Database, Vol. 2010, Article ID baq013, doi:10.1093/database/baq013

Original article

The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information

Tsute Chen[1,2,*], Wen-Han Yu[1], Jacques Izard[1,2], Oxana V. Baranova[1], Abirami Lakshmanan[1] and Floyd E. Dewhirst[1,2]

[1]The Forsyth Institute, Boston, MA 02115, USA and [2]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, MA 02115, USA

***Corresponding author:** Tel: +1 617 892 8359; Fax: +1 617 262 5200; Email: tchen@forsyth.org

Submitted 25 January 2010; Revised 28 May 2010; Accepted 20 June 2010

The human oral microbiome is the most studied human microflora, but 53% of the species have not yet been validly named and 35% remain uncultivated. The uncultivated taxa are known primarily from 16S rRNA sequence information. Sequence information tied solely to obscure isolate or clone numbers, and usually lacking accurate phylogenetic placement, is a major impediment to working with human oral microbiome data. The goal of creating the Human Oral Microbiome Database (HOMD) is to provide the scientific community with a body site-specific comprehensive database for the more than 600 prokaryote species that are present in the human oral cavity based on a curated 16S rRNA gene-based provisional naming scheme. Currently, two primary types of information are provided in HOMD—taxonomic and genomic. Named oral species and taxa identified from 16S rRNA gene sequence analysis of oral isolates and cloning studies were placed into defined 16S rRNA phylotypes and each given unique Human Oral Taxon (HOT) number. The HOT interlinks phenotypic, phylogenetic, genomic, clinical and bibliographic information for each taxon. A BLAST search tool is provided to match user 16S rRNA gene sequences to a curated, full length, 16S rRNA gene reference data set. For genomic analysis, HOMD provides comprehensive set of analysis tools and maintains frequently updated annotations for all the human oral microbial genomes that have been sequenced and publicly released. Oral bacterial genome sequences, determined as part of the Human Microbiome Project, are being added to the HOMD as they become available. We provide HOMD as a conceptual model for the presentation of microbiome data for other human body sites.

Database URL: http://www.homd.org

Introduction

The human oral microbiome is the most studied human microflora, due to the fact that it is easily sampled and is strongly associated with important oral infectious diseases such as tooth decay (dental caries) and gum disease (periodontitis). Approximately 600 prevalent bacterial species have been identified in human oral cavity (1) based on early cultivation studies and later culture independent 16S rRNA gene clonal analyses.

The investigation of such complex biofilms is confronted by two issues: to properly identify which bacteria are in the biofilm, and to understand their genetic potential in human health and disease. International efforts have launched the microbiome era—the systemic study of a microbial community as a functional entity (2–5). The National Institutes of Health Human Microbiome Project (HMP) is a major contributor to this international effort (2). A key aim of the HMP is the production of 1 000 microbial reference genomes, including about 300 from the oral cavity (6). As a community resource project, the HMP is also committed to rapid data release and has established a Data Analysis and Coordinating Center (DACC) (http://www.hmpdacc.org) to coordinate the storage, raw data distribution and sharing

Database, Vol. 2010, Article ID baq013, doi:10.1093/database/baq013

of all the conclusions generated by HMP-funded programs (6). HMP investigators looking at skin, oral, gastrointestinal or genitourinary tract also confront the issues of taxonomic classification of many newly identified organisms, as well as needing to provide resources for comprehensive analysis of body site-specific data.

Based on version 10 of the HOMD 16S rRNA gene reference sequences (described below), 47% of taxa are validly named species, 18% are unnamed isolates and 35% are unnamed and uncultured phylotypes known primarily from 16S rRNA sequence information. Sequence information tied solely to obscure isolate or clone numbers, and usually lacking accurate taxonomic anchor, is a major impediment to working with human oral microbiome data. The lack of association between a clone number in a sequence repository, the site from which it was isolated and the disease status of the subject further impedes research advancement. Many of the reference genome sequences being obtained by the HMP are not yet validly named species. Thus, there has been an urgent need to create a provisional taxonomic scheme and database for the unnamed members of the oral microbiome.

Our research over the past 20 years has focused on defining the breadth and diversity of the oral microbiome by obtaining 16S rRNA gene sequence information for both cultivable and as yet uncultivated oral bacteria (7). We have also been participating in genome sequencing of oral bacterial isolates (8–10), and are now continuing as contributors to the HMP of both reference DNA and strains of oral organisms. In this report, we describe the implementation of the Human Oral Microbiome Database (HOMD) specifically designed to provide a provisional naming scheme where each oral taxon is given an oral taxon number linked to comprehensive information and tools for examining and analyzing each taxon in the human oral microbiome at both the taxonomic and genomic level. Organisms of the human oral cavity are organized in a taxonomy hierarchy, which leads to individual pages for every oral taxon with comprehensive information and links. The genomic component of HOMD contains both static and dynamically updated annotations as well as bioinformatics analysis tools for all the genomic sequences, and curated 16S rRNA gene reference sequences for all the human oral microbes. HOMD may serve as an example of a body site-specific tool for other communities. The HOMD web site can be publicly accessed at http://www.homd.org.

Database creation and setup

The HOMD database and associated web site were created under contract U01 DE016937, 'A Foundation for the Oral Microbiome and Metagenome', from the National Institute of Dental and Craniofacial Research. The goal of the contract was to create a stable taxonomic structure for the unnamed oral taxa and to provide tools for analyzing 16S rRNA sequence data and oral genome data.

Database creation

The HOMD describes information linked to oral microbe species. For bacteria or archaea, which have not been validly named, there is no definition of 'species'. Molecular methods to identify novel species generally have used 16S rRNA sequencing of isolates or 16S rRNA-based analysis of clone libraries. These strains or clones can then be clustered into phylotypes or taxa based on their 16S rRNA sequences. Phylotype can be defined for any similarity cutoff, but we chose to use 98.5% similarity as that was the approximate level that defined species level clusters for most oral bacteria. Each validly named species and novel phylotype cluster was given a unique Human Oral Taxon (HOT) ID number (starting from 001). Every item in HOMD is keyed to HOT IDs.

16S rRNA reference set and provisional taxonomy

The basic list of oral bacteria came from the literature, primarily reports from Forsyth Institute investigators (11–15) and from Lillian Holderman Moore and Ed Moore (16–18) formerly at the Anaerobe Laboratory at the Virginia Polytechnic Institute. 16S rRNA sequences for these named species came either from sequences obtained in our laboratory or from GenBank. Over the past 20 years, our laboratory constructed and sequenced over 600 16S RNA gene libraries and obtained over 35 000 clone sequences. The samples for these libraries came from healthy subjects and subjects with over a dozen disease states such as caries, periodontal disease, endodontic infections and oral cancer. Clones were initially sequenced with a 519–533 reverse primer and those that differed from known reference sequences were fully sequenced using multiple primers on each strand. Several hundred apparently novel full 16S rRNA sequences were the starting point for defining novel oral taxa. The cloning, sequencing, aligning, treeing and clustering methods used to create HOMD are described in detail in a manuscript submitted elsewhere (F. E. Dewhirst *et al*., submitted for publication). Briefly, sequences were manually aligned in a secondary structure-based database using the program RNA (19). Distance matrices and neighbor joining trees were generated to determine the clustering of sequences. Sequences $>$98.5% similar were grouped together into single taxa. Sequences were extensively checked for chimeras and several sequences and some provisional taxa were removed. This analysis led to the creation of 619 Human Oral Taxa in the initial release of the HOMD database. The 753 reference 16S rRNA gene sequences upon which this analysis was done have been released publicly for download on the HOMD web site as version 10.

Genome analysis tools

A second goal of the NIH contract was to obtain survey sequences (~400 000 bp) for 100 phylogenetically representative oral taxa (8). Tools for display and analysis of the resulting sets of contigs for each organism did not exist. The genome analysis tools of HOMD were created by expanding those previously developed at Forsyth (20). After completing 12 survey sequences (8), the NIH contract was redirected to providing DNA to the HMP sequencing centers to obtain high coverage genomes. HOMD now has annotations for survey, high coverage and full-length oral genomes.

Overview of software and hardware

Software—web service, common gateway interface and database

The computational services provided by HOMD can be categorized into four major categories: web, database, file storage and computation pipeline. The primary software backbone is the LAMP system: Linux, Apache, MySQL and PHP, which provides operating system, web service, relational database and dynamic web page rendering, respectively. PHP also serves as the common gateway interface (CGI) to the relational databases. Load balancing of the LAMP services is provided by the IP-based Virtual Service (IPVS) of the Linux Virtual Server (LVS) project (http://www.linuxvirtualserver.org). File storage is provided by the software-based Distributed Replicated Block Device (DRBD, http://www.drbd.org) to ensure the data integrity by two independent fail-over servers carrying two hard drives that mirror each other. The highly available computational services are provided by the Heartbeat program of the Linux HA project (http://www.linux-ha.org). The TORQUE resource management and the MAUI cluster scheduler (http://www.clusterresources.com) are used to control batch job submission and distribution among the cluster node. All software providing these major services are open source and GNU compliant.

Hardware—computer servers and infrastructure

The HOMD computer servers comprise an array of interacting computer servers. The LAMP system as described above is run on two load-balancing eight-core Xeon servers. The relational database system and the high availability file storage system DRBD are hosted on two fail-over and load-balancing Intel quad-core servers. A cluster of eight Intel quad-core personal computers (a total of 32 CPU cores) is managed by a head node using the TORQUE-MAUI cluster software for handling the high computational demand of automatic genomic annotation pipeline as well as BLAST requests from web users.

Database schema

The taxonomy and the genomics relational databases are linked by taxon IDs and the genome IDs, respectively (see below). A peripheral database storing the individual genome annotations is linked by the genome IDs. The overall database schema and table structures can be viewed and downloaded from the HOMD web site, as described below. The contributors and curators accounts, as well as the user's grouping and privilege control are managed with a separate database based on the open-source PostNuke content management (http://www.postnuke.com).

Curation

Curation of the HOMD is carried out by the project investigators T. Chen, F.E. Dewhirst, J. Izard, B.J. Paster, A.C.R. Tanner and W.G. Wade. Each item on the Taxon Description pages, such as synonyms, descriptions and bibliography, is written or reviewed by these project investigators. The placement of novel taxa in the hierarchical taxonomy is based upon locating the position of the taxa in a tree of oral and non-oral reference sequences.

Summary of current database content

At the time of writing, HOMD contains a total of 619 described and curated human oral taxa. Of these, 151 taxa have at least one genomic sequence and annotation available and an additional 65 have a genome-sequencing project in progress. HOMD lists a total of 747 human microbial genomes in the Sequence Meta Information pages. Of the 747 genomes listed, 50 have dynamic annotation, 477 have static annotation and the remaining are in progress. Of these 527 annotated genomes, 181 have complete genomic sequences, 19 have only partial survey sequences and 327 are high-coverage genomic sequences. Metadata accompanying genomes are obtained from the original sequence repository and are provided as is.

Web representation and tools

Overview

HOMD contains various types of information on the human oral microorganisms including taxonomy, genomics and bibliography. The purpose of the HOMD web site is to provide interfaces to search, retrieve and navigate among these different types of information. HOMD also provides web-based software tools for data-mining and analysis. Different types of data for the same organism are linked together by unique HOT IDs, which appear in all the web interfaces and results of the analytic tools. The interfaces and tools are listed in a left-side menu on the HOMD home page as well as in a drop-down menu on top of every HOMD web page. Detailed descriptions for these interfaces and tools are provided in this section.

Database, Vol. 2010, Article ID baq013, doi:10.1093/database/baq013

Taxon Table

The Taxon Table provides the tabularized view of all the human oral microbial taxa defined and curated by HOMD. The table consists of five columns of information: Oral Taxon ID (HOT), Genus, Species, Taxon Link and Genome Link. The list can be sorted by the first three columns. The taxon and genome links lead to the detailed taxon description and genome description if available. The table page contains a search box for all the taxa. An alphabetic index is located on top of the table for quick access to a specific taxon. The table displays the total number of taxa in the database or the number of search result. Users can choose to display all result in one page or 100, 50 or 20 taxa per page. Options are also available to showing only named species, unnamed cultivated species or uncultured phylotypes.

Taxonomic hierarchy

The human oral microbial taxa are also arranged in the taxonomic hierarchy, i.e. from domain, phylum, class, order, family, genus, to species levels. The hierarchical tree is fully expanded by default and can be dynamically collapsed at any given level. The link, at the species level, brings users to the detail Taxon Description page. The designation of each level is followed by two numbers enclosed in the square brackets indicating the number of taxa and taxa genome sequences. For example, 'Phylum Proteobacteria [107, 144]' indicates that in the phylum Proteobacteria, 107 taxa were identified in the oral cavity and 144 strains have genome sequences available at HOMD. If a species has been sequenced by multiple groups, we provide each sequence when available for that particular species.

Taxonomic level

The Taxonomic level page provides a list of taxa and the number of taxa at the next lower level for each of the 7 taxonomic levels: Domain, Phylum, Class, Order, Family, Genus and Species. For example, at the Genus level, it lists the 169 genera in HOMD and for each lists the number of species as well as an up pointer to the family for each genus.

Taxon description

The HOMD Taxon Description Page provides comprehensive information for a specific human oral microbial taxon. Information provided can be summarized in four categories: taxonomic hierarchy, biological characteristics, references and community comments. Throughout the page, clickable dynamic cross-links are provided for additional information. The taxon page can be edited and curated by designated curators upon logged-in. The page also allows input and comments provided by the users in the research community. The information provided for each taxon are as follows.

HOT ID. The HOT ID is a unique numeric ID representing this particular taxon. The taxon can be unambiguously referred to from other source of scientific literature. The taxon can be accessed on the web with an easy universal resource locator (URL) format: http://www.homd.org/taxon=NNN, where NNN is the HOT ID.

Status. A taxon can be either a validly named cultivated species, an unnamed cultivated species or an unnamed uncultured phylotype. This status is shown in this field and will be updated upon the change of actual status of the taxon.

Type strain/reference strain. If the taxon's status is validly named cultivated species, the type strain is listed here; if the taxon is an unnamed isolate the strain information will be listed as reference strain. If no cultivated strain is available yet, the Reference Strain field will be listed as 'None, not yet cultivated'.

Classification. The Taxon Description page lists the nomenclatures of each taxonomic level from Domain to Species. The classification defined by HOMD may be different from the NCBI taxonomy. The NCBI taxonomy can be accessed using a dynamic link. The HOMD taxonomy is based on the analysis of where each taxon falls in phylogenetic trees generated using several treeing methods and including over 100 non-oral reference taxa identified by searching greengenes. For example, in HOMD, an organism such as *Eubacterium saburreum,* is placed in the family *Lachnospiraceae* (because that is where it falls phylogenetically), rather than in the family *Eubacteriaceae* (because its incorrect genus name 'Eubacterium' has not yet been revised). Synonyms of the taxon that are currently in use or were used before in the literature or publications are also provided.

16S rRNA gene sequence. Accession number and links to one or more 16S rRNA gene sequences for that taxon.

16S rRNA gene sequence alignment. HOMD provides the clone sequences preliminarily aligned to the reference sequence to which the clones belong. These alignments were generated automatically by computer search of GenBank and were not manually examined. The clone alignments are provided in Clustal format with the reference sequence(s) on top that are used as the template for alignment. To view the alignment in color format and for further adjustment, third-party alignment viewing software may be used, such as SeqView (http://pbil.univ-lyon1.fr/software/seaview.html) and BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Because some pairs of clone sequences may be nonoverlapping (i.e. 500-base

sequences at opposite end of the molecule) this file must be used with caution for tree construction.

Phylogeny. A phylogenic tree showing the position of this taxon among related HOMD taxa can be viewed or downloaded.

Prevalence by molecular cloning. The number of clones found for this taxon in an analysis of approximately 35 000 clones (F. E. Dewhirst *et al.*, submitted for publication). Based on the number of clones found, the rank abundance of the taxon (out of 619) is given.

Synonyms. Lists previous names for the organism if validly named. Isolates or clones designations are given as synonyms when they have appeared in the literature as 'names' for the taxon, such as 'BU063' (21).

NCBI taxonomy. For validly named species there is a link to the NCBI taxonomy. NCBI has no taxonomy for unnamed species, hence the reason HOMD was created.

PubMed search. The number of hits when the name (genus plus species) of this taxon is used in the PubMed search. HOMD automatically and periodically updates this hit number every 2 weeks. To get a most up-to-date search, simply click the 'PubMed Link' to pull up the search result live from NCBI PubMed site. In general there are no results for unnamed taxa, hence the need for HOMD. When articles referencing these taxa (often through clone numbers) are found by HOMD curators or community members, they are manually added to the taxon description.

Nucleotide search. Similar search as above using NCBI Entrez 'nucleotide' as reference database.

Protein search. Similar search as above using NCBI Entrez 'protein' as reference database.

Genomic sequence. Number of genomes that have been sequenced is indicated here with a link to a detail list of these genomes.

Hierarchy structure. An expandable/collapsible view of a dynamically displayed taxonomy tree indicating the position of the taxon on the page.

General information. Generic information regarding this taxon.

Cultivability. Conditions and media for growing strains of this taxon, if available.

Phenotypic characteristics. Generic phenotypic description of the taxon if the taxon has cultivated member(s).

Prevalence and source. Describes the frequency and source of clones and isolates from different oral sites and states of health or disease when known.

References. Literature and publications referencing this taxon. These references are manually curated with up to 10 key references that may also include older references not indexed in PubMed.

Community comments. Registered and logged-in users can provide their feedbacks related to this taxon. The comment requires the approval of the HOMD curators before it is shown to the public.

Curation and editorial information. When and by whom this page was created or last modified.

Identification of 16S rRNA gene sequence by BLAST search

One of the key features of HOMD is the comprehensive collection of the 16S rRNA reference gene sequences for each of the 619 taxa defined based on version 10 of the sequences. Since a phylotype can include members with up to 1.5% sequence divergence (23 bases for a full 1500 base sequence), multiple reference sequences have been selected where we have sequences diverging by more than 10 bases within a taxon. Thus, at the time of writing, there are 747 reference sequences for the 619 taxa. This set of sequence (the HOMD 16S rRNA RefSeq version 10) is available for download and searching against using the BLAST search tool provided. HOMD provides a customized BLAST search tool for identifying unknown 16S rRNA gene sequences for the closest match(s). The HOMD 16S rRNA BLAST search tool allows submission of multiple sequences in a single upload. The search is piped to the back-end computer cluster servers (described above) and the results are presented back to submitter in a tabularized format. Results containing up to 20 top matches for each query sequence can be downloaded in text or Excel file formats. Original full BLAST results including the alignments can also be accessed from the result page. The match identity is presented as straight BLAST results and as an adjusted percent identity (API) calculated as:

$$\mathrm{API} = \frac{100 \times \mathrm{M}}{\mathrm{M} + \mathrm{MM}}$$

where M is the matched (identical) and MM the mismatch sequence length between the query and the reference sequence, respectively. This calculation excludes any gaps introduced during the alignment process of the BLAST search. We have found that this correction gives much better values for single primer sequence reads where the sequence adjacent the primer often includes indels. The top hits are ordered by their API and sequences with alignment smaller than 95% of query sequence are excluded from

Database, Vol. 2010, Article ID baq013, doi:10.1093/database/baq013

ranking. The top four matched reference sequences are listed by this methods and the table shown on the web page contain links to the original BLAST results as well as to the taxon description pages for reference sequences. The results for the 20 top matches can be downloaded as plain text or Microsoft Excel format.

Dissemination of taxonomic naming scheme

The HMP Data Analysis and Coordination Center (DACC; accessible at http://www.hmpdacc.org/) is using HOT numbers to designate taxonomic identity isolates of the oral cavity with URLs cross-referenced to HOMD. These URLs will be embedded in the data provided by DACC so that user can track down to the more comprehensive information for individual genome. In our recent submission of approximately 35 000 16S rRNA clone sequences to GenBank, a hyperlink was provided in each sequence for cross-referencing back to the HOMD database.

Genomics tools overview

Complimentary to the taxonomy information, the HOMD also provides comprehensive information and tools for studying genomes of the human oral microbes. HOMD genomics database serves as the curated repository for the molecular sequences of human oral microbiome, including complete and partial genomics sequences, as well as 16S rRNA mentioned in the previous section. Genomic sequences available at HOMD can be either fully assembled genomes, high coverage genomes or genome surveys. HOMD also keeps tracks of the status of ongoing genome sequencing projects for human oral microorganisms. A sequence meta information page is created to hold relevant genomics and sequence meta information if a sequencing project for a human oral microbe is announced and available in the NCBI Genome Project Database. The genome project status is updated frequently based on information collected from the NCBI Genome Projects Database with an automatic query script. Once genomic sequences are publicly released, they are dynamically annotated by HOMD (dynamic annotation). Annotation done by other data center, if available, is termed 'static annotation' and is viewable in a separate panel in the Genome Viewer (described below). Relevant tools are provided for viewing and searching the annotation. These tools were first developed as part of the Bioinformatics Resource for Oral Pathogens (BROP: http://www.brop.org) (20). The programs and the data-mining schemes used in HOMD are designed for both finished and unfinished (collections of multiple contigs) genome sequences. The tools are integrated with the HOMD web site and are conveniently accessible by users. Icons or links to available tools pertaining to a specific genome are automatically presented on relevant page to users. Important genomic data and bioinformatics tools provided by HOMD are described below. Additional information is also available in the previous publication (20).

Genome table

HOMD organizes genomes in three viewing options: Taxa with Annotated Genomes, Taxa with Genomes in Progress and View All Genomes. The first option lists the oral taxa with annotated (static or dynamic) genomic information and provides links to all the genomes available for each taxon. The View Genome button links to the Genome Table showing all the available genomes of a specific taxon. The Genome Table shows the Oral Taxon ID (HOT), the Genus and Species names, Strain Culture Collection, HOMD Sequence ID (SEQ ID), Number of contigs and singlet, combined sequence length and links to available tools and information. The second option (Taxa with Genomes in Progress) lists those oral taxa with genomic sequencing project still in progress but no sequences are yet available. The third option shows all the genomes in the alphabetical order and provides searching and sorting function for easier navigation. Each genome listed will have a link to the Sequence Meta Information page described next.

Sequence Meta Information

The Sequence Meta Information page provides detailed biological, molecular biological, genetic, genomic and taxonomic as well as annotation information for a particular strain that has been, is being or will be sequenced. Information on these pages is semi-automatically updated. Updated information from both Genomes Online and NCBI Genomic Project database are retrieved frequently and compared with the existing database automatically. New or modified Genomic Project Information are then added to the Sequence Meta Information pages with confirmation by curators. The Sequence Meta Information pages contains the following human curated information related to the target organism: Oral Taxon ID, HOMD Sequence ID (SEQ ID), Organism Name (Genus, Species), Culture Collection Entry Number, Isolate Origin, Sequencing Status, NCBI Genome Project ID, NCBI Taxonomy ID, Genomes Online Goldstamp ID, NCBI Genome Survey Sequence Accession ID, JCVI (previously TIGR) CMR ID, Sequencing Center, Number of Contigs and Singlets, Combined Length (kbp), GC Percentage, DNA Molecular Summary, ORF Annotation Summary and 16S rRNA Gene Sequence. In addition, original external information such as NCBI Genome Project Database, NCBI Taxonomy Database, Genome Online Database and rRNA in NCBI Nucleotide Database, if available, are parsed into separate tables below the Sequence Meta Information for convenient referencing.

Full and high coverage genomes

Full genomes are the oral microbial genomes that have been fully assembled, while the high coverage genomes are not fully assembled but represent most of the genome coverage. Both types of genomes are annotated and deposited in the public database such as GenBank. HOMD aims to provide more frequently updated genomic annotation for bacterial oral isolates (see below). In addition, HOMD provides graphical genomic viewing for static annotations done by other public data centers such as NCBI or JCVI.

Genome surveys

One of the original major goals of the NIH funded project 'A Foundation for the Oral Microbiome and Metagenome' was to partially sequence up to 100 representative human oral microbial species. A total of 12 low-coverage partial genomic sequences were sequenced and deposited in NCBI and active annotation is being maintained by HOMD (8). Since the launch of the HMP, the HOMD team has been providing the genomic DNA of human oral microbes to the four HMP sequencing centers for high coverage rather than survey sequencing (8).

Dynamic annotation of genomic sequences

One of the major features of the HOMD Genomic Database is the automatic and frequent updating genomic annotation pipeline for genomes of oral isolates. Although the amount of sequence data is still growing rapidly, the computational power needed for bioinformatic analysis is catching up and the cost and energy consumption per CPU decreasing due to the availability of multi-core CPU formats. The lower cost of computational power has made it feasible for us to setup a computational cluster dedicated to the annotation of human oral microbial genomes. As described in the Hardware section, HOMD recruited a cluster of multi-core multi-node computer servers to frequently update the annotation. Current HOMD genome annotation algorithms include: (i) BLASTP (http://www.ncbi.nih.gov/BLAST/) (22) search against weekly updated NCBI non-redundant protein data (ftp://ftp.ncbi.nih.gov/blast/db); (ii) BLASTP search against Swiss-Prot protein data (http://us.expasy.org/sprot/) (23) and (iii) InterProScan search (http://www.ebi.ac.uk/Tools/InterProScan/) (24) against ScanRegExp, BlastProDom, ProfileScan, HMMPfam, superfamily, HMMTigr, Seg, Coil, HMMPIR, FPrintScan and HMMSmart databases (http://www.ebi.ac.uk/interpro/databases.html). To provide data on functional potential of genomes, BLASTP search result against Swiss-Prot are further processed for the construction of KEGG metabolic pathways and Gene Ontology trees, because the well-annotated Swiss-Prot protein sequence descriptions contains interlinks to the ENZYME (25) and Gene Ontology (26). The dynamic genome annotation is being repeated continuously based on NCBI's weekly update of non-redundant protein database. Additional genomes are being added to the annotation pipeline as more sequences are made available by other public sequencing projects such as the HMP (http://www.hmpdacc.org). A live update status of the genome annotation is provided on the HOMD home page indicating the latest genome annotated or updated. HOMD aims to maintain frequent and dynamic computer annotation for genomic sequence of at least one isolate from each oral taxon whenever sequences are made publicly available, as well as static annotation of all annotated releases.

Genome Explorer

Genome Explorer is the centralized web interface that inter-connects all the genomics resources in HOMD. The front end of Genome Explorer is a user-friendly interface that allows investigators to navigate among all the genomics information provided at HOMD. HOMD Genomics Tools can be accessed either by selecting the tool or the genome first. If the user chooses the desired tool first, the user is then directed to the Genome Explorer interface for selecting genomes. Once a target genome is chosen, the interface dynamically presents all the tools, including linked external databases, available for the selected genome. Currently available tools include Genome Viewer, Dynamic Annotation, BLAST, Annotator, EMBOSS, KEGG pathways (27), Gene Ontology Tree (28), Genomewide ORF Alignment and Sequence Download. The back-end of Genome Explorer is a searchable annotation database that integrates all the results generated from the dynamic annotation pipeline mentioned. The search result is presented in a paginated and sortable table that also provides web links to (i) a summary page for individual ORF, (ii) Genome Viewer to show the exact location of the target ORF in the genome and (iii) to the original BLAST or InteProScan results. The summary page provides all the information and tools available for a specific ORF, including all the data-mining results mentioned above, as well as convenient links to other web tools for performing new search and analysis. In short, Genome Explorer is a one-stop interface for all the genomic information available for each target genome or gene.

Genome Viewer

Genome Viewer is a unique graphical genomic sequence viewer developed originally for the BROP project (20). The Genome Viewer was designed to alleviate the inconvenience encountered when comparing two different sets of annotations for the same genome. Genome Viewer provides a graphical, six-frame translational view of the same region of the genome with individual panels showing different sets of annotations. It has easy navigating features including zooming, centering and searching by gene ID. For

Database, Vol. 2010, Article ID baq013, doi:10.1093/database/baq013

example, the genome *Porphyromonas gingivalis* W83 has been annotated by JCVI (TIGR), Los Alamos National Laboratory and NCBI separately. These different annotations can be viewed and compared side-by-side in the Genome Viewer (http://www.homd.org/index.php?name=GenomeExp&org=pgin&gprog=gview).

Meta-database search

Meta-database search engine searches across four major databases—Taxonomy, Genomes, User documents and Dynamic Genome Annotation. The Hit box will show the number of matches found in each databases and provides links to the results. For the annotated genomes, counts of search hits are shown for each individual genome, which are linked to the Genome Explorer showing corresponding matches.

Batch database content download

HOMD provides batch database content download in both tab-delimited text (viewable in a web browser or downloadable as file to user's computer) and Excel format for all the data in HOMD. These downloads can be accessed from the 'Tools and Download' top menu, leading to a page listing all the downloadable contents of HOMD. Currently the downloads include the following six primary categories: (i) Taxon tables; (ii) 16S rRNA gene sequences; (iii) Genomic sequences and dynamic annotation results; (iv) Genome meta informaton; (v) Database schema and (vi) Database table structures. The download page can be accessed by navigating from the top or side-panel menu, or through a direct URL access to the download page-http://www.homd.org/download.

User documentation

The HOMD user's guide (i.e. the help documentation) was designed to help users to use the tools, navigate the information and interpret the results provided by HOMD. The user's guide is accessible through the top navigation menu on every tool page and is dynamically linked to the relevant guide for each different tool. For example, when users are viewing the Taxon Table page, the 'How to use this page' menu item shown in the top navigation menu will lead directly to the page that explains the use of the Taxon Table. Alternatively users can also browse the entire user documentation (along with the 'general documentation') by the 'Table of content' tab shown op top of each documentation page. Every document of HOMD can be searched either through the search box located at the bottom of the table of content of the documentation page or through the Meta Database Search box located at the top-right part of the home page.

Usage tracking

HOMD uses the AWStats system (http://awstats.sourceforge.net) to track the usage of the web site. AWStats provides comprehensive web usage statistics graphically. It summarizes hourly, daily, weekly and monthly usages and aggregates the statistics by geographic locations such as countries or individual hosts. Major search engines such as Google, Yahoo and MSN are filtered for better tracking of true user visits (i.e. non-machine visits).

Discussion

The diverse and enormous amount of bioinformatics data available have presented to us with great challenges in terms of data usage and management. Concerns and suggestions for solving this problem have been addressed by several scientists (29–31). Overall, the issues related to the data management can be summarized in the following three categories: (i) standardization: how different formats of the same type of data (e.g. the transcriptomic data derived based on different microarray technologies) can be standardized and compared; (ii) incorporation: how to incorporate different types of data (e.g. transcriptomic and proteomic data) in the framework of the systems biology and (iii) integration: how to combine relevant data dispersed at different locations (e.g. annotations of the same genome done by different institutes) to gain a complete yet non-redundant view of all the available information. The solution to these issues is 'integration, integration, integration' (32) and scientists have actually begun to address some of these issues. For example, data standards have been gradually established for the microarray (33, 34) and proteomics data (34); various software are also available to help integrate the diverse and standalone bioinformatics software tools for a simpler workflow (35–40); and DDBJ, GENBANK and EMBL, the three major molecular databanks have been exchanging and synchronize their data on a daily basis (41–43). Nevertheless, to date the issues are far from over—scientists are still overwhelmed by many non-standard, multiform and disperse data. To exploit all the available bioinformatics data, one still needs to visit multiple websites, retrieve data in different formats and use incompatible software tools to gain a fullest picture of all the information. The complexity increases when multiple organisms are involved in biofilms for a single-body site, like the oral cavity.

To solve the data dispersion issues one is tempted to think of a 'centralized' resource center that is capable of doing everything. This however is not politically and financially feasible and such a centralized center can easily become monopolistic and consequently stifle the progress of science. Contrarily, localized (decentralized) standalone resource centers with focused topics can still coexist and

should be encouraged as long as they agree to the data standardization and provide exchangeability to other resources. For example, there are organism-oriented data centers such as NCBI's 'Human Genome Resource' (http://www.ncbi.nlm.nih.gov/genome/guide/human), the Mouse Genome Database (44), yeasts (45,46), etc.; function-oriented data centers such as the promoter (47) and protein interaction (48,49) databases; and the disease-oriented data centers such as cancer gene databases (50,51), STD sequence database (http://www.stdgen.lanl.gov), etc. The common feature for these specialized databases is that they provide integrated information for focused biological topics or themes.

HOMD has been designed as a body site-specific database with the specific biological theme focused on the human oral microbiome, providing integrated information and tools. The HOMD Taxonomy database serves as a referencing point for human oral taxa with easy URL web access codes. It allows investigators to conveniently access cross-linked genetic, phenotypic, clinical and bibliographic information. Integrated bioinformatics tools are available for studying both taxonomic and the genomic sequences.

Genomic sequences contain a plethora of information and have profoundly advanced our understanding of biology. Genome sequencing technologies have become more efficient and affordable, more and more genomes have been or are being sequenced by many institutes (http://www.genomesonline.org). While this is all very encouraging, this information avalanche often proves daunting to biologists for there are great difficulties encountered in searching, retrieving, interpreting or managing the data. The multiple sources of the data representing the same genomic entity, as described in this report, make the task even tougher. The HOMD Genomics Database includes a suite of software tools that were originally designed based on the daily practical needs of a group of biologists who study the oral microorganisms. The integrated resource provided by HOMD Genomics Database gathers useful tools and data that are otherwise scattered elsewhere and will help biologists to access the information and study the genomes more conveniently. Although the focus of HOMD is on human oral microorganisms, the integrated infrastructure that has been developed can be readily applied to other human body sites.

Summary

The HOMD is a body site-specific public database providing the scientific community with comprehensive information on prokaryote species that are present in the human oral cavity. This dynamic database provides a curated taxonomy of oral prokaryotes, a curated set of full-length 16S rRNA reference sequences, and BLAST tools that allow the identification of unknown isolates or clones based on their 16S rRNA sequence; additionally, phenotypic, bibliographic, clinical and genomic information are linked for each taxa. The web-based interfaces and software tools are implemented to facilitate the query and analysis of this comprehensive dataset. As oral taxa are sequenced as part of effort to obtain reference genomes in the HMP, the genomic sequences will be added to HOMD. The information of HOMD was organized based on a taxonomic structure built upon a well-curated 16S rRNA phylogeny. The organization, integration and presentation of the HOMD data can serve as a model for microbiome data from other human body sites such as gut, skin, vagina, which are being actively studied by international efforts. The database is publicly accessible at http://www.homd.org.

Funding

National Institute of Dental and Craniofacial Research (contract U01 DE016937 and grant DE017106, partial); American Recovery and Reinvestment Act of 2009 (DE016937, supplement). Funding for open access charge: U01 DE016937.

Conflict of interest. None declared.

References

1. Aas,J.A., Paster,B.J., Stokes,L.N. *et al*. (2005) Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.*, **43**, 5721–5732.
2. Turnbaugh,P.J., Ley,R.E., Hamady,M. *et al*. (2007) The human microbiome project. *Nature*, **449**, 804–810.
3. Verberkmoes,N.C., Russell,A.L., Shah,M. *et al*. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J.*, **3**, 179–189.
4. Kurokawa,K., Itoh,T., Kuwahara,T. *et al*. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.*, **14**, 169–181.
5. Gill,S.R., Pop,M., Deboy,R.T. *et al*. (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355–1359.
6. Peterson,J., Garges,S., Giovanni,M. *et al*. (2009) The NIH Human Microbiome Project. *Genome Res.*, **19**, 2317–2323.
7. Paster,B.J., Boches,S.K., Galvin,J.L. *et al*. (2001) Bacterial diversity in human subgingival plaque. *J. Bacteriol.*, **183**, 3770–3783.
8. Izard,J.; The Forsyth Metagenomic Support Consortium. (2010) Building the genomic base-layer of the oral 'omic' world. In: Sasano,T. and Suzuki,O. (eds.). *Interface Oral Health Science 2009: Proceedings of the 3rd International Symposium for Interface Oral Health Science*. Springer, New York.
9. Nelson,K.E., Fleischmann,R.D., DeBoy,R.T. *et al*. (2003) Complete genome sequence of the oral pathogenic bacterium *Porphyromonas gingivalis* strain W83. *J. Bacteriol.*, **185**, 5591–5601.
10. Downes,J., Vartoukian,S.R., Dewhirst,F.E. *et al*. (2009) *Pyramidobacter piscolens* gen. nov., sp. nov., a member of the phylum '*Synergistetes*' isolated from the human oral cavity. *Int. J. Syst. Evol. Microbiol.*, **59**, 972–980.
11. Dzink,J.L., Socransky,S.S. and Haffajee,A.D. (1988) The predominant cultivable microbiota of active and inactive lesions of destructive periodontal diseases. *J. Clin. Periodontol.*, **15**, 316–323.

12. Dzink,J.L., Tanner,A.C., Haffajee,A.D. and Socransky,S.S. (1985) Gram negative species associated with active destructive periodontal lesions. *J. Clin. Periodontol.*, **12**, 648–659.
13. Socransky,S.S. and Haffajee,A.D. (1994) Evidence of bacterial etiology: a historical perspective. *Periodontology 2000*, **5**, 7–25.
14. Tanner,A.C., Haffer,C., Bratthall,G.T. *et al*. (1979) A study of the bacteria associated with advancing periodontitis in man. *J. Clin. Periodontol.*, **6**, 278–307.
15. Tanner,A., Maiden,M.F., Macuch,P.J. *et al*. (1998) Microbiota of health, gingivitis, and initial periodontitis. *J. Clin. Periodontol.*, **25**, 85–98.
16. Moore,W.E., Holdeman,L.V., Cato,E.P. *et al*. (1983) Bacteriology of moderate (chronic) periodontitis in mature adult humans. *Infect. Immun.*, **42**, 510–515.
17. Moore,W.E., Holdeman,L.V., Smibert,R.M. *et al*. (1982) Bacteriology of severe periodontitis in young adult humans. *Infect. Immun.*, **38**, 1137–1148.
18. Moore,W.E. and Moore,L.V. (1994) The bacteria of periodontal diseases. *Periodontology 2000*, **5**, 66–77.
19. Paster,B.J. and Dewhirst,F.E. (1988) Phylogeny of campylobacters, wolinellas, *Bacteroides gracilis*, and *Bacteroides ureolyticus* by 16S ribosomal ribonucleic acid sequencing. *Int. J. Syst. Bacteriol.*, **38**, 56–62.
20. Chen,T., Abbey,K., Deng,W.J. and Cheng,M.C. (2005) The bioinformatics resource for oral pathogens. *Nucleic Acids Res.*, **33**, W734–W740.
21. Zuger,J., Luthi-Schaller,H. and Gmur,R. (2007) Uncultivated *Tannerella* BU045 and BU063 are slim segmented filamentous rods of high prevalence but low abundance in inflammatory disease-associated dental plaques. *Microbiology*, **153**, 3809–3816.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Boeckmann,B., Bairoch,A., Apweiler,R. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
24. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
25. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
26. Camon,E., Magrane,M., Barrell,D. *et al*. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662–672.
27. Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–101; discussion 101–103, 119–128, 244–252.
28. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
29. Cassman,M. (2005) Barriers to progress in systems biology. *Nature*, **438**, 1079.
30. Aderem,A. (2005) Systems biology: its practice and challenges. *Cell*, **121**, 511–513.
31. Liu,E.T. (2005) Systems biology, integrative biology, predictive biology. *Cell*, **121**, 505–506.
32. Chicurel,M. (2002) Bioinformatics: bringing it all together. *Nature*, **419**, 751, 753, 755 passim.
33. Andersen,M.T. and Foy,C.A. (2005) The development of microarray standards. *Anal. Bioanal. Chem.*, **381**, 87–89.
34. Ravichandran,V. and Sriram,R.D. (2005) Toward data standards for proteomics. *Nat. Biotechnol.*, **23**, 373–376.
35. Hoon,S., Ratnapu,K.K., Chia,J.M. *et al*. (2003) Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res.*, **13**, 1904–1915.
36. Leo,P., Marinelli,C., Pappadà,G. *et al*. (2004) BioWBI: an integrated tool for building and executing bioinformatic analysis workflows, *Bioinformatics Italian Society Meeting (BITS 2004)*, Padova.
37. Lu,Q., Hao,P., Curcin,V. *et al*. (2006) KDE Bioscience: Platform for bioinformatics analysis workflows. *J. Biomed. Inform.*, **39**, 440–450.
38. Oinn,T., Addis,M., Ferris,J. *et al*. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.
39. Tang,F., Chua,C.L., Ho,L.Y. *et al*. (2005) Wildfire: distributed, Grid-enabled workflow construction and execution. *BMC Bioinformatics*, **6**, 69.
40. Finak,G., Godin,N., Hallett,M. *et al*. (2005) BIAS: Bioinformatics Integrated Application Software. *Bioinformatics*, **21**, 1745–1746.
41. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J. *et al*. (2006) GenBank. *Nucleic Acids Res.*, **34**, D16–D20.
42. Cochrane,G., Aldebert,P., Althorpe,N. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.*, **34**, D10–D15.
43. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.*, **34**, D6–D9.
44. Bult,C.J., Blake,J.A., Richardson,J.E. *et al*. (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res.*, **32**, D476–D481.
45. Guldener,U., Munsterkotter,M., Kastenmuller,G. *et al*. (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
46. Hirschman,J.E., Balakrishnan,R., Christie,K.R. *et al*. (2006) Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **34**, D442–D445.
47. Schmid,C.D., Praz,V., Delorenzi,M. *et al*. (2004) The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res.*, **32**, D82–D85.
48. Gong,S., Park,C., Choi,H. *et al*. (2005) A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **6**, 207.
49. Pagel,P., Kovac,S., Oesterheld,M. *et al*. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.
50. Akagi,K., Suzuki,T., Stephens,R.M. *et al*. (2004) RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.*, **32**, D523–D527.
51. Levine,A.E. and Steffen,D.L. (2001) OrCGDB: a database of genes involved in oral cancer. *Nucleic Acids Res.*, **29**, 300–302.