Synonymous codon usage in forty staphylococcal phages identifies the factors controlling codon usage variation and the phages suitable for phage therapy

Arghya Kamal Bishal$^\$$, Soumya Saha$^\$$ & Keya Sau*

Department of Biotechnology, Haldia Institute of Technology, ICARE Complex, HIT Campus, P O- HIT, Haldia, Dist - Purba Medinipur-721657, WB, India; Keya Sau – Email: keya_sau@yahoo.co.in; *Corresponding author
$- Authors contributed equally to work

Abstract:
The immergence and dissemination of multidrug-resistant strains of *Staphylococcus aureus* in recent years have expedited the research on the discovery of novel anti-staphylococcal agents promptly. Bacteriophages have long been showing tremendous potentialities in curing the infections caused by various pathogenic bacteria including *S. aureus*. Thus far, only a few virulent bacteriophages, which do not carry any toxin-encoding gene but are capable of eradicating staphylococcal infections, were reported. Based on the codon usage analysis of sixteen *S. aureus* phages, previously three phages were suggested to be useful as the anti-staphylococcal agents. To search for additional *S. aureus* phages suitable for phage therapy, relative synonymous codon usage bias has been investigated in the protein-coding genes of forty new staphylococcal phages. All phages appeared to carry A and T ending codons. Several factors such as mutational pressure, translational selection and gene length seemed to be responsible for the codon usage variation in the phages. Codon usage indeed varied phage to phage. Of the phages, phages G1, Twort, 66 and Sap-2 may be extremely lytic in nature as majority of their genes possess high translational efficiency, indicating that these phages may be employed in curing staphylococcal infections.

Keywords: Staphylococcal phage, Synonymous codon usage, Translational selection, Mutational bias, Phage therapy.

Background:

Staphylococcus aureus, a gram-positive bacterium, causes various mild as well as life-threatening diseases to human and other animals [1]. Numerous virulence factors are synthesized in this organism by the co-ordinated action of multiple regulators. To exterminate the *S. aureus* infections, several antibiotics (that block the synthesis of cell wall, DNA, RNA, and protein) had been administered from 1943. Sadly, *S. aureus* strains with the resistance to multiple antibiotics were evolved gradually worldwide. Currently, vancomycin, linezolid, daptomycin, and tigecycline are being used in clinics to eradicate the multi-drug resistant *S. aureus* strains [2]. These agents are, however, either not completely safe for use or associated with the development of staphylococcal resistance. The *S. aureus* vaccines, so far developed by different groups, failed in the clinical trials though they yielded promising results in the animal models [3]. Bacterial viruses or bacteriophages (or simply phages) that usually kill the host bacteria have been presumed to substitute antibiotics as these agents showed promising results in the treatment of various bacterial infections including staphylococcal infections [4]. Over 250 staphylococcal phages have been reported so far [5]. Of the staphylococcal phages, relatives of phages K, G1, and Twort exhibited therapeutic potentials [6]. All these phages belong to *Myoviriade* family and

are virulent in nature. Therapeutic potentials of other staphylococcal phages, particularly those belonging to the families *Podoviridae* and *Siphoviridae* have not been demonstrated yet [7].

Synonymous codon usage biases had been determined in numerous living organisms and in viruses [8-12]. The results unanimously demonstrated that codon usage is species-specific and even varies gene to gene of the same species. Codon usage is usually positively co-related with the GC base composition of the genome. Several other factors such as mutational pressure, translational selection, secondary structure of protein, replication-translation selection, environmental factor, etc. were reported to influence the codon usage bias. Codon usage analysis has enriched both the basic and applied biological sciences in number of ways. It has assisted in understanding the expression levels of the genes, provided clues about the evolution of the genes and genomes, maximized protein expression in the heterologous system, enhanced immunogenicity of vaccines, etc. [13-15]. In addition, codon usage analysis also identified virulent phages having prospects in therapy [16-18]. The complete genome sequences of the fifty one staphylococcal phages (including phages K, G1 and Twort) are available in different viral databases (as of December 2011). Previously, codon usage biases were investigated in the twelve of these phage genomes [16]. The results suggested that phages K, 44AHJD, and P68 are extremely virulent in nature and may be useful in therapy. Currently, very little is known about the codon usage bias and the lytic ability of the rest thirty nine phages. In the present study, we performed codon usage analysis in the genes of forty staphylococcal phages (including phage K) at length. Our results revealed that phages G1, Twort, 66 and Sap-2 are extremely virulent in nature, indicating that these phages could be used in curing staphylococcal infections.

Methodology:
Complete genome sequences of forty staphylococcal phages were downloaded from different databases [available in NCBI (USA) and EMBL (UK) web pages]. The forty phages whose genomes were downloaded are 187, 2638A, 29, 47, 52A, 53, 69, 71, 85, 88, 92, 96, X2, 37, 3A, 42e, 55, 66, 80alpha, CNPH82, EW, G1, K, P954, PH15, PT1028, ROSA, SAP-2, SAP-26, Twort, phiETA2, phiETA3, phiMR11, phiMR25, phiNM1, phiNM3, phiPVL-CN125, phiPVL108, phiSauS-IPLA35 and phiSauS-IPLA88. Phage K was included in the study in order to compare its codon usage bias with those of other staphyloccal phages, particularly, phages G1 and Twort. The basic characteristics of all forty phages were presented in **Table 1 (see supplementary material)**. While phages PH15 and CNPH82 infect *S. epidermidis*, other phages infect *S. aureus*. Except phages P954, PT1028, ROSA, and phiPVL-CN125, all other phages are distributed across the three major families, namely, *Myoviridae*, *Podoviridae* and *Siphoviridae*. Gene number in the 40 phages varies from 20 to 214. Genome sequence of *S. aureus* USA300 [19] were also downloaded for a comparative study. Coding sequences of all forty phages were extracted by a program in in EMBOSS 6.4 [20] designated 'Coderet'. Coding sequences carrying less than 50 codons were not considered in this study. Hence a total of 2722 protein coding genes from 40 phages were utilized for all analysis. All coding sequences of *S. aureus* USA300 were also extracted by similar manner. The relative synonymous codon usage (RSCU) in all forty phages was

determined to understand the overall codon usage variation among the genes and genomes. RSCU is defined as the ratio of the observed frequency of codons to the anticipated frequency if all the synonymous codons for those amino acids are utilized uniformly [21]. The RSCU values greater than 1.0 denote that the corresponding codons are more often used than expected, whereas, the opposite is true for the RSCU values less than 1.0. G_{3s}, A_{3s}, T_{3s}, and C_{3s} are the frequencies of G, A, T, and C and GC_{3s} is the frequency of (G+C) at the synonymous third positions of codons. N_c, the effective number of codons used by a gene, is usually used to measure the bias of synonymous codons and independent of amino acid compositions and codon number [22]. The values of N_c vary from 20 (when one codon is used per amino acid) to 61 (when all the codons are used equally). The highly-biased genes are generally expressed highly [23]. As gene expression levels of the phages are not known with certainty, we have considered the highly-biased genes as the highly expressed in case of staphylococcal phages under study. The correspondence analysis [9] and the determination of all the parameters (mentioned above) were performed by the program CodonW 1.4 [24]. The copy number of tRNA species in *S. aureus* USA300 and corresponding anticodon sequences were determined by a program, namely, tRNAscan-SE [25]. The *Sau*3AI cut site in the genomic DNA was determined by 'Webcutter 2.0' program [26].
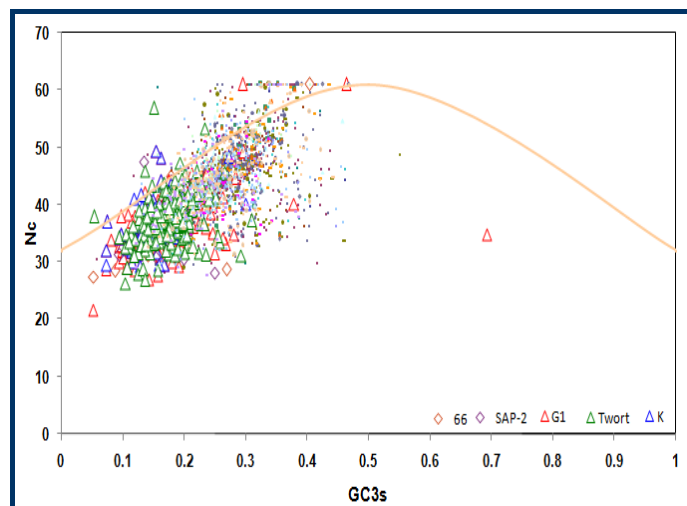

Figure 1: Nc plot of 40 *S. aureus* phage genes. The genes are represented as different colored dots or as indicated colored shapes.

Discussion and Conclusion:
Synonymous codon usage variation in 40 staphylococcal phages
To study the codon usage variation in the forty staphylococcal phages, the overall RSCU values in 2722 protein-coding genes of these phages were determined (data not presented). As expected from the AT-rich phage genomes, A and T ending codons appeared to be predominant in all the phages (data not shown). To determine the possible codon usage variation in the above phage genes, both N_c and GC_{3s} were determined and found to vary from 36.52 to 48.44 (with a mean of 44.39 and standard deviation 3.061) and from 0.159 to 0.29 (mean 0.244 and standard deviation 0.033), respectively. Such results are indicative of the marked variations in synonymous codon usage in the genes of all 40 staphylococcal phages. Several factors

including mutational bias seem to be responsible for the codon usage alternation in the staphylococcal phages.

Effect of mutational pressure on the codon usages variations
To determine the factors responsible for the codon usage variation in different organisms, both N_c plot (a plot of N_c versus GC_{3s}) and correspondence analysis (CA) were employed extensively [11]. Genes were suggested to fall on the continuous curve between N_c and GC_{3s}, if GC_{3s} were the only determinant of N_c. Secondly, comparison of the actual distribution of genes with the expected distribution under no selection could indicate whether codon usage bias is influenced by the factors other than mutational bias [22]. The N_c plot reveals that a small fraction of 40 staphylococcal phage genes lie on the expected curve towards GC poor regions **(Figure 1)**, suggesting their origination from the extreme mutational bias. In contrast, majority of the genes with low N_c values lie well below the expected curve **(Figure 1)**, indicating that these phage genes possess an additional codon usage bias which does not depend on GC_{3s}. Interestingly, genes of phages K, G1 and Twort are mostly overlapped, lie far away from the expected curve and tend to be separated from the rest phage genes. The data suggest that the effect of mutational bias on codon usage variation in the phages G1, K, and Twort is very weak.
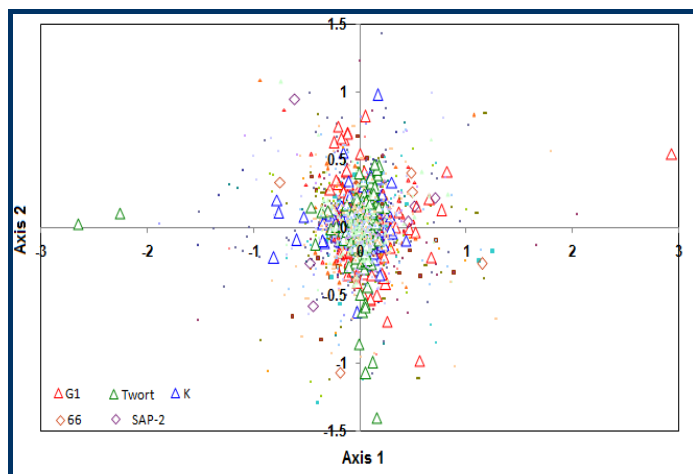

Figure 2: Positions of the genes of 40 staphylococcal phages along the two major axes of variation in the correspondence analysis on RSCU values. The genes are presented by the same symbols as in Figure 1.

To verify the suggestions made from the N_c plot, correspondence analysis of the RSCU values of the 2722 protein-coding genes of the above 40 phages were performed as described in Methodology. The distributions of the genes along the first two major axes were shown only as they accounted for 6.64% and 5.44% of the total variation, respectively **(Figure 2)**. An analysis between the positions of genes along first as well as second major axis (generated by CA) with the nucleotide composition at the third codon position revealed that the first major axis is negatively correlated with A_{3s} and T_{3s} but positively correlated with G_{3s} and C_{3s} **Table 2 (see supplementary materials)**. Interestingly, second major axis does not show any correlation with A_{3s}, T_{3s}, G_{3s} and C_{3s}. The results suggest that A and T ending codons are clustered on the negative side, whereas, G and C ending codons are located on the positive side of first major axis. To determine the differences

between these two clusters of genes, codon usage of 10% genes located at the extreme right side of axis 1 was compared with the 10% of genes located at the extreme left side of axis 1. As evident from **Table 3 (see supplementary material)**, there are 24 over-represented codons in the genes located at the positive side of the first major axis. Out of 24 predominant codons, there are 10 A ,13 T ending codons and 1 C ending codon, which actually represent 41.67% A, 54.17% T, and 4.17% C ending codons. The results suggest that genomic GC composition has a powerful effect in separating the genes along first major axis according to their RSCU values. The data from correspondence analysis therefore supports the data from of the N_c plot and together they indicate that mutational bias plays a major role in shaping the codon usage variation in the genes of 40 staphylococcal phages.
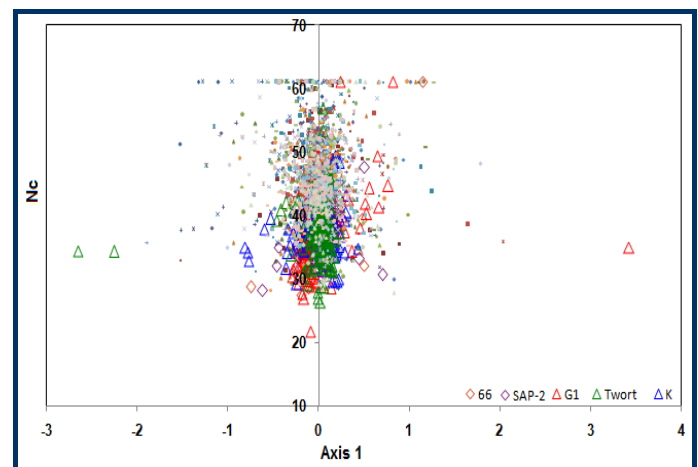

Figure 3: Scatter plot of *S. aureus* phages and Nc values. The genes are presented by the same symbol as in Figure 1.

Effects of translational selection and gene length on the codon usage variation
To ascertain whether translational selection or gene length also influence synonymous codon usage bias of the forty staphylococcal phages, correlation coefficients were determined between the positions of the genes along the first major axes with both N_c and gene length **Table 2 (see supplementary material)**. Apparently, only first major axis is positively correlated with N_c. A scatter plot **(Figure 3)** between the position of the genes along the first major axis and N_c values also revealed that several genes of the phages SAP-2, 66, K, G1 and Twort possess lower N_c values and seemed to be separated from the other phage genes. The phage genes with lower N_c values therefore carry comparatively AT rich codons. The data together suggest that a balance between mutational bias and translational efficiency is strongly contributing to the codon usage variation in the genes of staphylococcal phages. Besides, phages SAP-2, 66, K, G1, and Twort are seemed to be composed of the highly expressed genes primarily. Synonymous codon usages in the highly expressed genes are usually influenced by the cellular tRNA abundance [16, 27-29]. In most organisms, cellular tRNA abundance was found to be directly proportional to the corresponding tRNA copy number. To see whether there is any positive correlation between host tRNA abundance and synonymous codon usage of the phages G1, K, Twort, Sap-2 and 66, a comparative analysis was made between the tRNA copy numbers of *S. aureus* USA300 and the over-represented

synonymous codons of these phages separately **Table 4 (see supplementary material)**. Of the 18 abundant synonymous codons of phage Sap-2 or 66, 10 codons were found to be paired by those USA300-specific tRNAs whose copy numbers are greater than two or more per cell. Considering the standard base pairing rules of codon-anticodon, 7 additional over-represented codons of sap-2 or 66 appeared to be interacted by the abundant host tRNAs **Table 4 (see supplementary material)**. In contrast, 13 out of 18 over-represented codons of G1 or Twort or K were recognized by the over-represented tRNAs in USA300. Most of the genes of phages G1, Twort, K, 66 and Sap-2 therefore seemed to possess a high translational efficiency. Further analysis also shows that first axis in CA is negatively correlated with the gene length of the protein coding genes of the 40 phages **Table 2 (see supplementary material)**. This indicates that gene length influences codon usage variation to some extent in the above phages.
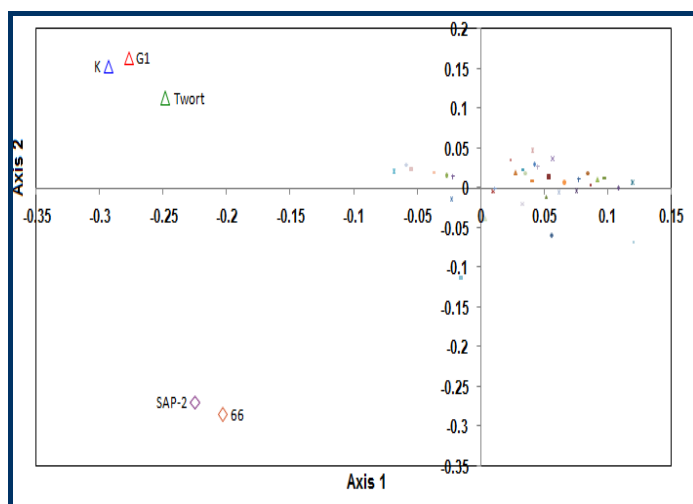

Figure 4: Position of the genes along the first two major axes produced by correspondence analysis on RSCU values of 40 staphylococcal phage genomes. The genomes are represented as different colored dots or as indicated colored shapes.

Synonymous codon usages of 40 staphylococcal phages are distinct

The data presented above clearly show that the synonymous codon usage of phages G1, K, Twort, 66 and Sap-2 are different from other staphylococcal phages being studied here **(Figures 1-3)**. To determine whether synonymous codon usages of all the 40 staphylococcal phages differ from each other, correspondence analysis was also carried out on RSCU values on genes grouped according to their corresponding genomes. The data reveal that the positions of the genes along the first major axis clearly separated the genomes according to their genomic AT composition **(Figure 4)**. Additional statistical analysis (using *F* statistics) on the RSCU values of all 18 amino acids, which are coded by two or more synonymous codons of the 40 staphylococcal phages, revealed that usages of each 59 synonymous codons differed significantly from genome to genome. The data together indicate that codon usages of the staphylococcal phages vary phage to phage. To study the variation in the codon usage of the 40 staphylococcal phages more accurately, a cluster analysis was performed on overall codon usage data of these phages by D-squared (D^2) statistic approach **[30]**. D^2 statistic values were determined by the

following equation: D^2 = sum of 64 codons of: (frequency $_{(codon, Table 1)}$ −frequency $_{(codon Table 2)})^2$; where frequency indicates the relative synonymous codon usage values. A lower D^2 value indicates a strong resemblance in the codon usage. A matrix containing the D^2 value of each set was undertaken to generate clusters. The clustering (dendrogram) produced by UPGMA (unweighted pair group method using arithmetic averages) procedure **[9]** shows that phages belonging to the *Myoviridae* (such as G1, K, and Twort) and *Podoviridae* (e.g., Sap-2 and 66) families are clustered in two distinct branches, whereas, *Siphoviruses* and unclassified viruses are clustered in different other branches **(Figure 5)**. Such distribution of the phages not only supports the correspondence analysis data **(Figure 4)** but also reveals that the synonymous codon usage is dissimilar even among the phages of each branch.
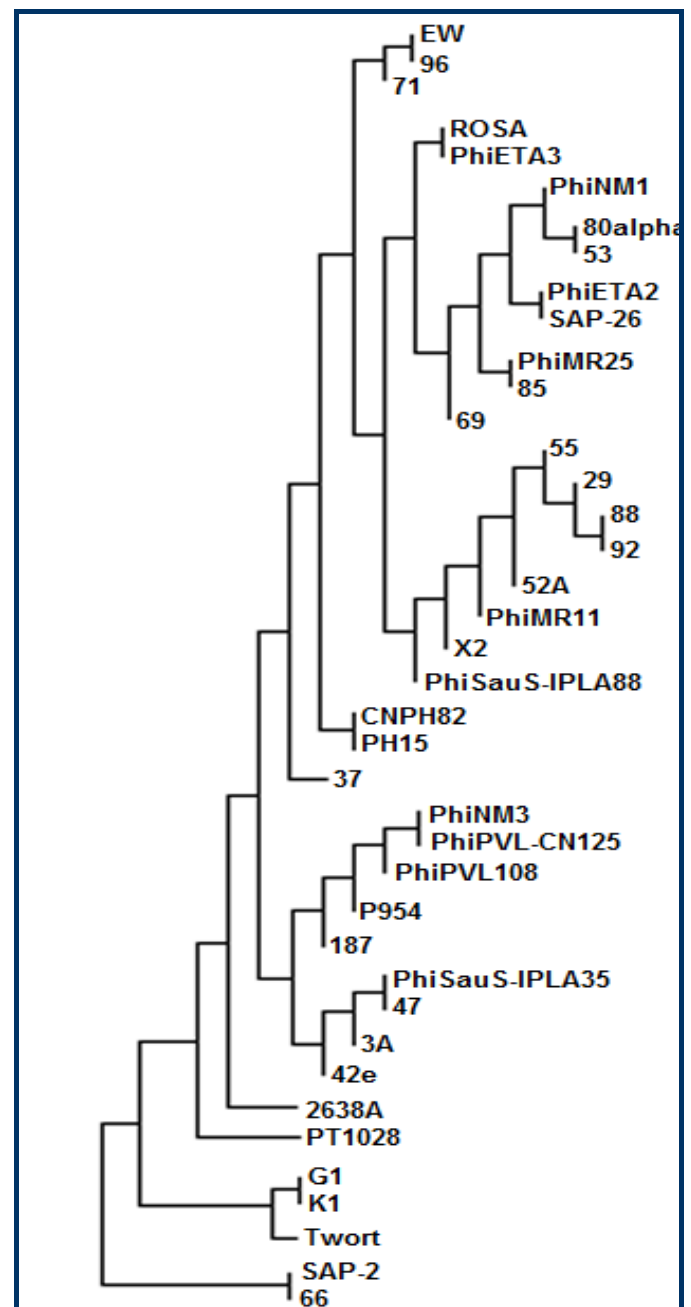

Figure 5: A dendrogram showing the codon usage variation of 40 staphylococcal phages constructed by the UPGMA method.

BIOINFORMATION

Identification of the staphylococcal phages possessing therapeutic potentials

To identify the staphylococcal phages suitable for therapy, we carried out an additional analysis of the data presented in (**Figure 3**) and in **Table 4 (see supplementary material)**. It appears that the N_c values of the genes of phages G1, K, Twort, Sap-2 and 66 mostly vary from ~20 - 40. Secondly, phages G1, K, and Twort harbor nearly 72% highly expressed genes, whereas, phages 66 and Sap-2 carry about 94% highly expressed genes. The results together suggest that the genes of the above five phages would be expressed rapidly by the translational machinery of the pathogenic *S. aureus* strain USA300. Based on such translational efficiency, two *podoviruses* (66 and Sap-2) seem to possess higher lytic ability than those of the three *myoviruses* (G1, K, and Twort). Interestingly, all *podoviruses* or *myoviruses* here might be similarly lytic in nature. These five phages carry no or very few GATC sites and also do not harbor any pathogenic factor-encoding / antibiotic-resistance gene. The results together indicate that all the five phages may be inducted in clinics for curing staphylococcal infections. Phages similar to G1, K, and Twort were already shown to have medicinal properties **[5, 6]**. Very little have been done yet to test the efficacy of the podoviruses as the remedies of staphylococcal diseases.

Acknowledgement:

References:

[1] Somerville GA & Proctor RA, *Microbiol Mol Biol Rev.* 2009 **73**: 233 [PMID: 19487727]
[2] Stryjewski ME & Corey GR, *Curr Opin Crit Care.* 2009 **15**: 403 [PMID: 19561492]
[3] Otto M, *Expert Opin Biol Ther.* 2010 **10**: 1049 [PMID: 20528609]
[4] Rafee Y *et al. BMC Infect Dis.* 2012 **12**: 45 [PMID: 22348549]
[5] Kwiatek M *et al. Arch Virol.* 2012 **157**: 225 [PMID: 22045271]
[6] Abedon ST *et al. Bacteriophage.* 2011 **1**: 66 [PMID: 22334863]
[7] O'Flaherty S *et al. J Bacteriol.* 2004 **186**: 2862 [PMID: 15090528]
[8] Anhlan D *et al. RNA.* 2011 **17**: 64 [PMID: 21068184]
[9] Hershberg R & Petrov DA, *Annu Rev Genet.* 2008 **42**: 287 [PMID: 18983258]
[10] Liu XS *et al. Virol J.* 2012 **9**: 68 [PMID: 22416942]
[11] Sharp PM *et al. Philos Trans R Soc Lond B Biol Sci.* 2010 **365**: 1203 [PMID: 20308095]
[12] Sau K & Deb A, *In Silico Biol.* 2009 **9**: 1 [PMID: 19537157]
[13] Crameri A *et al. Nat Biotechnol.* 1996 **14**: 315 [PMID: 9630892]
[14] Deml L *et al. J Virol.* 2001 **75**: 10991 [PMID: 11602739]
[15] Geddie ML *et al. J Biol Chem.* 2004 **279**: 26462 [PMID: 15069062]
[16] Sau K *et al. Virus Res.* 2005 **113**: 123 [PMID: 15970346]
[17] Sau K, *J Microbiol Immunol Infect.* 2007 **40**: 24 [PMID: 17332903]
[18] Ranjan A *et al. In Silico Biol.* 2007 **7**: 423 [PMID: 18391235]
[19] Diep BA *et al. Lancet.* 2006 **367**: 731 [PMID: 16517273]
[20] http://www.ebi.ac.uk/Tools/emboss
[21] Sharp PM & Li W, *Nucleic Acids Res.* 1987 **15**: 1281 [PMID: 3547335]
[22] Wright F, *Gene.* 1990 **87**: 23 [PMID: 2110097]
[23] Sharp PM & Cowe E, *Yeast.* 1991 **7:** 657 [PMID: 1776357]
[24] http://www.codonw.sourceforge.net
[25] http://www.lowelab.ucsc.edu/tRNAscan-SE
[26] http://www.rna.lundberg.gu.se/cutter2
[27] Grantham R *et al. Nucleic Acids Res.* 1981 **9**: r43 [PMID: 7208352]
[28] Ikemura T, *Mol Biol Evol.* 1985 **2**: 13 [PMID: 3916708]
[29] Kanaya S *et al. J Mol Evol.* 2001 **53**: 290 [PMID: 11675589]
[30] Pouwels PH & Leunissen JA, *Nucleic Acids Res.* 1994 **22**: 929 [PMID: 8152923]

BIOINFORMATION

Supplementary material:

Table 1: Basic characteristics of the staphylococcal phages used in the study

Phage Name	Family	Host	Accession No.	Nucleotide (bp)	No of genes	%GC
187	*Siphoviridae*	*S. aureus*	NC_007047	39620	77	34.3
2638A	*Siphoviridae*	*S. aureus*	NC_007051	41318	57	37.2
29	*Siphoviridae*	*S. aureus*	NC_007061	42802	67	35.7
47	*Siphoviridae*	*S. aureus*	NC_007054	44777	65	33.5
52A	*Siphoviridae*	*S. aureus*	NC_007062	41690	60	35.7
53	*Siphoviridae*	*S. aureus*	NC_007049	43883	74	34.2
69	*Siphoviridae*	*S. aureus*	NC_007048	42732	69	34.4
71	*Siphoviridae*	*S. aureus*	NC_007059	43114	67	35.5
85	*Siphoviridae*	*S. aureus*	NC_007050	44283	71	34.6
88	*Siphoviridae*	*S. aureus*	NC_007063	43231	66	35.7
92	*Siphoviridae*	*S. aureus*	NC_007064	42431	64	35.8
96	*Siphoviridae*	*S. aureus*	NC_007057	43576	74	35.2
X2	*Siphoviridae*	*S. aureus*	NC_007065	43440	77	36
37	*Siphoviridae*	*S. aureus*	NC_007055	43681	70	35.3
3A	*Siphoviridae*	*S. aureus*	NC_007053	43095	67	33.6
42e	*Siphoviridae*	*S. aureus*	NC_007052	45861	79	34
55	*Siphoviridae*	*S. aureus*	NC_007060	41902	77	36.1
66	*Podoviridae*	*S. aureus*	NC_007046	18199	27	29.9
80alpha	*Siphoviridae*	*S. aureus*	NC_009526	43864	73	34.2
CNPH82	*Siphoviridae*	*S. epidermidis*	NC_008722	43420	65	34.9
EW	*Siphoviridae*	*S. aureus*	NC_007056	45286	77	36.6
G1	*Myoviridae*	*S. aureus*	NC_007066	138715	214	30.9
K	*Myoviridae*	*S. aureus*	NC_005880	127395	115	31.2
P954	Unclassified	*S. aureus*	NC_013195	40761	69	34.2
PH15	*Siphoviridae*	*S. epidermidis*	NC_008723	44041	68	35
PT1028	Unclassified	*S. aureus*	NC_007045	15603	22	32.1
ROSA	Unclassified	*S. aureus*	NC_007058	43155	74	35.1
SAP-2	*Podoviridae*	*S. aureus*	NC_009875	17938	20	29.6
SAP-26	*Siphoviridae*	*S. aureus*	NC_014460	41207	63	34.2
Twort	*Myoviridae*	*S. aureus*	NC_007021	130706	195	30.7
PhiETA2	*Siphoviridae*	*S. aureus*	NC_008798	43265	69	34.5
PhiETA3	*Siphoviridae*	*S. aureus*	NC_008799	43282	68	35.1
PhiMR11	*Siphoviridae*	*S. aureus*	NC_010147	43011	67	35.8
PhiMR25	*Siphoviridae*	*S. aureus*	NC_010808	44342	70	34.6
PhiNM1	*Siphoviridae*	*S. aureus*	NC_008583	43128	64	34.3
PhiNM3	*Siphoviridae*	*S. aureus*	NC_008617	44061	65	33.6
PhiPVL-CN125	Unclassified	*S. aureus*	NC_012784	44492	65	33.5
PhiPVL108	*Siphoviridae*	*S. aureus*	NC_008689	44857	59	33.9
PhiSauS-IPLA35	*Siphoviridae*	*S. aureus*	NC_011612	45344	62	33.4
PhiSauS-IPLA88	*Siphoviridae*	*S. aureus*	NC_011614	42526	60	35.2

Table 2: Correlation coefficients between the positions of the genes along the first two major axes with both Nc and gene length

Axis	A3s	T3s	G3s	C3s	GC3s	Nc	Gene length bp)
First	-0.872**	-0.958**	**0.941****	0.899**	0.973**	0.982**	-0.554**
Second	-0.113	-0.073	0.125	-0.116	-0.002	0.097	0.718**

**Correlation is significant at the 0.01 level (2-tailed)

Table 3: Relative synonymous codon usage (RSCU) values of each codon for the two groups of genes located at the extreme ends of the first major axis as determined by CA.

AA	Codon	RSCUᵃ	Nᵃ	RSCUᵇ	Nᵇ	AA	Codon	RSCUᵃ	Nᵃ	RSCUᵇ	Nᵇ
Leu	UUA*	3.74	5935	2.87	1962	Pro	CCU*	2.58	2133	1.57	574
	UUG	0.35	562	1.1	751		CCC	0.15	122	0.23	86
	CUU	0.84	1332	0.8	548		CCA	1.09	903	1.38	504
	CUC	0.08	132	0.22	149		CCG	0.18	153	0.82	300
	CUA*	0.9	1431	0.79	537	Gly	GGU*	2.17	3652	1.69	1311
	CUG	0.09	136	0.22	152		GGC	0.19	322	0.96	743
Arg	CGU	1.24	868	1.46	551		GGA*	1.3	2182	1	775
	CGC	0.1	67	0.54	203		GGG	0.34	579	0.35	272
	CGA	0.37	258	0.77	290						
	CGG	0.05	34	0.16	61	Ile	AUU*	1.64	4672	1.35	1718
	AGA*	3.68	2569	2.46	927		AUC	0.28	789	0.74	934
	AGG	0.57	396	0.61	228		AUA*	1.09	3111	0.91	1160
Ser	UCU*	1.84	2393	1.13	609	Tyr	UAU*	1.45	4396	1.32	1514

AA	Codon	RSCU	count	RSCU	count	AA	Codon	RSCU	count	RSCU	count
	UCC	0.22	291	0.18	96		UAC	0.55	1678	0.68	787
	UCA	1.63	2123	1.54	826	His	CAU*	1.55	1453	1.16	547
	UCG	0.09	119	0.52	277		CAC	0.45	426	0.84	393
	AGU	1.85	2408	1.71	919	Gln	CAA	1.65	3501	1.65	1822
	AGC	0.36	462	0.93	500		CAG	0.35	739	0.35	381
Val	GUU	1.51	2801	1.6	1324	Asn	AAU*	1.47	5956	1.14	1931
	GUC	0.14	268	0.46	385		AAC	0.53	2121	0.86	1467
	GUA*	2.11	3918	1.36	1125	Lys	AAA	1.54	8324	1.5	3624
	GUG	0.23	428	0.59	486		AAG	0.46	2496	0.5	1205
Ala	GCU*	1.86	2196	1.44	1132	Asp	GAU*	1.52	6121	1.27	2121
	GCC	0.23	268	0.25	200		GAC	0.48	1940	0.73	1211
	GCA*	1.77	2092	1.54	1211	Glu	GAA*	1.5	7134	1.45	2851
	GCG	0.14	170	0.77	609		GAG	0.5	2365	0.55	1070
Thr	ACU*	1.55	2735	1.28	1045	Cys	UGU*	1.69	578	1.34	181
	ACC	0.19	342	0.25	200		UGC	0.31	107	0.66	90
	ACA*	2.03	3578	1.84	1503	Phe	UUU*	1.51	3577	1.32	1492
	ACG	0.23	403	0.63	517		UUC	0.49	1149	0.68	768

Each group contains 10% of genes at the two extreme ends of the first major axis. The '*', 'N', 'a', 'b' and 'AA' indicate the codons whose occurrences are significantly (p<0.01) higher in the group of genes on left side than the genes on the right side of the first major axis, number of codons, genes on the left side, genes on the right side, and amino acid, respectively.

Table 4: Correlation between RSCU values of phages K, Twort, G1, 66 and Sap-2 S. aureus-specific tRNA copy number

AA	Codon	RSCU					tRNA copy no in USA300*	Anticodon*
		K	Twort	G1	66	Sap-2		
Phe	UUU	1.45	1.62	1.45	1.56	1.56		
	UUC[1]	0.55	0.38	0.55	0.44	0.44	2	GAA
Leu	UUA	3.79	3.66	3.69	4.42	4.41	2	UAA
	UUG	0.35	0.35	0.36	0.41	0.39	1	CAA
	CUU	0.84	0.86	0.86	0.54	0.52		
	CUC	0.07	0.09	0.07	0.13	0.16	1	GAG
	CUA	0.89	0.95	0.92	0.37	0.46	1	UAG
	CUG	0.06	0.09	0.11	0.13	0.06		
Ile	AUU	1.59	1.7	1.58	1.79	1.85		
	AUC	0.29	0.24	0.28	0.47	0.41	2	GAU
	AUA	1.13	1.06	1.14	0.73	0.73		
Met	AUG[2]	1	1	1	1	1	4	CAU
Val	GUU	1.6	1.37	1.57	1.44	1.56		
	GUC	0.14	0.13	0.16	0.38	0.19		
	GUA	2.08	2.23	2.08	1.54	1.59	3	UAC
	GUG	0.17	0.27	0.19	0.64	0.65		
Ser	UCU	1.84	2	1.79	0.91	1.1		
	UCC	0.2	0.29	0.2	0.15	0.08	1	GGA
	UCA	1.72	1.35	1.71	2.79	2.66	3	UGA
	UCG	0.07	0.12	0.07	0.2	0.19		
Pro	CCU	2.82	2.33	2.75	1.16	1.35		
	CCC	0.09	0.24	0.1	0.17	0.18		
	CCA	0.89	1.34	0.92	2.23	1.98	2	UGG
	CCG	0.19	0.09	0.24	0.43	0.48		
Thr	ACU	1.59	1.58	1.58	0.99	0.83		
	ACC	0.18	0.23	0.17	0.15	0.14		
	ACA	2.07	1.9	2.07	2.37	2.48	3	UGU
	ACG	0.16	0.29	0.18	0.49	0.56		
Ala	GCU	1.88	1.88	1.86	1.27	1.59		
	GCC	0.24	0.19	0.24	0.24	0.23		
	GCA	1.79	1.71	1.79	2.09	1.93	3	UGC
	GCG	0.09	0.22	0.11	0.4	0.25		
Tyr	UAU	1.46	1.42	1.46	1.52	1.55		
	UAC	0.54	0.58	0.54	0.48	0.45	2	GUA
His	CAU	1.57	1.51	1.55	1.6	1.61		
	CAC	0.43	0.49	0.45	0.4	0.39	2	GUG
Gln	CAA	1.64	1.66	1.63	1.81	1.84	2	UUG
	CAG	0.36	0.34	0.37	0.19	0.16		
Asn	AAU	1.52	1.44	1.51	1.32	1.31		
	AAC	0.48	0.56	0.49	0.68	0.69	3	GUU
Lys	AAA	1.55	1.5	1.54	1.78	1.77	3	UUU
	AAG	0.45	0.5	0.46	0.22	0.23		
Asp	GAU	1.55	1.48	1.55	1.39	1.35		
	GAC[1]	0.45	0.52	0.45	0.61	0.65	4	GUC

Glu	GAA	1.51	1.5	1.49	1.65	1.63	3	UUC
	GAG	0.49	0.5	0.51	0.35	0.37		
Cys	UGU	1.71	1.65	1.72	1.61	1.65		
	UGC	0.29	0.35	0.28	0.39	0.35	1	GCA
Trp	UGG[1]	1	1	1	1	1	1	CCA
Arg	CGU	1.15	1.37	0.99	3.06	3.06	2	ACG
	CGC	0.06	0.12	0.08	0.22	0.33		
	CGA	0.33	0.46	0.31	0.47	0.37		
	CGG	0.02	0.07	0.06	0.03	0.04	1	CCG
Ser	AGU	1.8	1.94	1.83	1.65	1.75		
	AGC	0.37	0.3	0.41	0.3	0.23	1	GCU
Arg	AGA	4	3.16	4.07	2.08	2.1	1	UCU
	AGG	0.43	0.82	0.49	0.13	0.11		
Gly	GGU	2.16	2.12	2.14	2.73	2.84		
	GGC	0.18	0.19	0.19	0.33	0.31	2	GCC
	GGA	1.37	1.21	1.38	0.77	0.61	4	UCC
	GGG	0.29	0.47	0.29	0.18	0.23		

*The tRNA copy numbers and anticodon in the *S. aureus* USA300 genome sequence were determined by tRNAscan-SE program. The codons marked by '1' and '2' are also recognized by K and G1 and K, G1, and Twort-specific-tRNAs, respectively.