have been distinguished (Chen et al. 2012).

15.3 Application of Bioinformatics in Personal Medicines and Vaccines

Bioinformatics tools may provide better diagnoses in genomic level with earlier detection of disease and better targeted therapy through efficient personalized medicine development. Omics analysis provides a great assistance in the development of personalized medicine (Fig. 15.4). Bioinformatics tools helps in diagnosis, intervention, drug development, therapy, and personalized vaccination. Personalized vaccine means for an optimized prevention of disease with minimized reactogenicity and side effect. Personalized vaccines are developed to take care of haplotypes and polymorphism can become risk of an adverse vaccine reaction. Variation in gender, racial, ethnic, gene polymorphisms and other environmental factors affect the in

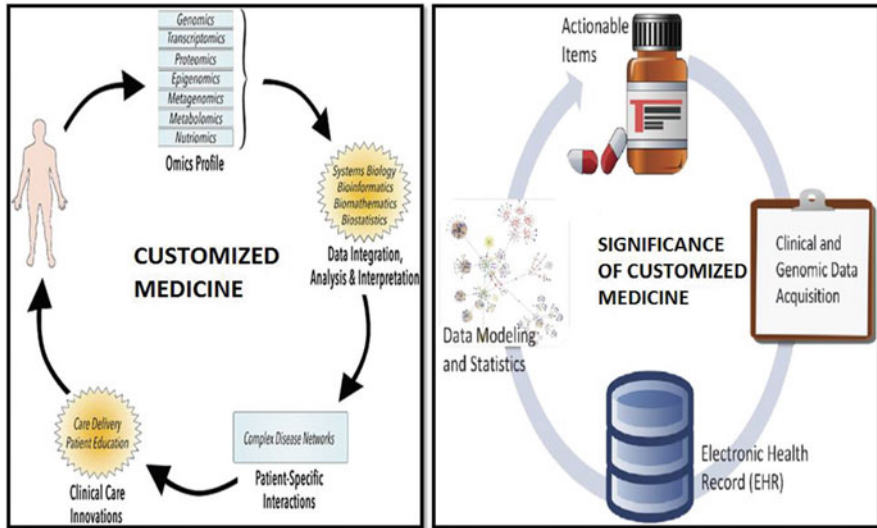


Fig. 15.4 Customized or personal medicine based on omics analysis

immune response to a particular vaccine where we required a personalized vaccine and drugs (Atsaves et al. 2019; Poland et al. 2011). Insilico designing helps vaccines and vaccine adjuvants design for immunologically different groups (Piasecka et al. 2018; Oli et al. 2020). The identification of Human Leukocyte Antigen (HLA) complex and stable polymorphism and effective vaccine for the each individual is possible through bioinformatics analysis of HLA class I and II molecules and predict suitable peptide for HLA binding (Gfeller and Bassani-Sternberg 2018). Due to the expression differences of MHC- HLA allele to viral proteins the T cell responses varies (Clemens et al. 2018; Auladell et al. 2019). The artificial neural network algorithms and datasets made possible to develop different epitope predicting tools. The tools help to predict the epitope peptides for a particular MHC-HLA binding of an individual (Chandra and Yadav 2016). Peptide motifs and MHC ligands databases obtained from epitope peptide prediction servers (Lundegaard et al. 2010; Glutting and Reinherz 2002). Immunoinformatics prediction of Immunodominant epitopes (SSNLYKGVY) from AA41-49 of glycoprotein 1 of *Lassa fever* virus can induce of humoral and cell-mediated immunity African populations and endemic country (Hossain et al. 2018) and in *Oropouche* virus (Adhikari et al. 2018). These approaches help individualized vaccination and prevent endemic diseases. The diversity of HLA regions suspected for the generation vaccine immune response in each individual (Kaifu and Nakamura 2017). The genetic variation in male and female may leads to differences in immune responses against influenza (Voigt et al. 2019; Fink et al. 2018), rubella (Mitchell et al. 1992), and measles immunization (Fischinger et al. 2019). The new cutting edge technology like vaccinomics a combination of immunogenomics, bioinformatics and immunogenetics could be helpful in the personalized vaccine development (Majumder

2015). The personalized vaccinology and medicine developed through international HapMap and that of the Human Genome Project. The variation in gene level, linkage disequilibrium maps, and single nucleotide polymorphism (SNP), have significant roles in immune responses (Brodin and Davis 2017; Cotugno et al. 2019). The sequencing technologies, bioinformatics analysis tools, genotypic and phenotypic data bases advances the immune response prediction of drugs, vaccines, insecticides and diagnostics (Gunawardena and Karunaweera 2015). Immunoinformatics studies were found to be successful in predicting epitope based vaccines for SARS-Cov2 (Joshi and Kaushik 2020; Akhtar et al. 2020), Dengue virus (Sunil Krishnan et al. 2020), and Nipah virus (Kaushik 2019), even the rarest bacterium like *Tropheryma whippelii* causing lipodystrophy could also be successfully targeted by epitope based vaccine formulations (Joshi et al. 2020).

15.4 Advantages and Disadvantages of PM

The upsides of PM would be relevant in the uncommon and complex ailment the board by refining patients and care suppliers, quicken exploration, and supporting vital changes in strategy and guideline. The new bioinformatics explores have been planning apparatuses and test pipelines to investigate singular affliction circumstance. The progressing customized medication has been in understanding consideration relevant to cardiovascular sicknesses, Mendelian problems, malignant growths, Kabuki condition and hereditarily heterogeneous issues (schizophrenia, irregular mental imbalance and range issues) (Table 15.1).

Phases involved in genomics analysis (Fig. 15.5) are sequencing by deploying next-generation approach like illumina solexa, 454 pyrosequencing, Ion torrent, etc. After sequencing genetic sets are analyzed to detect epigenetic relationships, to determine phylogenetic expressions involved to accumulate information in databases that can be used for personalized medication formulations.

Table 15.1 Personal medicine's advantage and disadvantages

S. no.	Advantages	Disadvantages
1	Minimize the incidence of adverse effect of treatment	Expensive and not accessible to everyone
2	Understanding of the individual patient or population treatment need	Economically impossible to target small patient populations
3	Interpret genetic information	Technology not licensed
4	Advancing personalized medicine in patient care	Fear of data leakage
5	Greater precision in diagnosis and more targeted drug development	Service providers are not common
6	For rare and complex diseases	Not in all cases successive
7	Increasing the accuracy of diagnosis	Need more tools to be developed to interpret the data

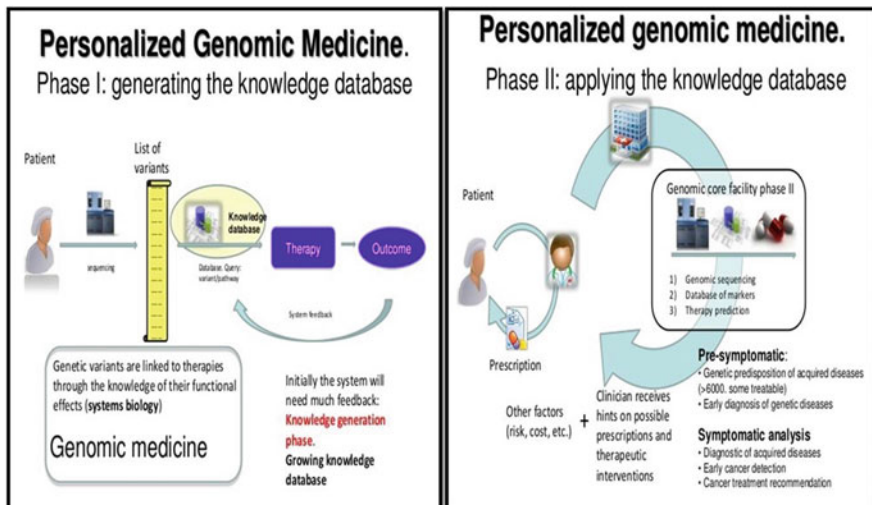


Fig. 15.5 Phases involved in genome based personalized medicine development

15.5 Bioinformatics Prerequisites Challenges for Personal Medicine Design

Computational and sequencing Infrastructure, availability of individual genomic data, Data quality, bioinformatics data analysis tools, computational pipelines, interpretation, and validation of biomarkers. The processed data analysis has five analytical steps like quality assessment, alignment, variant identification, variant annotation, and visualization. For the advancement of PM many challenges have to be overcome. The availability patient's genomic data are consulted only for a little treatment plans and hardly few medical centers used for treatment (Yngvadottir et al. 2009). Bioinformatics tools would help the diverse genomics data for PM design for individual patients. The challenge includes availability of computational or sequencing infrastructure, error rate in individual genomic data (1000 Genomes Project Consortium et al. 2010), data quality, bioinformatics data analysis tools, computational pipelines for large data processing, and validation of biomarkers. The processing such large amounts of genetic data obtained from next-generation sequencing (NGS) requires bioinformatics dataprocessing . High amount of data and its accuracy challenges for the analysis and interpretation. Through NGS the detection of Copy number variants (CNVs) and structural variants (SVs) are more difficult (Shendure and Ji 2008). Bioinformaticians have developed new algorithms for tools like BLAST (Altschul et al. 1990), BLAT (Kent 2002), BWA (Li and Durbin 2009), and MOSAIK, developed by the Marth Lab (Michael Stromberg, Boston University) to address these problems in different times.

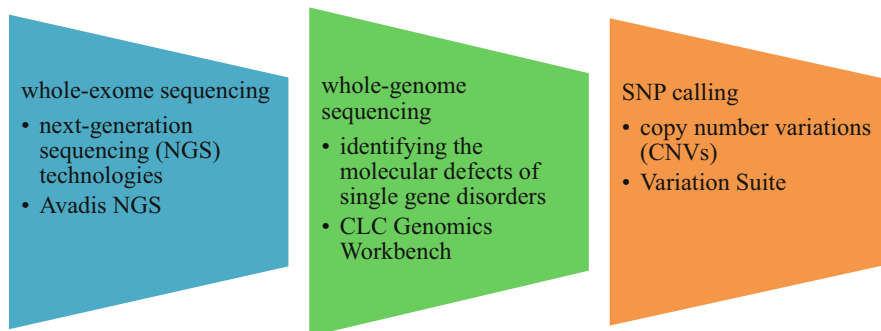


Fig. 15.6 General Sequencing and bioinformatics tools used in the PM design

The whole-genome and whole-exome data interpretation has an important role in the experimental success (Schadt et al. 2010). General Sequencing tools used in the PM design summarized in Fig. 15.6. The appropriate choice of tools, data handling and tool compatibility programs for variant analysis of NGS data. Intellectual property rights, reimbursement policies, patient privacy, data bases and confidentiality as well as regulatory oversight. The Vaccine, therapeutics or drug development process, and regulatory requirement need to be changed for targeting smaller patient populations with rare diseases.

15.6 Advanced Methods Involved in Personalized Medicine Designing

Coordinating a lot of information originating from high-throughput advances towards customized medication and diagnostics cannot be conceivable without utilizing computational ways to deal with sort out the unpredictability of handling and associating numerous factors at the “omics” level. Bioinformatics is an interdisciplinary field of science that is centered on applying computational methods for the investigation and separating data from information originating from biomolecules. Typically, it coordinates methods from the fields of informatics, software engineering, sub-atomic science, genomics, proteomics, arithmetic, and measurements. In spite of the fact that it began as a field completely committed to essential exploration in advancement and hereditary qualities, it has been advancing in corresponding with high-throughput strategies bringing about the improvement of numerous techniques and apparatuses that encourage the translation of “omics” information. In bioinformatics, high-throughput information is prepared and broke down methodically from crude information to the outcomes utilizing pipelines of examination utilizing the maximum capacity of PCs. Bioinformatics pipelines typically contain various strides for information quality evaluation, include extraction, measurement decrease, biomarker recognition, and results age. This arrangement of examination is

completely robotized where the client has no obstruction except for can assume the part of “caretaker” to check the approval of the yields (results).

With the advancement of the computational force, bioinformatics picked up the possibility to handle huge information and incorporate a lot of information a lot quicker than it is delivered, turning into an answer applying high-throughput methods in clinical diagnostics and customized medication. For instance, a few examinations have shown that bioinformatics pipelines produced for the investigation of MALDI-ToF mass spectra can extricate symptomatic data from pee, blood, and undeveloped organism culture media quicker than its ability of being created. In genomics, a few bioinformatics pipelines of examination for NGS, RNAseq, and microarrays have been additionally evolved to remove analytic data out of sequencing of infection, obsessive microorganisms, and malignant growth biopsies. In addition, bioinformatics instruments for preparing “omics” have likewise been fruitful in the revelation of novel medication focuses for malignant growth treatment. Bioinformatics can additionally improve clinical research facilities proficiency and expenses by sparing time and HR on the investigation and answering to centers and patients. This should be possible by creating pipelines of examination with computerized revealing and APIs completely committed to giving constant online access, encouraging the correspondence between labs, clinicians, and patients. Also, persistent chronicled information and metadata ought to be secure and sorted out in an organized manner (information “stockrooms”) with the end goal that it very well may be additionally pulled efficiently to bioinformatics pipelines. This would permit going past in integrative examination of patients by having their information as a component of time permitting a more customized checking of patients indicative and permitting better prognostics.

Apparatuses with direct significance to customized medication

1. Biomarker-driven medication: multi-omics, IT, approval, reproducibility, clinical utility.
2. Genomics information translation, in addition to phenotypes.
3. Man-made consciousness, Machine Learning, Simulation.
4. Resident Science, Biobanks, Health Data Cooperatives.
5. European frameworks for customized medication (for example, open science cloud).

The advancement of numerical models and calculations that produce strong expectations is a hard undertaking and requires thorough approval techniques before an indicator is fit to be dispatched into the market. Relatively few indicators for diagnostics are accessible to be utilized or can be adjusted to a given clinical lab setting. Accordingly, model turn of events and enhancement for every lab would be the ideal situation. Incorporating prescient displaying work processes in bioinformatics pipelines likewise encourage model advancement by organizing the cycle of approval and model determination utilizing the information and metadata. A few kinds of models can be utilized to settle on indicative expectations and the decision relies upon the information accessible, innovation, and the idea of the issue.

Measurable models dependent on known appropriations of biomarkers are normal to be utilized in the demonstrative of a specific illness. These are anything but difficult to actualize in bioinformatics pipelines and fill in as correlative data for clinicians. Execution of example acknowledgment, AI, and man-made brainpower (AI) calculations into bioinformatics pipelines are critical to enhance numerical models towards meeting more exact forecasts. Critically, the utilization of AI and AI calculations are fundamental for customized medication since they empower the fitting of conventional models of illness to every patient situation and body science. Deterministic models, for example, the coherent and dynamic demonstrating structures can likewise be utilized for reenactment of physiological situations and making powerful forecasts with clinical applications.

For instance, reproduction of the tumor miniature condition utilizing a consistent organization model of the guideline of cell attachment properties permitted to build up relations between malignancy de-guidelines and the metastatic potential. This has a tremendous potential for the future improvement of bioinformatics apparatuses that permit the expectation of the metastatic potential and propose the best treatment for each case dependent on the tumor biopsy. Dynamic models, then again, can possibly be more exact and produce a persistent scope of expectation esteems. Notwithstanding, their boundary assessment is perplexing and requires AI calculations to adjust them to a specific physiological framework. These kinds of models are phenomenal for portraying the digestion and can be valuable in as future apparatuses in customized medication (Dodin 2017).

15.7 Conclusions

The enhancement of bioinformatics tools and databases development for new diseases would be helpful for the advancement of PM. In future looking for more coverage, affordability in genome data processing, accuracy in data interpretation, fast genetic data processing, development of more bioinformatics tools the understanding of disease at the molecular level, and bioinformatics advancement for data interpretation. This would help the elevated success rate in this new young health care field. Discovery of bioinformatics tools would help to integrate the huge genomics data analysis and speed up the PM research. As the environment of partners attempts to progress customized medication, cooperation with government controllers and policymakers is important to empower inescapable utilization of these new devices and technology advances. The administrative cycle must develop in light of advances that are focused to littler patient populaces dependent on hereditary profiles, and arrangements and enactment must be sanctioned that give motivating forces to inventive exploration and selection of new advances. Together, progress in the exploration, clinical concern, and strategy empowering customized personal therapy can possibly improve the nature of patient consideration and to help contain medical services costs.

References

- Adhikari UK, Tayebi M, Rahman MM (2018) Immunoinformatics approach for epitope-based peptide vaccine design and active site prediction against polyprotein of emerging oropouche virus. *J Immunol Res* 2018:1–22
- Akhtar N, Joshi A, Singh B, Kaushik V (2020) Immuno-informatics quest against COVID-19/ SARS-COV-2: determining putative T-cell epitopes for vaccine prediction. *Infect Disord Drug Targets* 20:32957905. <https://doi.org/10.2174/1871526520666200921154149>
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410
- Atsaves V, Leventaki V, Rassidakis GZ, Claret FX (2019) AP-1 transcription factors as regulators of immune responses in cancer. *Cancer* 11(7):1037
- Auladell M, Jia X, Hensen L, Chua B, Fox A, Nguyen TH, Kedzierska K (2019) Recalling the future: immunological memory toward unpredictable influenza viruses. *Front Immunol* 10
- Brodin P, Davis MM (2017) Human immune system variation. *Nat Rev Immunol* 17(1):21–29. <https://doi.org/10.1038/nri.2016.125>
- Chandra H, Yadav JS (2016) Human leukocyte antigen (HLA)-binding epitopes dataset for the newly identified T-cell antigens of mycobacterium immunogenum. *Data Brief* 8:1069
- Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148(6):1293–1307
- Clemens EB, Van de Sandt C, Wong SS, Wakim LM, Valkenburg SA (2018) Harnessing the power of T cells: the promising hope for a universal influenza vaccine. *Vaccine* 6(2):18
- Cotugno N, Ruggiero A, Santilli V, Manno EC, Rocca S, Zicari S, Amodio D, Colucci M, Rossi P, Levy O, Martinon-Torres F, Pollard AJ, Palma P (2019) OMIC technologies and vaccine development: from the identification of vulnerable individuals to the formulation of invulnerable vaccines. *J Immunol Res* 2019:8732191. <https://doi.org/10.1155/2019/8732191>
- Daly AK, Day CP (2012) Genetic association studies in drug-induced liver injury. *Drug Metab Rev* 44(1):116–126
- Dodin G (2017) Personal genomics: new concepts for future community data banks. *bioRxiv*. <https://doi.org/10.1101/230516>
- Fink AL, Engle K, Ursin RL, Tang WY, Klein SL (2018) Biological sex affects vaccine efficacy and protection against influenza in mice. *Proc Natl Acad Sci* 115(49):12477–12482
- Fischinger S, Boudreau CM, Butler AL, Streeck H, Alter G (2019) Sex differences in vaccine-induced humoral immunity. In: *Seminars in Immunopathology*. Springer, Berlin Heidelberg, pp 239–249
- Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061
- Gfeller D, Bassani-Sternberg M (2018) Predicting antigen presentation—what could we learn from a million peptides? *Front Immunol* 9:1716
- Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. *Annu Rev Med* 63:35–61
- Gunawardena S, Karunaweera ND (2015) Advances in genetics and genomics: use and limitations in achieving malaria elimination goals. *Pathogens Global Health* 109(3):123–141. <https://doi.org/10.1179/2047773215Y.0000000015>
- Hossain MU, Omar TM, Oany AR, Kibria KK, Shibly AZ, Moniruzzaman M, Islam MM (2018) Design of peptide-based epitope vaccine and further binding site scrutiny led to groundswell in drug discovery against Lassa virus. *3 Biotech* 8(2):81
- Joshi A, Kaushik V (2020) In-silico proteomic exploratory quest: crafting T-cell epitope vaccine against Whipple’s disease. *Int J Pep Res Therapeut* 18:1–11. <https://doi.org/10.1007/s10989-020-10077-9>
- Joshi A, Joshi BC, Mannan MA, Kaushik V (2020) Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. *Inform Med* 19:100338. <https://doi.org/10.1016/j.imu.2020.100338>

- Kaifu T, Nakamura A (2017) Polymorphisms of immunoglobulin receptors and the effects on clinical outcome in cancer immunotherapy and other immune diseases: a general review. *Int Immunol* 29(7):319–325
- Kaushik V (2019) Silico identification of epitope-based peptide vaccine for Nipah virus. *Int J Pept Res Ther* 26(2):1147–1153. <https://doi.org/10.1007/s10989-019-09917-0>
- Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Res* 12(4):656–664
- Li H, Durbin R (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25(14):1754–1760
- Louca S (2012) Personalized medicine—a tailored health care system: challenges and opportunities. *Croat Med J* 53(3):211–213
- Lundegaard C, Lund O, Buus S, Nielsen M (2010) Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology* 130(3):309–318
- Majumder PP (2015) Genomics of immune response to typhoid and cholera vaccines. *Philos Trans Royal Soc London* 370(1671):20140142. <https://doi.org/10.1098/rstb.2014.0142>
- Mancinelli L, Cronin M, Sadée W (2000) Pharmacogenomics: the promise of personalized medicine. *AAPS PharmSci* 2(1):29–41
- Mitchell LA, Zhang T, Tingle AJ (1992) Differential antibody responses to rubella virus infection in males and females. *J Infect Dis* 166(6):1258–1265
- Oli AN, Obialor WO, Ifeanyichukwu MO, Odimegwu DC, Okoyeh JN, Emechebe GO, Ibeanu GC (2020) Immunoinformatics and vaccine development: an overview. *ImmunoTargets Ther* 9:13
- Piasecka B, Duffy D, Urrutia A, Quach H, Patin E, Posseme C, Hasan M (2018) Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc Natl Acad Sci* 115(3):E488–E497
- Poland GA, Kennedy RB, Ovsyannikova IG (2011) Vaccinomics and personalized vaccinology: is science leading us toward a new path of directed vaccine development and discovery? *PLoS Pathog* 7(12):e1002344
- Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol* 63(9):701–709
- Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. *Nat Rev Genet* 11(9):647–657
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26(10):1135–1145
- Sunil Krishnan G, Joshi A, Kaushik V (2020) T cell epitope designing for dengue peptide vaccine using docking and molecular simulation studies. *Mol Simul* 46(10):787–795. <https://doi.org/10.1080/08927022.2020.1772970>
- Voigt EA, Ovsyannikova IG, Kennedy RB, Grill DE, Goergen KM, Schaid DJ, Poland GA (2019) Sex differences in older adults’ immune responses to seasonal influenza vaccination. *Front Immunol* 10:180
- Yngvadottir B, MacArthur DG, Jin H, Tyler-Smith C (2009) The promise and reality of personal genomics. *Genome Biol* 10(9):237



Amit Joshi, Jitendra Sasumana, Nillohit Mitra Ray, and Vikas Kaushik

Abstract

Neural networks play very significant role when it comes to analysis of proteins and nucleic acid sequences. Many of the pattern recognition software are based on neural networks for prediction of biological patterns. Modern sequencing advancement fuels up the collection of data related to DNA, RNA, and protein sequences. The complexity and enormous size of this data require best computational algorithms for analysis and interpretation. This information will assist in developing useful insight for biomolecular structural predictions and prediction of interactions between such molecules. A neural system investigation framework is a succession of computations that attempts to see concealed associations in a lot of data through a technique that imitates the way where the human mind works. In this sense, neural frameworks suggest systems of neurons, artificial in nature. Vectors and matrices based linear algebra and topology designs supported various types of neural architectures. Neural frameworks can conform to advancing info; so the framework makes the best result without hoping to refresh the yield rules. The possibility of neural frameworks, which has its basic establishments in man-made consciousness, is rapidly getting ubiquity in the progression of in silico designing systems. Here, we talk about and sum up the uses of Neural Networks in computational biology, with a specific spotlight on applications in protein and Nucleic acid bioinformatics. We concluded with giving basic insights of neural networks in multiple domains of life sciences like gene prediction, protein structure prediction, epitope prediction, expression, co-expression, protein–protein interaction, and many other domains.

A. Joshi · J. Sasumana · N. M. Ray · V. Kaushik (✉)

Domain of Bioinformatics, School of Bioengineering and Biosciences, Lovely Professional University, Phagwara, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_18

351

KeywordsProtein · DNA · Bioinformatics · Neural networks

18.1 Introduction

Bioinformatics is an amalgamation of biotechnology and computer science, to provide better understanding of biomolecular interactions. From the beginning of Human genome project it was realized that the storing sequencing data will assist in future medicinal developments (Collins et al. 2003). DNA, RNA, Protein sequences of different organism serve as biological data. Enormous amount of biological data requires analysis to generate applicable information like genomic and proteomic interactions. Such bioinformatics analysis assisted not only in finding similarities between sequential stretches of nucleotides or amino acids but also in determining expression levels and control for genomic functional sets. In the past decade, determination of protein based vaccine candidates for different viruses and bacterial organisms become very easy for computational biologists. Neural network is based on cognitive learning of system with statistical probabilistic approach to predict the outcomes in a similar fashion like millions of interconnected functional neurons does (Hopfield 1982). Prediction based modeling of bimolecular structures, and determination of their functionality was found as greater achievement of artificial neural networks in denovo studies. These neural artificial intelligence systems can be utilized for prescient displaying, versatile control and applications where they can be prepared through a dataset. Self-evolving occurs because of experience that exists inside such systems, which can get inferences from a complex and apparently random arrangement of data. Propagation of theories related to neural networks were credited to Alexander Bain (Bain 1873), suggested that interconnections and electrical activity between neuron were responsible for cognitive learning and behavioral actions (Evans 1990). A neural framework is synaptic organization of counterfeit neuronal units that represent an artificial scientific or programming based coded model for information processing. Artificial neural network (ANN) is a flexible structure that changes its structure subject to outside or inside information that travels in the wholesome framework. In this chapter we observe fundamentals of algorithms behind ANN based web servers used in in silico methodologies. Many mathematical modeling tools along with coding developments lead to design fast and effective software and algorithms that evolve with more input data (Biological sequences like DNA, RNA, and proteins). Also observe tools that are deployed in proteomics and genomics for describing biomolecular structure, properties, and functional interactions. Firstly we will introduce biochemical background then summaries the type of neural networks algorithm that were commonly used in recent software's/servers designing for accurate structural and functional predictions. Also details of applications in genomics, transcriptomics, and proteomics are given to develop understanding of neural networks. Machine learning in silico tools based on artificial neural networking are very useful in bioinformatics and provide ease in

genomic as well as proteomic analysis. This will resolve big data analysis problems too. Aim of this chapter considered useful, as enormous data should not become problem for investigator rather it will act as boon for life sciences and medical world to develop better and fast regimens for several diseases and this will also interconnect health sector globally.

18.2 Biochemistry and Bioinformatics Background

Proteins and nucleic acids are integral part of each and every cellular entity of all living organisms. These biomolecules participate in almost all functional activities within cells from metabolic reactions to genetic expression of transcriptomes. Many life sustaining processes like cell cycle, apoptosis, enzymatic catalysis, cell signaling, adhesion, and central dogma are always depending on protein–protein and DNA–protein interactions. Proteins and nucleic acids are macromolecular heteropolymers of amino acids and nucleotides, respectively (Giorgini et al. 2020). Different proteins have different amino acids sequences, peptide bond forms between two amino acids due to bonding between amino and carboxyl group of adjacent amino acids (Fig. 18.1b and 18.1c). Similarly in nucleic acid (DNA or RNA) phosphodiester bonds exist between two adjacent nucleotides (Deoxyribose or Ribose sugar, Nitrogenous base (A,T,C,G), and phosphate group) (Fig. 18.1a). Sequencing studies produce enormous data regarding DNA, RNA, and protein sequences. These sequences are stored in databases like NCBI Genbank, DDBJ, and EMBL. Similarly structural information of proteins and ligands interacting to

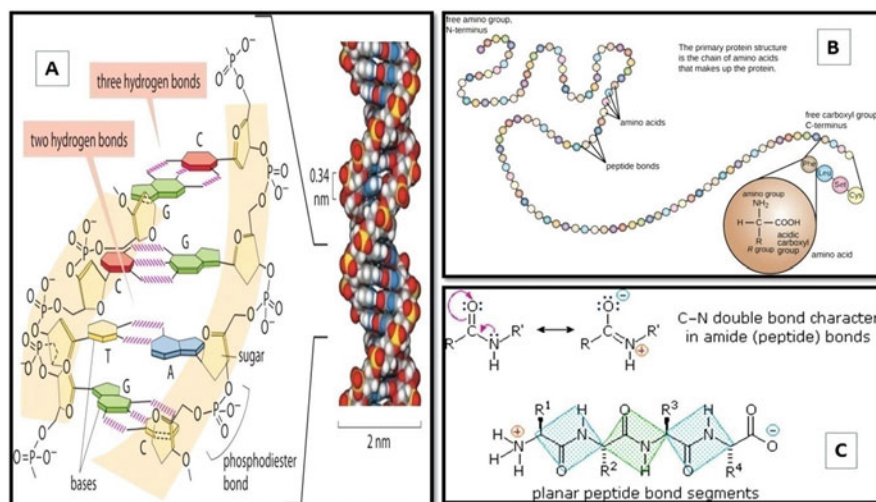


Fig. 18.1 (a) DNA structure: revealing phosphodiester bonds and hydrogen bond (b) protein chain: primary structure (c) peptide bond between amino acids

them are also submitted in databases like Pubchem, Chembridge, Maybridge, and Protein Data Bank (RCSB-PDB) by X-ray crystallographic experiments.

Bioinformatics not only provide platform to analyze the structures of protein in primary, secondary, and tertiary forms but also allow user to bring structural changes according to the influence of molecules in its vicinity. This means that protein structures show bonding or interactions with other biomolecules, which can impart inducible changes as in case with induced fit model analogy for enzymatic actions (Morgat et al. 2020). Computational studies created applications for deep neural networks to assist in prediction of protein–protein binding pockets or interaction sites (Zeng et al. 2020). Even protein contacts can also be predicted by metagenomic sequence data and residual neural networks (Wu et al. 2020). Such all recent studies indicate the importance of neural network bioinformatics and biochemistry together to generate a big informative picture of biomolecules functional aspect with accuracy and precision. These studies also suggest that the unsolved structure for known sequences could be easily determined by deploying neural networks architectures in biochemistry and medicine studies.

18.3 Neural Networks and Its Types

Neural networks are part of Artificial intelligence and Machine learning. These networks works like brain neurons, these networks are dependent on weights as we increase loads the networks learn more to predict suitable results or outputs. Deep learning becomes more advanced with the increase in data. Therefore, a neural network also adjusts its performance to greater extant as they grows bigger and deals with enormous flow of information. Neural networks are best in comparison to other machine learning tools that reach a plateau after a point. Activation functions play crucial role in switching on and off artificial nets that connects artificial neural elements. This allows systematic flow of information and deep learning in software based systems (Khan 2020). Each neural element receives multiple inputs and randomized loads and adds them to static bias of each neural element, then directs them to activator function that finally brings output of desired neural element of network (Fig. 18.2a). Activator function can be of linear (simplest), heviside step, and sigmoid function (complex) type. When final neural layer generates output, loss functions are calculated on the basis of inputs and outputs and back propagation conducted to bring alterations in loads that lead to loss minimization (Fig. 18.2b). To determine overall loads adjustment is main or central criteria of neural network architectures (Galushkin 2007).

Neural network architectures can be divided in many subtypes: based on framework, Datum transfer, Counterfeit neurons with weighted-density, multiple-layering and activation functions (Amato et al. 2013). Common types of neural networks are Feed forward network, Multi-layer perceptron, Convolutional neural network, Radial basis neural network, Recurrent neural network, modular networks, etc. (Fig. 18.2c and Fig. 18.3).

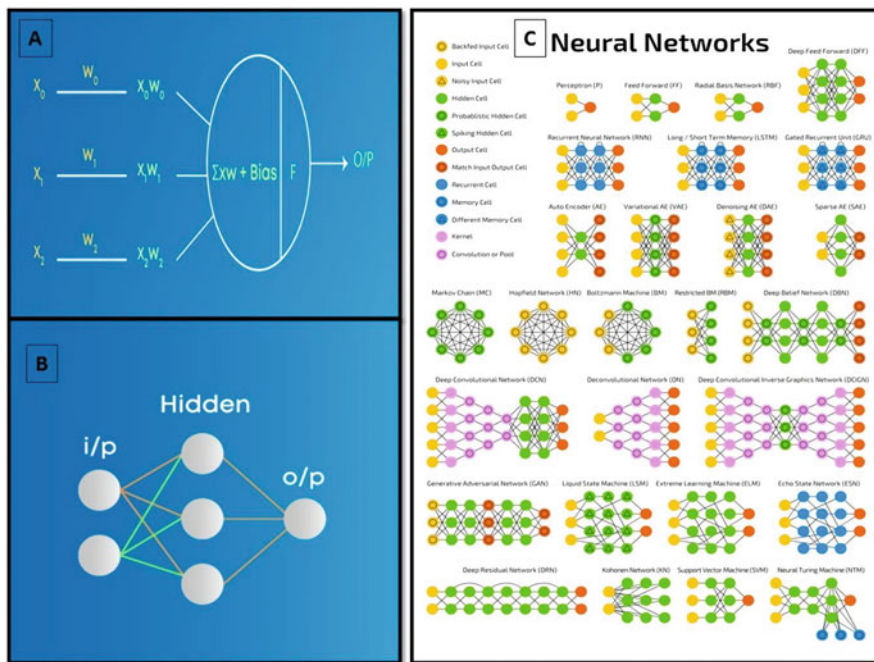


Fig. 18.2 (a) Load (numeric) values product with input data in back track to minimize loss, and the relation with activation function “F” to generate output. (b) Input layer “i/p” exhibits dimension of input vector, Hidden layer shows intermediary nodes separating input spaces with boundary limits that consider input load sets to synthesize information by activation function. Information as output “o/p” layer shows final information via ANN system architecture. (c) Various types of neural network architectures

Feed forward neural network was simplest first ANN to be deployed in bioinformatics. In such system one way flow of information exist (from input to hidden to output). No loops or closed cycles exist in such architecture.

The Feed forward network does not possess backward propagation. These systems have static loads (Shao 2020). Mostly step activation function is used here with 0 to 1 criteria ($f(v) = 1$ iff $v \geq a$; & $f(v) = 0$ iff $v < a$; where $v = \sum w_i x_i$, and $a =$ threshold). The neuron is actuated in the event that it is above edge (typically 0) and such counterfeit neuron generate 1 (informative yield). Counterfeit neuron is not enacted on the off chance that it is beneath edge (typically 0) which is considered as - 1. They are genuinely easy to keep up and are furnished with to manage information which contains a great deal of commotion. Significant for analysis as simple to design, fast, and also generate good responsiveness to noise. Only disadvantage is that it cannot be deployed in AI-processing tasks because of no deep stratifications and reverse tracking.

Multilayer-supervised model is advancement in Feed forward neural network. Each and Every single node is interconnected. Input and output layers are found in between of multiple hidden Layers (Heidari et al. 2020). It involves forward and

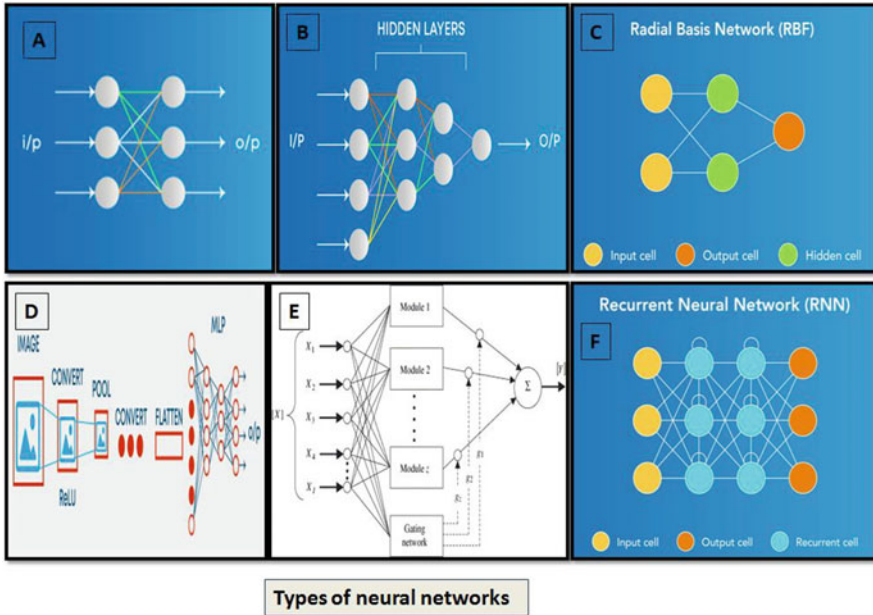


Fig. 18.3 Types of neural networks: (a) Feed forward network, (b) multi-perceptron network, (c) radial basis network, (d) convolution network, (e) modular network, (f) recurrent neural network

backward preparative tracking. Sources of info multiplied with loads and afterward exposed for activation function calculations along with reverse tracking; each informative node shows such alterations, so it can decrease the mislaying in information. Loads are machine taken in values from Neural Networks. Loads (W_i) are competent to self-modify contingent upon the contrast between anticipated yields versus preparing inputs. Nonlinear-initiation activator function is conveyed here, which makes them complex and best for deep learning tasks. Disadvantage includes comparatively slow functionality in huge data analysis.

Convolution neural network contains a 3D course of action of neurons, rather than the standard 2D arrangement. The principal layer is known as a convolution filter with activation mapping of counterfeit neurons. Every neuron in the convolution filter only processes the data from a little piece of the image related data (Chen et al. 2020). Information highlights are taken in clump astute like a channel. The system comprehends the pictures in parts and can figure these activities on various occasions to finish the full picture preparing. Handling includes change picture standards between RGB and dark scaling. Promoting adjustments for CRT screen dots worth assist with identifying corners, such pictures will be easily characterized. Engendering or tracks are following one-direction flow and convolution neural network holds at least one convolution filter accompanied by amalgamation of information and dual-directional when yield of convolution filters transfers information towards associated neural system for ordering the pictures as appeared in the

Fig. 18.3. Channels are utilized to remove certain pieces from picture. In Multilayer-supervised model the sources of info are duplicated with loads and subjected to the activation function. Convolution utilizes nonlinear enactment function followed by softmax. Convolution neural systems show viable outcomes in picture and video acknowledgment, semantic parsing and reword discovery. It is deployed for machine learning analysis. Disadvantage includes complexity in designing and slow functionality.

Radial Basis Function Network comprises an info vector followed by a layer of RBF neurons and a yield layer with one hub for each classification (Zaji et al. 2020). Characterization is performed by estimating the info's similitude to information focuses from the preparation set where every neuron stores a model. This will be one of the models from the preparation set. At the point when another info vector [the n-dimensional vector that you are attempting to classify] should be arranged, every neuron computes the Euclidean separation between the information and its model. Each RBF neuron looks at the info vector to its model and yields a worth running which is a proportion of similitude from 0 to 1. As the info equivalents to the model, the yield of that RBF neuron will be 1 and with the separation develops between the information and model the responses tumbles off exponentially towards 0. The plot created out of neuron's responses tends towards a typical the bell shaped plot. The yield layer comprises of a lot of neurons [one per category]. Its applications are found in power restoration.

Recurrent Neural framework fed back to info or data to provide assistance for anticipating results for each layer. Primary stratified division ordinarily show feed forward architecture accompanied intermittent counterfeit framework strata that holds data (past time-step), so recollected by storage assemblies acting as memory-units (Smyl 2020). Onward tracks executed for such situations. It holds the knowledge relevant for its potential use. On the off chance that the expectation is not right, the learning rate is utilized to roll out little improvements. Consequently, stepwise increment towards making the correct forecast during the back track. Its focal points are Model consecutive information where each example can be thought to be subject to verifiable ones, it is utilized with convolution layers to expand the pixel viability. Significant detriments of such network architecture is Gradient disappearing and detonating issues, preparing repetitive neural frameworks act as troublesome undertaking, hard to info-processing for successive information utilizing rectified-linear-units as initiating set. LSTM (Long short term memory) systems are a kind of RNN that utilizes exceptional units notwithstanding standard units. LSTM units incorporate a "memory cell" that can keep up data in memory for significant stretches of time. A lot of doors is utilized to control when data enters the memory when its yield, and when's it slipped it's mind. There are three types of gates, viz., Input gate (Info door), output gate (yield entryway), and forget gate (overlook entryway). Info door chooses what number of data from the last data set will be kept in memory; the yield entryway manages the measure of information went to the following layer, and overlook entryways control the tearing pace of memory put away. Such architecture lets them learn longer-term dependencies.

A modular neural system has various systems that work autonomously and perform sub-undertakings. The various systems do not generally collaborate with or signal each other during the calculation procedure (Li et al. 2020). They work autonomously towards accomplishing the yield. Therefore, an enormous and complex computational procedure is done essentially quicker by separating it into free segments. The calculation speed increments in light of the fact that the systems are not collaborating with each other but at last associated with one another. It is robust and efficient neural network, but sometimes has moving target problems. Commonly used by stock exchange market for predictions, and biological studies for compression of high level input data, and character recognition.

18.4 Application of Neural Networks

Many *in silico* tools, servers, and algorithms (Table 18.1) are currently used in both proteomic and genomic analysis. Structural and functional aspect of reacting biomolecules within cellular domains can be easily accessed by neural network algorithms. Neural networks have multiple applications in bioinformatics:

1. Protein and peptide structure prediction, including primary, secondary, and tertiary structures. All related estimations like biochemical properties including Ramachandran plot assessment. Stability investigations. Comparative or homology as well as *ab-initio* both type of model can easily predicted by deploying artificial neural networks.
2. In modern era fast protein–ligand interaction studies are conducted by using neural networks. Neural networks assists in determining binding pockets for ligand molecules, primarily in computer aided drug discovery.
3. Molecular docking and Molecular simulation studies are also based on neural networks architecture to give precise trajectories for interacting molecules (DNA–Protein as well as Protein–Protein).
4. Genome annotations and alignment of DNA or protein sequences, also uses neural architectures.
5. RNA-Seq or Whole genome Sequence analysis studies are also using neural networks in differential gene expression analysis.
6. In cancer studies, for prediction of pathogenicity of DNA variants.

18.4.1 Prediction of Structure for Proteins

Now a days, dual-direction recurrent neural architectures, PSI-BLAST-derived profiles, and enormous non-redundant guiding sets deployed in tools like PSIPRED (McGuffin et al. 2000) produces two new predictors: (a) SSpro program for secondary structure classification into three categories sheets, helix, and loops and

Table 18.1 List of various modern in silico tools/techniques based on deep learning or neural networks

Tools based on NN	Source	Function
Net MHC server	Lundegaard et al. (2011)	Epitopes selection and prediction from bacterial and viral proteins used in vaccine designing
NeuRiPP	De los Santos (2019)	Identification of genetic clusters to reveal ribosomally synthesized and post-translationally modified proteins
DeepGoPlus	Kulmanov and Hoehndorf (2020)	Protein function prediction
DEEPscreen	Rifaioğlu et al. (2020)	Prediction of drug targets
RONN	Yang et al. (2005)	Identification of disordered regions of proteins
RESCUE	Pons and Delsuc (1999)	NMR spectral assignment to proteins
DeepQA	Cao et al. (2016)	Estimation of single protein model
DeepInteract	Patel et al. (2017)	Protein–protein interaction analysis
ProLanGO	Cao et al. (2017)	Protein functionality assessment
DeepDrug3D	Pu et al. (2019)	Drug or ligand binding pocket analysis and identification with in proteins or enzymes
EpiDock	Atanasova et al. (2013)	Molecular docking tool based on MHC class II interactions with epitopes
DeepLNC	Tripathi et al. (2016)	A long non coding RNA elements identification
DeepRibo	Clauwaert et al. (2019)	Gene annotation for prokaryotes based on ribosome profiling signals and binding site patterns
Afann	Tang et al. (2019)	Alignment free genetic sequence comparisons
SECLAF	Szalkai and Grolmusz (2018)	Biological sequence classification
SpliceFinder	Wang et al. (2019)	Prediction of splice sites using convolutional neural network architecture
DeepImpute	Arisdakessian et al. (2019)	Impute single cell RNA-seq data
DanQ	Quang and Xie (2016)	Quantification of DNA functions
RNAsamba	Camargo et al. (2020)	Assessment of translational potential of RNA sequence
REVEL	Ioannidis et al. (2016)	Prediction of pathogenicity of rare missense DNA variants. Assist in cancer biology

(b) SSpro8 program for secondary structure classification into the eight classes produced by the DSSP (dictionary of secondary structure of proteins) program, types include 3/10 helix, alpha helix, pi helix, extended strand in parallel and/or anti-parallel β -sheet conformation, isolated β -bridge, hydrogen bonded turn, bend, and coil. 8-state secondary structure is frequently amassed into 3-state auxiliary structure (Pollastri et al. 2002). Predicting protein structural disorders can be estimated by using feed forward neural networks (Li et al. 1999). Artificial neural

networks are also used protein functional determination likely emulsification, and foaming for assisting food industry (Arteaga and Nakai 1993).

18.4.2 Binding Patterns and Epitope Selection: Immuno-Informatics Application

Binding or interaction between receptor and ligand molecules can be easily predicting by deploying neural networks, for example, K_{DEEP} is a fast machine learning tool that uses convolutional neural network architecture for protein to ligand binding (Jiménez et al. 2018). Molecular docking studies with known protein and ligand structure in Pdb format can assist in predicting interaction between their constituents, with proper binding scores, atomic contact energies and RMSD (root mean square deviation) values. Binding pockets within receptor protein is made up of reactive amino acids and ligand amino acids interact with it to exhibit perfect fitting. Even in drug discovery convolutional neural network architecture is mostly deployed to produce perfect results, for example, DeepDrug3D (Pu et al. 2019). Neural network architecture is also used in quality appraisal of protein and ligand interactional domains prediction, for example, FunFOLD-QA (Roche et al. 2012).

One of the studies utilized artificial Neural Network method for developing potential vaccine candidates against mumps virus. This involved a novel concept known as reverse vaccinology in which prediction of peptide epitopes was done which would potentially elicit an immune response in human body by B cells and T cells. Hemagglutinin-neuraminidase (HN) surface glycoproteins are the main antigenic structures present in mumps virus which served as the source of the candidate peptide epitopes. 593 HN glycoprotein sequences were retrieved from NCBI database. Then neural network was used to study binding of these peptide candidates to MHC class I molecules to determine the minimum inhibitory concentration (IC50). Percentile ranks of as low as 0.1 were obtained showing high binding affinity between the candidate epitope and the human MHC class I allele, indicating potential use of the epitope as a peptide vaccine against mumps virus (Babiker et al. 2020).

Prediction of continuous and linear B-cell epitopes and T-cell epitopes for antigens is basis for immunoinformatic analysis to craft rapid vaccination against pathogens (Fig. 18.4); recurrent neural architecture was successfully deployed in such studies (Saha and Raghava 2006). NetMHC server (Lundegaard et al. 2011), NetCTLpan server (Stranzl et al. 2010), and BepiPred (Jespersen et al. 2017) are some of the common tools that are mostly used in predicting epitopes. These servers are based on artificial neural networks and assist user in determining epitopes interacting with MHCI and II HLA alleles. After confirmation with molecular docking as well as molecular dynamic simulation trajectory analysis users can rapidly determine immunogenic properties of humans against deadly viruses like corona viruses (Joshi et al. 2020) and even in the rarest pathogenic bacterium, such as *Tropheryma whipplei* (Joshi and Kaushik 2020).

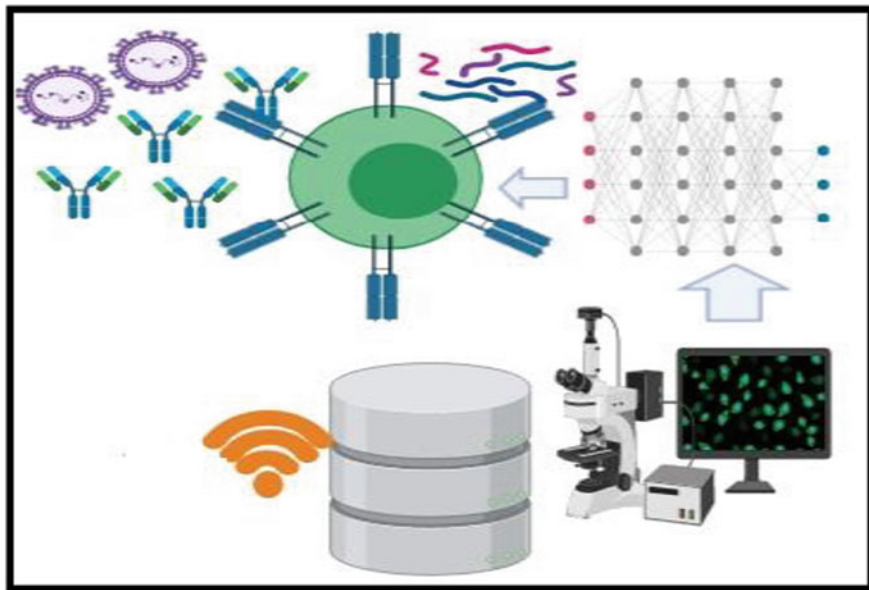


Fig. 18.4 Neural networks in epitope based vaccine crafting for viral pathogens

18.4.3 Role in Genomics and Transcriptomics

Eukaryotic and prokaryotic organisms have complex genomic expression, and to understand the mysteries behind it. Convolutional neural networks assisted users in predicting translational sites, regulatory mechanisms, and splicing domains in genetic elements (DNA or RNA) (Pedersen and Nielsen 1997). Sequential regulatory activity can be predicted across the chromosomes with convolutional neural architecture (Kelley et al. 2018). Deep neural network investigation also opens the gate of opportunities in for gene ontology and annotation (Chicco et al. 2014). Deep neural architecture plays crucial role in modeling RNA structures, and to conduct sequential alignment and comparisons (Wu and McLarty 2012). Modern sequencing studies need analysis of data generated for different organisms, to assist it deep neural networks play very crucial role. In WGS, and RNA-seq analysis, neural network tools like DanQ, RNAsamba tools were used along with Linux based freeware. Genome and transcriptome analysis was always data centric and to make better choices in selecting gene of interest to develop understanding about physiological or biochemical functionality was always primary feature that would lead scientific groups to triumph in the field of medicine.

18.5 Conclusion

Institutional computing facilities were improved lot in the past decade. Amalgamation of neural networks with advanced servers will assist rapid drug discovery, effective error free vaccine crafting, speedy alignments, structural predictions, and physiochemical analysis of biomolecules, etc. Modern world should not starve for better food security, medicinal treatments. To fulfill this broad socialistic view neural networks have intensified power to integrate, to access, and to analyze big data related to agriculture, animal husbandry, medicine, and physiology. Neural networks are constituent of deep learning domain of Artificial intelligence and machine learning; it holds significance in analyzing relationship about the integral features of IoT and bigdata (Mohammadi et al. 2018). Neural networks, as the name suggests it is the network or spider-web of interconnected artificial neurons joining input layer to output layer. Multiple types of neural networks assist users to develop insight about biomolecular structures and functions. Modern fast sequencing techniques generated enormous amount of data related to biological sequences, it was neural networks in bioinformatics who assisted researchers to bring fruitful outcomes in the field of agriculture as well as in medicine. It is ongoing research journey as neural networks are still evolving and linking to upgrading modern computing facilities to show its power of deep learning towards data analysis within the roots of big data and IoT.

References

- Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J (2013) Artificial neural networks in medical diagnosis. *J Appl Biomed* 11(2):47–58
- Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX (2019) DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol* 20(1):1–14
- Arteaga GE, Nakai S (1993) Predicting protein functionality with artificial neural networks: foaming and emulsifying properties. *J Food Sci* 58(5):1152–1156
- Atanasova M, Patronov A, Dimitrov I, Flower DR, Doytchinova I (2013) EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. *Protein Eng Des Sel* 26(10):631–634
- Babiker EAA, Almofti YA, Abd-Elrahman KA (2020) Novel T-lymphocytes vaccine candidates against human mumps virus via reverse vaccinology. *Eur J Biomed* 7(1):45–63
- Bain A (1873) *Mind and body: the theories of their relation*, vol 4. Henry S. King, London
- Camargo AP, Sourkov V, Pereira GAG, Carazzolle MF (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom Bioinform* 2(1):lqz024
- Cao R, Bhattacharya D, Hou J, Cheng J (2016) DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinform* 17(1):495
- Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z (2017) ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22(10):1732
- Chen Y, Tang L, Yang X, Bilal M, Li Q (2020) Object-based multi-modal convolution neural networks for building extraction using panchromatic and multispectral imagery. *Neurocomputing* 386:136–146

- Chicco D, Sadowski P, Baldi P (2014) Deep autoencoder neural networks for gene ontology annotation predictions. In Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics, pp. 533–540
- Clauwaert J, Menschaert G, Waegeman W (2019) DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic Acids Res* 47(6):e36–e36
- Collins FS, Morgan M, Patrinos A (2003) The human genome project: lessons from large-scale biology. *Science* 300(5617):286–290
- de los Santos EL (2019) NeuRiPP: neural network identification of RiPP precursor peptides. *Sci Rep* 9(1):1–9
- Evans RB (1990) William James, “the principles of psychology,” and experimental psychology. *Am J Psychol* 103(4):433–447
- Galushkin AI (2007) *Neural networks theory*. Springer, Berlin
- Giorgini E, Biavasco F, Galeazzi R, Gioacchini G, Giovanetti E, Mobbili G et al (2020) Synthesis, structural insights and activity of different classes of biomolecules. In: *The First Outstanding 50 Years of “UniversitàPolitecnicadelle Marche”*. Springer, Cham, pp 463–482
- Heidari AA, Faris H, Mirjalili S, Aljarah I, Mafarja M (2020) Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. In: *Nature-inspired optimizers*. Springer, Cham, pp 23–46
- Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci* 79(8):2554–2558
- Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S et al (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 99(4):877–885
- Jespersen MC, Peters B, Nielsen M, Marcatili P (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res* 45(W1):W24–W29
- Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G (2018) K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J Chem Inf Model* 58(2):287–296
- Joshi A, Joshi BC, Mannan MAU, Kaushik V (2020) Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. *Inform Med Unlocked* 19:100338
- Joshi A, Kaushik V (2020) In-Silico proteomic exploratory quest: crafting T-cell epitope vaccine against Whipple’s disease. *Int J Pept Res Ther* 27:169–179
- Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* 28(5):739–750
- Khan E (2020) *Neural fuzzy based intelligent systems and applications*. In: *Fusion of neural networks, fuzzy systems and genetic algorithms*. CRC Press, Boca Raton, FL, pp 105–140
- Kulmanov M, Hoehndorf R (2020) DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 36(2):422–429
- Li W, Li M, Qiao J, Guo X (2020) A feature clustering-based adaptive modular neural network for nonlinear system modeling. *ISA Trans* 100:185–197
- Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C- and internal regions. *Genome Inform* 10:30–40
- Lundegaard C, Lund O, Nielsen M (2011) Prediction of epitopes using neural network based methods. *J Immunol Methods* 374(1–2):26–34
- McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404–405
- Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M (2018) Deep learning for IoT big data and streaming analytics: a survey. *IEEE Commun Surv Tutor* 20(4):2923–2960
- Morgat A, Lombardot T, Coudert E, Axelsen K, Neto TB, Gehant S et al (2020) Enzyme annotation in UniProtKB using Rhea. *Bioinformatics* 36(6):1896–1901

- Patel S, Tripathi R, Kumari V, Varadwaj P (2017) DeepInteract: deep neural network based protein-protein interaction prediction tool. *Curr Bioinform* 12(6):551–557
- Pedersen AG, Nielsen H (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proc Inst Conf Intell Syst Mol Biol* 5:226–233
- Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins Struct Funct Bioinform* 47(2):228–235
- Pons JL, Delsuc MA (1999) RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. *J Biomol NMR* 15(1):15–26
- Pu L, Govindaraj RG, Lemoine JM, Wu HC, Brylinski M (2019) DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput Biol* 15(2):e1006718
- Quang D, Xie X (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 44(11):e107–e107
- Rifaoglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Doğan T (2020) DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem Sci* 11(9):2531–2557
- Roche DB, Buenavista MT, McGuffin LJ (2012) FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. *PLoS One* 7(5):e38219
- Saha S, Raghava GPS (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins Struct Funct Bioinform* 65(1):40–48
- Shao C (2020) A quantum model of feed-forward neural networks with unitary learning algorithms. *Quantum Inf Process* 19(3):102
- Smyl S (2020) A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *Int J Forecast* 36(1):75–85
- Stranzl T, Larsen MV, Lundegaard C, Nielsen M (2010) NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62(6):357–368
- Szalkai B, Grolmusz V (2018) SECLAF: a webserver and deep neural network design tool for hierarchical biological sequence classification. *Bioinformatics* 34(14):2487–2489
- Tang K, Ren J, Sun F (2019) Afann: bias adjustment for alignment-free sequence comparison based on sequencing data using neural network regression. *Genome Biol* 20(1):1–17
- Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK (2016) DeepLNC, a long non-coding RNA prediction tool using deep neural network. *Network Model Anal Health Inform Bioinform* 5(1):21
- Wang R, Wang Z, Wang J, Li S (2019) SpliceFinder: ab initio prediction of splice sites using convolutional neural network. *BMC Bioinform* 20(23):652
- Wu CH, McLarty JW (2012) Neural networks and genome informatics. Elsevier, Amsterdam
- Wu Q, Peng Z, Anishchenko I, Cong Q, Baker D, Yang J (2020) Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics* 36(1):41–48
- Yang ZR, Thomson R, Mcneil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21(16):3369–3376
- Zaji AH, Bonakdari H, Khameneh HZ, Khodashenas SR (2020) Application of optimized artificial and radial basis neural networks by using modified genetic algorithm on discharge coefficient prediction of modified labyrinth side weir with two and four cycles. *Measurement* 152:107291
- Zeng M, Zhang F, Wu FX, Li Y, Wang J, Li M (2020) Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 36(4):1114–1120

RESEARCH

Open Access



Codon usage studies and epitope-based peptide vaccine prediction against *Tropheryma whipplei*

Amit Joshi, Sunil Krishnan and Vikas Kaushik*

Abstract

Background: The *Tropheryma whipplei* causes acute gastroenteritis to neuronal damages in *Homo sapiens*. Genomics and codon adaptation studies would be helpful advancements of disease evolution prediction, prevention, and treatment of disease. The codon usage data and codon usage measurement tools were deployed to detect the rare, very rare codons, and also synonymous codons usage. The higher effective number of codon usage values indicates the low codon usage bias in *T. whipplei* and also in the 23S and 16S ribosomal RNA genes.

Results: In *T. whipplei*, it was found to hold low codon biasness in genomic sets. The synonymous codons possess the base content in 3rd position that was calculated as A3S% (24.47 and 22.88), C3S% (20.99 and 22.88), T3S% (21.47 and 19.53), and G3S% (33.08 and 34.71) for 23s and 16s rRNA, respectively.

Conclusion: Amino acids like valine, aspartate, leucine, and phenylalanine hold high codon usage frequency and also found to be present in epitopes KPSYLSALS AHLNDK and FKSFNYNVAIGVRQP that were screened from proteins excinuclease ABC subunit UvrC and 3-oxoacyl-ACP reductase FabG, respectively. This method opens novel ways to determine epitope-based peptide vaccines against different pathogenic organisms.

Keywords: *Tropheryma whipplei*, Synonymous codons, Ribosomal RNA, Gastroenteritis and codon usage

Background

Tropheryma whipplei is an actinobacteria pathogen causing Whipple's disease in *Homo sapiens*. This pathogenic problem was discovered and found to be associated with gastroenteritis, endocarditis, and neuronal damages in Caucasian individuals [1]. Regardless of this, its lethal impact was additionally seen in canines [2]. The credit for its name and disclosure was connected with honorable Nobel laureate G. H Whipple, who performed many explorations for lipodystrophy (malfunctioned lipid biosynthesis and ingestion) brought about by *T. whipplei* [3] has a broad-spectrum infection. Caucasian populaces, kids, sewage, and farming specialists were discovered to be generally influenced by this illness. The bacterium

causes immunomodulation with an extended IL-16 discharge, IL-10 synthesis, and dysregulation of mucosal T-helper cells. Further immunological irregularities were depicted because of Whipple's disease's multifaceted nature [4]. Clinical side effects of this infection were seen as extreme looseness of the bowels, loss of body weight, and weakness among patients [5]. *T. whipplei* assaults lamina propria of the gastrointestinal tract and targets macrophages for its replication [6]. Sequencing of two strains of *T. whipplei* (Twist and TW 08/27) was effectively led by the French researchers that already open scope for genomic examination and improvement of better treatment procedures for this lethal sickness; in their investigation, it was discovered that this actinobacterium has low GC content (46%) in correlations with other relatives of a similar order [7].

Current medicines like doxycycline, hydroxychloroquine, and trimethoprim/sulfamethoxazole must be used

*Correspondence: vikas.14664@pu.co.in
School of Bioengineering and Biosciences, Lovely Professional University,
Phagwara, Punjab, India

for almost 2 years and lifetime follow-up for patients [8, 9]. Later in silico concentrates on epitope-based vaccine design can become conceivable prophylaxis for Whipple's illness [10]. This actinobacterium has a huge encoding of surface proteins, while some are additionally connected with the enormous substance of noncoding redundant DNA. This genome additionally shows the fluctuation in genomic sets, including phase variations causing the modifications of cell proteins; this shows the importance of immune bypass and association with the host genome [1, 7]. Such uncommon genomic trademark highlights of bacterium open wide scope in discovering codon utilization patterns to uncover characteristic and mutational determination. Codons contained 3 nucleotides in sequence and coded for a particular amino acid or as a STOP codon for translation. The differences in codon usage are differences defined in codon usage bias. Equivalent codon utilization in numerous prokaryotic unicellular life forms is consistently connected with the directional mutational inclination and translational choice [11]. Other elements like replication-translation determination, protein hydrophathy, can likewise have a critical impact [12]. In some microbial pathogen species, mutational predisposition was discovered to be strand explicit, and those living beings show differed interchangeable and nonequivalent codon utilization [13]. This examination not just give experiences about characteristic and mutational determination pressures acting at genomic levels of *T. whipplei* yet besides offer a superior cognizance of transformative improvements in this host-versatile bacterium. This computational examination uncovered the data concerning profoundly translated proteins and enzymes of this bacterium, and the conceivable amino acids that can be considered in epitope-based prophylaxis plan to get the inhibitory effect on bacterial action on its host or to create a better conceivable treatment like in immunoinformatics-based recent studies [14, 15]. Ribosomal RNA (16S and 23S) codon usage patterns were analyzed here to determine the changes associated with evolutionary or phylogenetic patterns of the bacterium. In this study, we also revealed epitope-based peptide vaccine candidates against *Tropheryma whipplei*. The aim of the study is to determine codon usage patterns in *T. whipplei*, and on the basis of that we predicted epitope-based vaccine candidate by deploying latest bioinformatics tools.

Methods

Codon data retrieval

To measure the codon usage bias, retrieved codon usage tables from codon and codon pair usage tables (CoCoPUTs) database. This database showed the relative frequency that different codons are used in genes in *T.*

whipplei RefSeq data. Similarly, codon-pair usage tables displayed the counts of each codon pair in the CDSs of *T. whipplei* genomic data (RefSeq) and calculated codon-pair usage bias.

Retrieval of genomic data and codon usage table

The complete nucleotide sequences of *T. whipplei* strains. The selected FASTA sequences of Twist 16S ribosomal RNA and 23S ribosomal RNA were retrieved from the NCBI Refseq database (<https://www.ncbi.nlm.nih.gov/nucleotide>). The codon usage dataset was retrieved from the Codon Usage Database (<http://www.kazusa.or.jp/codon/>).

Genomic sequence optimization

All codons in the original sequence of *T. whipplei* strains are replaced with the corresponding redundant codon having the highest codon usage frequency. ATGme tool [16] was used to identify rare codons and accordingly optimize genomic sequences (<http://www.atgme.org/>). Genomic sequences in FASTA format pasted in the search box, and codon usage table pasted in the respective interface and processed the data for analysis of rare codons and sequence optimization.

Codon usage measurements

From the identified genomic sequences of ribosomal RNA, nucleotide composition was computed. The G + C composition of 1st, 2nd, and 3rd positions and GC1s, GC2s, and GC3s in the codons were discovered for the frequency and mean frequency identification. The

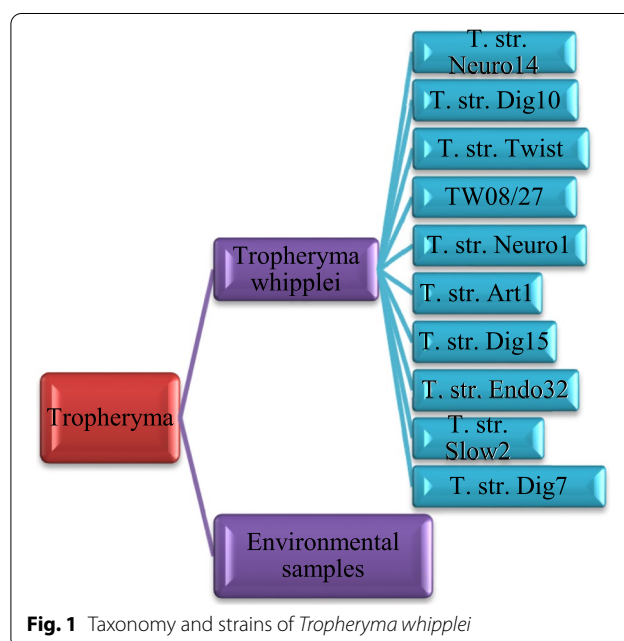
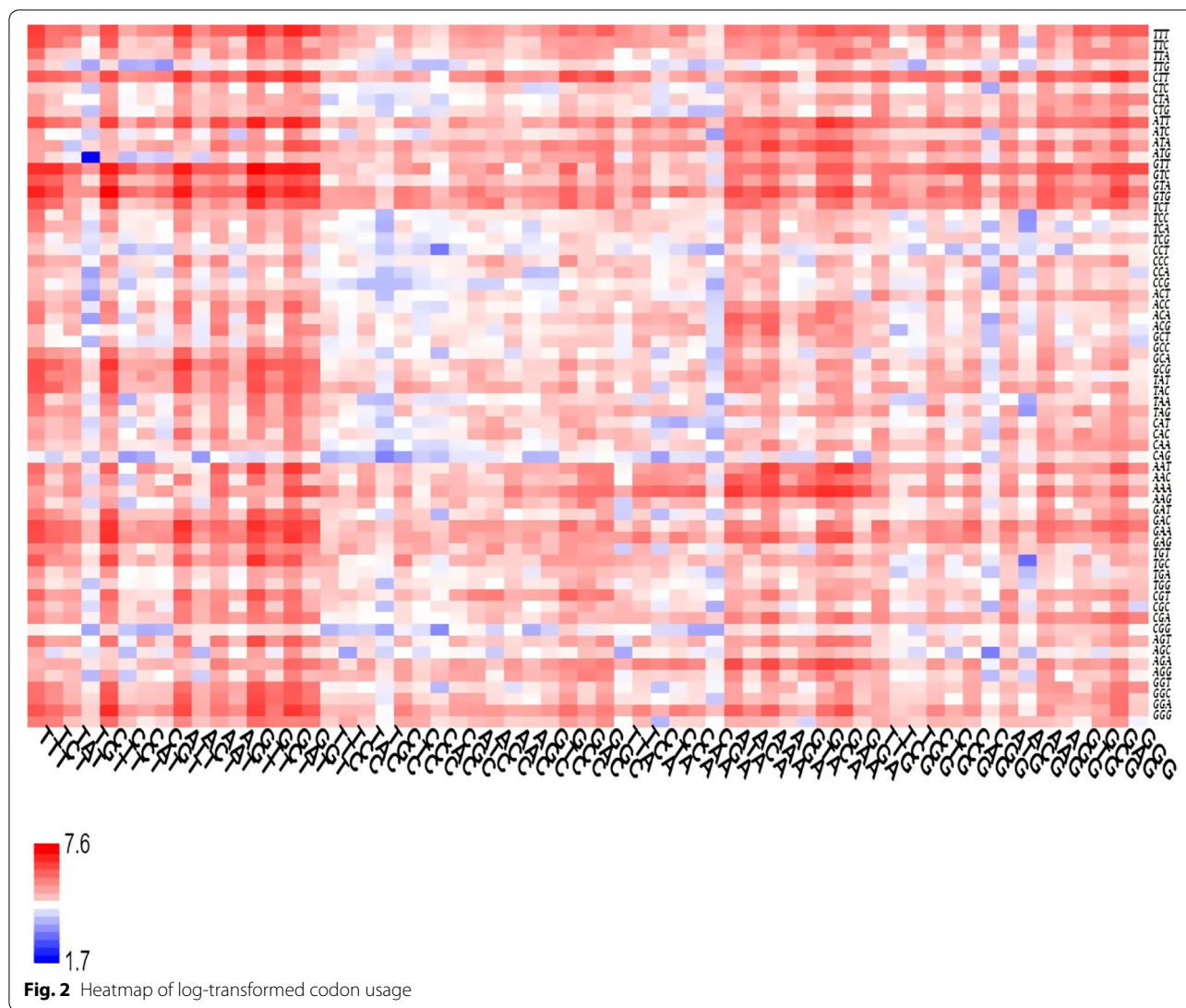


Fig. 1 Taxonomy and strains of *Tropheryma whipplei*



frequency of synonymous third position codon and percentage, i.e., A3, T3, G3, and C3 and %A3s, %C3s, %T3s, and %G3s, respectively, was calculated. To measure the bias of synonymous codons, the effective number of codons (ENC) was identified. Additionally, codon usage, codon usage per thousand, and **relative synonymous codon usage (RSCU)** were also calculated using “CAIcal” tool availed from <https://ppuigbo.me/programs/CAIcal/>.

Epitope-based vaccine prediction

Proteomic data for *Tropheryma whipplei* was accessed from NCBI GenBank database, and then allergenicity was estimated by deploying AllergenFP server [17]. Net-MHCIIpan-4.0 server [18] was used to screen epitopes from selected proteins that can interact with human leucocyte antigen (HLA) proteins. Vaxijen 2.0 tool [19] was

used to reveal antigenicity of screened epitopes. Epitopes structure was predicted by using PEP-FOLD 3.5 [20], and HLA allelic determinant HLA DRB1_0101 (PDB-ID:1AQD) was retrieved from RCSB-PDB database. Biochemical properties for epitopes were calculated by using ProtParam tool of ExPASy web server.

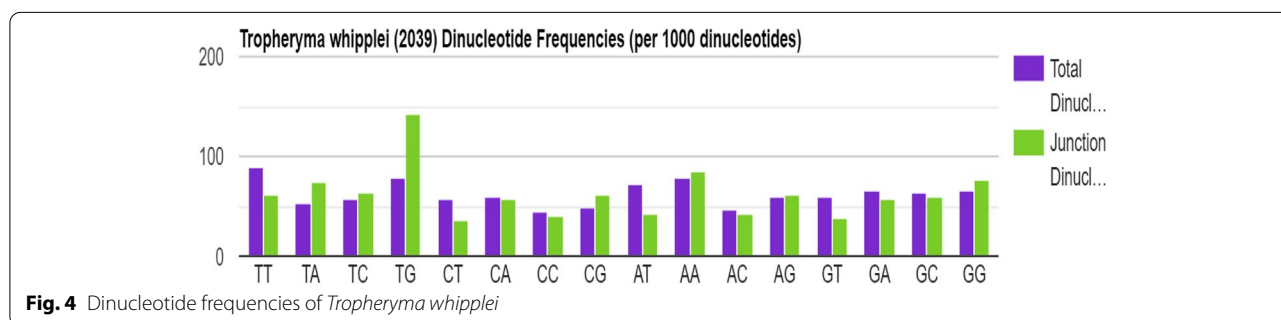
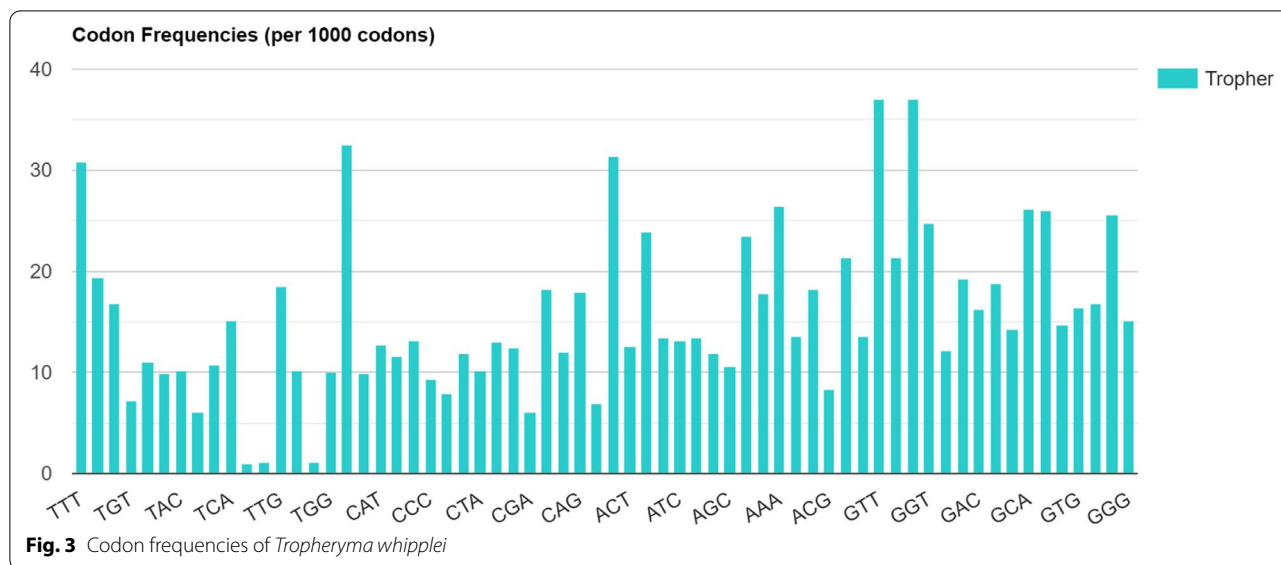
Molecular docking between epitopes and HLA determinants was done by using PatchDock [21], FireDock, and DINC web tool [22]. These tools not only assist in docking in user-friendly approach but also calculate

Table 1 Effective number of codon pairs for each *T. whipplei*

ENc	ENcp	ENc (GC corrected)	ENcp (GC corrected)	Genetic code
56.138	54.026	57.212	54.910	Standard code

Table 2 *Tropheryma whipplei* RefSeq codon table contains 88597 CDSs (28006357 codons)

Codon	Usage frequency	No. of codons	Codon	Usage frequency	No. of codons	Codon	Usage frequency	No. of codons	Codon	Usage frequency	No. of codons
TTT	30.88	(864933)	TCT	19.41	(543618)	TAT	16.82	(470999)	TGT	7.23	(202580)
TTC	11.07	(309980)	TCC	9.94	(278337)	TAC	10.24	(286912)	TGC	6.09	(170524)
TTA	10.79	(302319)	TCA	15.06	(421837)	TAA	1.00	(27960)	TGA	1.11	(31013)
TTG	18.49	(517778)	TCC	10.24	(286768)	TAG	1.14	(31981)	TGG	10.09	(282675)
CTT	32.53	(910922)	CCT	9.92	(277793)	CAT	12.79	(358300)	CGT	11.66	(326630)
CTC	13.10	(366773)	CCC	9.27	(259711)	CAC	7.92	(221772)	CGC	11.85	(331835)
CTA	10.11	(283081)	CCA	13.02	(364568)	CAA	12.40	(347268)	CGA	6.07	(169892)
CTG	18.23	(510468)	CCG	12.02	(336672)	CAG	17.96	(503036)	CGG	6.98	(195455)
ATT	31.40	(879451)	ACT	12.53	(350913)	AAT	23.82	(667060)	AGT	13.41	(375656)
ATC	13.10	(366940)	ACC	13.40	(375396)	AAC	11.87	(332502)	AGC	10.61	(297274)
ATA	23.41	(655721)	ACA	17.74	(496749)	AAA	26.46	(741084)	AGA	13.59	(380649)
ATG	18.25	(511118)	ACG	8.30	(232504)	AAG	21.39	(598923)	AGG	13.55	(379598)
GTT	37.06	(1037894)	GCT	21.35	(597805)	GAT	37.03	(1037065)	GGT	24.77	(693723)
GTC	12.15	(340402)	GCC	19.26	(539448)	GAC	16.30	(456523)	GGC	18.74	(524935)
GTA	14.34	(401675)	GCA	26.13	(731733)	GAA	25.97	(727374)	GGA	14.66	(410657)
GTG	16.39	(459063)	GCG	16.86	(472115)	GAG	25.54	(715420)	GGG	15.16	(424597)



different parameters like global energy, atomic contact energy, and binding energy for docked complexes.

Results

Identified codons and calculated usage bias

The codon-pair usage table and dinucleotide usage data were identified from the CoCoPUTs database [23, 24]. The *T. whipplei* taxonomy ID or taxid (2039) was verified by NCBI’s taxonomy tool, and the taxonomy was illustrated in Fig. 1. The log-transformed codon-pair frequency heat map was discovered from the data analysis as illustrated in Fig. 2. The degree of ENC values ranges from 20 to 61 [25]. If the value is 20, then one codon coding for each amino acid and value ranged to 61 means all the synonymous codon was used for each amino acid. The ENC value computed in our analysis was 56.138, which means more than one codon was used for each amino acid. The ENC value should be ≤ 35 for significant codon bias [26]. So, the higher ENC value indicates

the low codon usage bias in *T. whipplei*. The ENC value details are demonstrated in Table 1.

The codon usage details are summarized in the Table 2, and the codon usage frequency per 1000 codons is illustrated in Fig. 3. The RefSeq ($n = 859$) of *T. whipplei* had 88597 CDSs and 28006357 codons. Table 2 illustrated the CDS and its codon pair. The codons GTT (37.06), GAT (37.03), CTT (32.53), and TTT (30.88) were identified as the highest usage frequency (frequency value shown in bracket). Dinucleotide frequencies per 1000 dinucleotide are demonstrated in Fig. 4.

***Tropheryma whipplei* str. Twist codon usage table**

Tropheryma whipplei strain Twist complete sequence of 23S and 16S ribosomal RNA genes were composed of 3102 base pairs and 1521 base pairs, respectively. *Tropheryma whipplei* Twist strain’s CDS, codons, frequency per thousand, and the number of codons details are summarized in Tables 3 and 4. These codon usage tables were used for the identification of rare codons and sequence optimization.

Table 3 *Tropheryma whipplei* str. Twist 808 CDS' (266294 codons) codons, frequency per thousand, and in bracket number of codons

Codon	Frequency (no. of codon)	Codon	Frequency (no. of codon)	Codon	Frequency (no. of codon)	Codon	Frequency (no. of codon)
UUU	30.5 (8121)	UCU	19.7 (5246)	UAU	17.2 (4590)	UGU	7.3 (1938)
UUC	11.5 (3066)	UCC	10.1 (2690)	UAC	10.5 (2790)	UGC	6.1 (1626)
UUA	10.9 (2906)	UCA	15.4 (4100)	UAA	0.9 (250)	UGA	1.1 (281)
UUG	18.4 (4894)	UCG	9.9 (2643)	UAG	1.0 (277)	UGG	10.2 (2710)
CUU	31.8 (8461)	CCU	10.6 (2826)	CAU	12.8 (3409)	CGU	11.6 (3079)
CUC	13.1 (3492)	CCC	9.8 (2620)	CAC	7.9 (2111)	CGC	11.6 (3085)
CUA	10.6 (2832)	CCA	13.5 (3588)	CAA	12.5 (3316)	CGA	6.0 (1585)
CUG	18.3 (4871)	CCG	11.6 (3095)	CAG	18.4 (4889)	CGG	6.9 (1832)
AUU	30.6 (8157)	ACU	12.7 (3392)	AAU	23.7 (6313)	AGU	13.1 (3497)
AUC	13.2 (3503)	ACC	14.2 (3776)	AAC	11.9 (3179)	AGC	10.7 (2855)
AUA	23.3 (6209)	ACA	19.6 (5223)	AAA	26.2 (6970)	AGA	13.6 (3613)
AUG	18.0 (4784)	ACG	8.2 (2176)	AAG	21.2 (5642)	AGG	13.2 (3516)
GUU	36.7 (9774)	GCU	21.3 (5660)	GAU	36.3 (9679)	GGU	24.9 (6640)
GUC	12.2 (3247)	GCC	19.4 (5172)	GAC	16.1 (4283)	GGC	18.8 (5007)
GUA	14.7 (3916)	GCA	26.1 (6939)	GAA	25.2 (6702)	GGA	14.8 (3952)
GUG	16.6 (4431)	GCG	16.3 (4340)	GAG	24.7 (6586)	GGG	14.8 (3942)
GC percent information				Coding GC 46.46%	1st letter GC 54.59%	2nd letter GC 42.30%	3rd letter GC 42.48%

Table 4 *Tropheryma whipplei* TW08/27783 CDSs and 261028 codons, frequency per thousand, and in bracket number of codons

Codon	Frequency (no. of codon)	Codon	Frequency (no. of codon)	Codon	Frequency (no. of codon)	Codon	Frequency (no. of codon)
UUU	30.4 (7947)	UCU	19.8 (5158)	UAU	17.4 (4531)	UGU	6.9 (1813)
UUC	11.4 (2984)	UCC	10.3 (2683)	UAC	10.5 (2743)	UGC	5.7 (1496)
UUA	10.7 (2802)	UCA	15.6 (4063)	UAA	1.0 (251)	UGA	1.0 (265)
UUG	17.7 (4611)	UCG	9.8 (2567)	UAG	1.0 (267)	UGG	10.0 (2603)
CUU	31.9 (8314)	CCU	10.6 (2762)	CAU	12.8 (3343)	CGU	11.5 (2996)
CUC	13.4 (3509)	CCC	9.8 (2560)	CAC	7.8 (2034)	CGC	11.5 (3008)
CUA	10.8 (2829)	CCA	13.8 (3610)	CAA	12.6 (3276)	CGA	5.8 (1513)
CUG	18.2 (4741)	CCG	11.5 (3014)	CAG	18.4 (4793)	CGG	6.7 (1747)
AUU	30.7 (8013)	ACU	12.8 (3352)	AAU	23.7 (6193)	AGU	13.1 (3413)
AUC	12.9 (3377)	ACC	14.6 (3803)	AAC	12.1 (3149)	AGC	10.7 (2781)
AUA	23.6 (6166)	ACA	20.1 (5243)	AAA	26.2 (6829)	AGA	13.6 (3546)
AUG	17.9 (4662)	ACG	8.1 (2108)	AAG	21.2 (5533)	AGG	13.1 (3409)
GUU	36.9 (9638)	GCU	21.3 (5567)	GAU	36.7 (9577)	GGU	25.0 (6521)
GUC	12.2 (3193)	GCC	19.6 (5111)	GAC	16.2 (4218)	GGC	18.7 (4884)
GUA	14.7 (3835)	GCA	26.1 (6821)	GAA	25.2 (6578)	GGA	14.9 (3879)
GUG	16.3 (4256)	GCG	16.2 (4239)	GAG	24.9 (6488)	GGG	14.6 (3813)
GC percent information				Coding GC 46.41%	1st letter GC 54.66%	2nd letter GC 42.27%	3rd letter GC 42.29%

Rare and very rare codons

The analysis resulted from usage data, original sequence, and optimized sequence. *Tropheryma whipplei* strain Twist 23S ribosomal RNA gene sequence analyzed usage data predicted GTT and GAT (36.7% and 36.3 %) had the high frequency in codon usage. TAA, TAG, and TGA code as “STOP” had the lowest usage frequency percentage ((0.9 %, 1.0 % and 1.1 %) and found these are the very rare codons. The rare codons are CGA, TGC, CGG, TGT, CAC, ACG, CCC, and TCG. The stop codons are terminating the protein translation process [27]. The details of rare codons and very rare codons (code as, count, and percentage of usage frequency) of 23s and 16S rRNA were summarized in Tables 5 and 6.

Codon measurement

The calculated compositional properties for the coding sequences of the *Tropheryma whipplei* Twist strain are overall frequency of nucleotides A% (25.11 and 23.54),

Table 5 *Tropheryma whipplei* strain Twist 23S ribosomal RNA gene

Codon	Codes as	Usage frequency ‰	Count
TAA	STOP	0.9	14
TAG	STOP	1	26
TGA	STOP	1.1	14
CGA	Arg	6	31
TGC	Cys	6.1	12
CGG	Arg	6.9	22
TGT	Cys	7.3	19
CAC	His	7.9	8
ACG	Thr	8.2	15
CCC	Pro	9.8	21
TCG	Ser	9.9	16

Table 6 *Tropheryma whipplei* str. Twist 16S ribosomal RNA

Codon	Codes as	Usage frequency ‰	Count
TAA	STOP	0.9	8
TAG	STOP	1	3
TGA	STOP	1.1	5
CGA	Arg	6	5
TGC	Cys	6.1	8
CGG	Arg	6.9	15
TGT	Cys	7.3	3
CAC	His	7.9	6
ACG	Thr	8.2	7
CCC	Pro	9.8	6
TCG	Ser	9.9	10

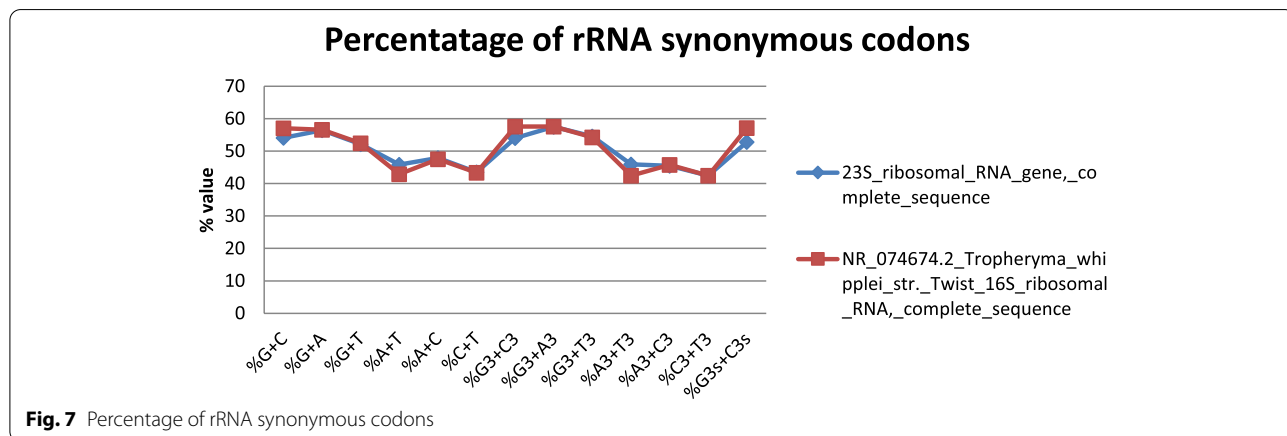
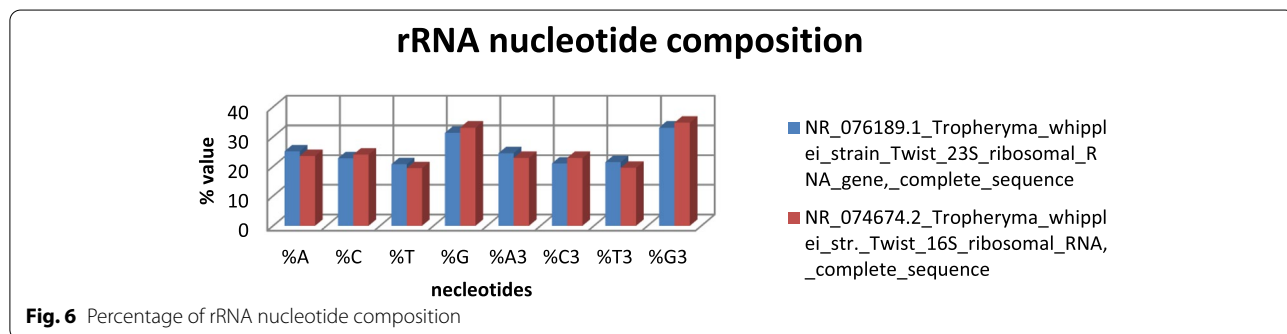
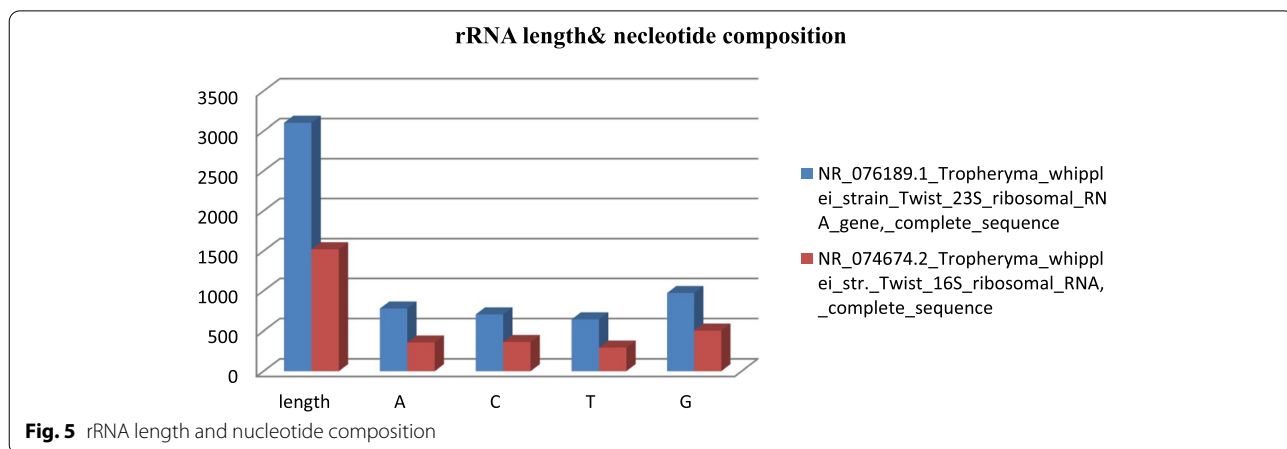
C% (22.76 and 24.0), T% (20.76 and 19.4), and G% (31.37 and 33.07) in 23s and 16s ribosomal RNA gene, respectively. The synonymous codons had the base content in 3rd position were calculated as A3S% (24.47 and 22.88), C3S% (20.99 and 22.88), T3S% (21.47 and 19.53), and G3S% (33.08 and 34.71) for 23s and 16s rRNA, respectively. GC3S% (52.85 and 57.85) is the third synonymous codon position in GC content of 23s and 16s rRNA, respectively. Figures 5 and 6 show rRNA characteristic features like length and nucleotide composition. In Fig. 7, rRNA synonymous codons percentage is given, while in Fig. 8, codon measurements were indicated.

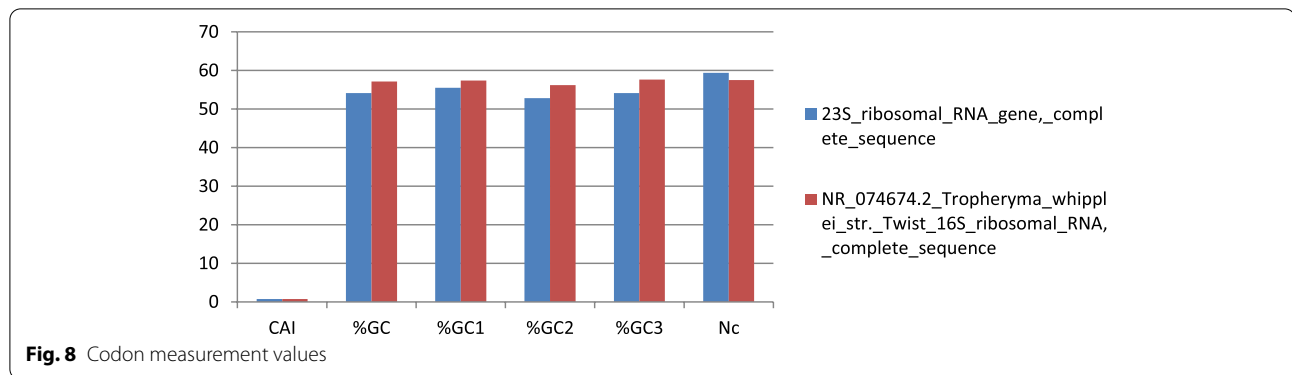
Epitope-based vaccine prediction: application of codon usage studies

The in silico analysis reveals two epitopes of 15 amino acid residues (i.e., KPSYLSALSAHLNDK and FKS-FNYNVAIGVRQP) that hold perfect interaction with HLA-DRB-0101 (MHC class II allelic determinant). In Table 7, retrieved sequences were shown with accession numbers, and allergenicity was also presented by deploying Allergen FP tool (this tool generates Tanimoto similarity index). Epitopes were determined by using NetMHCIIpan-4.0 server that gathers core information from IEDB database and uses artificial neural networks (ANN) to access interaction of peptidal stretches to HLA allelic determinants. Amino acids like valine, aspartate, leucine, and phenylalanine hold high codon usage frequency and also found to be present in these screened epitopes from excinuclease ABC subunit UvrC and 3-oxoacyl-ACP reductase FabG. In Table 8, all 10 peptides are holding good Vaxijen score, and NetMHCIIpan-4.0 scores are provided, but there were a total of 2151 epitopes discovered. Vaxijen score indicates antigenicity for peptides. ProtParam results reveal only two finalized epitopes to be stable (Table 9). Epitopes structure was predicted by using PEP-FOLD 3.5 [20], and HLA allelic determinant HLA DRB1_0101 (PDB-ID:1AQD) was retrieved from RCSB-PDB database to perform molecular docking analysis. Molecular docking of selected epitopes with HLA-DRB0101 shows perfect interaction (Table 10). Figure 9 indicates docked complexes of selected epitopes with HLA-DRB-0101 visualized in PyMOL software.

Discussion

The *Tropheryma whipplei* causes acute gastroenteritis to neuronal damages in *Homo sapiens*. Genomics and codon adaptation studies would be helpful advancements of disease evolution prediction, prevention, and treatment of disease. The codon-pair usage table and dinucleotide usage data were identified from the CoCoPUTs database [23, 24]. The ENC value computed in our





analysis was 56.138, which means more than one codon was used for each amino acid. The ENC value should be ≤ 35 for significant codon bias [26]. *Tropheryma whipplei* Twist strain’s CDS, codons, frequency per thousand, and the number of codons; for identification of rare codons and sequence optimization. The ratio of observed codon frequency to the expected synonymous codons usage for the amino acid i.e., relative synonymous codon usage (RSCU) [28]. The degree of bias towards estimated, i.e., Codon Adaptation Index, value was 0.73 and 0.725 for 23s and 16s rRNA respectively. The value ranged between 0 and 1; higher values indicate stronger

bias in codon usage and high gene expression level. In previous studies, membrane proteins were considered to be associated with considerable biasness [29], while in current study, we recognized rare codon biasness associated with entire genome of *T. whipplei*. The major requirement of codon biasness study assists in determining amino acids expressed patterns that can be linked to epitope-based vaccine predictions. In recent studies, for SARS-CoV2 [30, 31], dengue [32, 33], Nipah [34], Candida fungus [35], Canine circovirus [36], and Zika virus [37], vaccine predictions were found to be successful. So, codon usage pattern determination can be considered as

Table 7 AllergenFP score and proteins considered for *Tropheryma whipplei*

Proteins/no. of amino acid residues	GenBank-accession no.	Function	Allergen FP score	Inference
Prolipoprotein diacylglyceryl transferase (<i>Tropheryma whipplei</i>) 272 aa protein	WP_042506957.1	Catalyzes the transition of the diacylglyceryl group from phosphatidylglycerol to the sulfhydryl group of the N-terminal cysteine of a prolipoprotein, the first step in the development of mature lipoproteins	0.87	Non-allergen
Excinuclease ABC subunit UvrC (<i>Tropheryma whipplei</i>) 607 aa protein	WP_042506954.1	DNA excision repair	0.82	Non-allergen
Holliday junction resolvase RuvX (<i>Tropheryma whipplei</i>) 145 aa protein	WP_042506082.1	Nuclease activity, rRNA processing	0.82	Non-allergen
Exodeoxyribonuclease VII large subunit (<i>Tropheryma whipplei</i>) 404 aa protein	WP_042506175.1	Degrades single-stranded DNA bidirectionally, first into massive acid-insoluble oligonucleotides, then into small acid-soluble oligonucleotides	0.82	Non-allergen
Isoprenyl transferase (<i>Tropheryma whipplei</i>) 249 aa protein	WP_042506056.1	Isopentenyl diphosphate (IPP) condensation with allylic pyrophosphates is catalyzed, resulting in a number of terpenoids.	0.80	Non-allergen
3-oxoacyl-ACP reductase FabG (<i>Tropheryma whipplei</i>) 238 aa protein	WP_011096407.1	Catalyzes the NADPH-dependent reduction of beta-ketoacyl-ACP substrates to beta-hydroxyacyl-ACP products, the first reductive step in the elongation cycle of fatty acid biosynthesis	0.82	Non-allergen
ABC transporter permease subunit (<i>Tropheryma whipplei</i>) 332 aa protein	WP_206536426.1	Transmembrane transportation of molecules	0.90	Non-allergen

Table 8 Peptides showing interaction to HLA-DRB0101, NETMHCII PAN 4.0 server results, and VaxiJen score

Pos	Peptide	ID	Score	Rank	VaxiJen score	Inference
39	NRRFIVLTGNREFTA	WP_042506957.1	0.958934	0.16	-0.4516	Nonantigenic
316	KPSYLSALS AHLNDK	WP_042506954.1	0.978324	0.06	0.7208	Antigenic
384	LQKYLNLNSLPVRIE	WP_042506954.1	0.968518	0.11	1.1646	Antigenic
580	IEDISALPGFGVKTA	WP_042506954.1	0.960251	0.15	0.7039	Antigenic
227	RDKIQA AQT VLSRSA	WP_042506954.1	0.805061	0.85	0.1459	Antigenic
77	EFSRFLVSSGVQVRF	WP_042506082.1	0.651559	1.60	0.4449	Antigenic
235	KTPLISAIGHEADRP	WP_042506175.1	0.966542	0.12	-0.0952	Nonantigenic
231	DDFWAALRAYSGRSR	WP_042506056.1	0.960550	0.15	0.2368	Antigenic
24	FKSFNYNVAIGVRQP	WP_011096407.1	0.916978	0.35	0.7126	Antigenic
3	PARFFFV SPLSCVKP	WP_206536426.1	0.691033	1.40	0.6685	Antigenic

Table 9 ProtParam results: biochemical properties of epitopes

Peptides	Molecular mass	pI	Gravy score	Aliphatic index	Instability index	Half life mammalian reticulocytes
KPSYLSALS AHLNDK	1643.86	8.51	-0.553	91.33	5.83	1.3 h
LQKYLNLNSLPVRIE	1800.13	8.59	-0.147	149.33	86.04	5.5 h
IEDISALPGFGVKTA	1517.74	4.37	0.573	110.67	62.39	20 h
FKSFNYNVAIGVRQP	1739.99	9.99	-0.180	71.33	24.99	1.1 h
PARFFFV SPLSCVKP	1695.06	9.57	0.673	71.33	61.23	> 20 h

the preliminary step before deploying any ANN (artificial neural networking)-based web server/tool like NetMHC server for screening essential epitopes of small peptidal length (8–12 amino acids). The calculated compositional properties for the coding sequences of the *Tropheryma whipplei* Twist strain overall frequency of nucleotides A% (25.11 23.54), C% (22.76 24.0), T % (20.76 19.4), and G% (31.37 and 33.07) in 23s and 16 s ribosomal RNA gene respectively. In silico analysis reveals two epitopes of 15 amino acid residues (i.e., KPSYLSALS AHLNDK and FKSFNYNVAIGVRQP) that hold perfect interaction with HLA-DRB-0101 (MHC class II allelic determinant); future scope holds linkers and adjuvants to be connected and solid-phase synthesis of these epitopes to

further test these epitopes in model organisms. Recent developments in immunoinformatics show novel ways to predict epitope-based vaccine candidates and therapeutics against many harmful pathogens like *Candida auris* [35] and human cytomegalovirus [38]. Similarly, drug repurposing was made easy against harmful pathogens by deploying bioinformatic approaches [39]. Similarly, for animal models, viral pathogenic proteomes were screened for vaccine designing by deploying immunoinformatics [33, 36, 40]. This study is unique in terms of saving time and money for peptide-based vaccine crafting.

Conclusions

Considerable biases in codon usage and amino acid usage indicate clearly that *T. whipplei* has a low codon bias. The synonymous codons had the base content in 3rd position were calculated as A3S% (24.47 and 22.88), C3S% (20.99 and 22.88), T3S% (21.47 and 19.53), and G3S% (33.08 and 34.71) for 23s and 16s rRNA, respectively. Also, codon-usage patterns clearly indicate that there will be less chances of variational or evolutionary alterations in

Table 10 ACE VALUE, global energy, and binding energy for selected docked complexes (epitopes to HLA DRB0101)

Epitope	ACE value (Kcal/Mol)	Global energy (Kcal/Mol)	Binding energy (Kcal/Mol)
KPSYLSALS AHLNDK	-6.59	-36.93	-2.80
FKSFNYNVAIGVRQP	-3.79	-1.19	-3.40

Molecular Docking Results

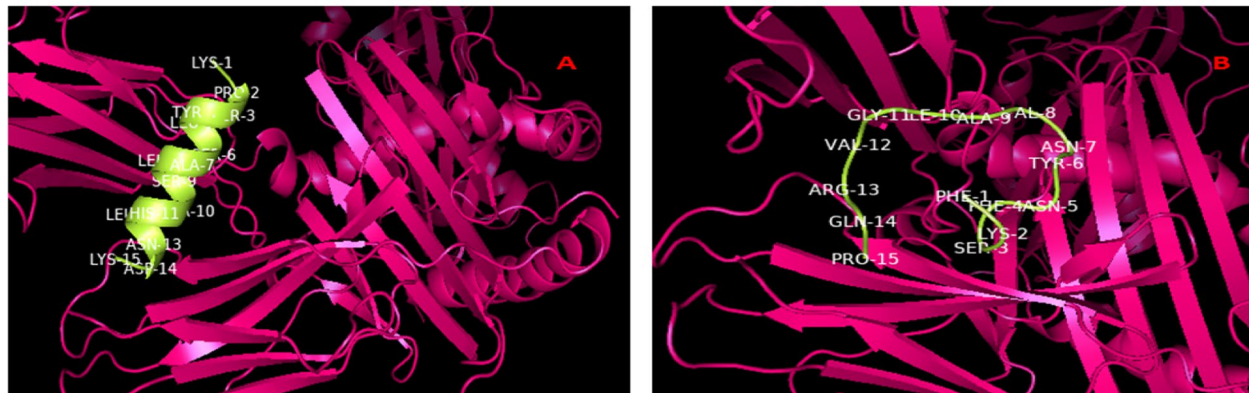


Fig. 9 Molecular docking results of epitopes with HLA-DRB-0101. **A** KPSYLSALSALHNDK from protein excinuclease ABC subunit UvrC and **B** FKSFNYNVAIGVRQP from protein 3-oxoacyl-ACP reductase FabG

T. whipplei genomic sets. The analysis could be targeted for disease evolution prediction, developing drugs, or vaccine candidates. We also found KPSYLSALSALHNDK and FKSFNYNVAIGVRQP, two epitopes, can possibly act as vaccine candidates against *T. whipplei*. A future development requires wet-lab validations for these epitopes that are highly expressed in this bacterium and have therapeutic peptide formation capability.

Abbreviations

IEDB: Immune epitope database; CAI: Codon Adaptation Index; RNA: Ribonucleic acid; NCBI: National Center for Biotechnology Information; HLA: Human leukocyte antigen; RSCU: Relative synonymous codon usage; MHC: Major histocompatibility complex.

Acknowledgements

All the authors are thankful towards the school of bioengineering and biosciences, Lovely Professional University, Phagwara, Punjab, India.

Authors' contributions

AJ and VK, peptide identification using codon bias studies. VK, conception of idea of this article and gap identification in existing studies and editing of the paper. AJ and SKG, molecular dynamic simulation study and analysis. The authors read and approved the final manuscript.

Availability of data and materials

All data is provided in manuscript.

Declarations

Ethics approval and consent to participate

Not applicable. There is no impact on ethical standards in this study, and there is no human or animal involvement.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 3 December 2021 Accepted: 22 February 2022

Published online: 07 March 2022

References

1. Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM (2003) *Tropheryma whipplei* Twist: a human pathogenic actinobacteria with a reduced genome. *Genome Res* 13(8):1800–1809 <http://www.genome.org/cgi/doi/10.1101/gr.1474603>
2. Keita AK, Raoult D, Fenollar F (2013) *Tropheryma whipplei* as a commensal bacterium. *Future Microbiol* 8(1):57–71. <https://doi.org/10.2217/fmb.12.124>
3. Dolmans RA, Boel CE, Lacle MM, Kusters JG (2017) Clinical manifestations, treatment, and diagnosis of *Tropheryma whipplei* infections. *Clin Microbiol Rev* 30(2):529–555. <https://doi.org/10.1128/CMR.00033-16>
4. Moos V, Schmidt C, Geelhaar A, Kunkel D, Allers K, Schinnerling K, Ignatius R (2010) Impaired immune functions of monocytes and macrophages in Whipple's disease. *Gastroenterology* 138(1):210–220. <https://doi.org/10.1053/j.gastro.2009.07.066>
5. Lagier JC, Lepidi H, Raoult D, Fenollar F (2010) Systemic *Tropheryma whipplei*: clinical presentation of 142 patients with infections diagnosed or confirmed in a reference center. *Medicine* 89(5):337–345. <https://doi.org/10.1097/MD.0b013e3181f204a8>
6. Gorvel L, Al Moussawi K, Ghigo E, Capo C, Mege JL, Desnues B (2010) *Tropheryma whipplei*, the Whipple's disease bacillus, induces macrophage apoptosis through the extrinsic pathway. *Cell Death Dis* 1(4):e34–e34. <https://doi.org/10.1038/cddis.2010.11>
7. Bentley SD, Maiwald M, Murphy LD, Pallen MJ, Yeats CA, Dover LG et al (2003) Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet* 361(9358):637–644. [https://doi.org/10.1016/S0140-6736\(03\)12597-4](https://doi.org/10.1016/S0140-6736(03)12597-4)
8. Lagier JC, Fenollar F, Lepidi H, Raoult D (2011) Evidence of lifetime susceptibility to *Tropheryma whipplei* in patients with Whipple's disease. *J Antimicrob Chemother* 66(5):1188–1189. <https://doi.org/10.1093/jac/ckr032>
9. Fenollar F, Rolain JM, Alric L, Papo T, Chauveheid MP, van de Beek D, Raoult D (2009) Resistance to trimethoprim/sulfamethoxazole and

- Tropheryma whipplei. *Int J Antimicrob Agents* 34(3):255–259. <https://doi.org/10.1016/j.ijantimicag.2009.02.014>
10. Joshi A, Kaushik V (2021) In-silico proteomic exploratory quest: crafting T-cell epitope vaccine against Whipple's disease. *Int J Pept Res Ther* 27:169–179. <https://doi.org/10.1007/s10989-020-10077-9>
 11. Zavala A, Naya H, Romero H, Musto H (2002) Trends in codon and amino acid usage in *Thermotoga maritima*. *J Mol Evol* 54(5):563–568. <https://doi.org/10.1007/s00239-001-0040-y>
 12. Lafay B, Atherton JC, Sharp PM (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* 146(4):851–860. <https://doi.org/10.1099/00221287-146-4-851>
 13. Romero H, Zavala A, Musto H (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res* 28(10):2084–2090. <https://doi.org/10.1093/nar/28.10.2084>
 14. Sharma P, Sharma P, Ahmad S, Kumar A (2022) Chikungunya virus vaccine development: through computational proteome exploration for finding of HLA and cTAP binding novel epitopes as vaccine candidates. *Int J Pept Res Ther* 28(2):1–15. <https://doi.org/10.1007/s10989-021-10347-0>
 15. Joshi A, Ray NM, Singh J, Upadhyay AK, Kaushik V (2022) T-cell epitope-based vaccine designing against Orthohantavirus: a causative agent of deadly cardio-pulmonary disease. *Netw Model Anal Health Inform Bioinform* 11(1):1–10. <https://doi.org/10.1007/s13721-021-00339-x>
 16. Daniel E, Onwukwe GU, Wierenga RK, Quaggin SE, Vainio SJ, Krause M (2015) ATGme: open-source web application for rare codon identification and custom DNA sequence optimization. *BMC Bioinform* 16(1):1–6
 17. Dimitrov I, Naneva L, Doytchinova I, Bangov I (2014) AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* 30(6):846–851
 18. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 48(W1):W449–W454
 19. Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinform* 8(1):1–7
 20. Thévenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tuffery P (2012) PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Res* 40(W1):W288–W293
 21. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) Patch-Dock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33(suppl_2):W363–W367
 22. Antunes DA, Moll M, Devaurs D, Jackson KR, Lizée G, Kaviraki LE (2017) DINC 2.0: a new protein-peptide docking webserver using an incremental approach. *Cancer Res* 77(21):e55–e57
 23. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN et al (2019) Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *J Mol Biol* 431(13):2434–2441
 24. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U et al (2017) A new and updated resource for codon usage tables. *BMC Bioinform* 18(1):1–10
 25. Wright F (1990) The 'effective number of codons' used in a gene. *Gene* 87(1):23–29
 26. Butt AM, Nasrullah I, Qamar R, Tong Y (2016) Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg Microbes Infect* 5(1):1–14
 27. Seligmann H (2019) Localized context-dependent effects of the "ambush" hypothesis: more off-frame stop codons downstream of shifty codons. *DNA Cell Biol* 38(8):786–795
 28. Sharp PM, Li WH (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* 14(19):7737–7749
 29. Das S, Paul S, Dutta C (2006) Evolutionary constraints on codon and amino acid usage in two strains of human pathogenic actinobacteria *Tropheryma whipplei*. *J Mol Evol* 62(5):645–658
 30. Joshi A, Joshi BC, Mannan MAU, Kaushik V (2020) Epitope based vaccine prediction for SARS-COV-2 by deploying immuno-informatics approach. *Inform Med Unlocked* 19:100338. <https://doi.org/10.1016/j.imu.2020.100338>
 31. Akhtar N, Joshi A, Singh B, Kaushik V (2020) Immuno-informatics quest against COVID-19/SARS-COV-2: determining putative T-cell epitopes for vaccine prediction. *Infect Disord Drug Targets*. <https://doi.org/10.2174/1871526520666200921154149>
 32. Krishnan S, Joshi A, Akhtar N, Kaushik V (2021) Immunoinformatics designed T cell multi epitope dengue peptide vaccine derived from non structural proteome. *Microb Pathog* 150:104728. <https://doi.org/10.1016/j.micpath.2020.104728>
 33. Krishnan S, Joshi A, Kaushik V (2020) T cell epitope designing for dengue peptide vaccine using docking and molecular simulation studies. *Mol Simul* 46(10):787–795. <https://doi.org/10.1080/08927022.2020.1772970>
 34. Kaushik V (2019) In silico identification of epitope-based peptide vaccine for Nipah virus. *Int J Pept Res Ther* 1–7. <https://doi.org/10.1007/s10989-019-09917-0>
 35. Akhtar N, Joshi A, Kaushik V, Kumar M, Mannan MAU (2021) In-silico design of a multivalent epitope-based vaccine against *Candida auris*. *Microb Pathog* 155:104879. <https://doi.org/10.1016/j.micpath.2021.104879>
 36. Jain P, Joshi A, Akhtar N, Krishnan S, Kaushik V (2021) An immunoinformatics study: designing multivalent T-cell epitope vaccine against canine circovirus. *J Genet Eng Biotechnol* 19(1):1–11. <https://doi.org/10.1186/s43141-021-00220-4>
 37. Sharma P, Kaur R, Upadhyay AK, Kaushik V (2020) In-silico prediction of peptide based vaccine against Zika virus. *Int J Pept Res Ther* 26(1):85–91. <https://doi.org/10.1007/s10989-019-09818-2>
 38. Akhtar N, Joshi A, Singh J, Kaushik V (2021) Design of a novel and potent multivalent epitope based human Cytomegalovirus peptide vaccine: an immunoinformatics approach. *J Mol Liq* 116586. <https://doi.org/10.1016/j.molliq.2021.116586>
 39. Joshi A, Krishnan GS, Kaushik V (2020) Molecular docking and simulation investigation: effect of beta-sesquiphellandrene with ionic integration on SARS-CoV2 and SFTS viruses. *J Genet Eng Biotechnol* 18(1):1–8. <https://doi.org/10.1186/s43141-020-00095-x>
 40. Joshi A, Pathak DC, Mannan MAU, Kaushik V (2021) In-silico designing of epitope-based vaccine against the seven banded grouper nervous necrosis virus affecting fish species. *Netw Model Anal Health Inform Bioinform* 10(1):1–12. <https://doi.org/10.1007/s13721-021-00315-5>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)