Summary

Summary

The human gut is densely populated by microbes which have reportedly incredible effects on the development and health of the host. The gut microbial community structure is known to be altered in many diseased conditions. Differences in gut microbial community structure have also been reported in subjects belonging to different geographical regions. Diet is one of the major factors that influences the gut microbiome composition. This is understandable considering that a major part of the human diet consists of carbohydrates and that humans do not synthesize the enzymes required to metabolize these complex carbohydrates but are dependent on their gut symbionts for their metabolism. The gut microbes with their impressive array of over 15,000 CAZymes help in the metabolism of these complex carbohydrates. Glycan binding proteins encoded by the gut microbes also facilitate adherence and host colonization. **Chapter 1** is a literature review, introducing various topics related to this thesis - metagenomics, gut structure and function, human gut glycan landscape, mucin, human microbiome, gut microbes and glycans, carbohydrate-binding proteins, phage display library, biopanning, biopanning ligands and glycan arrays. It also outlines the objectives of the thesis.

India is the second most populous country with diverse ethnicities, cultures, dietary habits and environmental conditions. Little knowledge about the gut microbiota of Indian population was available during the inception of this study. Besides, the gut microbial metagenome is a veritable treasure house that can be exploited for the discovery of novel proteins and enzymes. So, this study was designed with the aims of understanding the microbial diversity of the gut microbiota in Indian subjects, and exploring and exploiting the human gut metagenome for novel carbohydrate binding proteins with potential for use in various glycobiology applications.

Summary

Chapter 2 deals with the first objective of this study, which was to gain knowledge about the gut microbial diversity in healthy subjects from three different rural regions of India (Ladakh, Jaisalmer and Khargone), varying vastly in environment, culture, dietary habits, and life-style. We isolated metagenomic DNA from the feces and sequenced 31 samples (Khargone, n = 12; Jaisalmer, n = 10; Ladakh, n = 9). The sequencing was performed on the Illumina platform. The reads were analyzed by One Codex and MetaPhlAN to understand microbial diversity and community composition in the subjects from different regions of India. We observed that *Prevotella*, *Bifidobacterium*, *Bacteroides*, *Eubacterium* and *Faecalibacterium* were the five most abundant genera in the Indian subjects from all the three regions studied. *Prevotella* was the most abundant genus in all the three regions of the Indiam, constituting over 40% of all the genera identified. *Listeria* was the only genus which showed significant variation across the three regions, being more abundant in Ladakh subjects as compared to the Khargone and Jaisalmer.

When we compared the microbial community structure of the Indian subjects with subjects of other nationalities we found that the Indian subjects had very distinct gut microbial community composition as compared to healthy subjects of China, Denmark, France, Germany, USA (HMP), and Spain. We observed significant differences in the abundance of 56 genera across the different countries. *Lactobacillus* and *Prevotella* were the two genera which were more abundant in Indian subjects as compared to the subjects of all other nationalities.

Next, we analyzed CAZyme and CBM profiles. We identified a total of 105 Glycoside hydrolases (GHs), 19 Polysaccharide Lyases (PLs), 15 Carbohydrate Esterases (CEs), 76 Glycosyltransferases (GTs), and 6 enzymes of Auxiliary Activities (AAs) in Indian subjects. The CAZyme and CBM profiles looked similar across all the three regions. When we compared the CAZymes and the CBMs of Indian subjects with those of subjects from other nations, we

observed that the repertoire of CAZymes and CBMs were similar across the subjects but there was a significant difference in the relative abundance of 96 CAZymes and 24 CBMs.

Chapter 2 deals with the second objective, which was to mine novel glycan binding proteins from the human gut microbiome. So, the second objective of this thesis dealt with the construction of a gut metagenomic DNA T7 phage display library, optimization of the screening protocol, and screening and identification of sequences binding to mucin, glycogen, PAA-sugars, and many microbial and plant derived glycans. To do this, we first generated a metagenomic phage display library having $1x10^7$ recombinant phages accommodating around 10 Gbp of metagenomic DNA, almost equivalent to ~2000 bacterial genomes. Then, in order to optimize our biopanning procedure to minimize false-positives and non-recombinant phages after screening, we generated a "positive control" – a recombinant fucose binding T7 10-3b phage (SrNaFLD-T7 phage). Subsequently, to optimize and validate the biopanning protocol, we mixed different ratios of recombinant phage (SrNaFLD-T7 phage; which binds to the ligand) and non-recombinant phage (T7 10-3b phage; which is not expected to bind to the ligand) and biopanned these mixtures against mucin. Starting biopanning with 100 µl of a mixture of $1x10^3$: $1x10^{11}$ pfu/ml of *Sr*NaFLD-T7 phage: T7 10-3b phages with a multiplicity of screening of just 100, we found that 6 rounds of biopanning were required for enrichment amounting to at least one true recombinant binder out of ten phages. Using phage mixtures with higher proportions of recombinant phages resulted in more enrichment, as expected. We also observed that the natural ligand was a better elution agent than SDS.

Following this validation and optimization of the biopanning protocol, we screened the metagenomic phage display library against 25 glycoconjugates. We were able to sequence 32 unique clones that were enriched following biopanning rounds against various glycans. Many of

these sequences shared similarity with proteins that were unannotated with respect to function, i.e., hypothetical proteins or proteins with putative functions. Some had remote similarities to proteins involved in carbohydrate metabolism. Thus, our function-based metagenomics approach helped in the identification and functional assignment of many novel carbohydrate-binding proteins.

Chapter 3 describes the cloning, expression, purification, and analysis of the glycan binding specificities of the polypeptides obtained following biopanning against mucin. This was done to serve two purposes. First, to affirm that the polypeptides with unknown functions obtained during screening were true glycan-binders. Second, to understand the glycan specificities of these novel mucin-binders, as mucin has complex N-linked and O-linked glycans. We were able to successfully clone, express and purify the four mucin glycan binding polypeptides (MG1, MN3, MU1, and MU3). To find the glycan specificities of the polypeptides, glycan microarray analyses were performed. We found that MG1 bound to a total of 39 glycans out of the 600 glycans on the glycan microarray. The majority of the glycans that bound to MG1 were complex and biantennary with an N-linked core structure and LacNAc extensions on both arms, and with or without core-fucosylation. MN3 bound to a total of 42 glycans. MN3 displayed binding to glycans with terminal LacNAc (Galβ1-4GlcNAc) units terminating with galactose ("i" blood group antigen) and preferred short LacNAc units. MU1 bound to a total of 64 glycans and displayed preferential binding to glycans with LacNAc (Galβ1-4GlcNAc) units terminating with galactose ("i" blood group antigen). It preferred long LacNAc chains. MU3 bound to a total of 11 glycans and displayed preferential binding to glycans with complex, multi-antennary, short N-glycan type structures (with a single LacNAc unit on each arm) with terminal sialic acids on all arms in α2-6 linkages. MU3 did not bind to glycans with terminal sialic acids in α2-3 linkage,

and did not also bind to biantennary N-glycans with an α2-3 linked sialic acid on one arm and an α2-6 linked sialic acid on the other arm. Thus we validated that the clones obtained following biopanning against mucin were indeed glycan binding proteins, and also determined detailed glycan binding specificities of these mucin binding clones.

To conclude, the study described in this thesis employed both sequence-based and function-based metagenomics approaches to study the human gut microbiota of Indian subjects. We provide insights into the distinctness of the gut microbial community structure of Indian subjects from different geographic regions as compared to the rest of the world, and analyze carbohydrate-active enzymes and carbohydrate-binding modules in the gut microbiota that facilitate adherence to host glycoconjugates and dietary polysaccharides. We also present 32 novel glycan-binding proteins mined from the metagenome of the gut microbiome by biopanning a metagenomic phage display library against various glycoconjugates, and we provide the detailed glycan binding specificities of four of these clones. It is hoped that this work will pave the way for future studies aimed at better understanding the numerous microbes that call our body home and make us veritable super-organisms.