Polyphasic taxonomy revolutionized the field of microbial taxonomy by providing 'gold standards' for characterizing different organisms (Thompson et al., 2015). The reduction in the cost of the genome sequencing technologies resulted in the increase in the number of genome sequences in public repositories. The taxonomic characterization of microbes for their evolutionary role, clinical aspect, and epidemiology is important as they can undergo a process of recombination that leads to inter-species and intra-species diversity that is not easily traceable (Fraser et al., 2007; Hanage, 2013). In such cases fast and accurate methods of taxonomic characterization are required wherein the laborious and time consuming polyphasic taxonomic methods may not keep up the pace. There are cases where genotypic and phenotypic methods fail to describe the microbes, where different species of a genus cohabit in complex and share gene pool as depicted in cases of *Burkholderia* (Vandamme and Peeters, 2014), *Wolbachia* (Ellegaard et al., 2013), and *Pseudomonas* (Alvarez-Perez et al., 2013). The reduced cost of genome sequencing technologies significantly shifted the paradigm from classical taxonomic methods to the genome based taxonomic characterization of the species (Thompson et al., 2015). Whole genome sequencing also offered identification of protein sequence clustered at high resolution, identification of HGTs either in sympatric or allopatric species; which could not be identified using polyphasic taxonomic protocols (Thompson et al., 2015). This redefinition of clusters could help in providing the species definition and to resolve the species with extensive gene transfers creating fuzzy boundaries (Hanage, 2013; Hanage et al., 2005). The use of sequence data for the species characterization with the advent and advancement of sequencing technologies has changed the perspective of systematists in large to use the genomics information as an additional methodology for species characterization. The technique has been found to be useful in characterizing accurately various species and strains (Chelo et al., 2007; Devulder et al., 2005; Lodders et al., 2005; Naser et al., 2005; Paradis et al., 2005; Richert et al., 2007; Richter et al., 2006; Thompson et al., 2005). To use genomics as a tool for the species characterization up to the strain level of an organism, we performed similar tests to achieve the proposed objectives in this study.

The first objective was to develop a workflow for improving genome assembly and annotation where sequencing and assembly of more than sixty microbes including bacteria, yeast, and fungi (Chapter 2) were performed. Various assemblies using several software *viz*. Velvet, SOAPdenovo, CLCbio, SPAdes were performed with different parameters to find out the most optimum assembly results in terms of N50 value and number of contigs. Adding up the data from different sequencing technologies (Illumina+Roche or Illumina+Iontorrent or IlluminaPE+PacBio) or multilibrary data (IlluminaPE+IlluminaMP or RochsSG+RochePE) and performing hybrid assembly do improve the assembly quality significantly. The hybrid assemblies for *S. boulardii* strains EDRL, and biocodex, and *Grimontia indica* AK16 improved significantly in comparison to their original assemblies. We found that the coverage of shotgun reads had a significant effect on the assembly results and the assembly statistics improved significantly with the increase of data. But there are not significant improvements in the assembly and hence, the shotgun data at 200x coverage is sufficient enough to represent an accurate bacterial genome.

Scaffolding and Gap-filling are the methods by which using PE information an assembly obtained for an organism can be improved significantly. Another method to improve the assembly is by using the assembly outputs from different assemblers to find the overlaps and thus link the contigs in draft assembly, a process called as reconciliation. The method is implemented in various software *viz*. Zorro, e-RGA, scaffold builder, CLC Microbial finishing tool and Minimus. Zorro, e-RGA and Minimus use nucmer as a tool to align the contigs and then fill the gaps whereas CLC Microbial finishing tool aligns the contigs using its mapping algorithm. Ten microbial assemblies were improved significantly using CLC Microbial finishing tool (Chapter 2). The nucmer based algorithms bridge the unknown gap using lower identity regions of contigs which might result in misjoins which was observed in case of *S. boulardii* assemblies. Thus, for this example, we manually checked for the contigs overlap from end to end and found that there was a significant reduction in the number of contigs but N50 value improved only marginally. Also, scaffold builder improved the assemblies of *S. boulardii* quite substantially with stringent parameters (Identity 100% and not adding any ambiguous bases to join the contigs).

One more significant factor on which the assembly results depend on is the genomic DNA isolation. Sometimes two or more organisms co-habiting or with same phenotype gets co-isolated. If such samples are sequenced using short-read sequencing technology then the assembly in such cases is challenging. In our study we sequenced two such cases where one was sequenced using short read technology Illumina HiSeq-1000 PE technology and one with long read PacBio technology. In the sample sequenced using Illumina HiSeq-1000 technology, *Fontibacter* sp. AK8 and *Marinilabila* sp. AK2 were identified which could not be resolved as individual assemblies for both of the organisms. In another case, *Bacillus subtilis* TO-A JPC was co-sequenced with *Enterococcus faecium* T110 strain using PacBio technology; we were able to retrieve the complete genome of *Bacillus subtilis* TO-A JPC and the draft genome of *Enterococcus faecium* T110. Both the genomes assembled were further confirmed for any putative contamination using Onecodex and proteome analysis. The contiguity of the complete genomes obtained was confirmed by comparing the assembly to the *B. subtilis* TO-A genome (CP005997) assembled by other group independently using Illumina shotgun reads.

After analyzing a large number of microbial assemblies we found that there is scope for the improvement of assembly results thus, we tried using both *de novo* and reference mapping methods to improve the assemblies. Single and multiple references assisted *de novo* assemblies were performed for the organisms where the complete genomes of related organisms were already available. The multiple reference assisted *de novo* assembly improved substantially in case of *Escherichia coli* assembly in comparison to its *de novo* assembly.

The second objective Genomics-enabled microbial systematics was framed to achieve the taxonomic characterization of microbes using whole genome information. This part was divided into the taxonomic characterization of bacteria (Chapter 3) and taxonomic characterization of yeast (Chapter 4). In chapter 3, characterization of 33 bacterial genomes was performed using 16S rRNA genes and eight Bacillaceae members were characterized using GyrB protein sequences. Of the bacterial genomes *Vibrio fluvialis* and *Burkholderia* sp. could not be correctly characterized by 16S sequences. We performed the MLSA, proteome hits to NR database, *in-silico* DDH values, and whole genome taxonomy to find the best taxonomic markers of the selected microbes. In the case of *Vibrio fluvialis* strains PG41 and I21563 we found that the *in silico* DDH values, MLSA using six housekeeping genes *ftsZ*, *mreB*, *pyrH*, *recA*, *rpoA* and *topA* and proteome hit based methods helped in the accurate characterization of the species (Khatri et al., 2013d). Similarly, *Burkholderia* sp.

AU4i was characterized as the strain of *B. vietnamiensis* with 16S and was identified closer to *Burkholderia cenocepacia* with RecA gene and MLSA based analysis, but the whole genome-based phylogenetic analysis using Gegenees and proteome hits to NR database revealed that the organism is closer to *B.* sp. 383 and hence was characterized as *Burkholderia* sp. AU4i (Devi et al., 2015).

We sequenced three probiotic members of the Bacillaceae family namely *Bacillus clausii*, *Bacillus subtilis* and *Bacillus coagulans*. The genus *Bacillus* has the members with varied physiological differences from pathogenicity to biological control agents to economically and medically important strains. Thus, phylogenetic analysis of all the members of genus *Bacillus*, *Clostridium* and *Listeria*, *Lactobacillus* and family Bacillaceae was performed using various molecular markers *viz.* GyrB, RpoB, Tuf, GroEL, and several combinations of these markers. Also, the whole genome and core set of 26 housekeeping proteins were used to build the phylogeny for these organisms. The phylogenetic tree obtained by 26 housekeeping genes could resolve the species-specific clades i.e. could resolve *Bacillus anthracis*, *Bacillus cereus* and *Bacillus thuringiensis* into separate clades. *Bacillus subtilis* TO-A JPC was present in *Bacillus subtilis* clade whereas *Bacillus coagulans* S-lac was present in a separate clade. In the phylogenetic tree *Bacillus clausii*, *Bacillus lehensis*, *Bacillus pseudofirmus*, *Bacillus gelatini*, *Bacillus macauensis* and *Bacillus methanolicus* are grouped with the members of Bacillaceae family which represent these members as the outgroup *Bacillus*. From all these case studies we can conclude that though the 16S rRNA based phylogeny can resolve the organism at genus level but it may not be a good molecular marker in cases where several different species shares large protein content e.g. in cases of *Burkholderia cepacia* complex (Baldwin et al., 2005) or *Bacillus anthracis-cereus-thuringiensis* clade (Helgason et al., 2000; La Duc et al., 2004).

The taxonomic characterization of fungi (Objective 2, Chapter 4) was performed using 5.8S-ITS rDNA region, D1/D2 26S rDNA region, RNA polymerase, β-tubulin, γ-actin, ATP synthase, and elongation factor EF-1α; single copy genes and their MLSA. ITS sequences of most of the fungal and yeast sequences characterized the species sequenced in this study. Our study was specific to the *S. boulardii* conspecific to *S. cerevisiae*, where the identification of molecular markers to demarcate the probiotic strains in phylogenetic tree as compared to brewer and baking yeasts is challenging. Thus, to trace the phylogenetic markers for identifying the separate clade of the probiotic yeast *S. boulardii* in comparison to brewer and baking strains of *S. cerevisiae* we performed phylogenetic analysis using various molecular markers and found that most of the phylogenies obtained were unresolved owing to >99% genome relatedness among the strain of *S. cerevisiae.* We concluded here that the phylogenetic study incorporating all the orthologs from all *S. cerevisiae*, *S. boulardii,* and outgroup strains presented a more accurate phylogenetic tree where the probiotic strains shared the clade with the wine strains of *S. cerevisiae.*

Objective 3 incorporating the study of genomic properties of probiotic yeast *S. boulardii* was framed to identify the genomic differences in probiotic strains as compared to the brewer and baking strains of *S. cerevisiae.* The complete genomes of *S. boulardii* biocodex and unique28 were assembled with the long PacBio reads and three draft genomes for *S. boulardii* strains EDRL, kirkman and unisankyo were determined. The genome sequencing with long read technology assisted identifying the complete MATa and MATα locus that represents diploidy; and also the plasmid sequences were identified which were closer to 2 μ circle plasmid of *S. cerevisiae* strain YJM993. The complete absence of Ty1/2 elements in *S. boulardii* was mentioned in previous studies (Edwards-Ingram et al., 2007) to be one of the distinguishing features in comparison to *S. cerevisiae* but our analysis revealed the presence Ty2 and Ty5 elements in the complete genomes of *S. boulardii.* Another distinguishing feature was the incapability of *S. boulardii* to utilize galactose as a source of carbon and palatinose as compared to *S. cerevisiae* (Lukaszewicz, 2012; McCullough et al., 1998; McFarland, 1996; Mitterdorfer et al., 2001; Sellick et al., 2008); but we found that all the galactose-metabolizing enzymes of Leloir pathway were present in the genome. The proteins for palatinose utilization IMA2, IMA3 and IMA4 proteins were found to be absent in all the strains of *S. boulardii* but IMA1 and IMA5 were present. The complete set of *S. cerevisiae* genes from yeastmine database (6604) (Balakrishnan et al., 2012) were mapped to the *S. boulardii* genomes and found that 144 genes (85 dubious ORFs, 32 uncharacterized genes, and 27 functionally verified genes) were absent. Most of these genes were either telomeric or subtelomeric. Our analysis also revealed that the PAU proteins, a member of the seripauperin multigene family, were present in 18-20 copies *i.e.* maximum in the genome, and gag-pol fusion genes were present in 15 copies in the whole genomes of *S. boulardii* strains biocodex and unique28. The variation in the number of copies of these genes may be related to the phenotypic and physiological differences in between probiotic and brewer yeasts (Landry et al., 2006). All flocculins except FLO5 and adhesins were present across all strains of *S. boulardii*, and we found that these have a larger number of repeats comparable to most of the *S. cerevisiae* strains. The 54 kDa protease (Castagliuolo et al., 1999), 63 kDa phosphatase (Buts et al., 2006), and 120 kDa proteins (Czerucka and Rampal, 1999) of *S. boulardii* were reported earlier to exhibit antimicrobial effect against pathogenic bacteria, and were found to be present in all strains of *S. boulardii* and *S. cerevisiae* (Khatri et al., 2013a).

We compared our five strains of *S. boulardii* to the publically available two strains of *S. boulardii* and 145 strains of *S. cerevisiae*. We found five *Z. bailii* ISA1307 proteins have introgressed and further undergone duplication in *S. boulardii* strains. Three of these proteins were annotated as uncharacterized proteins; one is a probable 5-oxoprolinase, and one is involved in allantoate transport. These proteins were also present in *S. cerevisiae strain* UFMG A-905; wine strains YJM339, RM11-1a, L1528, YS9, EC1118, Vin13, VL3, AWRI796 and LalvinQA23; and bioethanol producing strain *Sc* JAY291. The ASP3 locus absent from probiotic strains was also likewise absent in wine and bakery strains of *S. cerevisiae* but was present in industrial *S. cerevisiae* strains. The phylogenetic studies using core genes of the *S. cerevisiae*, *S, boulardii* and outgroup species revealed that *S. cerevisiae* strain UFMG A-905 extracted from cachaca (distilled spirit made from sugarcane) was always found close to *S. boulardii* strains in phylogenetic analysis and has been reported to exhibit the probiotic properties (Martins et al., 2011). From the phylogenetic analysis and the genomic properties, we predict that *S. cerevisiae* wine strain BC187 could be a putative probiotic strain. The study revealed that the probiotic yeast has no marked molecular and genetic differences as compared to its conspecies *S. cerevisiae* regarding the probiotic molecules identified previously but with no doubt, the physiological differences have been known for the organism.

Objective 4 incorporating the study of genomic properties of probiotic bacteria *B. clausii* ENTPro, *B. coagulans* S-lac and *B. subtilis TO-A JPC* extracted from probiotics sold in the market was framed to identify the genomic differences in probiotic strains as compared to the other strains of their similar species and probiotic strains of *Bacillus* and *Lactobacillus* genus. Several spore-forming strains of *Bacillus* are marketed as probiotics due to their ability to survive harsh gastrointestinal conditions and confer health benefits to the host. The complete genomes of three probiotic spore-forming, phylogenetically distinct bacteria, *Bacillus clausii*, *Bacillus subtilis* and *Bacillus coagulans* were determined. Also, a circular plasmid sequence was retrieved for *B. clausii* ENTPro. From the proteome analysis of all the three probiotic genomes we found that there were no putative HGTs and unique proteins in *B. subtilis* TO-A JPC whereas *B. coagulans* S-lac *and B. clausii* ENTPro has a large unique repertoire with 117 and 184 unique proteins, respectively. Our analysis revealed the presence of proteins that might play a role in probiotic function such as adhesins, sporulation proteins and stress-responsive proteins in these probiotic genomes. Two cyclized peptides were identified as bacteriocin clusters in *B. subtilis* TO-A JPC along with four CDS coding for subtilosin bacteriocins, all of which are well-characterized. We found two head to tail cyclized bacteriocins in *B. coagulans* S-lac, of which one is a novel bacteriocin as it did not match any sequence in the NR database. Novel bacteriocins belonging to lanthipeptide Class I was identified in *B. clausii* ENTPro. A larger numbers of MFS and other efflux transporters were identified in *Bacillus* probiotics as compared to *Lactobacillus* probiotics. Tetracycline MFS efflux and class A and class D domains of β-lactamases were absent from *B. coagulans* and *Lactobacillus* spp. whereas they were present across other *Bacillus* probiotics. Type II CRISPR/cas system with one CRISPR repeat and seven *cas* genes were identified in *B. coagulans* S-lac genome whereas only cas6 gene was present in *B. clausii* ENTPro. Instead of CRISPR system, the presence of RNase III system in *B. subtilis* suggests defense against invading foreign DNA. Our analysis revealed the presence of two copies of Type I RM system and one copy of Type III RM system in *B. coagulans* S-lac genome, Type I RM system in *B. clausii* ENTPro, whereas *B. subtilis* TO-A JPC has only the Type II RM system. We found a Type III RM system was present exclusively in the *B. coagulans* strains, S-lac and GBI-30, and in *L. casei*. We suggest a possible horizontal transfer of this Type III RM system from *Clostridium kluyveri* in the probiotic strains of *Bacillus coagulans*.

In our study, the primary objective was to utilize the genomic information for phylogenetic analysis of bacteria and yeast. We found that the highest accuracy and resolution of the species and conspecies in the bacterial clade could be achieved by either whole genome-based methods using *in silico* DDH value; or by a core set of housekeeping genes and conserved orthologs. The molecular markers used for the phylogenetic analysis differs for the species in question. For the conspecific yeast strains we found that only core set of orthologous proteins can achieve the resolution and accuracy to decipher completely a separate clade whereas the whole genome-based methods cannot be used in such cases due to very high similarities in the genomes which groups everything together.