HuMiChip: Development of a Functional Gene Array for the Study of Human Microbiomes

Qichao Tu[1], Ye Deng[1], Lu Lin[12], Jian Xu[2], Chris L. Hemme[1], Zhili He[1], Jizhong Zhou[1]

[1]The University of Oklahoma, Norman, OK, [2]Qingdao Inst. of BioEnergy and Bioprocess Technology, Qingdao, CHINA

MS-2476

IEG Institute for ENVIRONMENTAL GENOMICS http://ieg.ou.edu/

QIBEBT CHINESE ACADEMY OF SCIENCES Qingdao Institute of Bioenergy and Bioprocess Technology CHINESE

ABSTRACT

Microbiomes play very important roles in terms of nutrition, health and disease by interacting with their hosts. Based on sequence data currently available in public domains, we have developed a functional gene-based array, human microbiome chip (HuMiChip) to monitor both organismal and functional gene profiles of normal microbiota in human hosts. In this study, we have included 322 genomes of bacterial strains from different human body sites, and 27 human gut metagenomes. First, seed sequences were identified from KEGG databases, and used to construct a seed database (seedDB) containing 139 gene families in 20 metabolic pathways closely related to human microbiomes. Second, a mother database (motherDB) was constructed with the 322 bacterial genomes and 27 metagenomes, and used for gene selection and probe design. Gene prediction for metagenomes was carried out by the MG-RAST server. In total, there are 913,192 and 2,157,747 sequences for bacterial genomes and metagenomes, respectively. Then the motherDB was searched against the seedDB using the HMMer program, and gene sequences in the motherDB that were highly homologous with seed sequences in the seedDB were used for probe design by the CommOligo software.In addition, mannual inspection for the HMMer output was carried out, and keywords for each gene were designed and used. Different degrees of specific probes, including gene-specific, inclusive and exclusive group-specific probes were selected. All candidate probes were checked against the motherDB and NCBI databases for specificity. Finally, 36062 probes covering 47979 sequences were designed, and this HuMiChip is expected to be able to detect the diversity and abundance of functional genes, the gene expression of microbial communities, and potentially, the interactions of microorganisms and their hosts.

EXPERIMENTAL DESIGN

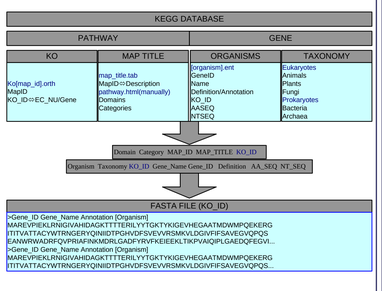
Sequence Collection

Full and partial genome sequences for human microbial bacteria were collected from several different public available domains. 266 draft genomes sequences were downloaded from NCBI human microbiome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/HUMAN_MICROBIOM/). 34 fully annotated and finished genome sequences reported as human microbiomes were collected from NCBI bacteria database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). In addition, 22 oral human oral specific genomes, which were not available in NCBI database yet then, were downloaded from HOMD (http://www.homd.org/) and oralgen database (http://www.oralgen.lanl.gov/). 27 human gut metagenome datasets were annotated and downloaded from MG-RAST server.

Seed database construction

Seed database was consisted with full length protein sequences with already known function and were be used to build HMMer seed models. Here the seed database was constructed based on KEGG databases including pathway database and gene database. Pathways and genes that play important roles in microorganism were manually selected for seed database construction. In total, 19 pathways, 136 genes and 13814 sequences were selected and used. The data retrieving and integration process is shown in Fig.1.

Fig. 1



MotherDB construction

Our motherDB included 322 genomes of bacterial strains from various human body sites, and 27 metagenomes. Gene prediction was performed by Glimmer3 for unannotated bacterial genomes, and by the MG-RAST server for metagenomes. In total, **913,192** and **2,157,747** gene sequences were identified for bacterial genomes and metagenomes, respectively. Then the motherDB was searched against the seedDB using the HMMer program, and gene sequences in the motherDB that were highly homologous with seed sequences in the seedDB were used for probe design by the CommOligo software.

Oligonucleotide Probe design

The computer program *CommOligo3.0* (Li et al., 2005) was used to design different degrees of specific probes, including gene-specific, inclusive and exclusive group-specific probes based on the following criteria: (i) gene-specific probes: <=90% sequence identity, <=20-base continuous stretch, and >=-35 kcal/mol free energy; (ii) group-specific probes: >=96% sequence identity, >= 35-base continuous stretch, and <=-60 kcal/mol free energy. Each gene sequence or a group of homologous sequences had up to three probes. All candidate probes were checked against the motherDB and NCBI databases for specificity.

RESULTS

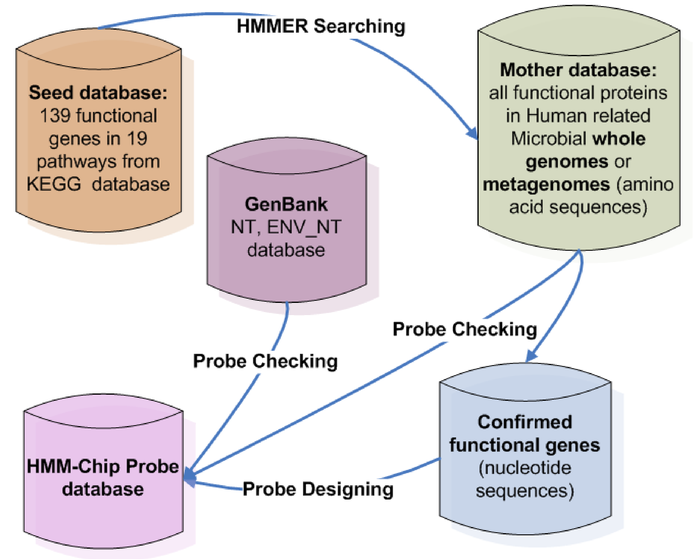


Fig. 2 Major steps for construction of a comprehensive HMM functional gene array. *CommOligo* is the core program to select gene-specific, inclusive and exclusive group-specific oligonucleotide probes. The designed probes are also checked with NCBI and MotherDB for specificity using BLAST.

REFERENCES

Li X*, He Z* and Zhou J (2005). *Nucleic Acid Res.* **33:** *6114-6123 (*Co-first authors).*

http://www.oralgen.lanl.gov/

http://www.homd.org

http://www.hmpdacc.org/

http://metagenomics.nmpdr.org/

ACKNOWLEDGEMENTS

This work was supported by supported by the Genomics:GTL program through the Virtual Institute of Microbial Stress and Survival (VIMSS; http://vimss.lbl.gov), and the Oklahoma Applied Research Support (OARS), Oklahoma Center for the Advancement of Science and Technology (OCAST), the State of Oklahoma through the Project AR062-034.

Work was supported, in part, under DOE Contract No. DE-AC02-05CH11231.

***MotherDB*. Complete or partially complete genomes were collected from various sources including NCBI human microbiome database, NCBI bacterial genome database, HOMD database and oralgen database. *Metagenome sequences were collected from MG-RAST server.**

Table 1. Summary of HuMiChip probes by covered microbial phylum and classes

Phylum	Class	Genome Num	Probe Num	Covered CDS
Actinobacteria	Actinobacteria	44	2919	2677
Bacteroidetes	Bacteroidia	46	5222	4006
	Flavobacteria	4	325	300
	Sphingobacteria	2	275	240
Euryarchaeota	Methanobacteria	3	122	90
Firmicutes	Erysipelotrichi	9	726	638
	Clostridia	64	6761	5695
	Bacilli	75	4046	3725
Fusobacteria	Fusobacteria	16	250	232
Proteobacteria	Gammaproteobacteria	21	1593	1461
	Alphaproteobacteria	1	135	132
	Deltaproteobacteria	1	86	74
	Epsilonproteobacteria	12	724	634
	Betaproteobacteria	12	625	578
Spirochaetes	Spirochaetes	4	200	193
Synergistetes	Synergistia	4	111	108
Tenericutes	Mollicutes	1	107	94
Unclassified	Unclassfied	3	52	55
Metagenome*	Human Gut*	27	25550	27124

Table 2. The summary of probes, covered genefamilies and gene sequences on HuMiChip

Functional process	No. of gene categories	No. sequences retrieved	No. of probes designed	No. CDS covered
Amino acid synthesis	36	31914	9002	12550
Amino acid transport and metabolism	26	17589	5373	7747
Central Carbon Metabolism	7	4967	1596	2053
Cofactor Biosynthesis	11	5009	1815	2473
Complex Carbohydrates	12	6611	2337	2847
Exotic Metabolisms	3	1591	637	794
Fatty Acid Biosynthesis and Metabolism	2	1659	572	839
Feeder Pathways to Glycolysis	10	6411	2266	2994
Glycan Biosynthesis and Metabolism	10	10021	3774	4606
Glycan structures degradation	2	3758	1476	1505
Glycerolipid Metabolism	1	1366	459	530
Glycosaminoglycan degradation	3	2714	1140	1490
Isoprenoid biosynthesis	3	2480	869	1336
N-Glycan degradation	1	656	270	405
Nitrogen Metabolism	3	601	274	327
Organic Acids	9	4820	1492	1865
Purine metabolism	6	6508	1624	2672
Pyrimidine metabolism	8	10518	2969	4042
Respiration	1	2580	581	747
Total	**139**	**113233**	**36062**	**47979**

CONCLUSIONS

1. **The developed HuMiChip contains 36062 probes covering 47979 bacterial sequences from different human body sites, targeting 19 most important pathways and 139 gene families for microbial bacteria.**
2. **This HuMiChip is able to detect the diversity and abundance of functional genes, the gene expression of microbial communities, especially in human gut, and potentially, the interactions of microorganisms and their hosts.**

LEGAL DISCLAIMER