The eventual objective of all genome assembly, annotation and validation studies is to utilize costs and time effective sequencing methods for microbial research. The present dissertation work was designed to deal with evaluation of performance of genome assemblers, whole genome assembly and annotation of novel. This practice was efficiently done and outcome of the studies is anticipated to be beneficial for the scientific community.

Benchmarking of genome assembly software has been performed to find out the best genome assembler and assembling parameters. In this study, following six genome assemblers have been evaluated; Velvet, SOAPdenovo, ABySS, Euler-sr, Edena and SSAKE. These assemblers were evaluated on simulated datasets and a real sequencing dataset of *Pseudomonas syringae* pv. *syringae* B728a (Farrer, Kemen, Jones, & Studholme, 2009). It was observed that Velvet and SOAPdenovo assemblers perform better than other genome assemblers.

Burkholderia sp. SJ98 is a bacterium responsible for the biodegradation of several nitroaromatic compounds. Detailed genome annotation of strain SJ98 let to the identification of genes responsible for the degradation of nitroaromatic compounds *i.e.* fluorobenzoate, chlorocyclohexane and chlorobenzene, xylene, dioxan, styrene, chloroalkane and chloroalkene, toluene, benzoate and ethylbenzene etc. A total of 37 *che* genes, including 19 methyl accepting chemotaxis proteins (MCPs), which involved in sensing of different attractants, have been identified in the genome of SJ98. Phylogenomic studies on the basis of *rpoB* gene followed by whole genome comparison with nearest strains provide the idea about the taxonomic position of stain SJ98. Chemotaxis gene clusters was also identified and compared with other related stains *i.e.* YI23, CCGE 1001, CCGE 1002 and CCGE 1003.

Rhodococcus imtechensis RKJ300 is another important microbe responsible for the biodegradation of nitroaromatic compounds *i.e.* p-Nitrophenol, 2-chloro-4-nitrophenol and 2, 4-dinitrophenol. Whole genome assembly and annotation of *Rhodococcus imtechensis* RKJ300 has been done to explore the potential of this microbe. Genome was sequenced using Illumina Hi-Seq 1000 platform, and assembled using SOAPdenovo software. In order to annotate RKJ300, PGAAP pipeline of NCBI was used to predict genes and their function. Genes involved in the catabolism of several biomolecules, like

cholate, vanillin and lipids etc. have been identified in the genome of strain RKJ300. Important genes involved in the metabolism of various storage compounds like triacylglycerols (TAG), wax esters, polyhydroxyalkanoates (PHA), glycogen and polyphosphate (PolyP) have also been identified during the genome annotation. Whole genome comparison provides the idea about nearest neighbors of strain RKJ300 *i.e.* *Rhodococcus opacus* M213, *Rhodococcus* jostii RHA1, *Rhodococcus opacus* PD630 and *Rhodococcus* sp. JVH1.

Whole genome sequencing and assembly of some other members of genus *Rhodococcus* (*i.e. Rhodococcus qingshengii* Strain BKS 20-40, *Rhodococcus triatomae* strain BKS 15-14 and *Rhodococcus ruber* strain BKS 20-38) have been done to study the different aspects related to each strain. *Rhodococcus qingshengii* strains BKS 20-40, *Rhodococcus triatomae* BKS 15-14 and *Rhodococcus ruber* BKS 20-38 are known to produce cholesterol oxidase and degrade cholesterol into 4-cholesten-3-one. Illumina Hi-Seq 1000 technology was used to sequence these strains. Velvet and SOAPdenovo software produced draft genomes of these microbes that were further annotated by RAST server and PGAAP pipeline of NCBI.

In addition to these bacteria, two fungal genomes (*Debaryomyces hansenii* var. *hansenii* MTCC 234 and *Rhodosporidium toruloides* MTCC 457) have been sequenced and assembled. *Debaryomyces hansenii* var. *hansenii* MTCC 234 is yeast of biotechnological importance and associated with cheese and meat processing also. Analysis of whole genome indicates the presence of 5,313 CDSs, 69 tRNAs and 3 rRNAs in the genome of strain MTCC 234. *Rhodosporidium toruloides* MTCC 457 is oleaginous yeast, having carotenoids, responsible for its red color. It is a lipid storing fungus stores nearly 75% of dry weight under certain conditions. Whole genome annotation of strain MTCC 457 is done with the help of transcripts as evidences. A total of 8,412 transcripts were generated after aligning the transcriptome data (Illumina sequencing data) on to the draft genome by using TopHat and Cufflinks software.

Whole genome sequencing and assembly of *Acinetobacter baumannii* Strain MSP4-16, *Amycolatopsis decaplanina* Strain DSM 44594^{T}, *Streptomyces gancidicus* Strain BKS 13-15, *Arthrobacter* sp. Strain SJCon, *Mycobacterium avium subsp. paratuberculosis* Strain S5 and *Citrobacter freundii* MTCC 1658 have also been performed. Further, annotation

of these mentioned microbes have been performed for different aspects by PGAAP pipeline and RAST server.

Tools for the analysis of NGS data have been developed and implemented in web servers and databases. An all-In-one pipeline named 'Genotrick', was made to, i) Filter the raw NGS data, ii) Assembling of genomes from reads and iii) Annotation of genomic contigs. In this pipeline, popular software like NGSQC toolkit, velvet and Prokka, were integrated for various tasks like raw filtering, assembly and annotation. For benchmarking of genome assemblers, various modules were prepared for both Linux and Windows operating systems and distributed freely through GenomeABC server (http://crdd.osdd.net/raghava/genomeabc/). All Perl based scripts have been implemented in genome assembly benchmarking server *i.e.* GenomeABC.

Whole genome, transcriptome and exome sequencing by NGS technologies generate data in form of short reads. Further, genomic fragments (*i.e.* contigs) and genes are analyzed for specific purposes. Modules to analyze short reads, contigs and genes have been integrated at databases; CancerDR (http://crdd.osdd.net/raghava/cancerdr/) and PCMDB (http://crdd.osdd.net/raghava/pcmdb/).

Present work is important in four aspects; 1) assessment of tools used for genome assembly, 2) analysis of NGS data generated for different genomes, 3) whole genome annotation and identification of important genes, gene-clusters and pathways and 4) development of tools for NGS data analysis and implementation at web servers/databases. In future, this work will be helpful; 1) for the selection of genome assembly software for newly sequenced organism, 2) identification and characterization of important genes from microbial genomes, sequenced in this study, 3) and development of user friendly software packages to simplify the analysis of NGS data.