Journal of
Data Mining in Genomics & Proteomics

Nair et al., J Data Mining Genomics Proteomics 2016, 7:2
DOI: 10.4172/2153-0602.1000192

Research Article Open Access

Genome Mining and Comparative Genomic Analysis of Five Coagulase-Negative Staphylococci (CNS) Isolated from Human Colon and Gall Bladder

Ramesan Girish Nair[1], Gurwinder Kaur[2], Indu Khatri[3], Nitin Kumar Singh[2], Sudeep Kumar Maurya[1], Srikrishna Subramanian[3], Arunanshu Behera[4], Divya Dahiya[4], Javed N Agrewala[1]* and Shanmugam Mayilraj[2]*

[1]*Immunology Laboratory, CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh -160036*
[2]*Microbial Type Culture Collection and Gene bank (MTCC), Institute of Microbial Technology, Sector 39-A, Chandigarh -160036, India*
[3]*Protein Science and Engineering, CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh -160036, India*
[4]*Department of Surgery, Postgraduate Institute of Medical Education and Research, Chandigarh, India*

Abstract

Coagulase-negative Staphylococci (CNS) are known to cause distinct types of infections in humans like endocarditis and urinary tract infections (UTI). Surprisingly, there is a lack of genome analysis data in literature against CNS particularly of human origin. In light of this, we performed genome mining and comparative genomic analysis of CNS strains *Staphylococcus cohnii* subsp. *cohnii* strain GM22B2, *Staphylococcus equorum* subsp. strain *equorum* G8HB1, *Staphylococcus pasteuri* strain BAB3 isolated from gall bladder and *Staphylococcus haemolyticus* strain 1HT3, *Staphylococcus warneri* strain 1DB1 isolated from colon. We identified 29% of shared virulence determinants in the CNS strains which involved resistance to antibiotics and toxic compounds, bacteriocins and ribosomally synthesized peptides, adhesion, invasion, intracellular resistance, prophage regions, pathogenicity islands. 10 unique virulence factors involved in adhesion, negative transcriptional regulation, resistance to copper and cadmium, phage maturation were also present in our strains. Apart from comparing the genome homology, size and G + C content, we also showed the presence 10 different CRISPR-cas genes in the CNS strains. Further, KAAS based annotation revealed the presence of CNS genes in different pathways involved in human diseases. In conclusion, this study is a first attempt to unveil the pathogenomics of CNS isolated from two distinct body organs and highlights the importance of CNS as emerging pathogens of health care sector.

Keywords: CNS; CLC; Staphylococcus; Comparative genomics

Introduction

The genus Staphylococcus is very well characterized consisting of fifty one species and twenty seven sub-species (www.bacterio.net/staphylococcus.html). The members are Gram-stain-positive with low G + C content 30-35 mol % [1,2]. Methicillin resistant *Staphylococcus aureus* (MRSA) and vancomycin resistant *Staphylococcus aureus* (VRSA) are some of the prominent pathogens that cause wide variety of infections in humans as well as animals [3-7]. The human body consists of a vast repertoire of bacteria, among which the genus *Staphylococcus* represents the proportion of bacteria that can cause severe infections to the host and majority of these colonize inside new born babies through mother's skin [8,9]. *Staphylococcus aureus* is a major pathogen in the genus that causes endovascular infections, pneumonia, septic arthritis, endocarditis, osteomyelitis, foreign-body infections and sepsis in hospitals and outpatients [10-13]. Second most important are those CNS that target neonates through intravascular catheters, prosthetic devices, post-operative sternal wound infections and immune-compromised hosts in the health care environment [14-16]. Interestingly, CNS are gradually developing drug resistance characteristics, which limits present therapies and poses a great threat to the health care system worldwide [8,14,17-21].

To study more insight in the pathogenicity of CNS isolated from humans, we sequenced the draft genomes of three CNS strains recovered from gall bladder and two from colon of human organs collected at Post Graduate Institute of Medical Education and Research, Chandigarh, India.

The homology and differences in the five CNS genomes were assessed using Mauve 2.3.1 and BRIG. Further, a comparative genomic strategy was employed using the published genome of *S. aureus* strain RF122 to characterize the pathogenic properties among the CNS isolates. We were able to identify and analyze the major virulence determinants between these strains, which included adhesion, resistance to antibiotic and toxic compounds, bacteroicins and ribosomally synthesized peptides, invasion and intracellular resistance, phages and prophages, CRISPR-cas proteins etc. Findings demonstrated the pathogenic potential of CNS isolated from two distinct body organs and identified them as emerging human pathogens.

Materials and Method

Bacterial strain isolation and identification

Out of five species of staphylococci, three strains *viz.*, *Staphylococcus cohnii* subsp. *cohnii* strain GM22B2, *Staphylococcus equorum* subsp. *equorum* strain G8HB1, *Staphylococcus pasteuri* strain BAB3 were isolated from gall bladder, whereas *Staphylococcus haemolyticus* strain 1HT3, and *Staphylococcus warneri* strain 1DB1 were isolated from the colon. All the five strains of CNS were isolated from five different patients. Patient 1: Strain G22B2, Patient 2: Strain G8HB1, Patient

***Corresponding authors:** Shanmugam Mayilraj, Microbial Type Culture Collection and Gene bank (MTCC), Institute of Microbial Technology, Sector 39-A, Chandigarh -160036, India, Tel: +0091-172-6665; E-mail: mayil@imtech.res.in

Javed N Agrewala, Immunology Laboratory, CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh 160036, India, Tel: +0091-172-6261; E-mail:javed@imtech.res.in

Received January 28, 2016; **Accepted** February 18, 2016; **Published** February 28, 2016

Citation: Nair RG, Kaur G, Khatri I, Singh NK, Maurya SK, et al. (2016) Genome Mining and Comparative Genomic Analysis of Five Coagulase-Negative Staphylococci (CNS) Isolated from Human Colon and Gall Bladder. J Data Mining Genomics Proteomics 7: 192. doi:10.4172/2153-0602.1000192

3: Strain BAB3, Laparoscopic cholecystectomy was performed for removal of gallstones. Patient 4: Strain 1HT3, gastric lipoma, sample for biopsy. Patient 5: Strain 1DB1, carcinoma cecum, terminal colon. The tissues samples were recovered during the course of surgery. They were cut into smaller pieces with sterile scissor and forceps. The tissue samples were homogenized in sterile 1X PBS and centrifuged at 4000 rpm for 2 minutes to remove debris. The supernatants were serially diluted and plated on tryptic soya agar (TSA; HiMedia, India), incubated at 37°C for 36 h and pure colonies were isolated. The selected strains were identified by 16S rRNA gene sequencing. Genomic DNA extraction and amplification was performed as previously described [22]. Identification of phylogenetic neighbours and the calculation of pairwise 16S rRNA gene sequence similarities were achieved using the EzTaxon server [23] and alignment was carried out using Mega version 6.0 [24]. Phylogenetic trees were constructed using the neighbour-joining as well as maximum likelihood and maximum parsimony algorithms. Bootstrap analysis was performed to assess the confidence limits of the branching (Figure 1).

Whole genome sequencing and annotation

The draft genome was sequenced at Genotypic Pvt. Ltd. (Bengaluru, India, http://www.genotypic.co.in). Library preparation was performed at Genotypic Technology's genomics facility following NEXTFlex DNA library protocol as per manufacturer's instructions. ~ 3 µg of genomic DNA was sonicated using Covaristo to obtain 500 to 700 bp fragment size. The size distribution was checked by running an aliquot of the sample on Agilent HS DNA Chip. The resulting fragmented DNA was cleaned up using HighPrep PCR clean up system as described by the manufacturer. Fragmented DNA was subjected to a series of enzymatic reactions that repaired frayed ends, phosphorylated fragments, and added a single nucleotide 'A' overhang then ligated adaptors using NEXTFlex DNA sequencing kit following the protocol as described by manufacturer. Sample cleanup was done using HighPrep PCR beads. After ligation-cleanup, ~ 500-800 bp fragments was size selected on 2% low melting agarose gel and cleaned using MinElute column, QIAGEN. PCR (cycles) amplification of adaptor ligated fragments was done and cleaned up using HighPrep PCR Clean-up beads. The prepared

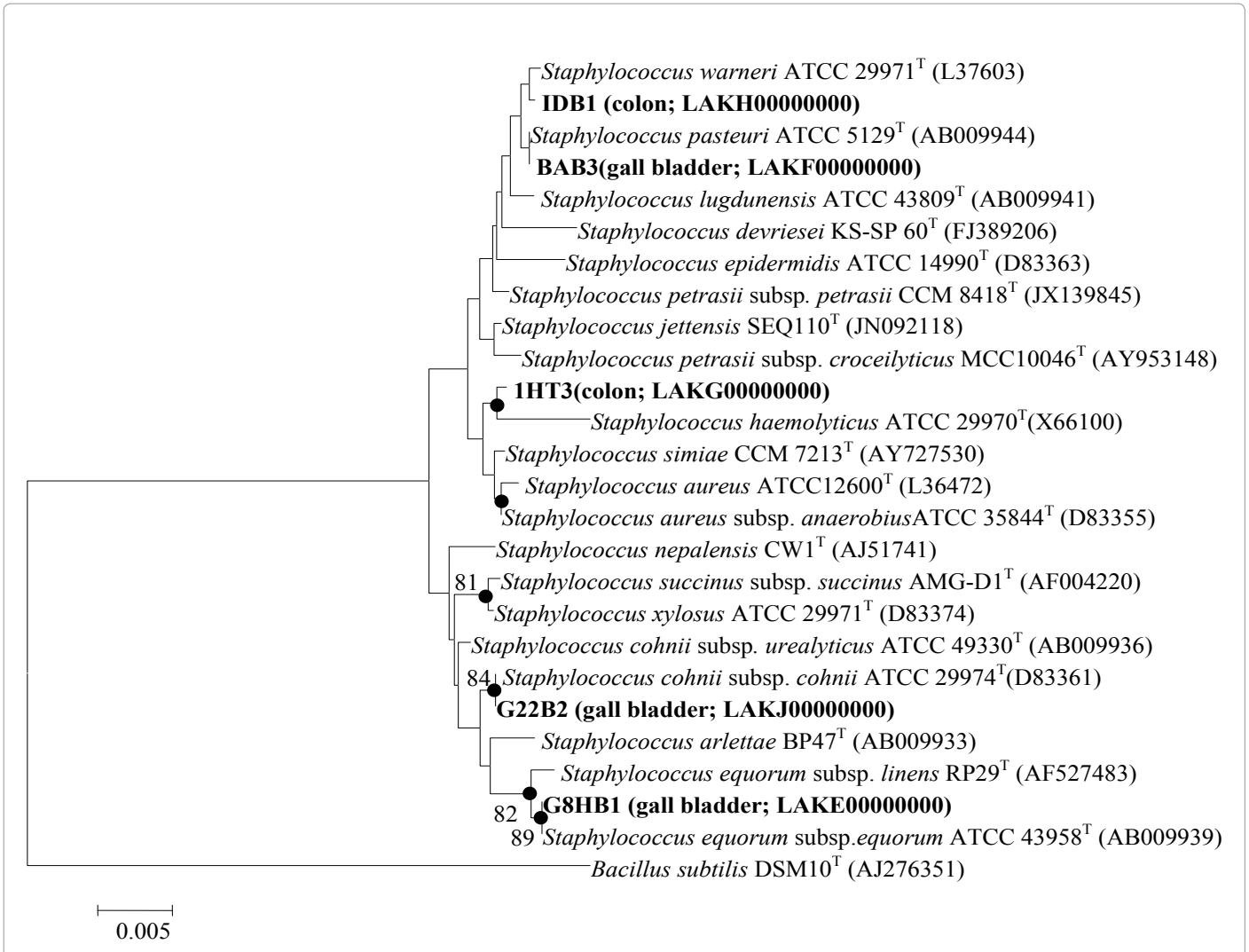

Figure 1: Neighbour-joining tree based on 16S rDNA sequences, showing the phylogenetic relationship between CNS strains and other related members of the genus *Staphylococcus*. *Bacillus subtilis* DSM10^T (AJ276351) was used as an out group Bootstrap values (expressed as percentage of 100 replications) greater than 70 % are given at the nodes. Filled circles indicate that corresponding nodes were also recovered in the tree generated with maximum parsimony and maximum likelihood algorithms. Bar 1 % sequence variation.

libraries were quantified using Qubit flurometer and validated for quality by running an aliquot on High Sensitivity Bio analyzer Chip, Agilent. Assembly was carried out with CLC Bio Workbench v6.0.4 (CLC Bio, Denmark).

Comparative genomics and pathogenomics

The automated genome annotation for all five *Staphylococcus* strains was accomplished using RAST [25-27]. The ribosomal RNA genes in the genomes were identified by RNAmmer 1.2 [28]. The tRNA and tmRNA genes were identified by ARAGON [29]. Prophage regions were identified by PHAST [30]. Insertion sequence (IS) elements were identified by the IS finder (http://www-is.biotoul.fr/) [31]. Genome sequence similarity among the five CNS along with reference strain *Staphylococcus aureus* RF122 was carried out using BRIG [32]. Multiple whole genome sequence alignment was performed using Mauve 2.3.1 which uses sum of pairs breakpoint score for rearrangement detection in the whole genomes. Even if two strains have unequal genome content, this method can predict genome rearrangement with high accuracy which makes it an important tool for evolutionary genomics [33,34]. CRISPR finder tool was used to identify the CRISPR genes in the genomes of the CNS strains [35]. Further, KEGG Automatic Annotation Server (KAAS, http://www.genome.jp/tools/kaas/) was employed to map the orthologous genes and their biological roles in various pathways.

Results

Genome features

The genomes were sequenced with Illumina Miseq and assembly was carried out with CLC bio workbench v6.0.4 (CLC Bio, Denmark) (www.clcbio.com). Among the CNS isolates the genome size of the *S. equorum* subsp. *equorum* G8HB1 was the largest (2.799 Mb) compared to other CNS species (ranging from 2.403 Mb to 2.776 Mb). Genomic G+C content of *S. equorum* subsp. *equorum* G8HB1 (33.08%) was highest among all CNS. *S. haemolyticus* 1HT3 (32.78%), *S. cohnii* subsp. *cohnii* G22B2 (32.28%) and *S. warneri* 1DB1 (32.55%) was higher than S. pasteuri BAB3 (31.50%). The variation in the genomic G + C content among the CNS species could be attributed to mutation and selection pressures [36,37] which may arise due to multiple factors like environment [38], symbiotic lifestyle [39], aerobiosis [40], and nitrogen fixing ability [41]. There was a slight variation in the rRNA operons and tRNA coding genes (Table 1) among the CNS species, which could be correlated with the strength of codon usage bias value known as the S value. The species of CNS growing rapidly would have more rRNA, tRNA genes and more codon usage bias [41]. Number of Insertion sequence (IS) elements were maximum in *S. haemolyticus* strain 1HT3, 33 IS elements belonging to 11 IS families, which indicated a genome-wide inversion and rearrangement (Table 1). Although, the total number of IS elements in the other CNS species were comparable but, the diversity and distribution in IS families were different.

Multiple whole genome alignment

The multiple whole genome alignment of the five CNS strains was carried out with reference strain *S. aureus* RF122. Extent of Local Collinear Blocks (LCB) connecting lines depicted more homology in the genomes. The genomes of strain 1DB1, 1HT3 and BAB3 showed more connecting lines therefore have a greater extent of homologous regions in their genomes. Conditional links in red colour between 1DB1, G22B2 and reference strain RF122 portrayed common stretch of sequence among each other (Figure 2).

Draft genome visualization using BRIG

Whole genome circular comparative map of five CNS strains against reference genome *S. aureus* RF122 was generated on the basis of BLAST sequence similarities using BRIG software [32] which mapped the whole genome in the form of concentric rings. Each genome was represented by a different colour and the darker areas in the circular genome displayed 100% sequence similarity with the reference genome, whereas the lighter (grey) areas showed 70% sequence similarity (Figure 3).

Genome comparison and identification of virulence determinants

Genome comparison among the five CNS strains with reference genome of *S. aureus* RF122 revealed several major categories of genes, among which two categories viz., 1. Virulence, disease and defense. 2. Phages, prophages, transposable elements and plasmids were further studied because of their utmost relevance in contributing pathogenesis in humans. Total genes present in the all the major categories that are responsible for virulence in the strains were 128. Significant numbers of unique virulence factors present in genomes of CNS strains were 10 (Table 2). Pie chart depicts the putative factors contributing towards pathogenesis in the CNS strains (Figures 4A and 4B).

Genes responsible for adhesion

Adhesion to the surface of the host cell is the primary step in the process of infection, which determines the pathogen survival and extent of pathogenicity. A total number of 23 genes were present in all the strains and divided into two sub-systems that were responsible for adhesion. One gene involved in functional role as chaperonin

Attributes	*S. aureus* RF122	*S. cohnii* subsp. *cohnii* G22B2	*S. equorum* subsp. *equorum* G8HB1	*S. pasteuri* BAB3	*S. haemolyticus* 1HT3	*S. warneri* 1DB1
Isolation source	*Bos taurus* (cow)	Gall Bladder	Gall Bladder	Gall Bladder	Colon	Colon
Genome Size (bp)	27,42,531	27,76,466	27,99,869	25,62,464	24,03,066	25,89,616
G + C content (%)	32.8	32.28	33.08	31.5	32.78	32.55
N50	2,742,531	48131	156794	302579	35235	318901
Number of contigs	1	101	22	38	99	83
Total no. of CDS	2609	2723	2706	2457	2318	2480
rRNA operons	16	6	7	6	6	8
tRNA coding genes	60	68	58	60	64	60
tmRNA coding genes	1	1	1	1	1	1
No. of IS elements	79	21	24	26	33	22
Prophage regions	4	2	2	1	1	4
Psuedogenes	152	79	38	33	35	37

Table 1: Genome features of five CNS strains.

Strain Name	Gene	Length	Function
G22B2	Fibronectin binding protein (PrtF)	396 bp	Group of adhesive glycoproteins that interact selectively and non-covalently with fibronectin
	Negative transcriptional regulator copper transport operon (TR)	495 bp	Negatively regulates transcription in order to maintain accurate concentration of copper in cell
	HNH homing endonucleases (HNH)	777 bp	Phage associated endonuclease
G8HB1	Copper chaperone (CopZ)	237 bp	Assists in delivery of copper ions to target proteins
	Cadmium-transporting ATPase (EC 3.6.3.3)	1512 bp	ATPase function in providing resistance against Cadmium
	Phage major capsid protein (MajCap)	975 bp	Structural subunit of phage capsid
1HT3	Copper resistance protein copD	1248 bp	Copper ion binding to maintain copper homeostasis
	Copper resistance protein copC	1248 bp	Copper ion binding to maintain copper homeostasis
1DB1	Adhesin of unknown specificity (SdrD)	3204 bp	Adhesion to host cell
	Phage head maturation protease	558 bp	Protease involved in maturation of phage heads

Table 2: Unique virulence determinants present in the CNS strains.

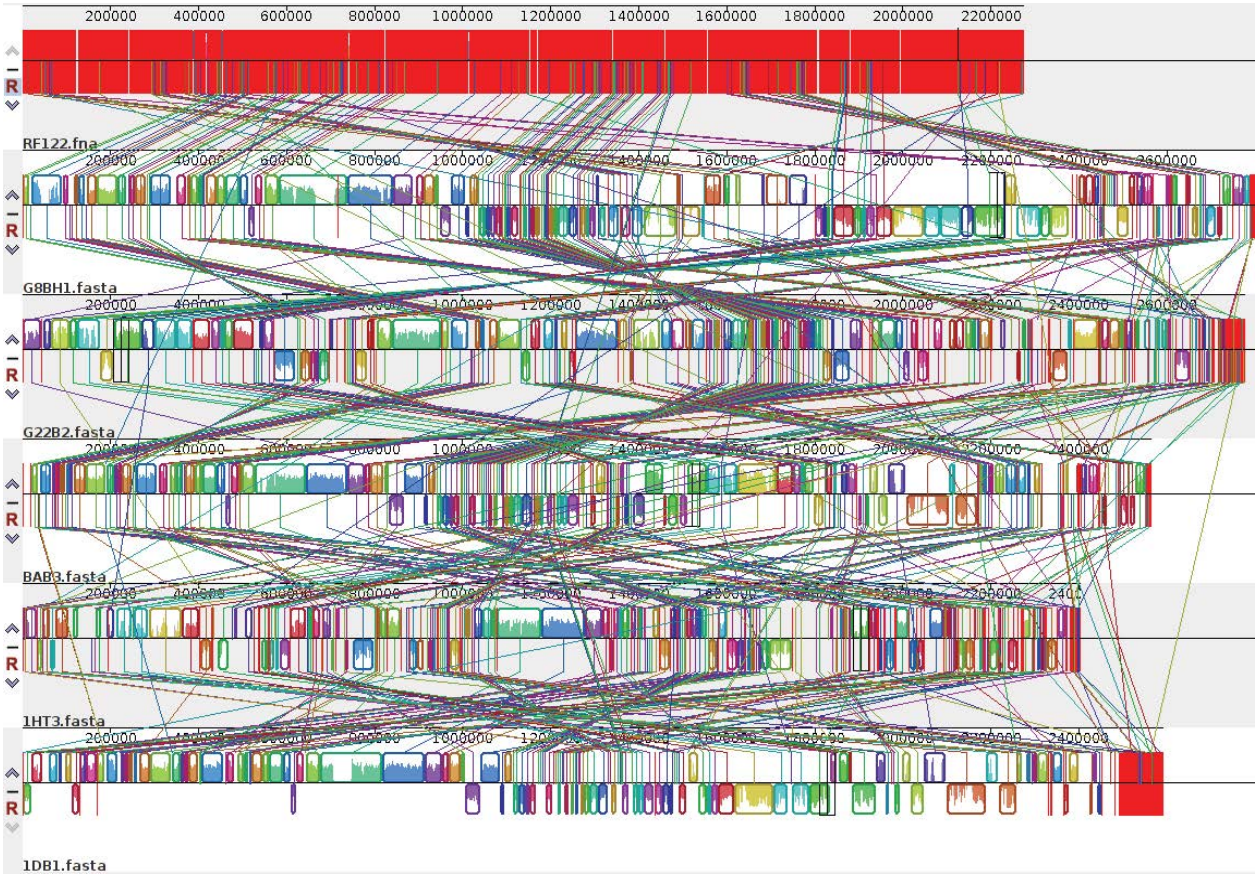

Figure 2: The topmost horizontal panel in red colour depicts the genome of reference strain RF122. Each of the CNS genome is separated by a single black horizontal center line, top of which name of strain is written. Coloured block outlines are known as Locally Collinear Blocks (LCBs) and are connected by corresponding coloured lines. LCBs represent the regions of similarity among the genomes that are homologous and have not undergone any rearrangement. Blocks lying above the center black horizontal line are in forward orientation relative to the first genome sequence. Blocks below the center line indicate regions that align in the reverse complement (inverse) orientation. Regions outside blocks show no homology among the input genomes. The similarity profile of the bacterial genome is drawn inside each block; the height of similarity profile corresponds to the average level of conservation in that region of the genome.

(heat shock protein 33) was common in all the strains (Figure 5). The number of genes responsible for adhesion varies to a great extent among the CNS strains compared to the reference genome of *S. aureus* RF122 which had most number of genes. *S. cohnii* subsp. *cohnii* strain G22B2 had only one gene and *S. equorum* subsp. *equorum* strain G8HB1 had no genes that could be related to poor adhesion capacity of these strains. Adhesion helps to predict the bio-film formation which is crucial in understanding pathogenicity.

Genes encoding toxins and super antigens

S. aureus species is well known to produce exfoliative toxins (ETs) and pyrogenic toxin super-antigens (PTSAgs) that cause diseases like staphylococcal scalded-skin syndrome (SSSS), staphylococcal toxic shock syndrome (TSS), and staphylococcal food poisoning (SFP) [10,18]. Super antigens of *S. aureus* can directly bind to the V-β region of T-cells and major histocompatibility complex (MHC) class II molecules of antigen presenting cells, thereby avoiding antigen

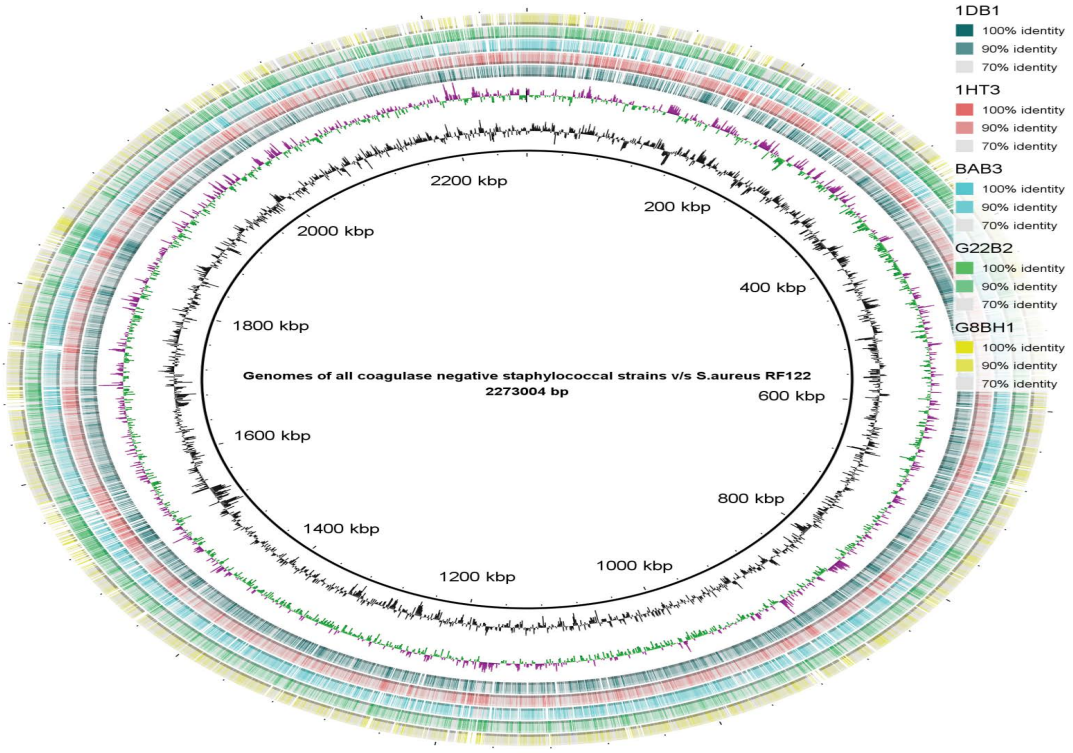

Figure 3: Comparison of five CNS strains viz., *S. cohnii* subsp. *cohnii* strain G22B2, *S. equorum* subsp. *equorum* strain G8HB1, *S. haemolyticus* strain 1HT3, *S. pasteuri* strain BAB3, *S. warneri* strain 1DB1 against *S. aureus* strain RF122. The innermost dark circle represents the reference genome of strainRF122 (genome size 2273,004 bp), the dark green circle surrounding reference genome signifies the genome of strain 1DB1 and the outermost yellow circle denotes genome of strain G8HB1, light green (G22B2), cyan (BAB3), red (1HT3). The circles lying in between the reference genome and the CNS genomes; black symbolise the GC content and dual coloured lines indicates GC skew.

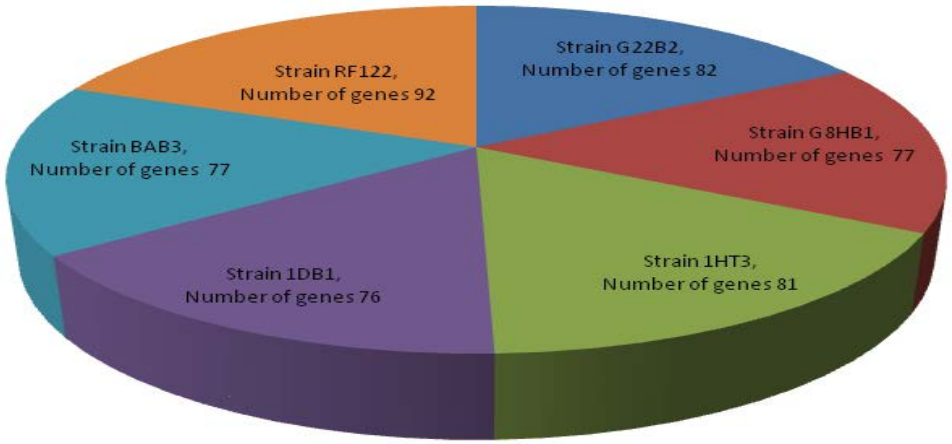

Figure 4(A): Genes involved virulence, disease and defence.

processing and presentation. This leads to direct activation of V- β expressing T-cells [15]. Genes for toxins and superantigens were absent in the CNS strains, except reference strain RF122.

Genes involved in production of bacteriocins and ribosomally synthesized antibacterial peptides

Bacteriocins are low molecular weight proteins that have lethal effect against bacteria by a fast acting mechanism, which forms pores in target membranes. The sub-category bacteriocins and ribosomally synthesized antibacterial peptides comprised of two subsystems; bacitracin stress response and colicinV bacteriocins production cluster, which consisted of a total of 12 genes. ColicinV bacteriocins production cluster was present in all the CNS strains. Strains G8HB1 and G22B2 had less genes involved in bacitracin stress response (Figure 6). These findings indicate that CNS strains are equally contributing towards antimicrobial activity as compared to *S. aureus* RF122.

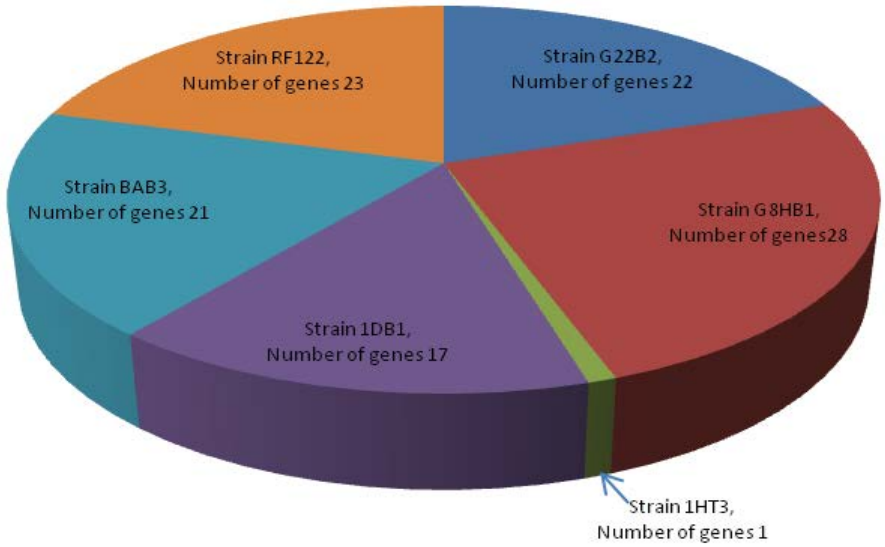

Figure 4(B): Genes involved phages, prophages, transposable elements and plasmids.

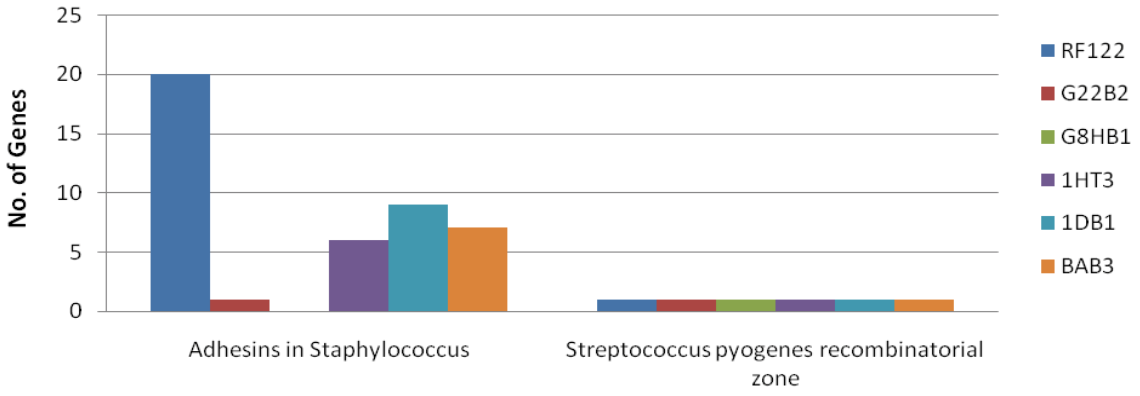

Figure 5: Comparison of genes responsible for adhesion in the Staphylococcus strains.

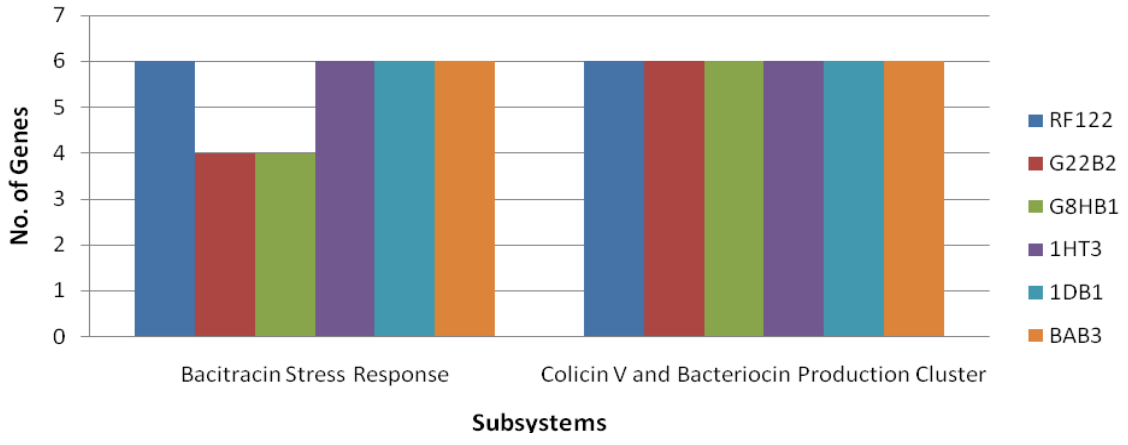

Figure 6: Comparison of genes involved in the production of bacteriocins and ribosomally synthesized antibacterial peptides present in the Staphylococcus strains.

Genes involved in resistance to antibiotic and toxic compounds

Resistance to drugs and toxic compounds is a prevalent feature in all CNS strains that necessitates for the development of a newer improved multipotent drug against the clinical CNS strains. A total number of 46 genes were present in all the staphylococcal strains under study, divided into fifteen subsystems that confer resistance to antibiotics and toxic compounds. 18 genes were commonly present in all the staphylococcal strains (Figure 7). It was observed that the CNS isolates have more number of resistance genes compared to the reference genome of *S. aureus* RF122. Strains G22B2 and G8HB1 showed maximum resistance against arsenic toxicity, cobalt-zinc cadmium toxicity, mercuric reductase and cadmium resistance. Strains 1HT3 and G8HB1 showed maximum resistance to copper haemostasis compared to *S. aureus* RF122. In beta lactamase resistance, strain 1DB1 had maximum number of genes. Genes for bile hydrolysis, multidrug resistance 2 protein versions were found in Gram-positive bacteria, tecioplanin resistance in *Staphylococcus* and resistance to fluroquinolones were equally present in all the strains. No resistance was observed against fosfomycin and chromium compounds in strains RF122, 1HT3, 1DB1 and BAB3 whereas strains G8HB1 and G22B2 showed resistance against fosfomycin and chromium. Aminoglycoside adenyltransferases resistance was absent in all the strains except G8HB1 and 1DB1. These data suggest that the clinically isolated CNS strains could be multi drug resistant.

Genes involved in invasion and intracellular resistance

This sub-category included two subsystems; Mycobacterium virulence operon involved in protein synthesis (SSU ribosomal proteins) and Mycobacterium virulence operon involved in protein synthesis (LSU ribosomal proteins), had a total number of 9 genes present in all the strains. In Mycobacteria SSU and LSU ribosomal proteins contribute in Tuberculosis infection. Our findings demonstrate that these proteins are also encoded by CNS strains, possibly conferring them pathogenicity. SSU proteins constitute the smaller subunit of ribosome and are named using Rv number for eg: Rv0682-Rv0686. Here Rv0682 is encoded by gene rpsL and forms Protein S12 which is involved in the translation initiation step. Similarly LSU proteins constitute the larger subunit of ribosome, Rv1641 is one such which is encoded by gene infC and functions as initiation factor -3 during the protein synthesis. These genes possibly originate from *Mycobacterium tuberculosis* and help in invading the host cell, supporting intracellular survival for longer periods; thereby evading host immune system.

Genes encoding phages and prophages

Several phages and prophage regions have been found in the genomes of *Staphylococcus* spp. that could contribute to virulence. This subcategory included 9 subsystems having 23 genes. Maximum numbers of genes were present in the subsystem phage packaging machinery in strain 1DB1 (Figure 8). Phage tail length tape measure

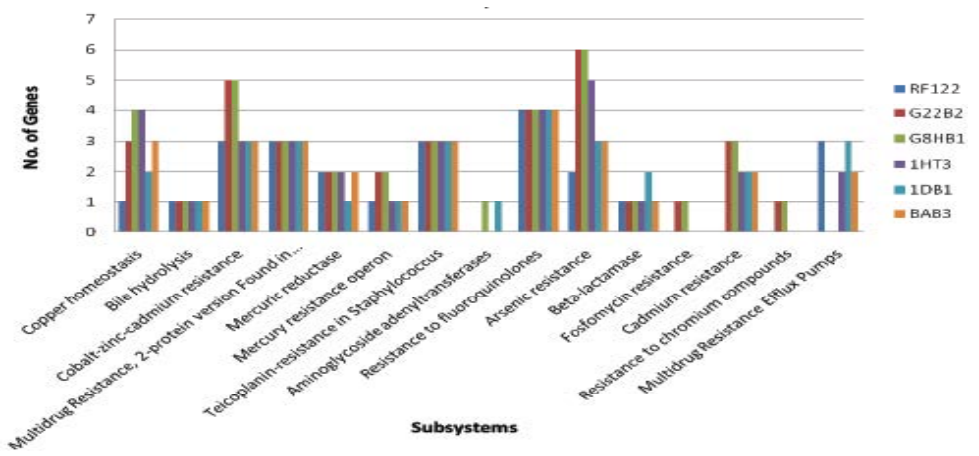

Figure 7: Comparison of genes present in the *Staphylococcus* strains involved in conferring resistance against antibiotics and toxic compounds.

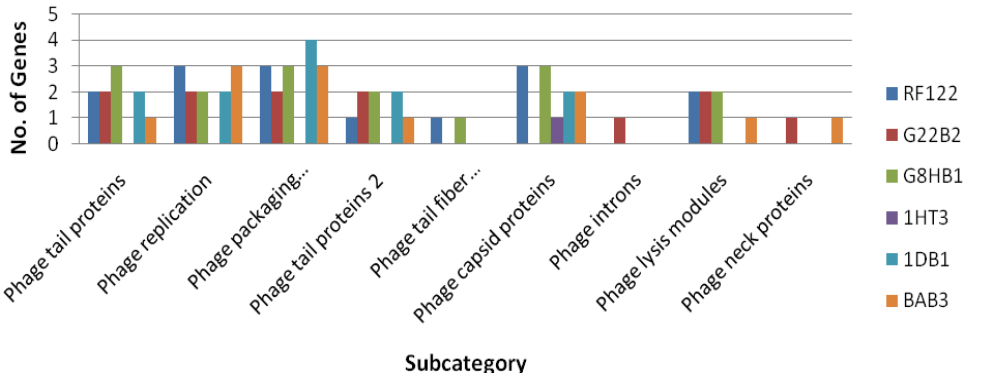

Figure 8: Comparison of genes present in the *Staphylococcus* strains encoding phages and prophages.

proteins were present in all the strains except 1HT3 which showed a possible horizontal transfer of tail fibre proteins among these strains.

Pathogenicity islands in staphylococci

Pathogenicity islands (PAIs) are one among the factors like plasmids and bacteriophages that is responsible for evolution of the pathogens [42]. In staphylococci PAIs were first identified in *S. aureus* known as SaPIs. They consisted of phage-related chromosomal islands that were highly mobile, encoding super antigens [43]. Only one PAI was found in all staphylococcal strains. Listeria pathogenicity island LIPI-1 extended contained two genes phosphatidylinositol-specific phospholipase C (EC 4.6.1.13) present only in the reference strain RF122 and zinc metalloproteinase precursor (EC 3.4.24.29) present in all the strains except 1HT3.

Analysis of prophage regions

Prophages regions comprise of phage DNA integrated into the bacterial genome. Phage DNA acts as a mobile genetic element and can be considered as a vector for lateral gene transfer among bacteria. Infact, a greater proportion of the bacterial virulent factors are encoded by phage [44]. Prophage regions were identified in all the six strains, including reference genome RF122. Five strains had intact prophage regions. There were total of twelve different types of prophage regions found in all the strains, including the common prophage region PHAGE_ Staphy_PT1028_NC_007045, present in strains 1DB1 and 1HT3. Also prophage region PHAGE_Staphy_StB20_NC_019915 was present in strains 1DB1 and G22B2. Diversity in the prophage regions contributes to the adaptation of lysogens to new hosts and responsible for pathogenicity in the CNS strains (Table 3).

Identification of CRISPR-cas proteins

The bacteria and archea also have defence mechanisms, which enable them to protect themselves against foreign bodies like viruses. One such system is the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and their associated CRISPR-associated sequence (CAS) proteins that provide adaptive immunity to the bacteria [45-47]. CRISPR-cas genes were detected in strains 1DB1, G8HB1 and 1HT3, whereas in strains G22B2 and reference strain RF122 were not detected. 1DB1 had two CRISPR-cas genes and was coding for hypothetical protein (*S. aureus*) in region the 1265562-1265698 bp. 1HT3 was having only one CRISPR-cas gene in the region 10866-10925 bp coding for hypothetical protein (*S. haemolyticus*). CRISPR-cas genes were found in two regions 354805-354864, 36501-36775bp of BAB3 genome. In G22B2, two CRISPR-cas genes were found in regions 440-521, 117647-117712 bp. Among all the strains, G8HB1 had maximum number of CRISPR-cas genes present in the regions 2255-2348, 219970-220100, 291655-291801 bp. Presence of these genes in the strains presumably indicate that they have evolved phage resistance mechanisms.

Gene candidates involved in the human diseases

Putative functional genes showing involvement in the pathways correlating human diseases were identified in the CNS strains using KASS [48]. In total, 357 putative genes were associated with different processes such as metabolism, genetic information processing, cellular processing and virulence properties (Figures 9A-9E). Functional pathways in relevance to the human diseases were identified in all strains except 1DB1 (Table 4). Protein named groEL with a chaperonin function was involved in three human diseases viz; Tuberculosis, Legionellosis and Type I diabetes mellitus. In tuberculosis, it stimulates the production of IL-18 via TLR-4 mediated myD88 signaling, whereas in legionellosis it mediates the legionella entry into the epithelial cell or macrophage and in Type I diabetes mellitus it causes the apoptosis of β cells in the pancreas via IL-2 and IFN-gamma release which targets the CD8 cytotoxic T-cells and macrophages. Gene regulation is under the tight control of MicroRNA's (miRNAs) which is small cluster of 21-23 nucleotides in length. The miRNA signatures have been observed in several cancers. Protein DNMT 1 is a target of microRNA miR-152, which blocks DNMT 1 involved in hepatocellular carcinoma. Other proteins involved in the cancer pathways included frmA (alcohol dehydrogenase), which functions in chemical carcinogenesis by conversion of chloral acetate to trichloroacetic acid during the synthesis of olefines which mey lead to renal tubule adenoma and carcinoma. Protein fum C performs the reversible conversion of fumarate to malate in the mitochondria thereby acting tumor suppressor by allowing the production HGH (hypoxia-inducible factor prolyl hydroxylase) which is negatively regulated by fumarate. HGH regulates the HIF-α (hypoxia-inducible factor 1 alpha) pathway by degrading HIF- α which mediates cell proliferation via TGF-β.

Strain	Region	Length	Status	#CDs	Putative phage	G+C%	Location compared to RF122
RF122	1	39.4 Kb	incomplete	15	PHAGE_Staphy_phiSauS_IPLA35_NC_011612	34.7	250921-290365
	2	39.4 Kb	questionable	36	PHAGE_Staphy_phiSauS_IPLA35_NC_011612	33.3	317413-356830
	3	8.5 Kb	incomplete	7	PHAGE_Mycoba_ScottMcG_NC_011269	34.5	639871-648459
	4	24.5 Kb	incomplete	28	PHAGE_Staphy_prophage_phiPV83_NC_002486	33.4	684411-708939
	5	63.3 Kb	intact	77	PHAGE_Staphy_phiSauS_IPLA88_NC_011614	33.1	1519517-1582895
	6	13.7 Kb	questionable	22	PHAGE_Staphy_phiMR25_NC_010808	32.8	1684908-1698665
1DB1	1	19.2 Kb	incomplete	21	PHAGE_Staphy_PT1028_NC_007045	30.26	32989-52214
	2	42.8 Kb	intact	66	PHAGE_Staphy_StB20_NC_019915	33.51	704755-747648
	3	27.9 Kb	incomplete	29	PHAGE_Staphy_PT1028_NC_007045	31.82	2251317-2279302
	4	6.5 Kb	incomplete	11	PHAGE_Staphy_phiN315_NC_004740	32.53	2558870-2565402
1HT3	1	14.4 Kb	incomplete	21	PHAGE_Staphy_PT1028_NC_007045	31.75	2240233-2254667
BAB3	1	60.9 kb	intact	65	PHAGE_Staphy_PH15_NC_008723	34.02	2301984-2362978
G8HB1	1	45.9 kb	incomplete	62	PHAGE_Staphy_187_NC_007047	36.87	661409-707377
	2	45.8 kb	intact	67	PHAGE_Staphy_phiRS7_NC_022914	34.75	2391127-2436954
G22B2	1	43.7 kb	intact	65	PHAGE_Staphy_StB20_NC_019915	34.76	2098998-2142756
	2	32.8 kb	questionable	50	PHAGE_Staphy_phiSa119_NC_025460	35.43	2736626-2769462

Table 3: Putative prophage regions present in CNS strains.

KEGG orthology (KO)	Protein name	Major pathways in human diseases	Function in pathway
K04077	groEL	Tuberculosis	chaperonin GroEL
K04077 K03596	groEL lepA	Legionellosis	chaperonin GroEL GTP-binding protein LepA
K04077	groEL	Type I diabetes mellitus	chaperonin GroEL
K01679	E4.2.1.2B	Renal cell carcinoma	fumarate hydratase, class II [EC: 4.2.1.2]
K01679	E4.2.1.2B	Pathways in cancer	fumarate hydratase, class II [EC: 4.2.1.2]
K00121	frmA	Chemical carcinogenesis	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC: 1.1.1.284 1.1.1.1]
K00370 K07704 K11625	narG lytS YdfJ	Two component systems	nitrate reductase alpha subunit [EC: 1.7.99.4] LytT family, sensor histidine kinase LytS [EC: 2.7.13.3] membrane protein YdfJ
K00558	DNMT1	MicroRNAs in Cancer	DNA (cytosine-5)-methyltransferase 1 [EC:2.1.1.37]

Table 4: Major CNS genes involved in the metabolic pathways related to human diseases.

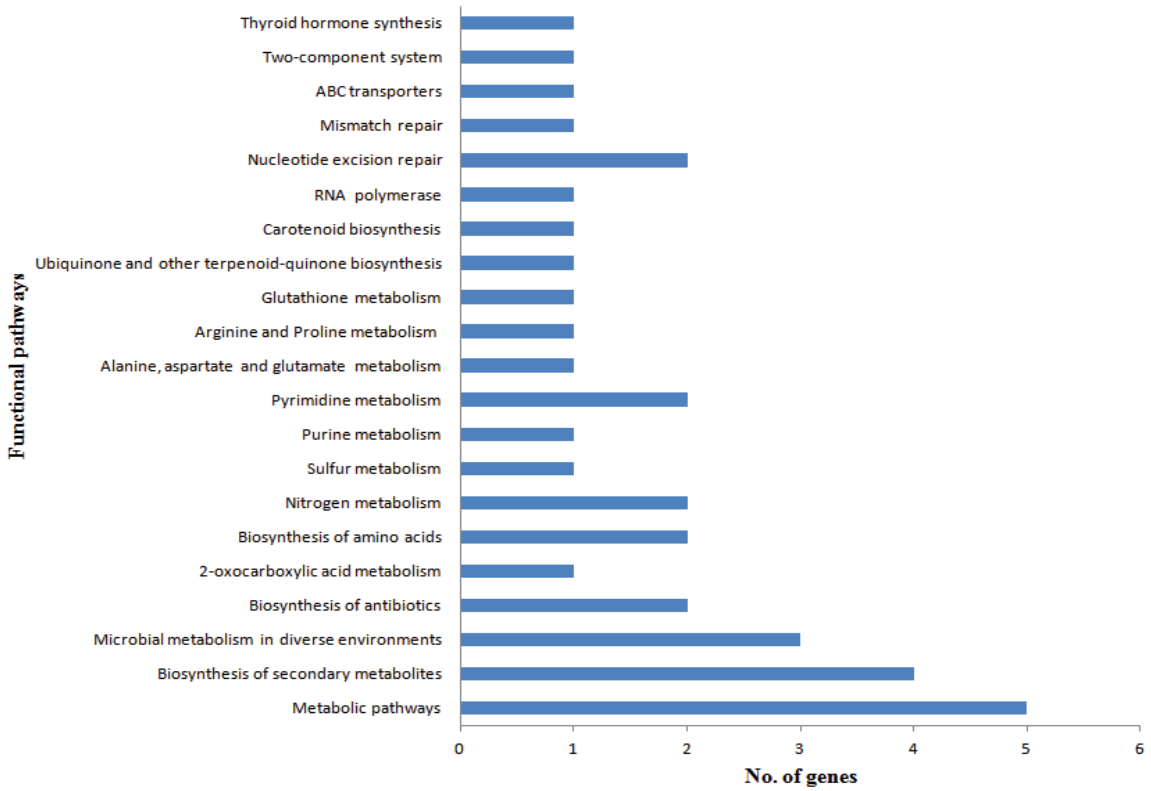

Figure 9 (A): Functional classification of the genes encoded by genome of CNS strain 1DB1. These functions were assigned using the KAAS (KEGG automatic annotation server).

Two component systems

In response to environmental cues bacteria can change their genes expression pattern through receptors which respond external chemical and physical stimuli, which constitutes the two component system. Each two-component system consists of a sensor protein-histidine kinase (HK) and a response regulator (RR). Phosphorylation of response regulator leads to change in its output domain which now can bind to DNA mediating the transcriptional control [49]. Three proteins involved in two component systems pathway were identified in CNS strains using KASS [48]. NarG (nitrate reductase alpha subunit) present in strains G8HB1, BAB3 and IDB1 was involved in nitrogen metabolism functioning in nitrate reduction. LytS (LytT family, sensor histidine kinase) of G22B2 mediated autolysis via generation of holin like proteins lrgA and lrgB. YdfJ (membrane protein) of strain 1HT3 was involved in signal transduction.

Discussion

CNS is the most prominent group associated with the hospital acquired infections, therefore they are increasingly becoming important in light of diagnostics and pathogenesis. Though these bacteria inhabit the human skin and mucous membranes, they are also isolated from a wide variety of habitats and are no longer considered as symbionts. In hospital settings, as the strategies of more invasive procedures like foreign polymer bodies catheters etc. increases, the risk of these bacteria to colonise the polymer surface by the formation of a thick, multi-layered biofilm increases. Newer antibiotic resistant strains of

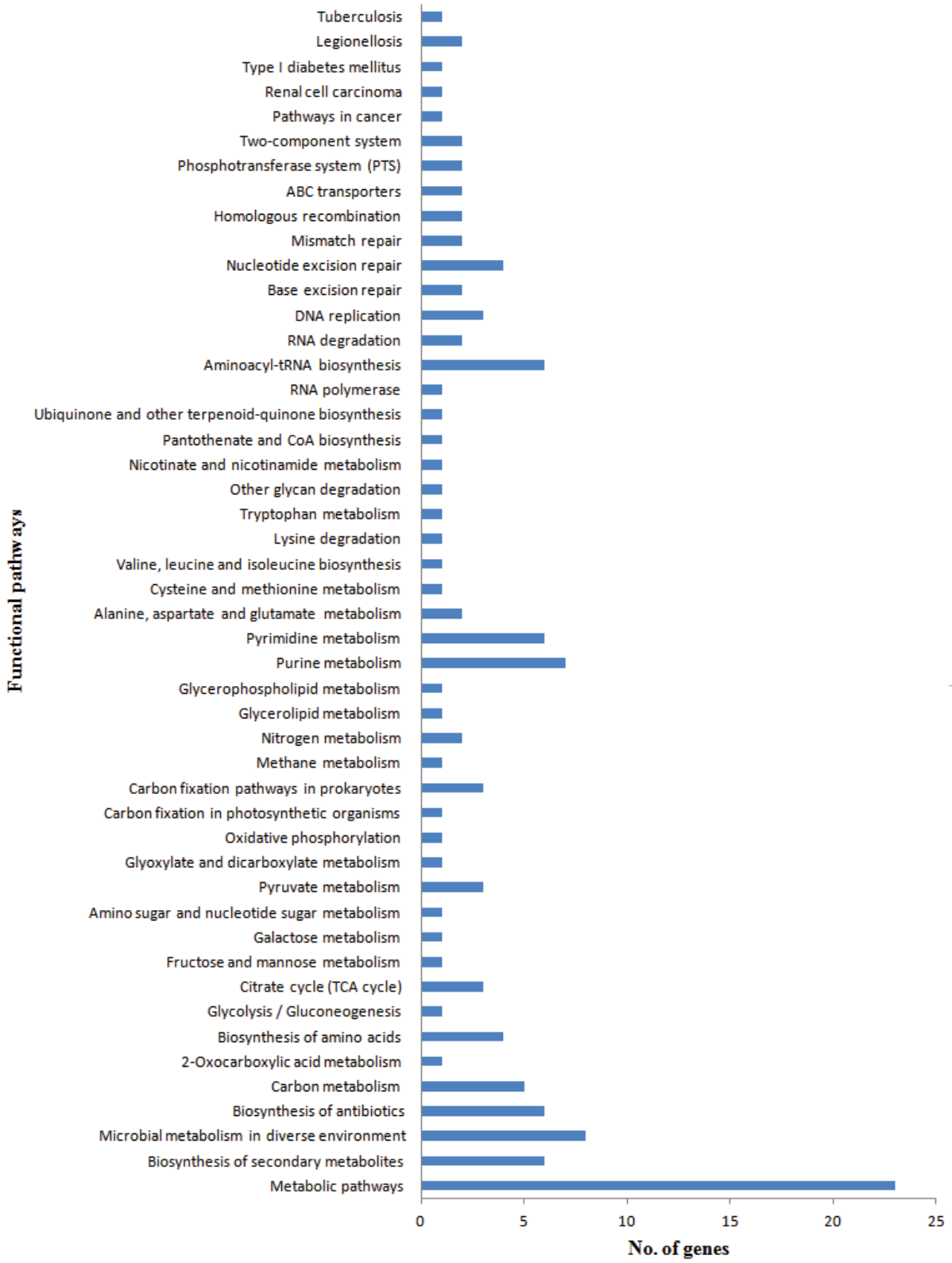

Figure 9 (B): Functional classification of the genes encoded by genome of CNS strain 1HT3. These functions were assigned using the KAAS (KEGG automatic annotation server).

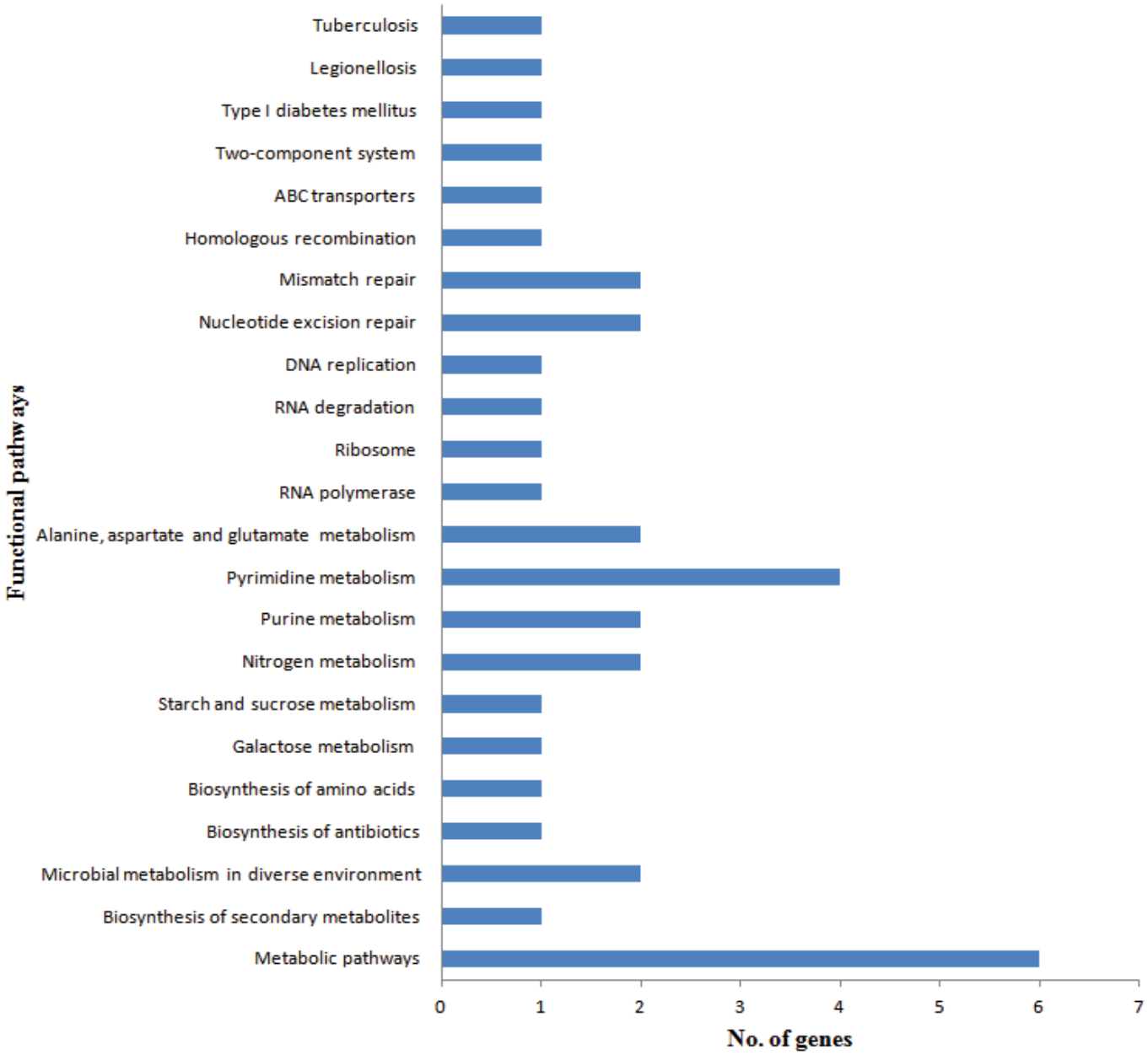

Figure 9 (C): Functional classification of the genes encoded by genome of CNS strain BAB3. These functions were assigned using the KAAS (KEGG automatic annotation server).

CNS are appearing due to continuous use of antibiotics in hospitals. CNS top's in the list of pathogens causing nosocomial infections at global level and it is imperative to understand its pathogenomics for an effective therapy. The standard microbiological and molecular biology approaches used to assess the virulence profile of pathogenic CNS, although effective is more time consuming. The limitation in these approaches could only be overcome by Whole Genome Sequencing (WGS), which is a powerful tool for studying bacterial genomics. Sequencing the entire genome provides a more detailed and robust information for comparison between two or more genomes.

In the present study using WGS and bioinformatics analysis we determined virulence factors in five different CNS strains isolated from two distinct organs, gall bladder and colon in humans. The study revealed numerous attributes which are responsible for conferring pathogenicity to the five CNS strains such as resistance to antibiotics and toxic compounds, adhesion, invasion, intracellular resistance, prophage regions etc., which were present in their genome. There were several traits that were conserved among all the CNS strains including the reference genome RF122, like genes encoding invasion and intracellular resistance, resistance to fluoroquinolones and teicoplanin, multidrug resistance 2-protein version found in Gram-positive bacteria, colicin-V and bacteriocin production cluster etc. which could be possibly explained on the basis of horizontal gene transfer. Fewer

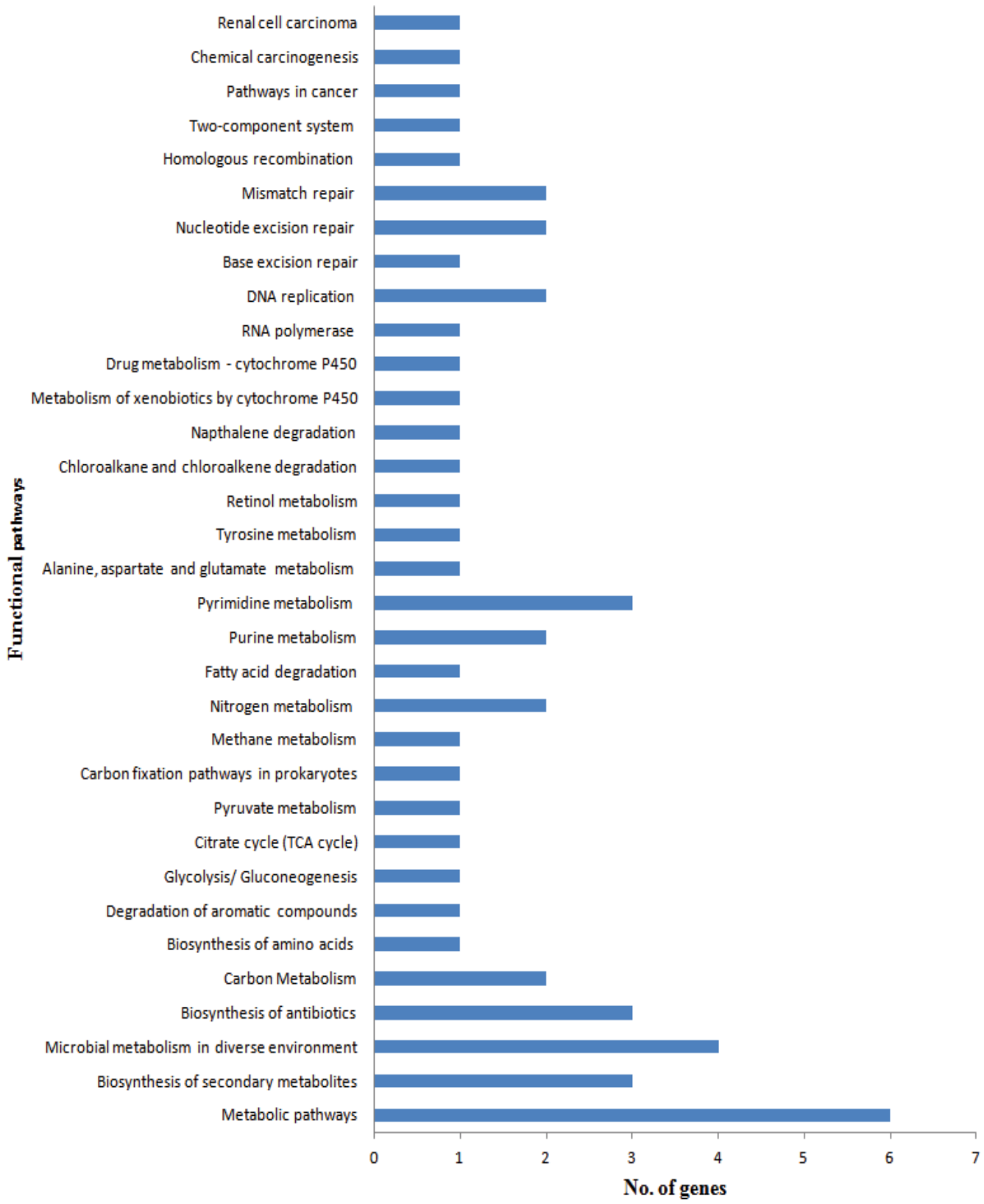

Figure 9 (D): Functional classification of the genes encoded by genome of CNS strain G8HB1. These functions were assigned using the KAAS (KEGG automatic annotation server).

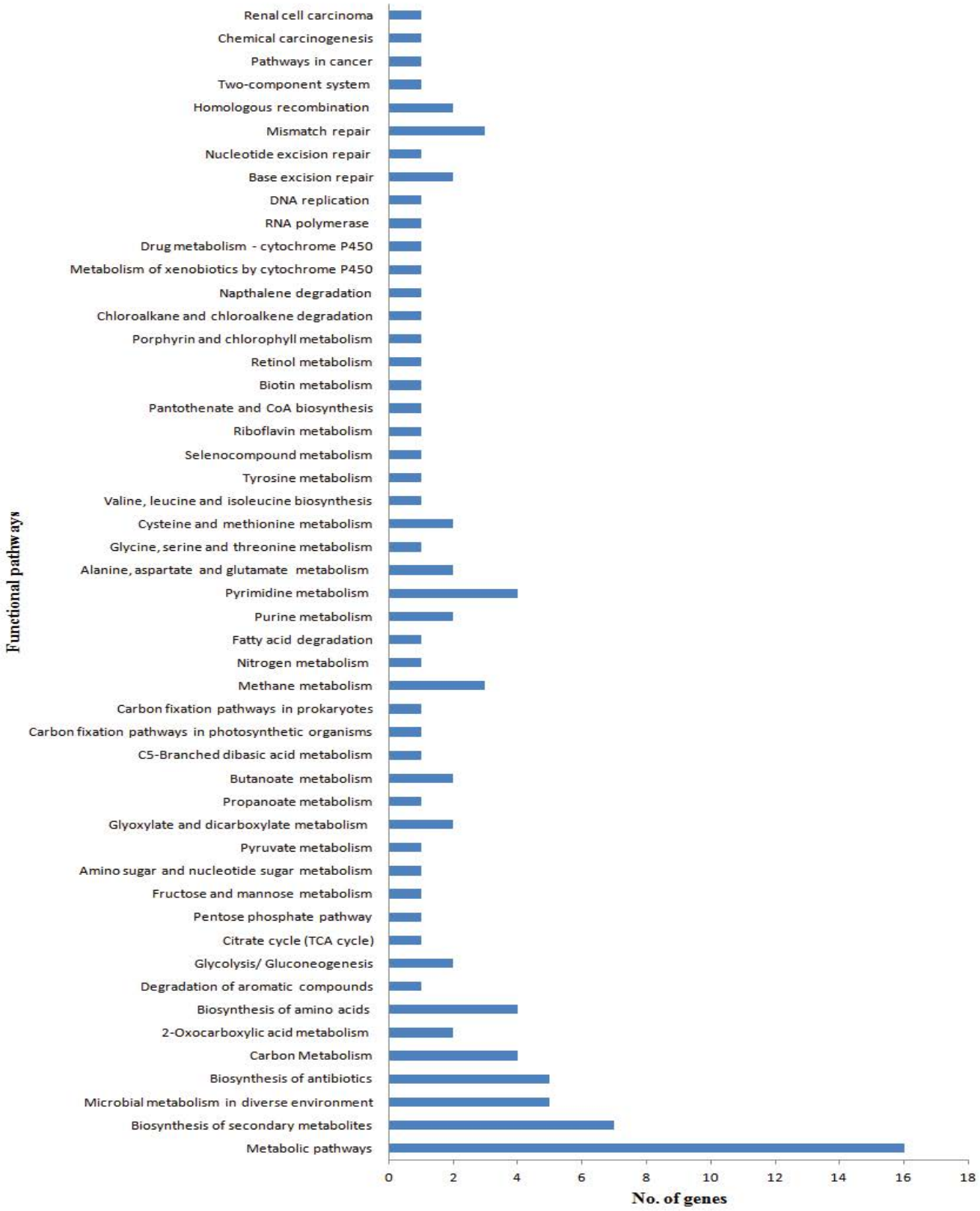

Figure 9 (E): Functional classification of the genes encoded by genome of CNS strain G22B2. These functions were assigned using the KAAS (KEGG automatic annotation server).

traits showed uniqueness among the CNS strains that were absent in the reference genome of RF122. These included arsenical resistance operon repressor, cadmium resistance protein, cadmium efflux system accessory protein and phage major tail protein etc. The acquisition of these new resistance genes in the CNS strains differentiates them from the other CNS species and clearly demonstrates their greater pathogenic potential. The genomes of all five CNS strains could be ordered on the basis of pathogenic potential starting from most pathogenic to least pathogenic; *S. equorum* subsp. *equorum* strain G8HB1, *S. cohnii* subsp. *cohnii* strain G22B2, *S. pasteuri* strain BAB3, *S. warneri* strain 1DB1 and *S. haemolyticus* strain 1HT3.

In conclusion, this study is a first attempt to map and understand the virulence profiles of different CNS species isolated from distinct human body organs. WGS is the most appropriate tool to analyse the pathogenic potential of CNS. Our analysis has provided new insights into the genome of CNS and offers a potential in developing therapies against emerging CNS pathogens.

Acknowledgments

We are grateful to Council of Scientific and Industrial Research, New Delhi, India for funding network project Man as a Super organism: Understanding the Human Microbiome (HUM) and Management of Infectious Diseases by Immunomodulation (BSC0119). RGN, GK, and SKM are thankful to Council of Scientific and Industrial Research for providing fellowships. IK and NKS are supported by a University Grant Commission (UGC) fellowship. We are thankful to the patients for participating in the study and allowing us to collect the tissues during their surgical intervention. This is IMTECH communication number **072/2015**.

Ethical Clearance

The study was ethically approved by the Institutional Ethics Committee of the Postgraduate Institute of Medical Education and Research, (Ref PGI/IEC/2013/1674) and Institutional Biosafety Committee of the CSIR-Institute of Microbial Technology (Ref/IBSC/2012-2/09), Chandigarh, India. Those samples were collected that were resected from the patients during the surgical intervention.

GenBank Accession Numbers

The genome sequence of strain(s) 1DB1, G22B2, 1HT3, G8HB1 and BAB3 were deposited in GenBank under the accession numbers LAKH00000000, LAKJ00000000, LAKG00000000, LAKE00000000 and LAKF00000000 respectively.

References

1. Rosenbach FJ, Bergmann JF (1884) Microorganismen bei den Wund-Infections-Krankheiten des Menschen. Wiesbaden 1-122.
2. Schleifer KH, Kloos WE (1975) Isolation and Characterization of Staphylococci from Human Skin. Int J Syst Bacteriol 25: 50-61.
3. Gould IM (2010) VRSA-doomsday superbug or damp squib? Lancet Infect Dis 10: 816-818.
4. Cuny C, Friedrich A, Kozytska S, Layer F, Nubel U, et al. (2010) Emergence of methicillin-resistant Staphylococcus aureus (MRSA) in different animal species. Int J Med Microbiol 300: 109-117.
5. Grundmann H, Aanensen DM, Wijngaard VD, Spratt CC, Harmsen BG, et al. (2010) Geographic distribution of *Staphylococcus aureus* causing invasive infections in Europe:a molecular-epidemiological analysis. PLoS Med 7: e1000215.
6. Plata K, Rosato AE, Wegrzyn G (2009) *Staphylococcus aureus* as an infectious agent: overview of biochemistry and molecular genetics of its pathogenicity. Acta Biochim Pol 56: 597-612.
7. Van Loo I, Huijsdens X, Tiemersma E, de Neeling A, van de Sande-Bruinsma, et al. (2007) Emergence of methicillin-resistant *Staphylococcus aureus* of animal origin in humans. Emerg Infect Dis 13: 1834-9.
8. Venkatesh MP, Placencia F, Weisman LE (2006) Coagulase negative staphylococcal infections in the neonate and child: an update. Semin Pediatr Infect Dis 17: 120-127.
9. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, et al. (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in new-borns. Proc Natl Acad Sci USA 107: 11971-11975.
10. Bohach GA, Fast DJ, Nelson RD, Schlievert PM (1990) Staphylococcal and streptococcal pyrogenic toxins involved in toxic shock syndrome and related illnesses. Crit Rev Microbiol 17: 251-272.
11. Fry DE, Barie PS (2011) The changing face of *Staphylococcus aureus*: a continuing surgical challenge. Surg Infect Larchmt 12: 191-203.
12. Lowyv FD (1998) *Staphylococcus aureus* infections. N Engl J Med 339: 520-32.
13. Nickerson EK, West TE, Day NP, Peacock S (2009) *Staphylococcus aureus* disease and drug resistance in resource-limited countries in South and East Asia. Lancet Infect Dis 9: 130-135.
14. Huebner J, Goldmann DA (1999) Coagulase-negative staphylococci: role as pathogens. Annu Rev Med 50: 223-236.
15. Llewelyn M, Cohen Jet (2002) Superantigens: microbial agents that corrupt immunity. Lancet Infect Dis 2: 156-162.
16. Zong Z, Peng C, Lu X (2011) Diversity of SCCmec elements in methicillin-resistant coagulase-negative Staphylococci clinical isolates. PLoS One 6: e20191.
17. Anderson-Berry A, Brinton B, Lyden E, Faix RG (2011) Risk factors associated with development of persistent coagulase negative staphylococci bacteremia in the neonate and associated short-term and discharge morbidities. Neonatology 99: 23-31.
18. Balaban N, Rasooly A (2000) Staphylococcal enterotoxins. Int J Food Microbiol 61: 1-10.
19. de Kraker ME, Davey PG, Grundmann H (2011) Mortality and hospital stay associated with resistant *Staphylococcus aureus* and *Escherichia coli* bacteremia: estimating the burden of antibiotic resistance in Europe. PLoS Med 8: e1001104.
20. Baptiste NJ, Benjamin DK, Wolkowiez CM, Fowler VG, Laughon M, et al. (2011) Coagulase negative staphylococcal infections in the neonatal intensive care unit. Infect Control Hosp Epidemiol 32: 679-686.
21. Marra AR, Camargo LF, Pignatari AC, Sukiennik T, Behar PR, et al. (2011) Nosocomial bloodstream infections in Brazilian hospitals: analysis of 2,563 cases from a prospective nationwide surveillance study. J Clin Microbiol 49: 1866-1871.
22. Mayilraj S, Saha P, Suresh K, Saini HS (2006) *Ornithinimicrobium kibberense* sp. nov., isolated from the Indian Himalayas. Int J Syst Evol Microbiol 56: 1657-1661.
23. Kim O, Cho YJ, Lee K, Yoon SH, Kim M, et al. (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA Gene sequence database with phylotypes that represent uncultured species. Int J Syst Evol Microbiol 62: 716–721.
24. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30: 2725-2729.
25. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.
26. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, et al. (2015) RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batch of genomes. Sci Rep 10: 8365.
27. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, et al. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42: D206–D214.
28. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100-3108.
29. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucl Acids Res 32: 11-16.
30. Zhou Y, Liang Y, Lynch K, Dennis JJ, Wishart DS (2011) PHAST: A Fast Phage Search Tool. Nucl Acids Res 39: W347-W352.
31. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34: D32–D36.
32. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12: 402.

33. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research 14: 1394-1403.

34. Darling AE, Mau B, Perna NT (2010) ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5: e11147.

35. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35: W52-W57.

36. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6:e1001107.

37. Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A 48: 582–592.

38. Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. EMBO Rep 6: 1208-1213.

39. Rocha EP, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18: 291-294.

40. Naya H, Romero H, Zavala A, Alvarez B, Musto H (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC %) in prokaryotes. J Mol Evol 55: 260-264.

41. McEwan CE, Gatherer D, McEwan NR (1998) Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas 128: 173-178.

42. Morschhauser J, Kohler G, Ziebuhr W, Oehler BG, Dobrindt U, et al. (2000) Evolution of microbial pathogens. Philosophical Transactions of the Royal Society B:Biological Sciences 355: 695–704.

43. Novick RP, Christie GE, Penadés JR (2010) The phage-related chromosomal islands of Gram-positive bacteria. Nat Rev Microbiol 8: 541–551.

44. Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H (2003) Prophage Genomics. Microbiol Mol Biol Rev 67:238–276.

45. Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64: 475–493.

46. Karginov FV, Hannon GJ (2010) The CRISPR system: small RNA-guided defense in bacteria and archea. Mol Cell 37: 7–19.

47. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archea. Nat Rev Microbiol 6: 181–186.

48. Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35: W182-W185.

49. Alexander Y, Mitrophanov Eduardo A, Groisman (2008) Signal integration in bacterial two-component regulatory systems. Genes Dev 22: 2601–2611.