Distinctive features gleaned from the comparative genomes analysis of clinical and non-clinical isolates of *Klebsiella pneumoniae*

Jina Rajkumari¹, Supriyo Chakraborty²*, Piyush Pandey¹*

¹Department of Microbiology, Assam University, Silchar 788011, Assam, India; ²Department of Biotechnology, Assam University, Silchar 788011, Assam, India; Piyush Pandey and Supriyo Chakraborty - Email: piyushddn@gmail.com, supriyoch_2008@rediffmail.com; RKJ – Email: jinalantham.micro@gmail.com

Received January 29, 2020; Revised March 10, 2020, Accepted March 15, 2020; Published March 31, 2020

DOI: 10.6026/97320630016256

Declaration on official E-mail:
The corresponding author declares that official e-mail from their institution is not available for all authors

Declaration on Publication Ethics:
The authors state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/.
The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Abstract:
It is of interest to describe the distinctive features gleaned from the comparative genome analysis of clinical and non-clinical isolates of *Klebsiella pneumoniae*. The core genome of *K. pneumoniae* consisted of 3568 genes. Comparative genome analysis shows that *mdtABCD*, toxin-antitoxin systems are unique to clinical isolates and catB, benA, and transporter genes for citrate utilization are exclusive to non-clinical isolates. We further noted aromatic compound degrading genes in non-clinical isolates unlike in the later isolates. We grouped 88 core genes into 3 groups linked to infections, drug-resistance or xenobiotic metabolism using codon usage variation analysis. It is inferred using the neutrality plot analysis of GC12 with GC3 that codon usage variation is dominant over mutation pressure. Thus, we document data to distinguish clinical and non-clinical isolates of *K. pneumoniae* using comparative genomes analysis for understanding of genome diversity during speciation.

Keywords: *Klebsiella pneumoniae*, Comparative genomics, Codon usage bias

Background
*Klebsiella pneumoniae* is a non-motile, gram-negative bacterium, which inhabits in diverse ecological niches ranging from soil to water and plants, and it is also opportunistic pathogen causing hospital-acquired disease in patients with the compromised immune system [1]. Human clinical isolates are considered as indistinguishable from environmental isolates with respect to their biochemical reactions and other attributes [2]. In plants, *K. pneumoniae* strains had been reported to fix nitrogen and promote growth of the plants [3]. Rajkumari et al. [4] had reported that
Klebsiella strains have the ability to degrade hydrocarbons, including polyaromatic hydrocarbons (PAH). Previously, Holt et al. [5] have determined the diversity of genetic variation in specific genes associated with virulence and antibiotic resistance to track the emergence of invasive K. pneumoniae infections.

The advent of whole-genome sequencing along with reliable bioinformatics workflow has made possible to study the distinct genome patterns that may reflect the variation due to evolutionary pressures. Identification of conserved genomic core genes and pan genes from a collection of bacterial genomes depends on classifying orthologous genes based on similar sequences [6]. It provides valuable information and deeper insights into bacterial genome evolution, genes associated with host adaptation, virulence, and pathogenesis [7]. In previous work, the whole genome sequence analysis of clinical and environmental K. pneumoniae suggested that they are closely related but antibiotic resistance and virulence factors were more frequent in clinical isolates. In fact, the phylogenomics analysis of K. pneumoniae whole genome failed to result in any distinct segregation of clinical and non-clinical clades of K. pneumoniae, as the genomes originating from either group were mixed throughout the tree [8].

Therefore, in this work, a genome analysis workflow was designed and used to identify the genes and their functions, which can be considered to resolve between clinical and non-clinical K. pneumoniae isolates (Figure 1). We used phylogenomics approach followed by the comparison of K. pneumoniae codon usage bias (CUB) pattern across the genes. In fact, the synonymous codons are known to be used non-randomly, and this unequal usage of the synonymous codon is called codon usage bias [9]. It is of interest to describe the distinctive features gleaned from the comparative genome analysis of clinical and non-clinical isolates of K. pneumoniae.

Table 1: Genomic features and comparison of K. pneumoniae genomes used for analysis

| Strain                        | Size (bp) | CDS | GC (%) | Source                   | References |
|-------------------------------|-----------|-----|--------|--------------------------|------------|
| K. pneumoniae AWD5 (AWD5)     | 4807409   | 4636| 58.18% | Non-clinical             | Rajkumari et al. 2017 |
| K. pneumoniae 342 (Kp342)     | 564239    | 5768| 56.87% | Non-clinical             | Fouts et al. 2008 |
| K. pneumoniae SKGH01 (SKGH01) | 6088457   | 5777| 56.54% | Clinical                 | Alfaresi 2018 |
| K. pneumoniae subsp. pneumonia PNDM01 (PNDM01) | 8512034 | 5529| 56.78% | Clinical                 | Doi et al. 2014 |
| K. pneumoniae J1 (J1)         | 5408666   | 5039| 57.24% | Non-clinical             | Pang et al. 2016 |
| K. pneumoniae strain KCTC-2242 (KCT242) | 5462423 | 5152| 57.28% | Non-clinical             | Shin et al. 2012 |
| K. pneumoniae subsp. pneumonia RJF293 (RJF293) | 5450593 | 5077| 57.20% | Clinical                 | Wang et al. 2018 |
| K. pneumoniae strain KP-1(KP-1) | 5131085 | 4755| 57.60% | Non-clinical             | Lee et al. 2013 |
| K. pneumoniae NTUH-K2044 (NK2044) | 5248520 | 5006| 57.7   | Clinical                 | Wu et al. 2009 |
| K. pneumoniae ATCC BAA-2146 (BAA2146) | 5608367 | 5552| 56.90% | Clinical                 | Hudson et al. 2014 |
| K. pneumoniae subsp. pneumonia KPNH10 | 5395263 | 563 | 57.14% | Clinical                 | NZCP007727 |
| K. pneumoniae subsp. pneumonia KPNH1 | 5390456 | 5654| 57.14% | Clinical                 | NZCP008827 |
| K. pneumoniae subsp. pneumonia strain KPNH33 | 5524202 | 5500| 57.21% | Clinical                 | NZCP009771 |
| K. pneumoniae subsp. pneumonia KPNH27 | 5524163 | 5932| 57.16% | Clinical                 | NZCP007371 |
| K. pneumoniae subsp. pneumonia KPNH124 | 5363164 | 5590| 57.12% | Clinical                 | NZCP008797 |
| K. pneumoniae subsp. pneumonia KPNH30 | 5306618 | 5346| 57.26% | Clinical                 | NZCP008972 |
| K. pneumoniae strain XH209 | 5118878 | 4881| 57.63% | Clinical                 | NZCP009461 |
| K. pneumoniae strain CAV1016 | 5367681 | 5391| 57.22% | Clinical                 | NZCP007394 |
| K. pneumoniae strain CAV1042 | 5354189 | 5500| 56.81% | Clinical                 | NZCP008671 |
| K. pneumoniae strain CAV1596 | 5402147 | 5450| 57.14% | Clinical                 | NZCP009467 |
| K. pneumoniae strain CAV1417 | 5208900 | 5373| 56.98% | Clinical                 | NZCP013852 |
| K. pneumoniae Kp_Goe_71070 | 5497083 | 5434| 56.95% | Clinical                 | NZCP018450 |
| K. pneumoniae Kp_Goe_827026 | 5573056 | 5662| 56.62% | Clinical                 | NZCP018707 |
| K. pneumoniae Kp_Goe_149473 | 5373055 | 5673| 56.62% | Clinical                 | NZCP018773 |
| K. pneumoniae Kp_Goe_153021 | 5373056 | 5683| 56.62% | Clinical                 | NZCP018713 |
| K. pneumoniae Kp_Goe_828304 | 5373056 | 5683| 56.62% | Clinical                 | NZCP018719 |
| K. pneumoniae MNCRE78 | 5454003 | 5584| 56.86% | Clinical                 | NZCP018428 |
| K. pneumoniae MNCRE53 | 5490693 | 5627| 56.87% | Clinical                 | NZCP018437 |
| K. pneumoniae Kp_Goe_154414 | 5159815 | 5618| 56.62% | Clinical                 | NZCP018357 |
| K. pneumoniae Kp_Goe_62269 | 5423372 | 5607| 57.16% | Clinical                 | NZCP018364 |
| K. pneumoniae Kp_Goe_33208 | 5497872 | 5429| 56.96% | Clinical                 | NZCP018447 |
| K. pneumoniae Kp_Goe_822917 | 5824741 | 5360| 57.21% | Clinical                 | NZCP018348 |
| K. pneumoniae strain CR04 | 5470889 | 5869| 56.78% | Clinical                 | NZCP013592 |
| K. pneumoniae Kp_Goe_121641 | 5478335 | 5390| 56.96% | Clinical                 | NZCP018735 |
| K. pneumoniae subsp. pneumonia KP0928 | 5309305 | 5286| 57.32% | Clinical                 | NZCP008831 |
| K. pneumoniae strain 34618 | 5313576 | 5487| 57.21% | Clinical                 | NZCP010392 |
| K. pneumoniae strain AR0049 | 5435743 | 5661| 56.98% | Clinical                 | NZCP018816 |
Methodology

Genome comparison and Sequence Data

Comparative genome analysis of a collection of *K. pneumoniae* (clinical and non-clinical genomes) was examined. A private project was created with standard *K. pneumoniae* genomes, comprising of non-clinical/environmental and clinical isolates. Phylogenetic relationships between thirty-eight *K. Pneumoniae* genomes were analyzed using EDGAR software platform (v.2.3.1) (http://edgar.computational.bio). EDGAR allows the calculation and identification of the core-pan genomes between different genomes. The nucleotide coding sequence (CDS) for eighty-eight genes, (24 drug resistant (DRGs), 16 infections related genes (IRGs) and 48 xenobiotic metabolism genes (XMGs) of five clinical and five non-clinical *K. pneumoniae* genomes having perfect start and stop codon were retrieved from IMG database (http://img.jgi.doe.gov) and gene details are given below.

The effective number of codons

The observed effective number of codons (ENC) for each coding sequence of gene sets of *K. pneumoniae* was calculated using the formula given by Wright [10]. ENC value shows an inverse relationship with the degree of codon bias. ENC values range from 20 to 61, where low ENC value (<35) indicates high codon usage bias and high ENC value indicates low codon usage bias [10].

\[
ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}
\]

where Fk (k= 2, 3, 4, 6) is the mean of Fk values for the k-fold degenerate amino acids

Nucleotide composition analysis

The overall nucleotide composition (A%, C%, T% and G%) and occurrence of overall frequency of the nucleotide (G+C) at first (GC1), second (GC2) and third (GC3) position of the synonymous codons were calculated in the coding sequences of the genes to quantify the extent of base compositional bias. The calculations were done using a Perl script developed by one of the authors (SC).

Neutrality plot

The neutrality plot is a scattered plot, which is used to determine the role of directional mutational pressure against selection pressure during evolution. It is the regression of GC12 on GC3, as the synonymous mutation occurs in the 3rd position of codon while non-synonymous mutations occur in the 1st and 2nd position. The non-synonymous mutation transforms the activity of the gene, which resulted from the alteration of amino acid sequence. In neutrality plot, if the regression line falls near the diagonal, it signifies weak external selection pressure and the role of mutation pressure is dominant.

Software and Statistical Analysis

Heat map of the specific clinical and non-clinical genes was generated by Expression Heatmapper using an average linkage method with Euclidean distance [11]. A network of genes was created for selected unique genes of clinical and non-clinical isolates by Cytoscape 3.4.0 with GeneMANIA plugin. The node degree distribution of the complex protein-protein interaction network was obtained from Cytoscape by Network analyzer [12]. A PERL program was developed to estimate the genetic codon usage bias indices and the selection pressure on the coding sequence of *K. pneumoniae* genes. Correlation analysis was performed to identify the degree of relationship between two parameters by Karl Pearson’s method. The significance of the correlation coefficient was tested by t-test for (n-2) degrees of freedom at p<0.01 or p<0.05. Statistical analyses were performed using IBM SPSS version 21.0 for windows.

Results:

Phylogenomic relationship of *K. pneumoniae* (clinical and non-clinical) genomes was examined from the deduced amino acid sequences of the core genomes and resolved the close relationship among non-clinical and clinical isolates. The main features of the genome sequences of *K. pneumoniae* non-clinical strain and clinical strain are summarized below (Table 1). The tree was built out of core genome taking 3568 genes per genome, 135584 in total. There was no clear separation of clades between clinical and non-clinical isolates (Figure 2). Phylogenomic analysis of the core genomes 48 clinical and 29 environmental *K. pneumoniae* isolates had demonstrated that the isolates were intermixed and failed to result in any distinct segregation of clinical and non-clinical clades of *K. pneumoniae* [8].
Based on the phylogenomics results, genomes (clinical and non-clinical) were selected for further analysis and orthologous genes in K. pneumoniae were calculated. Core genome analyses of the K. pneumoniae genomes showed that it consisted of 3568 conserved genes. While the pan-genome appeared to grow rapidly and the core genome was limited to less than 4000 genes. There are several Pan-genome analysis pipelines available [13, 14, 15, 16] however we have used EDGAR pipeline for studying genetic variation, and function enrichment analyses of the gene clusters. Core versus pan-genome development analysis of K. pneumoniae genomes revealed that 3568 formed the core genome while 11780 genes formed the pan-genome when K. pneumoniae AWD5 is used as the reference genome (Figure 3). The genes in AWD5 strain cover 89.45% of the coding genes in the genome. Average nucleic acid identities (ANI) of AWD5 with K. pneumoniae KP-1, K. pneumoniae ATCC-BAA 2146 and K. pneumoniae subsp. pneumoniae NTUH-K2044 revealed 99% sequence homologies and 94.02% with K. pneumoniae 342. The genomes of AWD5 and ATCC BAA-2146 (clinical strain) appear to be most similar by sharing 4529 orthologs whereas 4442 orthologs were found between AWD5 and the environment isolate KP-1. Pan-genome of six K. pneumoniae strains was reported to be consisted of 4,829 core genes, such a high percentage signifies a high rate of conservation among the strains. Previously, some studies had reported that phenotypic and genetic features of K. pneumoniae of environmental and clinical origin were similar and therefore, the isolates cannot be distinguished [2, 17].

**Figure 3:** Core vs pan-genome development plot of K. pneumoniae genomes (EDGAR 2.2 software platform)

**Gene content analysis of K. pneumoniae**

By investigating the presence and absence of genes in K. pneumoniae, most of the regions within the genomes were found to be conserved, including the virulence genes present in the clinical strains regardless of disease source. The orthologs include components of regulatory pathways such as basic transcriptional machinery, DNA relication, homologous recombination, mismatch repair, nucleotide excision repair, bacterial secretion system and protein export. Genes attributable to the production of indole-3-acetic acid (IAA) (ipdC), solubilization of phosphate (pqqABCDEF, phn and pho gene clusters, pstBACS), synthesis of siderophore (ent, fep gene clusters), acetoin and 2,3-butanediol (alsDSR, budC) are found to be conserved. It also revealed the presence of multicomponent nitrate or nitrite transport system. More than 13 genes involved in benzoate (ben genes), catechol (cat genes),
protocatechuate (pca genes) were conserved in the core genome of *K. pneumoniae*.

Further, to obtain the unique genes, the genomes were segregated into two groups i.e., clinical and non-clinical, taking five genomes for each group and analyzed separately. In clinical isolates, such core genes included virulence factor capsule assembly protein, multidrug transporter subunit (*mdtABCD*), type II toxin-antitoxin systems, and type VI secretion system (TVISS). TVISS were found to be shared among clinical and non-clinical isolates however, at least eight CDSs determined in clinical genomes involved in TVISS, were not found in non-clinical isolates. It has been reported that an environmental isolate Kp342 and clinical isolate MGH78578 seemed to share core components of TVISS [18]. The unique core genome of non-clinical isolate consisted of aromatic compound degrading genes *catB*, *benA* and several transporters including genes for citrate utilization. This was in agreement to the hypothesis that environmental isolates are more versatile for their catabolic processes, as the ability to degrade organic compounds govern the evolution of novel catabolic abilities of bacteria that can survive in different habitats [19]. We took the presence/absence of such individual genes as the categorical analyst parameter subjected to build a phylogenetic tree. And showed that the strains from similar clinical and environmental sources were not linked than those from different sources (Figure 4). The core genes discriminated the clinical and environmental genomes and accessory genes segregated the strains within the group. Based on the gene functions, several gene sets were experimented which may resolve clinical and non-clinical isolates. Accordingly, one set of unique core genes in non-clinical isolates and clinical isolates is given in the figure 4 on the basis of which genomes were resolved into two distinct categories, further few genes were useful for resolving the genomes within the particular group. Virulence factor (*galF*, *wcaI*, *wcaG*, *wzb*, *wzc*, *manBC*), allantoin utilization genes (*gcl*, *allABCDS*, *ybbW*), TVISS (*impL*, *vgrG*, *tssFGJ*) and other antibiotic resistance genes were also observed as accessory genes in ten *K. pneumoniae*. The clinical genomes possessed multidrug resistance gene *blaCTX-M-15* as accessory gene; this is an indication of horizontal gene transfer that might have occurred between virulent and multidrug-resistant *K. pneumoniae* strains [20]. This finding agrees with the previous study, which reported that resistance to multiple antibiotics was found to be more frequent in clinical origin than that of environmental origin [21]. The gene interaction network of specific core genes of clinical and non-clinical *K. pneumoniae* genomes was analyzed in cytoscape and GeneMANIA (Figure 5) and data is given in supplementary file. It was noted that the functional partners were not part of the core genome of the respective group; however, it surely had interactions, and hence influence the functions of these genes.

**Figure 4:** Dendogram showing the relationship between genomes based on the presence and absence of genes designated as group specific.

### Extent of Codon usage variation in *K. pneumoniae*

In order to analyze codon usage variation five *K. pneumoniae* genomes were selected considering their level of similarity. The data set was restricted to 5 genomes to avoid non-ambiguity. According to the EDGAR interface, genes were placed at intersections of the Venn diagram only if they were reciprocal matches. The analysis utilizes all CDS of the genomes and it is not restricted to the core genome. Among the *K. pneumoniae* strains, AWD5 chromosome shared 31 orthologous CDS with BA2146 and further 188 CDS conjointly with the strain KP-1. Moreover, 65 orthologous CDS are found to be shared conjointly by environmental isolates AWD5, KP-1 and Kp342 (Figure 6). Also, it indicated that AWD5 has 10 singleton genes. Besides, the singletons that resemble to the genes without reciprocal best hit to another
genome, as orthologs does not necessarily have to be a proper singleton.

Figure 5: Cytoscape and GeneMANIA interaction network of known (a) clinical and (b) non-clinical unique core genes with co-expression and genetic interaction of other related genes. Nodes represent proteins and edges represent interactions between the nodes. Power law node-degree distribution of the networks in genes of clinical and non-clinical genomes (c & d). The node degree (k) is represented on the x-axis and number of nodes with a particular k is represented on the y-axis. The Pearson correlation coefficient values (r²) and the probability of degree distributions P(k) are shown. The red lines indicate the power law. The R-squared is computed on log values.

Figure 6: Visualizing common gene pools in five K. pneumoniae strains by Venn diagrams (EDGAR 2.2 software platform)
The codon usage variation was studied for eighty-eight genes (24 DRGs, 16 IRGs, and 48 XMGs) from five K. pneumoniae genomes of non-clinical origin (Kp342, KP-1, AWD5) and clinical origin (NK2044, BA2146) core genome. Selected genes details are given in the supplementary file (Table 2).

To quantify the extent of variation in codon usage among different genomes of K. pneumoniae, the effective number of codon (ENC) values for three gene sets of each genome was calculated. The ENC is a non-directional measure of codon usage bias, widely used to measure for individual genes. The ENC values among five K. pneumoniae genomes was almost no variation of codon usage bias among the genes of K. pneumoniae. It also indicates conserved genomic composition among different K. pneumoniae genomes. Correlation analysis between codon usage bias and compositional properties of GC content were analyzed to understand the effect of base composition on codon usage bias. Negative correlation was observed between ENC and GC composition (Table 4). These results suggested that natural selection might have played an important role in codon usage pattern across the genes. Our results show that codon usage bias and gene expression among different K. pneumoniae genomes was lower, slightly biased in the genes [22].

### Table 2: Selected gene of K. pneumoniae and functions

| Sl.no | Gene | Protein function |
|-------|------|-----------------|
| 1     | lepA | GTP-binding protein |
| 2     | ureC | urease subunit alpha |
| 3     | groL | chaperonin GroEL |
| 4     | fimA | type 1 major fimbrial subunit precursor |
| 5     | fimD | putative export and assembly usher protein of type 1 fimbriae |
| 6     | fimG | type 1 fimbrial minor component |
| 7     | fimH | type 1 fimbrial adhesin precursor |
| 8     | cyaA | adenylate cyclase |
| 9     | sdiA | succinate dehydrogenase subunit A |
| 10    | tolC | outer membrane channel protein |
| 11    | norV | nitric oxide reductase |
| 12    | norW | anaerobic nitric oxide reductase flavonoid redoxin |
| 13    | RpoS | RNA polymerase, sigma 38 subunit |
| 14    | RpoN/SigL | RNA polymerase, sigma 54 subunit |
| 15    | yqiQ | putative protease |
| 16    | LuxS | 5-ribosylhomocysteine lyase/quorum-sensing autoinducer 2 (AI-2) reporter protein |
| 17    | arsA | UDP-4-amino-4-deoxy-L-arabinose formyltransferase |
| 18    | arsB | UDP-4-amino-4-deoxy-L-arabinose-oxoglutarate aminotransferase |
| 19    | arsC | undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase |
| 20    | arsD | undecaprenyl phosphate-alpha-L-arab4N synthetase subunit |
| 21    | arsT | 4-amino-4-deoxy-L-arabinose transferase |
| 22    | amiA | N-acetyl-L-muramoyl-L-alanine amidase |
| 23    | amiB | N-acetyl-L-muramoyl-L-alanine amidase |
| 24    | sapB | cationic peptide transport system permease protein |
| 25    | sapC | cationic peptide transport system permease protein |
| 26    | sapD | cationic peptide transport system ATP-binding protein |
| 27    | sapF | cationic peptide transport system ATP-binding protein |
| 28    | basS | two-component system, OmpR family, sensor histidine kinase |
| 29    | mraY | Phospho-N-acetyl-L-muramoyl-pentapeptide-transferase |
| 30    | oppA | oligopeptide transport system substrate-binding protein |
| 31    | oppB | oligopeptide transport system permease protein |
| 32    | oppC | oligopeptide transport system permease protein |
| 33    | oppD | oligopeptide transport system ATP-binding protein |
| 34    | oppF | oligopeptide transport system ATP-binding protein |
| 35    | ppiA | peptidyl-prolyl cis-trans isomerase A (cyclophilin A) |
| 36    | phoP | two-component system, OmpR family, response regulator |
| 37    | phoQ | two-component system, OmpR family, sensor histidine kinase |
| 38    | fsI | peptidoglycan synthetase |
| 39    | dnaK | molecular chaperone |
| 40    | marA | transcriptional regulator, AraC family |
| 41    | benA/ylX | benzoate/toluate 1,2-dioxygenase alpha subunit |
Table 3: Nucleotide composition analysis in the coding sequence of gene sets in K. pneumoniae strains

| Sl. No. | A (%) | T (%) | G (%) | C (%) | AT% | GC% | GC1% | GC2% | AT3% | GC12% | ENC |
|--------|-------|-------|-------|-------|------|-----|------|------|------|-------|-----|
| AM2    | 313.1 | 259.9 | 403.8 | 345.6 | 40.7 | 59.3 | 63.4 | 42.3 | 72.1 | 27.9  | 52.8 |
| KP1    | 308.8 | 260.9 | 406.6 | 341.2 | 40.5 | 59.5 | 63.8 | 42.5 | 72.1 | 27.8  | 52.8 |
| KP34   | 299.7 | 247.1 | 385.2 | 415.4 | 40.6 | 59.5 | 63.9 | 42.2 | 72.2 | 27.8  | 53.1 |
| BA2146 | 296.3 | 248.1 | 390.1 | 415.4 | 41.3 | 59  | 63.8 | 41.7 | 71.4 | 28.6  | 52.9 |
| NK2044 | 297.2 | 248.8 | 390.1 | 415.4 | 40.5 | 59.5 | 63.8 | 42.5 | 72.2 | 27.7  | 53.2 |
| Mean   | 303.02| 252.96| 394.5 | 421.78| 40.6 | 59.5 | 63.7 | 42.4 | 72   | 27.96 | 52.96|
| SD     | 7.501 | 6.827 | 9.96  | 10.85 | 0.207| 0.219| 0.195| 0.328| 0.339| 0.365 | 0.182|

**Infection-related gene**

| AM2     | 214.8 | 210.8 | 302.3 | 323.2 | 40.4 | 59.5 | 64.4 | 40.6 | 73.5 | 26.4  | 52.6 |
| KP1     | 210.8 | 210.8 | 302.3 | 323.2 | 40.4 | 59.5 | 64.4 | 40.6 | 73.5 | 26.4  | 52.6 |
| KP34    | 213.3 | 213.3 | 313.3 | 313.3 | 44.4 | 55.6 | 66.6 | 33.3 | 66.6 | 33.3  | 55.6 |
| BA2146  | 209.9 | 209.9 | 309.9 | 309.9 | 40.9 | 59.1 | 62.7 | 37.3 | 62.7 | 37.3  | 59.1 |
| NK2044  | 210.8 | 210.8 | 302.3 | 323.2 | 40.4 | 59.5 | 64.4 | 40.6 | 73.5 | 26.4  | 52.6 |
| Mean    | 211.7 | 211.7 | 311.7 | 311.7 | 44.4 | 55.6 | 66.6 | 33.3 | 66.6 | 33.3  | 55.6 |
| SD      | 3.001 | 3.001 | 3.001 | 3.001 | 0.207| 0.219| 0.195| 0.328| 0.339| 0.365 | 0.182|

**Drug resistance gene**

| AM2     | 214.8 | 210.8 | 302.3 | 323.2 | 40.4 | 59.5 | 64.4 | 40.6 | 73.5 | 26.4  | 52.6 |
| KP1     | 210.8 | 210.8 | 302.3 | 323.2 | 40.4 | 59.5 | 64.4 | 40.6 | 73.5 | 26.4  | 52.6 |
| KP34    | 213.3 | 213.3 | 313.3 | 313.3 | 44.4 | 55.6 | 66.6 | 33.3 | 66.6 | 33.3  | 55.6 |
| BA2146  | 209.9 | 209.9 | 309.9 | 309.9 | 40.9 | 59.1 | 62.7 | 37.3 | 62.7 | 37.3  | 59.1 |
| NK2044  | 210.8 | 210.8 | 302.3 | 323.2 | 40.4 | 59.5 | 64.4 | 40.6 | 73.5 | 26.4  | 52.6 |
| Mean    | 211.7 | 211.7 | 311.7 | 311.7 | 44.4 | 55.6 | 66.6 | 33.3 | 66.6 | 33.3  | 55.6 |
| SD      | 3.001 | 3.001 | 3.001 | 3.001 | 0.207| 0.219| 0.195| 0.328| 0.339| 0.365 | 0.182|
The nucleotide composition analyses of coding sequences of IRGs, DRGs, and XMGs (Table 3) showed mean percentage of GC and AT compositions was 59.44% and 40.5% in IRGs, 59.36% and 40.66% in DRGs and 61.66% and 38.34% in XMGs respectively. Genes were GC-rich, GC codon content at the third position was higher than at the first and second codon positions and the greatest difference of GC content was found between the second and the third codon positions. Hence, the overall nucleotide composition suggested that the nucleotide C and G occurred more frequently compared to A and T in the coding sequences and it is expected that G/C-ended codons might be preferred over A/T ended codons in the genomes. The difference in frequencies of A and T and to that of G and C were not the same which indicates that natural selection might have played a role in codon usage pattern [22].

A neutrality plot analysis of GC12 (average value of GC1 and GC2) versus GC3 (Figure 7) was drawn to characterize the correlation of the three codon position of GC [23], and to estimate the influence of selection and mutation pressure on codon usage bias of IRGs, DRGs and XMGs of five K. pneumonia strains AWDS, BA2146, KP-1, Kp342 and NK2044 were 0.0133, 0.0701, 0.0249, 0.0033 and 0.0099 indicating relative neutrality of 1.33%, 7.01%, 2.49%, 0.33%, and 0.99% respectively. The GC12 was influenced by mutation pressure and natural selection with a ratio of 1.33/98.67 = 0.014, 7.01/92.99 = 0.070, 2.49/97.51 = 0.0249, 0.33/99.67 = 0.003, and 0.99/99.01 = 0.009 for AWDS, BA2146, KP-1, Kp342 and NK2044 respectively. The subsequent correlation

Table 4: Summary of correlation analysis between ENC and various GC content

| ENC  | GC  | GC1 | GC2 | GC3 |
|------|-----|-----|-----|-----|
| AWDS | r   | -0.414* | -0.443* | -0.184 | -0.245 |
| p value | 0.044 | 0.03 | 0.388 | 0.249 |
| KP-1 | r   | -0.440* | -0.401* | -0.212 | -0.265 |
| p value | 0.031 | 0.032 | 0.319 | 0.211 |
| Kp342 | r   | -0.386 | -0.511* | 0.148 | -0.298 |
| p value | 0.057 | 0.009 | 0.481 | 0.148 |
| BA2146 | r   | -0.466* | -0.474* | -0.079 | -0.356 |
| p value | 0.022 | 0.019 | 0.715 | 0.087 |
| NK2044 | r   | -0.441* | -0.450* | -0.228 | -0.265 |
| p value | 0.031 | 0.027 | 0.284 | 0.211 |

Table 5: Summary of correlation analysis of GC12 Vs GC3

| ENC  | GC1 | GC2 | GC3 |
|------|-----|-----|-----|
| AWDS | Drug resistance gene r | 0.478* | 0.454* | 0.305 | 0.440* | 0.429* |
| p value | 0.018 | 0.026 | 0.138 | 0.031 | 0.037 |
| KP-1 | Infection related gene r | 0.038 | -0.069 | -0.009 | 0.222 | -0.026 |
| p value | 0.888 | 0.800 | 0.974 | 0.408 | 0.923 |
| Kp342 | Xenobiotic metabolism gene r | 0.478* | 0.482* | 0.483* | 0.499* | 0.503* |
| p value | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 |

*P < 0.05, **P < 0.001
analysis revealed positive correlation between GC12 and GC3 in DRGs and XMCs (Table 5). These results suggest that natural selection played a major role while mutation pressure played minor role in shaping the composition of coding sequence of *K. pneumoniae* [25].

**Figure 7**: Neutrality plot of infection-related genes, drug resistance gene and xenobiotic metabolism genes of *K. pneumoniae* genome AWD5 (a-c), KP-1 (d-f), Kp342 (g-i), BA2146 (j-l), NK2044 (m-o). Individual genes are plotted based on the mean GC content in the first and second codon position (GC12) versus GC content of the third codon position (GC3).
**Conclusion:**
A comparative genome analysis shows that clinical and non-clinical strains of K. pneumonia are similar and are not separated by phylogeny. However, gene level comparison help distinguish these isolates where mdtABCD, toxin-antitoxin systems are distinctive to clinical isolates and catB, benA, transporter genes for citrate utilization are limited to non-clinical isolates. Thus, these data help distinguish clinical and non-clinical isolates of K. pneumoniae towards the understanding of genome diversity during speciation.

**Supplementary Materials:**
The gene network of specific core genes of clinical and non-clinical isolates was analyzed in Cytoscape (Figure 5a & 5b). The gene network for the unique core genes of clinical isolates consisted of 64 nodes and 191 edges. The relationship of genes of interest in our study with functional partners was specified to find co-expression, genetic interactions, and physical interactions. A single cluster was identified. The relationship of the selected genes was correlated with other genes, for example, phosphoproteins that are assistant transporter of drug transmembrane transport (bacs, baeR, urvY) correlated to multidrug efflux system (mdtABCD). These two clusters were further connected closely to uracil catabolism system. Similarly, gene cluster of type II toxin-antitoxin module was connected to a group of genes with the phosphotransferase system and fructose specific subsystem. Hierarchical clustering led to the removal of seven outlier samples (malE, aslA, yaiZ, ushA), which belong to the unique core of clinical isolates. This is probably due to no interaction with the genes in the network.

We then expanded the analysis of the gene network for the unique core genes of non-clinical isolates that consisted of 36 nodes and 101 edges. Hierarchical clustering led to the removal of seven outlier samples (arsB, benA, betB, catB, ppx, sufE). Interestingly, two clusters were identified in which the cluster for the query gene ppdA with its functional partners showed no interaction with other genes. In the larger cluster, the relationship of the genes of interest was also correlated, such as an integral component of plasma membrane yifA; melB interacted with plasma membrane genes ubiD, ybd] with its gene cluster. The query gene of citrate utilization (citCDE) showed its interaction with functional partners of the gene cluster along with other genes. As well, the functional partners of the outliers’ catB, benA, and arsB were specified, which have a functional role in the catabolism of catechol, benzoate and arsenite transport. It was noted that the functional partners were not part of the core genome of the respective group; however, it surely had interactions, and hence influence the functions of these genes.

To understand the topological properties of these networks, the probability of the node degree distribution P(k) showed that each network satisfied scale-free topology following power law (r>0.6; in core genes of clinical genomes and r>0.2 in core genes of non-clinical genomes) (Figure 5c & d). The node distribution based on degree implied the presence of genes with centrality values. The distribution following the power law distribution and showing the nature of the scale-free network suggested a hierarchical organization in the network.

**References:**
[1] Jones RN. Clin Infect Dis 2010 51:581 [PMID: 20597676]
[2] Matsen JM et al. 1974 Appl. Microbiol 28: 672 [PMID: 4607526]
[3] Lee KWK et al. Genome Announc. 2013 6: e1082 [doi: 10.1128/genomeA.01082-13]
[4] Rajkumari J et al. 3 Biotech 2018 8 [doi.org/10.1007/s13205-018-1134-1]
[5] Holt KE et al. PNAS 2015 112(27): E3574 [doi: 10.1073/pnas.1501049112]
[6] Lefebure T et al. Genome biology and evolution 2010 2: 646 doi: 10.1093/gbe/evq048
[7] Segerman B. Frontiers in Cellular and Infection Microbiology 2012 2(116) [doi: 10.3389/ficmb.2012.00116]
[8] Runchaoren C et al. Genome Medicine 2017 9 [doi: 10.1186/s13073-017-0397-1]
[9] Bulmer M. J. Evol. Biol. 1988 1: 15 [doi.org/10.1046/j.1420-9110.1988.1010015.x]
[10] Wright F. Gene 1990 87: 23 [doi: 10.1016/0378-1119(90)90491-9]
[11] Babicki S et al. Nucleic acids Res. 2016 44:W147 [doi: 10.1093/nar/gkw419]
[12] Montojo J et al. F1000Research 2014 3 [PMID: 25254104]
[13] Chen X et al. 2018 Front. Microbiol. 9:1910 [doi: 10.3389/fmicb.2018.01910]
[14] Kulsum U et al. 2018 Adv Exp Med Biol. 1052:39 [doi: 10.1007/978-981-10-7572-8_4]
[15] Zhao Y et al. 2012 Bioinformatics 28:416 [PMID: 22130594]
[16] Zhao Y et al. 2018 BMC Genomics 36 [PMID: 29363431]
[17] Caputo A et al. Biology Direct 2015 10 [PMID: 26420254]
[18] Fouts DE et al. Plos Genetics 2008 4:e1000141 [doi:10.1371/journal.pgen.1000141]
[19] Chou HC et al. Infection and Immunity 2004 72:3783 [doi: 10.1128/IAI.72.7.3783-3792.2004]
[20] Bialek-Davenet S et al. Emerging Infectious Diseases 2014 20: 1812 [doi.org/10.3201/eid2014.114026]
[21] Podschun R. Zentralbl Hyg Umweltmed 1990 189: 527 [PMID: 2200423]
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article for FREE of cost without open access charges. Comments should be concise, coherent and critical in less than 1000 words.
