Al Ameen J Med Sci 2021; 14(1):11-19 ● US National Library of Medicine enlisted journal ● ISSN 0974-1143

REVIEW ARTICLE CODEN: AAJMBG

Functional genome Annotation in Bacterial infections: A Review

Sharique Ahmad[1], Shivani Singh[1], Shriya Arora[1] and Saurabh Srivastava[2*]

[1]Department of Pathology, Era's Lucknow Medical College and Hospital, Era University, Hardoi Road, Lucknow-226003, Uttar Pradesh, India and [2]Department of ENT, Era's Lucknow Medical College and Hospital, Era University, Hardoi Road, Lucknow-226003, Uttar Pradesh, India

Received: 03rd September 2020; ***Accepted:*** 20th December 2020; ***Published:*** 01st January 2021

Abstract: Functional annotation provides the information of gene about its biological identity and various types of molecular function and biological role like sub-cellular location and its expression. It functions by attaching biological information of gene and protein. The primary level of annotation is performed by using Basic logical Alignment Tool (BLAST) for finding similarity and then annotation of genes are performed based on this information. The function of gene and variation in gene functioning can be understand by the elucidation of genome annotation. Ensemble annotation tool gives fast computational annotation in eukaryotes derived from mRNA, RNAseq. The two groups HAVANA and RefSeq (human and vertebrate analysis and annotation and Reference Sequence respectively) give rise to wide genome manual transcript annotation and are the finest computational method which recognize 70% of manual annotated loci. The prokaryotic genomes are involved in various genome projects with utilization of Next Generation Sequencing (NGS) it leads to decrease in time and money consumption. Bacterial genome annotation is done by the help of annotation pipeline it provides NCBI resource of content including Nucleotide, Protein, BLAST by the use of software GLIMMER which helps in identification of long protein coding gene mainly in bacteria. A novel bacteriophage Abp2 infects the bacteria Acinetobacter baumannii a multi drug resistance (MDR). The genome annotation of this phage reveals that it is a circular DNA with 85 open reading frame (ORF) with putative genes in which 41 known function and 47 unknown function protein were recognized. This review article summarizes genome annotation function and structural aspects.

Keywords: Bacteria, Gene, DNA, Elucidation, Genome annotation, mRNA.

Introduction

The functional genomics helps to determine the function and interactions of genes and proteins by the use of genome-wide approaches in comparison of gene-by-gene approach of molecular biology techniques. It collects data from various processes related to DNA sequence, gene expression and protein function such as transcription of coding and non-coding sequence, translation interaction of protein-DNA, protein-RNA, and protein-protein interactions.

Functional annotation is a procedure of collecting information of genes about biological identity with various kinds of molecular function, biological role, sub-cellular location and its expression domain. Functional annotation is basically a process of attaching biological information to the genes or proteins. The basic level of annotation is performed by using sequence alignment tool that is basic logical alignment search tool (BLAST) for finding similarities, and then annotating genes or proteins based on that. More information of biological function is required for annotation system. The additional information helps in hand-operated annotation for differentiating between gene and proteins that have same annotation.

Beside the many genome sequence computational annotation characterizes proteins and genes. Functional annotations identify various variant functions based on information that variant loci are in known functional region or not which consist genomic or epigenomic signals. The non-coding variants function is extensive in terms of affected region of genome and they are involved in all the process from transcription to gene regulation and post translational level.

Automated and Computational annotation

The capacity of understanding the function of gene and effect of variation on gene function is dependent on its structure and this can be understood by genome annotation. It is started in its elementary form through gene predication which search putative genes structure, from genome like signal involved in transcription, splicing, and coding protein [1-4]. Ensembl is the automated gene annotation tool which gives quick computational annotation in genome of eukaryotes derived from known mRNA, RNAseq data protein sequence data [5-6]. Computational annotation provide important information and outline of the gene content in newly formed sequences in comparison to manual annotation which is still known to be gold standard for exact and comprehensive annotation [7].

The best computational method identified 70% manual annotated loci, only 3.2% of transcript predicated by computational method were found to validated. The HAVANA and RefSeq (human and vertebrate analysis and annotation and reference sequence) two groups gives manual transcript annotation of genome in wider scale [8]. RefSeq associated assembly of transcript and its proteins are annotated manually by NCBI in U.S.A, there are multiple RefSeq which are annotated through manual annotation but a significant proportion are not annotated [9]. In HAVANA transcript is annotated on the genome, on the other hand RefSeq annotate separately from genome based on mRNA sequence alone, it give rise problems in mapping the genome.

The GENECODE gains benefit by both of HAVANA and Ensembl from HAVANA manual annotation and from Ensemble automated annotation, two into single dataset combination is performed by Ensembl gene. This identified four functional categories of prime genes: pseudogene, protein coding gene, long-noncoding RNA and small RNA. The acquisition in these biotypes in gene and transcriptional level both has been enhanced through the annotation. The Coding sequence type, not containing 5′ or 3′ UTRs, are utilized in panels of exome with the full gene sets of Ref Seq and GENCODE which forms many of the target sequences in the panel of exome. The GENECODE genes improves Coding Sequence because it is enhanced by addition of alternatively spliced transcript at coding genes of protein also pseudogene lncRNA annotations are most detailed set of genes [10]. For presenting genome annotation in meaningful and useful perspective web based interfaces are provided publicly, both Ensembl and UCSC genome browser shows GENCODE types [11]. GENCODE genes have been rationalized two times in one year and consensus coding sequence (CCDS) is updated once in whole year.

The entire transcript has been assigned by a different kind stable identifier which changes only if transcript goes through with any type of structural changes making temporary tracking of sequence assay. The genome browser have provided the great functionality by graphical interface for displaying and interrogating the genome information. It is connected by other relevant biological database for recognizing variation of sequence and predicting its conserved sequence by the use of variant effect predictor (VEP). It helps in investigating phenotypic information and tissue specific gene expression also search related to sequence of genome by the use of BLAST.

Genome annotation is introduced by genome assembly using de novo approach in assembled genome annotation is started by identifying, masking RNA genes by the use of RNAmmer and tRNA Scan SE. Tools for finding gene such as Prodigal, Gene Mark are used for finding Open reading frame in genome sequence. These ORF are searched by Basic Local Alignment Search Tool (BLAST) it is used for searching similarity in sequences against GENEBANK database for identifying putative gene functions. Domains of protein are identified by Inter Pro Scan it is a sequence analysis application which combines different protein signature identification methods into one resource. Genome annotation describes the function of the product of a predicated gene.

All of this can be achieved by use of bioinformatics with specific features like signal sensors (TATA box, start, stop codon), content sensors (codon usage), similarity detection (protein from closely related organisms). Genome annotation is divided

into three categories. Nucleotide-level annotation, protein level, process level. Genome of prokaryotes are involved in many genome project by the utilization of Next generation sequencing (NGS) leading to decrease in time and money consumption in all projects. [12] Genome annotation in microbes is performed by automatic annotation with the help of pipeline it gives all the raw data from public repository and this is followed by curation of the results manually [13].

Utmost annotation pipeline provide content for NCBI resources including nucleotide, protein use methods of homology which transfer information to closely related reference genome to the newly formed sequence. Annotation of automatic pipeline type can result in poor annotation and further error can occur. This is the reason why manual curation step is used. Therefore, now it has become possible to sequence various genomes of microbes in one day at low cost. In spite of this, high quality of annotation can go beyond gene predicting software and annotation transfer between close relations.

It function by adding quality apart from coding sites like termination site, ribosomal binding site and other conserved sites. All these features will not only result in full annotation but also rectify the errors occurred earlier parts during the process of annotation. As ribosomal binding site, termination site will give us more clear information of gene beside gene prediction solitary. The various software tool are there for predicting these features [14-18]. RNAseq give better indication of protein role when incorporated experimentally. The idea of giving a quality tend to annotation will be seldom not novel [19].

Bacterial genome Annotation

Bacterial annotation is mostly done by annotation pipeline which employs gene prediction software they commonly use GLIMMER [20]. It helps in identifying long protein coding gene mainly in bacteria. Gene finding can also be performed extrinsically by identifying ORF directly by comparing to database of protein [21]. When the coding region are identified they either are lineup with reference genome annotation or the solely with Uni Prot [22]. RAST, BASys, WeGAS, MaGe/Microscope. These are bacterial annotation pipelines which are published also MIcheck which is used to check annotated sequence or any syntactic errors [23-26]. Beside all this hypothetical proteins there is also a term which we should be aware of, it is a protein consist gene which had been identified through software and it do not have any known function in database. There are various bacterial gene which do not have any known functionality are called y-gene found on the orthologs site of E. coli K-12 [27].

Gene annotation has progressed throughout in many strains of bacterial species. The gene Yab F is one of the example regarding the hypothetical protein of E. coli as Yab F function is glutathione regulated efflux of K+ system based and on the other hand Yab F is totally non-functional in all of the genomes. While performing annotation of genome the two important genes to study are orthologs and paralogs. Orhtologs are gene which arises by speciation and paralogs gene arise by duplication of genes. Both of the genes are involved in both evolutionary and functional relationship. As orthologs gene gain its similar function but paralogs diverge to perform distinct function [28].

Therefore, while transmission of functional annotation from sequence to newly formed genome, it is important to define orthologs correctly [29]. Orthologs and paralogs can be defined by phylogentic tree based approach but it will be impractical to construct phylogenetic tree for each gene. So an alternative for this is bidirectional or reciprocal bit hit approach which are determined by BLASTA or FASTA. These are two search tools of Bioinformatics for similar sequence of DNA and protein.

Evolution of techniques for Genome Annotation

The advancement in genomic technology has given unprecedented data of genome variation of various disease to the researchers [30]. The genome technology cannot be possible without genome sequencing which was mainly established by the Human genome project. Human genome project provides sequencing of approx 3 billion of base pair that human constitute. The project was officially launched in 1987.

The computational annotation process provides biological functionality to the genome illustrating 30-40 thousands of coding protein among which 22 pair of autosome and 1 pair of sex chromosome in 2.9 billion of base of genome [31]. Recent publication estimate that in 3.1Gb of genome only 20,000 of protein coding gene are present [32]. Now DNA sequencing has been utilized in huge parameter in various research programs like Deciphering Development Disorders (DDD) study [33].

The DNA sequence speed and amount that can be produced and genome numbers which sequenced have been broadly increased by (NGS). Such kind of advancement enables huge collaborative projects which looks at variation in a population such as 1000 genome projects [34]. Also several investigators who are investigating the medical value of WGS in 100,00 genome project [35]. This all will help in research, diagnosis and prognosis of disease.

The analysis of genome for the prognosis purpose is performed by investigating patients genome and then sequencing the genome and finally it is line up with the reference genome and analyzed for variation in the sequence. The Burrows-Wheeler Aligner (BWA) software is utilized for both types of read short and long [36]. Genome Analysis Toolkit (GATK) is used for sequence variation. Pathogenic variation of sequence can range from single variation in nucleotide and deletion or insertion of 50bp for large structural variation, classified as genomic variation region greater than 1 kb such as segmental duplications retrotransposon elements, and several genomic rearrangements [37].

Our knowledge about gene function and regulatory pathway is increased by profiling the abundance of transcript in various cells in different circumstances. The genome consist of most of the non coding region. For knowing the consequences of variation in sequence, non-coding region is classified as cis-regulatory elements like promoters and distal elements (enhancer). Comprehensive map of these regions has been done by the large collaboration of ENCODE Road Map Epigenetics [38-39]. Whether variant lie in this region or not is determined by Ensemble regulatory build and Variant effect predictor (VEP) [40-41].

But they are unable to find out the pathogenicity, tools which can find it have begun to rise out like Genomiser and Fun-seq. The non-coding RNA are of two types small non-coding RNA and long non-coding RNA. Small non-coding RNA are miRNA, piRNA, siRNA, snoRNA[42]. The small non-coding RNA can be predicted by using tool infernal and Rfam [43-44]. It makes the sequence interpretation and variation when compared with long noncoding RNA.

In past years RNA transcript was analyzed by Northern blotting or RT-PCR and revealed that they were restricted only in limited number of transcript. After that Serial analysis of gene expression (SAGE) was evolved it consist sequence of small tag corresponding to 3' fragment of mRNA. SAGE allows highly quantative analysis only by counting the number of tags which map to the specific gene. In spite of various improvements in the original protocol of SAGE it has not been utilized longer because of labor intensive feature and lower rate of successful delivery in comparison to newly developed technique next generation sequencing (NGS).

NGS have provided the power of analyzing broad range of genome wide in relation of gene expression and transcript profiles, which are known as RNA seq. Firstly the sample of RNA is read and mapped to the reference genome after that number of sequence read map of certain gene will correspond to the expression level of gene. RNASeq can also be used for analyzing the transcript boundaries intron, exon junction and discover novel transcript and alternative splicing variants. It can also be applied for profiling non-coding RNA, newly formed transcript and ribosome associated mRNA.

Annotation of Mobile antibiotic resistance Bacteria

The functional region of coding gene is designated as coding sequence and 5' and 3' UTR. The transcript of 5' UTR contains regulatory regions which are translated for producing protein that regulates the function of main coding sequence [45]. Coding sequence variants are well studied and understood areas for pathogen variation

sequence. The 3' UTR transcript has been the region for controlling regulatory protein like mRNA and protein involved in binding RNA. It has also been connected to the overall translation efficiency and stability of mRNA [46]. Both 3' and 5' UTR interact with one another for regulation translation by closed loop mechanism. The sequence motif which are important sequence for controlling expression of gene are promoter, silencer, enhancer. They are present in intergenic, intragenic, exonic regions. Therefore, there can be vast amount of transcript which is present in a specific cell and same transcript may not dominant in other place and if it is identified as dominant transcript it may not be functional.

Repetitive sequences and transposable elements are involved in more than 2/3 of the human genome and involved in diseases. These elements also have strong association with genomic Copy number variations CNVs [47]. Alu and Long interspersed nuclear elements are involved in genomic instability enhancement by non-allelic homologous recombination. This event will leads to pathogenic duplications and deletion. The capability of detecting accurately the repetitive sequences is important because it will create problem during assembly of sequence reads [48]. The analysis of human genome is mainly done by Repbase annotation, computational algorithms like hidden Markov model (HMM) derived database Dfam for repeats [49-50].

Bacteria have the capability of multi resistance this is due the acquisition of various antibiotic resistance genes captured from different source of organism which move as part of multiple mobile genetic elements. These mobile elements minor proportion consist coding sequence without promoter sequence named cassettes genes. This carries an attC recombination site and ORF often a resistance gene. Integron between attC and attI is a site specific recombination catalyzed by Int1 integron which leads to expression and capture of cassette-borne genes [51].

Insertion elements consist little bit more portion than transpose gene generally utilizing almost whole length of the flanked element designated by short Inverted repeats (IR) as IR_L and IR_R in relation to the tnp transcription direction. The pair of same Insertion sequence (IS) capture mediate resistance gene in composite transposon. The various mobile elements and their resistance gene travel horizontally in between bacteria also in other kinds of species. These all elements contain backbone which encodes function of plasmid in which accessory region are inserted. The insertion elements which do not cause any effect in plasmid function are called founder elements, following with non-disruptive insertion frequently leads to multiresistance region (MRR) formation. Comparative analysis and annotation of MRR provide understanding of evolution and relationship which complicates by insertion of one mobile element in another by deletions for rearrangements [52].

The non-specific annotation software is currently utilized for identifying potential gene from known function homology. There are several resources present for identifying resistance gene are Res Finder, CARD, ARG-ANNOT, SRST2 recognize the resistance gene short read data and SSTAR stands alone tool use other database. IS finder provides database and BLAST tool for the identification of insertion elements.

A transposon registry list gives Tn numbers which gives links to the sequences. VRprofile detect resistance genes and virulence genes and several mobile elements by utilizing these and other database. Integron finder finds attC sites and several other integron components. From all of these not any tool provides accurate annotation of resistance gene and mobile elements. Then Attacca developed utilizes (FDB) feature database BLAST and computational grammars for accurate and consistent annotation to recognize the patterns [53].

Multi Antibiotic Resistance Annotation (MARA) shows the location name of every feature and its orientation with the arrow which indicates the transcription of genes direction or transposase of genes. MARA is most suitable for analyzing multiresistance region (MRR) of fully assembled genome sequence or plasmid through Enterobacteriaceae, on other hand in slightly assembled genome sequence annotation of sequence will not be performed due to the errors or shorten features.

The entire genome of bacterial can be submitted for annotation or contig, because MARA can annotate any repeat fragment present at the end of contig. The annotation provides the information which can be utilized for PCR designing to confirm the association between contigs. The features included only in FDA can be recognized by MARA, though mobile elements and novel elements cannot be annotated. Further to identify novel features (MRR) gaps are needed to analyze like IS finder (1585bp of gap).

Likewise the nearly spaced fragments of related or same, Insertion elements provides the suggestion of sequence error or variant that are very difficult to annotate. The site of MARA give access for the Attacca automatic search engine annotation which allows simple quick annotation of MRR in DNA sequence through bacteria Enterobacteriaceae. The reliable features to Acinetobacter and Pseudomonas are already present and it has been exploring through FDB for annotation of MRR and resistance islands in these species.

Annotation of Multidrug resistance bacteria

Acinetobacterbaumannii bacteria are responsible for many health associated infections like wound or infected burn [54]. This is main cause of infection in urinary tract and respiratory tract sepsis and secondary meningitis. Few strains of this gram negative bacteria is found resistant to all of the known antibiotics so there is instant requirement of finding a different treatment for all of these infections [55]. Roach et al revealed in 1910s that bacteriophages were helpful and utilized in these kind of human infections [56].

The bacteriophage is a potential strategy for fighting against Multi drug resistance A. baumannii. Zhang et al performed a study which revealed that bacteriophage preserved earlier has a strong lytic capability in A. baumannii. Multi drug resistance bacteria was inoculated by the wastewater and ICU patients undergoing burn treatment the sequence analysis shown that the phage is different from previously reported in A. baumannii phage Abp1 and it was then named Phage Abp2 [57]. Genome annotation and characterization of the phage Abp2 will provide more understanding and information regarding treatment of bacterial infections. The high productivity sequence read was congregated into a genome sequence which was fully closed and circular by the use of SOAP denovo which is a novel short read assembly method which can built a large human-size genome in form of denovo draft. The circularity in genome of phage was established by restriction endonuclease mapping.

The genome (DNA) of the phage contains 45,373 bp, 37.84% GC content. The analysis by BLAST tool of entire genome revealed that sequence of phage Abp2 was completely different from previously detected phage shared 0% identity, which suggests that they were entirely different from each other. Therefore, Abp2 genome exhibit 93% nucleotide identity, coverage of 71% to A. baumannii, with number of ORF never differing significantly. Its genome contains 88 putative ORFs. The open reading frame containing the protein encoding power of known function are grouped into various groups in which one is associated with morphogenesis and the other associated with structure, Replication, recombination, repair, biological metabolism, transcription lysis assembly packaging, putative foreign proteins and homing endonucleases.

BLASTn and BLASTp reveals structure and assembly associated proteins including fiber protein of tail, Putative capsid protein, baseplate- associated protein, head protein, portal protein, DNA replication, recombination, repair and several gene encoding region and DNA binding protein (ORF-3), (ORF6), (ORF11), (ORF14) ,(ORF38 and 45), (ORF79), (ORF81) and (ORF85), respectively were present in entire genome of phage Abp2. All of these sequences of ORF 41 encodes proteins with higher level of similarity in functioning and 47 were undefined ORFs. The ORF identified were involved in various functioning like transcription, lysin, lysozme, packaging, assembly, transcription host interaction and various biological metabolism. But there were 4 ORF (46, 48, 80, and 87) found in encoding putative foreign proteins and the range of similarity in sequence was 24-25%.

The other bacteria like Streptococcus pneumonia is also responsible for bacterial

infection. Pneumonia which infects mainly lungs, gram-positive bacteria are main pathogen for this infection. The efficiency for bacterial pneumonia treatment is dependent on choice of antibiotic drugs. For accurate diagnosis and successful treatment of pneumonia differentiation between gram positive and negative pathogens is important. The bacterial pneumonia incidence had been markedly enhanced and prognosis became poor due to increased bacterial resistance to antimicrobial agents [58].

The increase understanding of mechanism for bacterial recognition and clearance through immune system had revealed that certain deregulated gene and bio-functional pathways in lungs and several organs are the location for primary infection. Therefore some studies were conducted for revealing differentially expressed gene (DEG) and bio-functional pathway in bacterial infections participated. This study was performed on gram-positive pneumonia by the use of gene expression datasets. The sample from peripheral blood mononuclear cells for DEGs between healthy control and pneumonia patients were recognized.

Further, for annotation DAVID (Database for Annotation, Visualization and Integrated Discovery) was used for identifying DEGs and analyze the genome which are enriched GO and KEGG (Gene Ontology and Kyoto Encyclopedia of Genes) pathways. The bioinformatics analysis of bacterial gene was performed to determine DEGs and its associated pathway, 2 datasets which are purely independent from each other were selected. The two databases identified that total 40 DEGs were found to associate with pneumonia between both of the databases. DEGs all 40 were subject to annotation by GO/KGG functional analysis by the use of (DAVID). The DEGs mainly involved are CCL4, TIMP1, ICAM1, PLAUR and CTSB. These were further mapped by PPI network. This study revealed that there were 5 key DEGs present in gram negative bacteria. The bioinformatics performed were rarely performed for studying any kind of the disease. These ICAM1, TIMP1 and CCL4 co-function Gram-positive bacterial pneumonia by the help of NF-kB cell signaling pathways [59].

Conclusion

Functional annotation has provided the accurate and complete structural and functional aspects of genome. With the abundance of information emerge from various studies it is also being a challenge to gather this information data in a single network and also find out transcripts, protein work to regulate biological process which reveals the cell function progression. This has lead to development of more computational annotation methods for network analysis.

All the tools help in prediction of protein function. There are multiple tools available for analyzing and determining the certain pathways over-represented in various biological process. This will ultimately lead to building improved models of biologically relevant interactions between all components of a cell. Beside this the genome annotation in bacterial infection has provided to explore the mechanism of Gram-positive bacteria in pneumonia. Another bacteria A. baumannii multi drug resistance (MDR) and provide the opportunity to reveal the bacteriophage Abp2 in treatment of MDR bacteria by its genomic annotation.

All of this concludes that knowledge regarding the genome and its mechanism in several diseases has provided improvement in the diagnosis and prognosis of several diseases.

Financial Support and sponsorship: Nil

Conflicts of interest: There are no conflicts of interest.

References

1. Burge CB, Karlin S. Finding the genes in genomic DNA. *CurrOpinStruct Biol* 1998; 8:346-354.
2. Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* 2000; 10:516-522.

3. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 2003; 19Suppl 2: ii215-225.
4. Mudge J, Harrow J. Methods for improving genome annotation. In: Alterovitz G, Ramoni MF, editors. Knowledge based bioinformatics: from analysis to interpretation. *Chichester, West Sussex: John Wiley & Sons* 2010;p.209-214.
5. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 2012; 13:329-342.
6. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 2011; 39(Database issue):D214–219.
7. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 2006; 7Suppl 1: S4.1-9.
8. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 2014; 42(Database issue):D756–763.
9. Frankish A, Uszczynska B, Ritchie GR, Gonzalez JM, Pervouchine D, Petryszak R, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* 2015; 16Suppl 8:S2.
10. Mudge JM, Frankish A, Harrow J. Functional transcriptomics in the post ENCODE era. *Genome Res.* 2013; 23:1961-1973.
11. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* 2017; 45(D1): D626-34.
12. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 2009; 7:287-296.
13. Stothard P, Wishart DS. Automated bacterial genome analysis and annotation. *Curr Opin Microbiol* 2006; 9:505-510.
14. Attwood TK, Bradley P, Flower DR, et al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 2003; 31:400-402.
15. Suzek BE, Ermolaeva MD, Schreiber M et al. A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics* 2001; 17:1123-1130.
16. Ermolaeva MD, Khalak HG, White O, et al. Prediction of transcription terminators in bacterial genomes. *J MolBiol* 2000; 301:27-33.
17. Sigrist CJ, Cerutti L, de Castro E et al. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 2010; 38:D161-166.
18. Finn RD, Mistry J, Tate J et al. The Pfam protein families database. *Nucleic Acids Res* 2010; 38:D211-222.
19. Janssen P, Goldovsky L, Kunin V et al. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005; 6:397-9.
20. Delcher AL, Harmon D, Kasif S et al. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999; 27:4636-4641.
21. Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *MolBiolEvol* 1999; 16:512-24.
22. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 2011; 39: D214-219.
23. Aziz RK, Bartels D, Best AA et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 2008; 9:75.
24. Van Domselaar GH, Stothard P, Shrivastava S, et al. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 2005; 33:W455-459.
25. Lee D, Seo H, Park C, etal. WeGAS: a web-based microbial genome annotation system. *Biosci Biotechnol Biochem* 2009; 73: 213-216.
26. Cruveiller S, Le Saux J, Vallenet D et al. MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res* 2005; 33:W471–479.
27. Rudd KE. Linkage map of Escherichia coli K-12, edition 10: the physical map. *MicrobiolMolBiol Rev* 1998; 62: 985-1019.
28. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 2005; 39:309-338.
29. Kristensen DM, Wolf YI, Mushegian AR, et al. Computational methods for Gene Orthology inference. *Brief Bioinform* 2011; 12:379-391.
30. Epi PM Consortium. A roadmap for precision medicine in the epilepsies. *Lancet Neurol.* 2015; 14:1219-1228.
31. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409:860-921.
32. GENCODE. Human GENCODE version 24. 2016. *http://www.gencodegenes. org/stats/current.html.* Accessed 14 Feb 2017.
33. Firth HV, Wright CF. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* 2011; 53:702-703.
34. 1000 Genomes Project Consortium Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010; 467:1061-1073.
35. 100K Genomes. Sequencing 100000 Genomes. 2014. *http://www. genomicsengland.co.uk/.* Accessed 14 Feb 2017. 17.
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows Wheeler transform. *Bioinformatics* 2009; 25:1754-1760.
37. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM et al. Copy number variation: new insights in genome diversity. *Genome Res* 2006; 16:949-961.
38. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489:57-74.
39. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015; 518:317-330.
40. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J et al. Ensembl 2017. *Nucleic Acids Res* 2017; 45(D1):D635-642.

41. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A et al. The Ensembl variant effect predictor. *Genome Biol.* 2016; 17:122.
42. Morris KV, Mattick JS. The rise of regulatory RNA. *Nat Rev Genet* 2014; 15:423-437.
43. Barquist L, Burge SW, Gardner PP. Studying RNA homology and conservation with infernal: from single sequences to RNA families. *CurrProtoc Bioinformatics* 2016; 54:12.13.1–12.13.25.
44. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* 2015; 43(Database issue):D130-137.
45. Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, et al. Before it gets started: regulating translation at the 5′ UTR. *Comp Funct Genomics.* 2012; 2012:1-8.
46. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* 2000; 10:1001-1010.
47. Zhou W, Zhang F, Chen X, Shen Y, Lupski JR, Jin L. Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms. *Hum Mol Genet* 2013; 22:2642-2651.
48. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 2011; 7: e1002384.
49. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015; 6:11.
50. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res* 2016; 44(D1):D81–89.
51. Hall RM, Stokes HW. Integrons: novel DNA elements which capture genes by site-specific recombination. *Genetica* 1993; 90:115-132.
52. Partridge SR. Analysis of antibiotic resistance regions in Gram-negative bacteria. *FEMS Microbiol Rev* 2011; 35:820-855.
53. Tsafnat G, Copty J, Partridge SR. RAC: repository of antibiotic resistance cassettes. *Database* 2011; 2011: bar05.
54. Howard A, O'Donoghue M, Feeney A, Sleator RD. Acinetobacterbaumannii: an emerging opportunistic pathogen. *Virulence* 2012; 3(3):243-250.
55. Gong Y, Shen X, Huang G, Zhang C, Luo X, Yin S et al. Epidemiology and resistance features of Acinetobacterbaumannii isolates from the ward environment and patients in the burn ICU of a Chinese hospital. *J Microbiol (Seoul, Korea)* 2016; 54(8):551-558.
56. Roach DR, Debarbieux L. Phage therapy: awakening a sleeping giant. *Emerg Top Life Sci.* 2017; 1(1):93-103.
57. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J ComputBiol* 2000; 7(1-2):203-214.
58. Burwen DR, Banerjee SN and Gaynes RP: Ceftazidime resistance among selected nosocomial gram-negative bacilli in the United States. National nosocomial infections surveillance system. *J Infect Dis.* 1994; 170:1622-1625.
59. Yang Z, Liu X, Shi Y, Yin S, Shen W, Chen J et al. Characterization and genome annotation of newly detected bacteriophage infecting multidrug-resistence Acinetobacterbaumannii. Arch. Virol 2019; 164:1527-1533.

Cite this article as: Ahmad S, Singh S, Arora S and Srivastava S. Functional genome Annotation in Bacterial infections: A Review. *Al Ameen J Med Sci* 2021; 14(1):11-19.

*All correspondences to: Dr. Saurabh Srivastava, Assistant Professor, Department of ENT, Era's Lucknow Medical College and Hospital Era University, Hardoi Road, Lucknow-226003, Uttar Pradesh, India. E-mail: saurabhsribhu@gmail.com