A Data Analysis and Coordination Center for the Human Microbiome Project

Michelle Gwinn Giglio

June 2009

The Human Microbiome Project

- There are 10 times as many microbial cells than human cells in the human body and therefore there are at least 1000 times as many microbial genes as human genes

- Changes in the composition of these communities of organisms (microbiomes) have been shown to correlate with human health and disease

- The HMP is an NIH Roadmap project that represents a significant investment ($148 million) in providing a standardized dataset that will form the foundation of subsequent human microbiome work

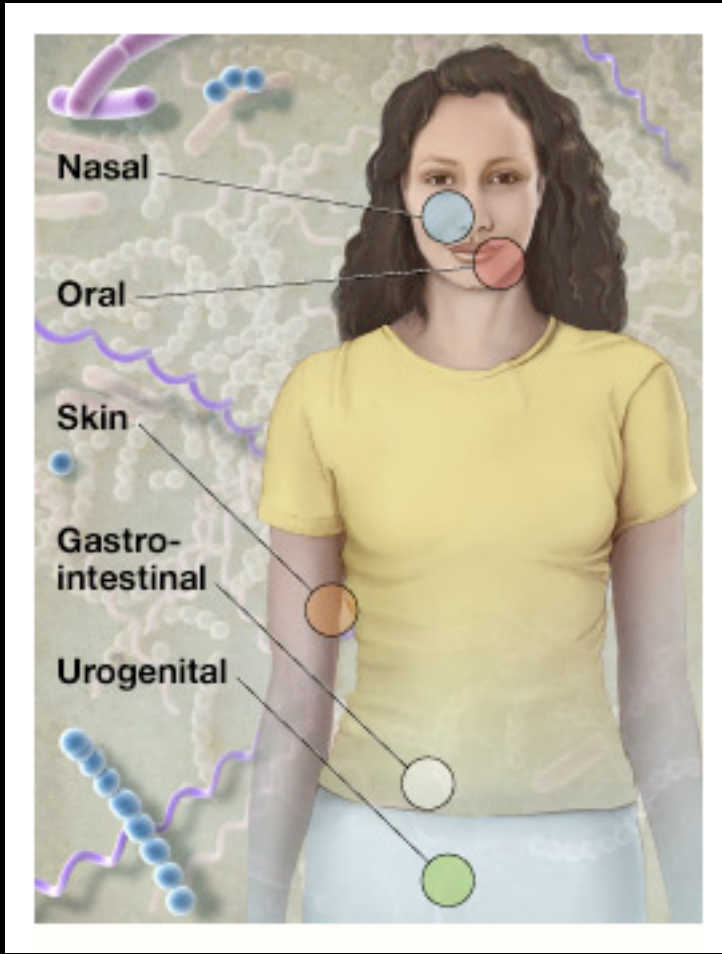


image from NIH HMP website http://nihroadmap.nih.gov/hmp/

Some of the questions

- Is there a core human microbiome that is shared by all humans?
- Is there a core set of microbial functions that is shared by all humans?
- To what extent does human phenotype or behavior impact the microbiome?
- Are changes in the microbiome causing disease?
- Can disease be treated by altering the microbiome?

International Human Microbiome Consortium

- The NIH project is part of a larger international effort
- Includes a European Consortium and seven additional countries throughout the world
- Data to be made available by parallel DACCs in the US and Europe as well as at NCBI.

www.human-microbiome.org

Some (of many) health correlations

Crohn's disease

Necrotizing enterocolitis

Obesity

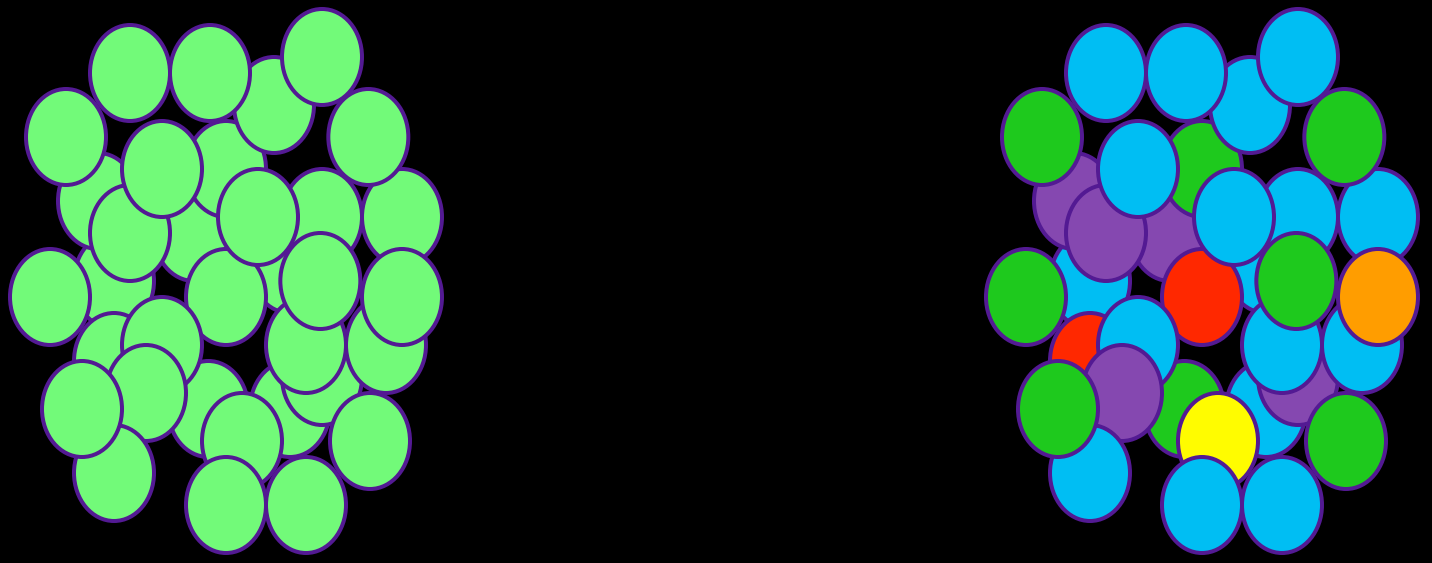
Vaginosis

Reflux disease/esophageal cancer

Skin diseases

Metagenomics

- Metagenomics is the sequencing and analysis of the DNA from a mixture of organisms, as opposed to traditional single isolate genomics
  - Provides a view of the unculturable organisms of the human microbiomes



Goals of metagenomic analysis

- Find out who is there
  - what organisms in what quantities
  - phylogenetic analysis of 16s RNA sequence
- Find out what functions are present
  - pathways/metabolites in the microbes
  - how might these interact with host pathways/metabolites
  - functional annotation of predicted genes in the sequence
  - whole metagenome shotgun sequencing
- Compare samples from different environments/subjects to find correlations with phenotype/disease/etc.

Initiatives of the HMP

- Jumpstart
  - Initial 1.5 years, 4 established centers with existing NHGRI funding
- Reference Data Set Generation
- Demonstration Projects
- Supporting Initiatives
- And of course us – the DACC

Focus Areas

- 16S ribosomal RNA sequencing
- Whole metagenome shotgun sequencing
- Reference Genome Sequencing
- Training and Outreach

Roll of the DACC

- Provide access to data
- Assist in standardization of data analysis
- Assist in standardization of pipelines
- Provide standardization in data storage and display (including metadata, sequence, annotation, analysis, much more)
- Provide access to SOPs
- Outreach/Training

Who is the DACC?

Overall PI: Owen White,
Institute for Genome Sciences
University of Maryland School of Medicine

Todd DeSantis, Gary Andersen, Rob Knight, Nikos Kyrpides, Victor Markowitz

Funded by the NIH Common Fund

Sequencing Centers of Jumpstart phase

- Baylor College of Medicine
- Broad Institute
- J. Craig Venter Institute
- Washington University Genome Center

Focus 1
Why Sequence Reference Genomes?

- They provide guideposts for the metagenomic analysis
- They are knowns in a world of unknowns

Which ones?

- A set of guidelines has been developed by the HMP Centers
  - Phylogenetic diversity
  - Clinical relevance
  - Abundance
  - Duplicates from different body sites
  - Opportunity to explore pangenomes
  - Whatever we can get our hands on

HMP Reference Genome Goals

- How many?
  - Jumpstart phase: sequence 200 genomes
  - Reference Data Set phase: sequence an additional 400 genomes
  - Overall goal: 1000 total genomes collected from human body sites gathered from HMP and other efforts
- These will mostly be in “draft” state
  - ~15% will be finished
- View entire list on the DACC HMP Catalog

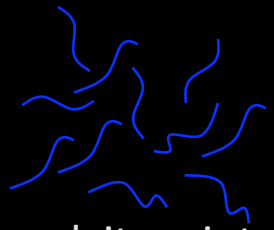
Draft standard metrics

- minimal draft standards have been established by the sequencing centers
  - so that “draft” means the same for all genomes

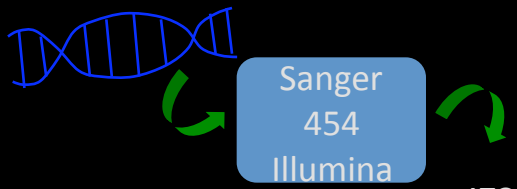
- Six metrics
  - Completeness: >90% of genome contained in contigs (where total size of genome is based on scaffold size)
  - Accuracy: >90% of bases greater than 5x read coverage
  - Continuity: N50, N75, N90 contig length minimums
  - Continuity: N50, N75, N90 scaffold length minimums
  - Choppiness: average contig length minimum
  - Completeness/Annotatability: >90% core genes present

whole genome sequencing

isolate genomic DNA



break it up into random fragments

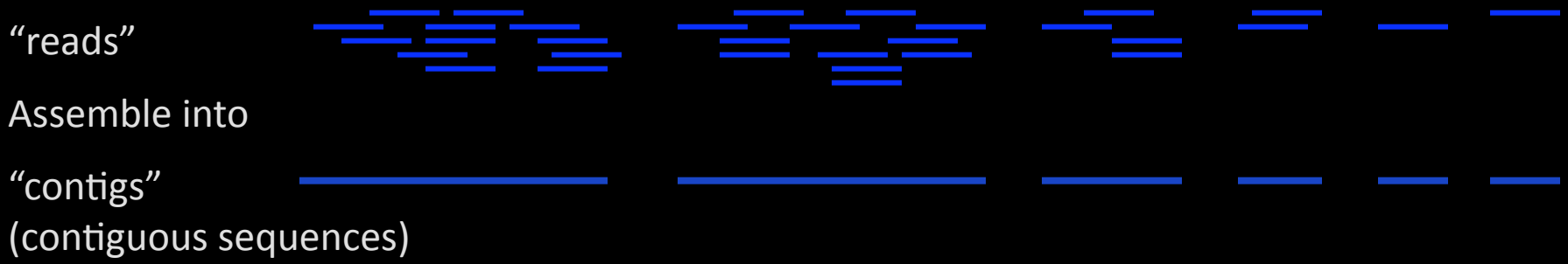


sequence

usually sequence 10x the length of the genome

ATGCGAAGTC
CTAGACCAGA
TTGAC……..

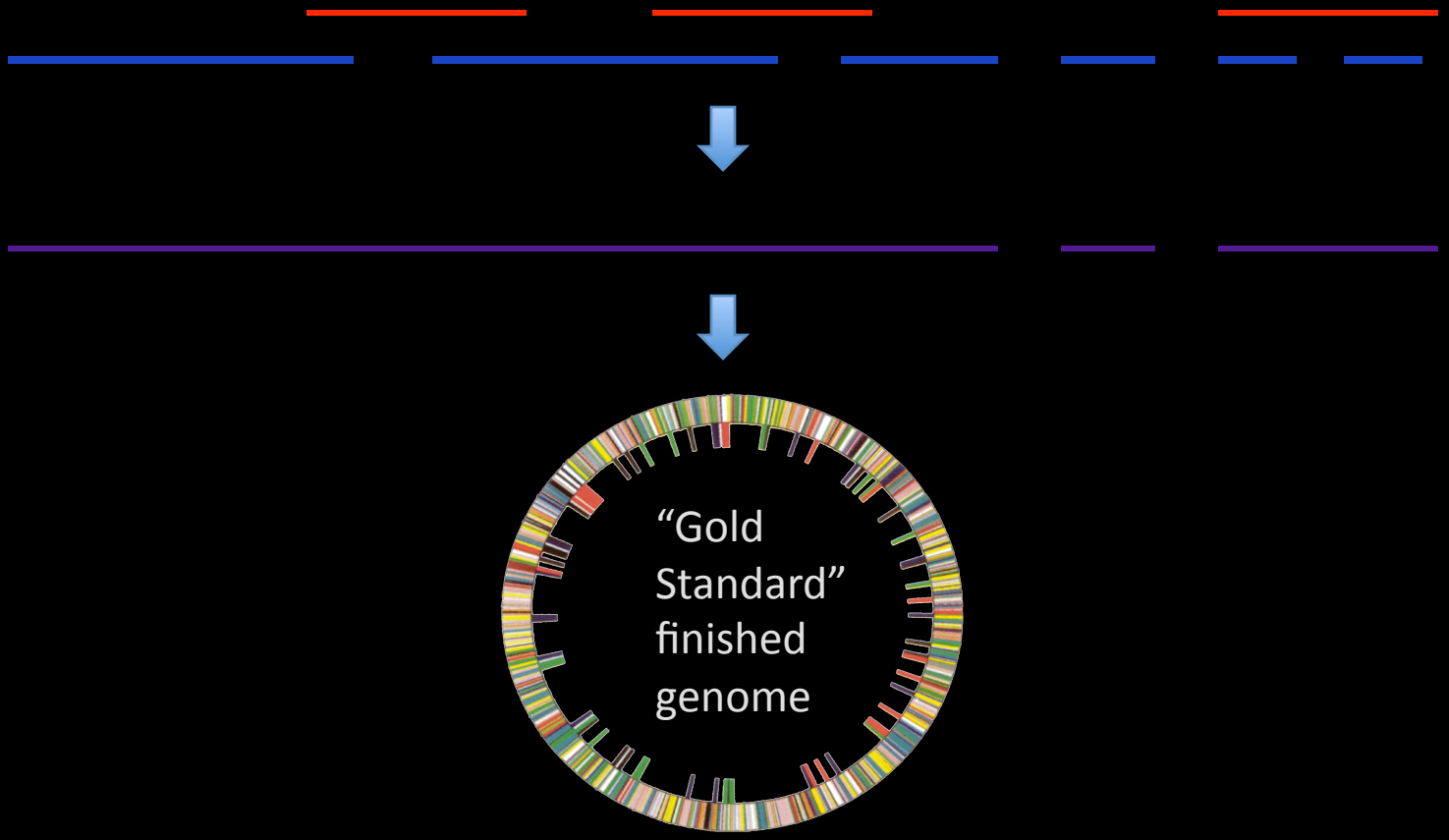
Read Lengths: Sanger >900, 454 ~400, Illumina 50-100



Genome in "draft" state

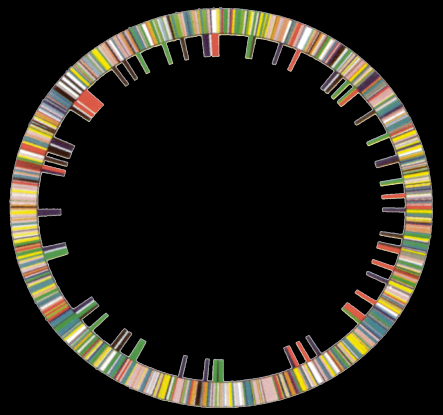
Finishing

- Apply finishing process:
  - Additional sequences from alternate platform
  - PCR
  - Manual editing of data
  - Etc.

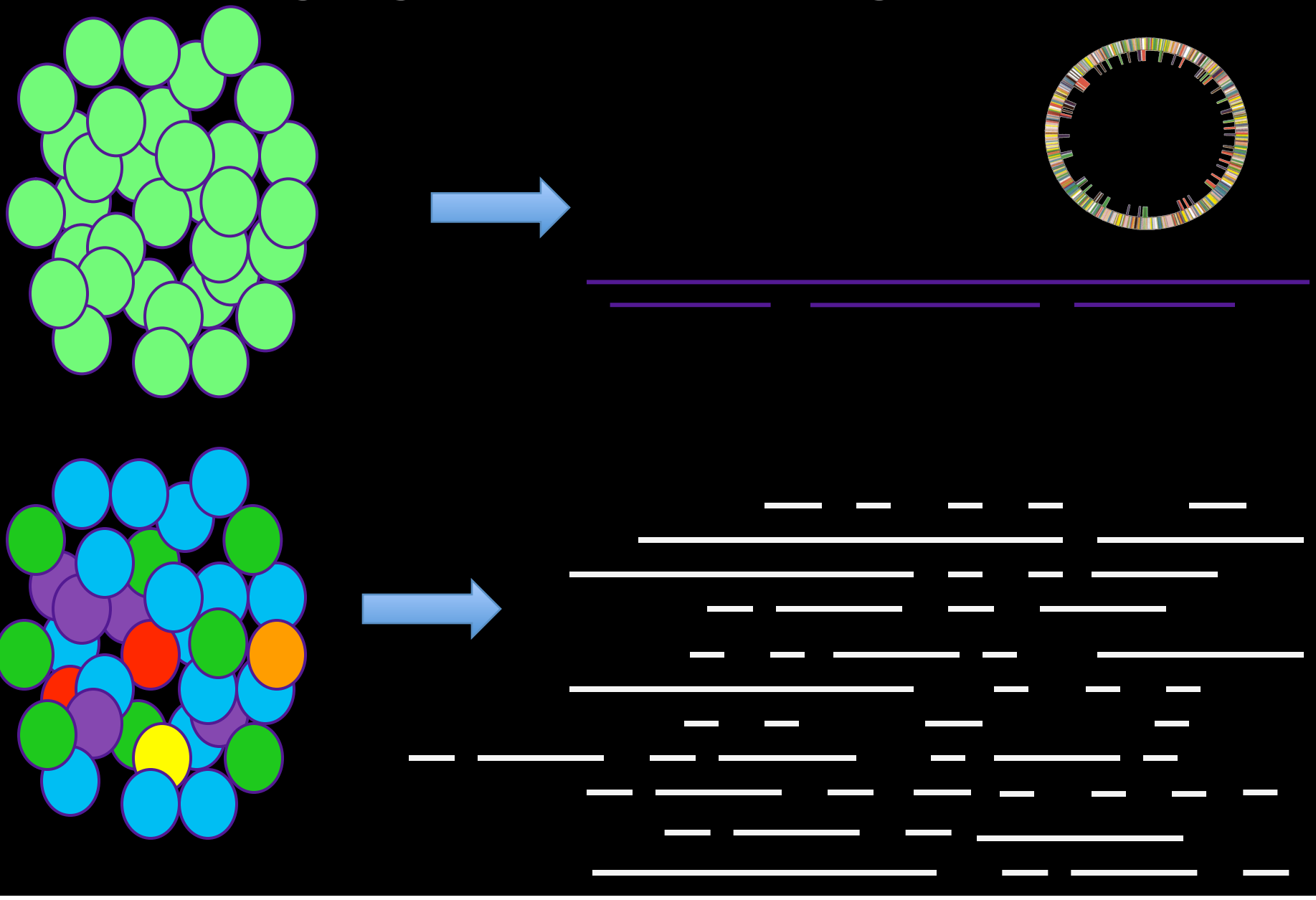


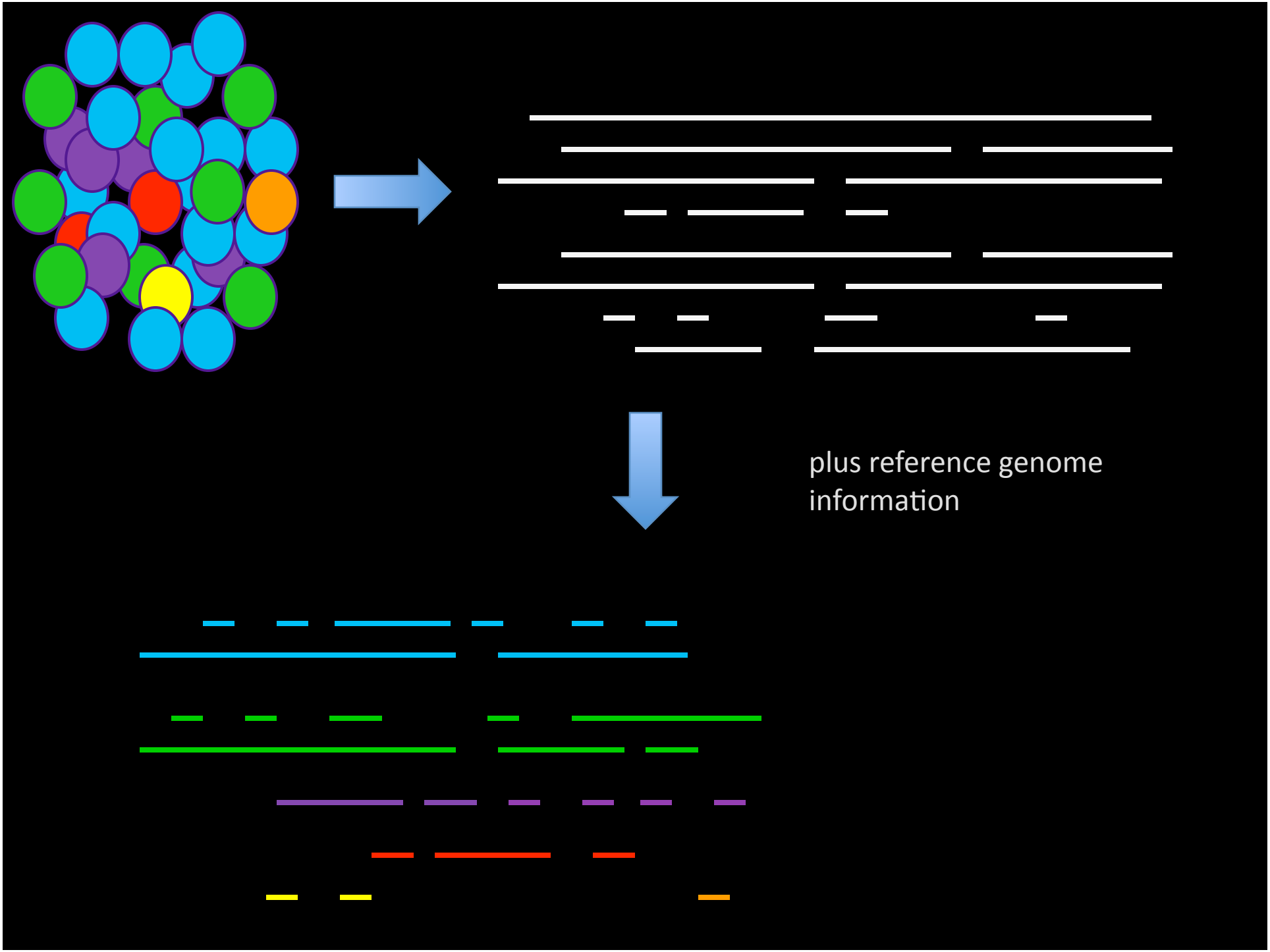
Finishing Standards

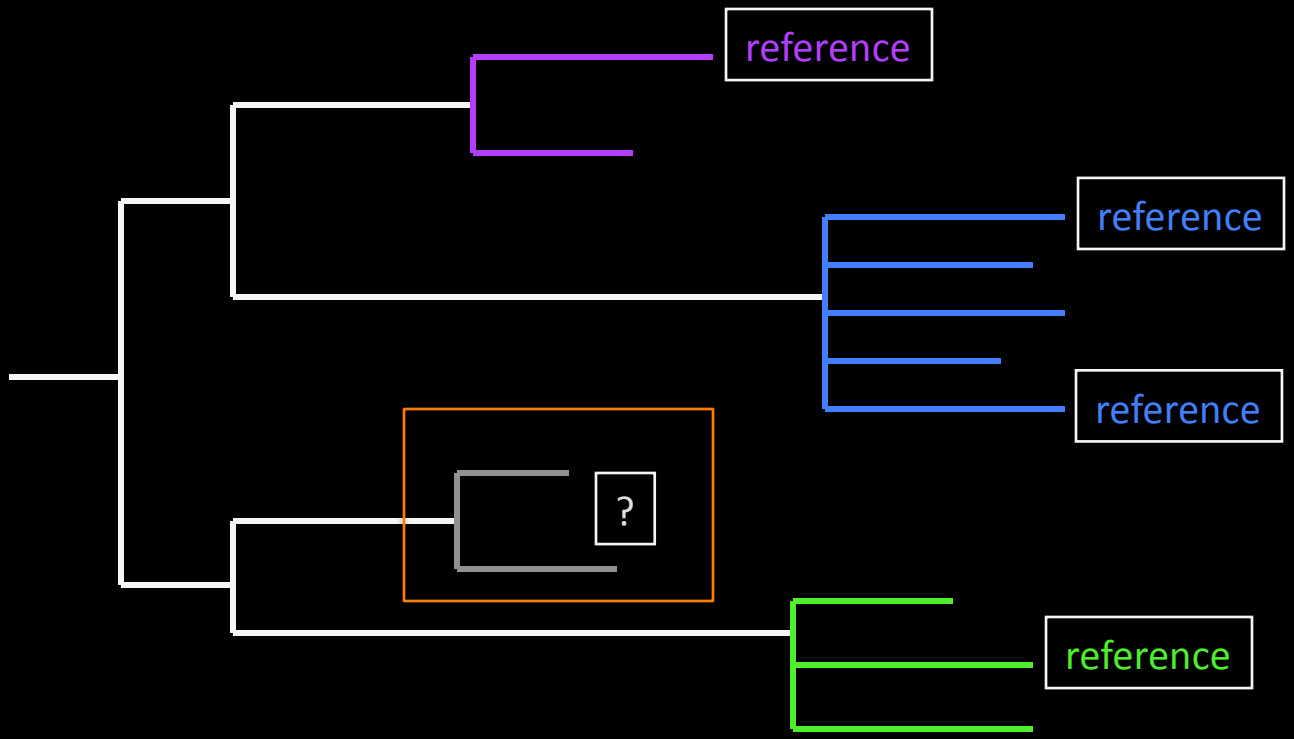
- What does each center mean by “finishing”?
- A set of categories ranging from “standard draft” to “gold standard” is being established by an international working group – these will be published soon.



single genome vs. metagenome

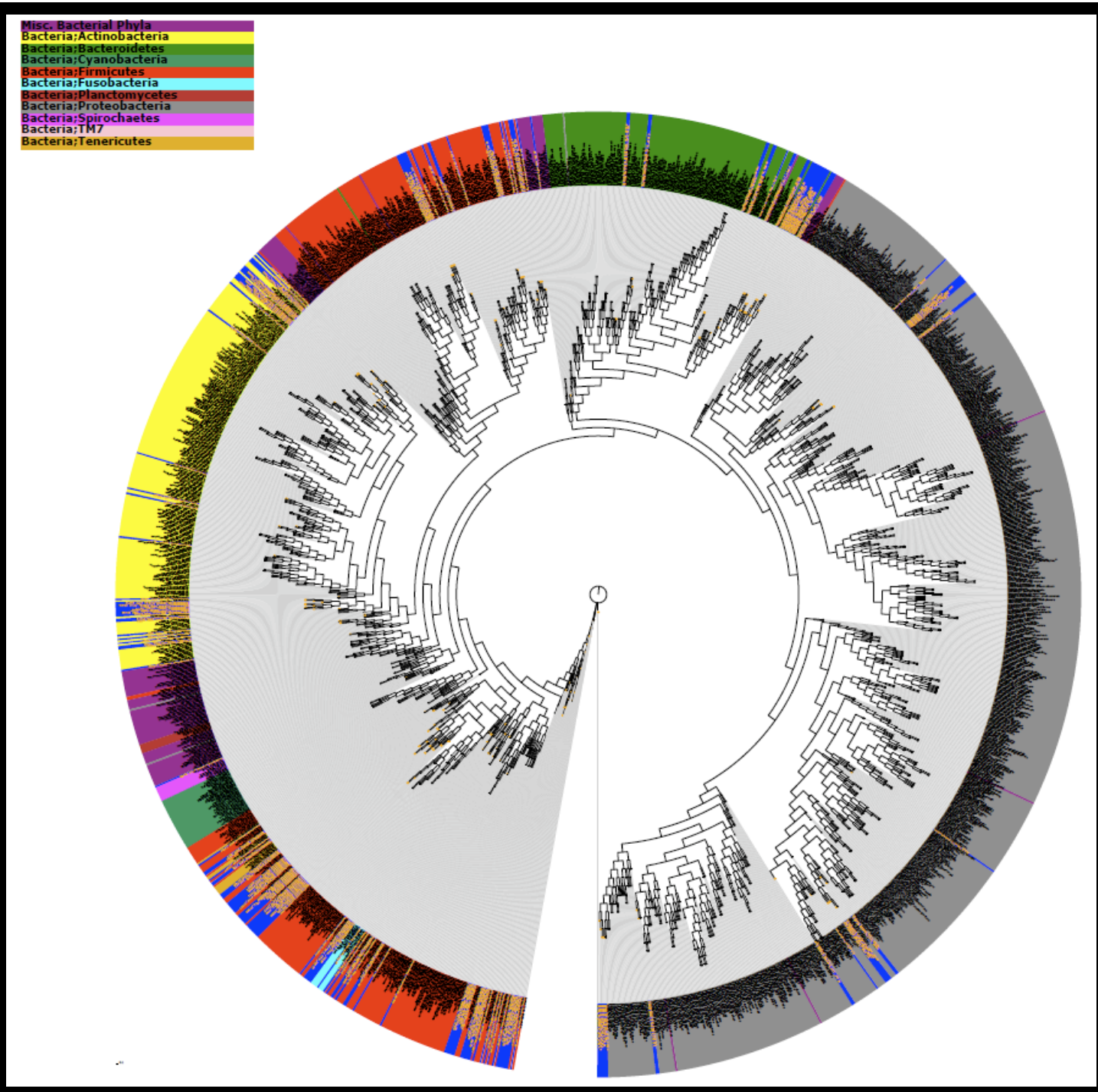



plus reference genome
information


reference
reference
reference
?
reference

Progress so Far

- Jumpstart Phase is finishing up
- Almost 200 genomes have been sequenced to at least the standard draft phase, some to higher finishing levels
  - All of these have sequence and annotation deposited at NCBI
- A paper is in preparation to describe the draft standard metrics and also describe:
  - How many new genes has this produced?
  - How has the tree of life been enriched?
  - What is the diversity in the new strains?
    - Overall
    - Between body sites
    - Pangenomes
  - Body site specific pathways?



Phlyogenetic tree of the reference genomes sequenced in the initial period of the HMP project (in royal blue) on top of tree of containing completed prokaryotic genomes color coded by clade

Figure by Jonathan Badger of JCVI

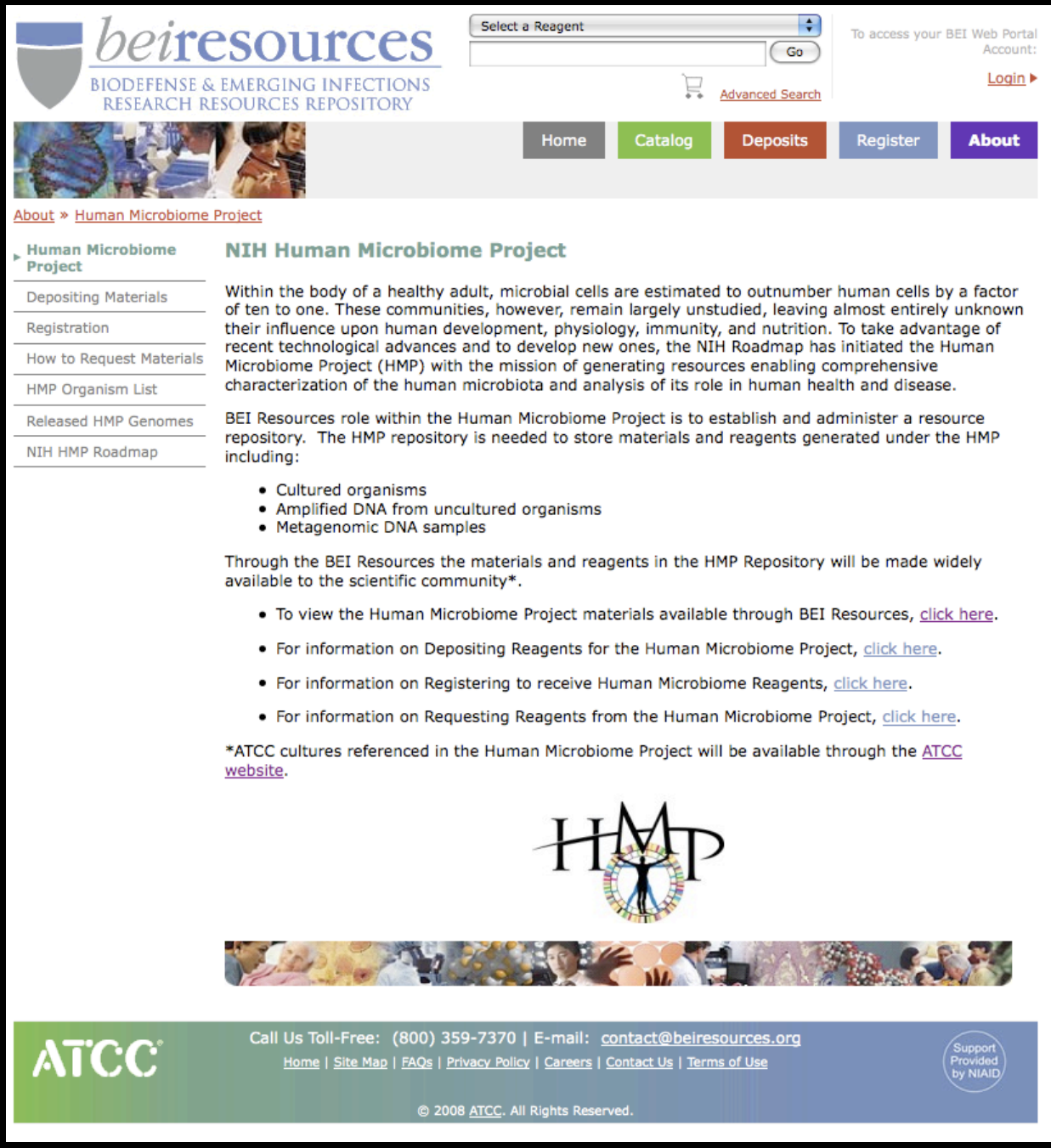
The Jumpstart Reference Genomes

Acidaminococcus sp. D21
Acinetobacter sp. ATCC 27244
Actinomyces coleocanis DSM 15436
Actinomyces urogenitalis DSM 15434
Actinomyces odontolyticus ATCC 17982
Alistipes putredinis DSM 17216
Anaerococcus lactolyticus ATCC 51172
Anaerococcus tetradius ATCC 35098
Anaerococcus hydrogenalis DSM 7454
Anaerofustis stercorihominis DSM 17244
Anaerostipes caccae DSM 14662
Anaerotruncus colihominis DSM 17241
Atopobium vaginae DSM 15829
Atopobium rimae ATCC 49626
Bacteroides caccae ATCC 43185T
Bacteroides capillosus ATCC 29799
Bacteroides cellulosilyticus DSM 14838
Bacteroides coprocola DSM 17136
Bacteroides coprophilus DSM 18228
Bacteroides dorei DSM 17855
Bacteroides eggerthii DSM 20697
Bacteroides finegoldii DSM 17565
Bacteroides intestinalis DSM 17393
Bacteroides ovatus ATCC 8483
Bacteroides pectinophilus ATCC 43243
Bacteroides plebeius DSM 17135
Bacteroides sp. 2_2_4
Bacteroides sp. 9_1_42FAA
Bacteroides sp. D1
Bacteroides sp. D4
Bacteroides stercoris ATCC 43183
Bacteroides uniformis ATCC 8492
Bifidobacterium longum subsp. infantis
Bifidobacterium adolescentis L2-32
Bifidobacterium angulatum DSM 20098
Bifidobacterium bifidum NCIMB 41171
Bifidobacterium breve DSM 20213
Bifidobacterium catenulatum DSM16992
Bifidobacterium dentium ATCC 27678
Bifidobacterium gallicum DSM 20093
Bifidobacterium longum subsp. Infantis
Bifidobacterium pseudocatenulatum
Blautia hansenii DSM 20583
Blautia hydrogenotrophica DSM 10507
Bryantella formatexigens DSM 14469
Campylobacter rectus RM3267
Capnocytophaga gingivalis JCVIHMP016
Capnocytophaga sputigena Capno
Catenibacterium mitsuokai DSM 15897
Chryseobacterium gleum ATCC 35910
Citrobacter sp. 30_2
Citrobacter youngae ATCC 29220
Clostridiales bacterium 1_7_47_FAA
Clostridium asparagiforme DSM 15981
Clostridium bartlettii DSM 16795
Clostridium bolteae ATCC BAA-613
Clostridium hiranonis DSM 13275
Clostridium hylemonae DSM 15053
Clostridium leptum DSM 753
Clostridium methylpentosum
Clostridium nexile DSM 1787
Clostridium ramosum DSM 1402
Clostridium scindens ATCC 35704
Clostridium sp. 7_2_43FAA
Clostridium sp. L2-50
Clostridium sp. M62/1
Clostridium sp. SS2/1
Clostridium spiroforme DSM 1552
Clostridium sporogenes ATCC 15579
Collinsella aerofaciens ATCC 25986
Collinsella intestinalis DSM 13280
Collinsella stercoris DSM 13279
Coprococcus comes ATCC 27758
Coprococcus eutactus ATCC 27759
Corynebacterium accolens ATCC 49725
Corynebacterium glucuronolyticum
Corynebacterium lipophiloflavum
Corynebacterium pseudogenitalum
Corynebacterium striatum ATCC 6940
Corynebacterium amycolatum SK46
Corynebacterium matruchotii ATCC 33806
Desulfovibrio piger ATCC 29098
Dorea formicigenerans ATCC 27755
Dorea longicatena DSM 13814
Eikenella corrodens ATCC 23834
Enterococcus faecalis ATCC 29200
Enterococcus faecalis HH22
Enterococcus faecalis TX0104
Enterococcus faecalis TX1332
Enterococcus faecium TX1330
Escherichia coli 83972
Escherichia sp. 1_1_43
Escherichia sp. 3_2_53FAA
Eubacterium biforme DSM 3989
Eubacterium dolichum DSM 3991
Eubacterium hallii DSM 3353
Eubacterium sireaum DSM 15702
Eubacterium ventriosum ATCC 27560
Faecalibacterium prausnitzii M21/2

continued

Finegoldia magna ATCC 53516
Fusobacterium mortiferum ATCC 9817
Fusobacterium sp. 2_1_31
Fusobacterium sp. 4_1_13
Fusobacterium sp. 7_1
Gardnerella vaginalis ATCC 14019
Helicobacter bilis ATCC 43879
Helicobacter canadensis MIT 98-5491
Helicobacter cinaedi CCUG 18818
Helicobacter pullorum MIT 98-5489
Helicobacter winghamensis ATCC BAA-430
Holdemania filiformis DSM 12042
Lactobacillus paracasei subsp. paracasei ATCC 25302
Lactobacillus plantarum subsp. plantarum ATCC 14917
Lactobacillus acidophilus ATCC 4796
Lactobacillus brevis subsp. gravesensis ATCC 27305
Lactobacillus buchneri ATCC 11577
Lactobacillus crispatus JV-V01
Lactobacillus fermentum ATCC 14931
Lactobacillus gasseri JV-V03
Lactobacillus hilgardii ATCC 8290
Lactobacillus jensenii JV-V16
Lactobacillus johnsonii ATCC 33200
Lactobacillus reuteri MM4-1
Lactobacillus reuteri SD2112
Lactobacillus reuterii CF48-3A
Lactobacillus rhamnosus LMS2-1
Lactobacillus ruminis ATCC 25644
Lactobacillus salivarius ATCC 11741
Lactobacillus ultunensis DSM 16047
Lactobacillus vaginalis ATCC 49540
Lactobacillus gasseri MV-22
Lactobacillus jensenii 115-3-CHN
Lactobacillus paracasei subsp. paracasei 8700:2
Lactobacillus reuteri MM2-3
Lactobacillus sakei subsp. Carnosus
Leuconostoc mesenteroides ATCC 19254
Listeria grayi DSM 20601
Methanobrevibacter smithii DSM 2374
Methanobrevibacter smithii DSM 2375
Mitsuokella multacida DSM 20544
Mobiluncus curtisii ATCC 43063
Mobiluncus mulieris ATCC 35243
Mollicutes bacterium D7
Neisseria cinerea ATCC 14685
Neisseria flavescens NRL30031/H210
Neisseria lactamica ATCC 23970
Neisseria mucosa ATCC 25996
Neisseria subflava NJ9703
Oribacterium sinus F0268
Oxalobacter formigenes HOxBLS
Oxalobacter formigenes OXCC13
Parabacteroides johnsonii DSM 18315
Parabacteroides merdae ATCC 43184
Parvimonas micra ATCC 33270
Porphyromonas endodontalis ATCC 35406
Porphyromonas uenonis 60-3
Prevotella copri DSM 18205
Propionibacterium acnes SK137
Proteus mirabilis ATCC 29906
Proteus penneri ATCC 35198
Providencia alcalifaciens DSM 30120
Providencia rettgeri DSM 1131
Providencia rustigianii DSM 4541
Providencia stuartii ATCC 25827
Rhodococcus erythropolis SK121
Roseburia intestinalis L1-82
Roseburia inulinivorans DSM 16841
Ruminococcus gnavus ATCC 29149
Ruminococcus lactaris ATCC 29176
Ruminococcus obeum ATCC 29174
Ruminococcus torques ATCC 27756
Sphingobacterium spiritivorum ATCC 33300
Staphylococcus aureus subsp. aureus MN8
Staphylococcus aureus subsp. aureus TCH60
Staphylococcus capitis SK14
Staphylococcus hominis SK119
Streptococcus infantarius subsp. Infantarius
Streptococcus salivarius SK126
Subdoligranulum variabile DSM 15176

All HMP strains will be available in a public repository



HUMAN MICROBIOME PROJECT

DACC

REFERENCE GENOMES | 16S RNA SEQUENCING | METAGENOMIC WGS | OUTREACH & TRAINING | DOWNLOADS | SOPs | RESOURCE REPOSITORY

www.hmpdacc.org

Welcome

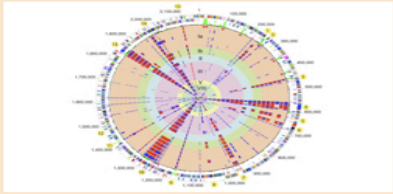
Welcome to the Data Analysis and Coordination Center (DACC) for the Human Microbiome Project (HMP). The HMP was launched by the National Institutes of Health Roadmap for Medical Research and is designed to fuel research into the multitude of microbes that live in the various environments of the human body. A major goal of the HMP is to look for correlations between changes in the microbiome and human health. More information about the project can be found on the NIH Roadmap website at http://nihroadmap.nih.gov/hmp.

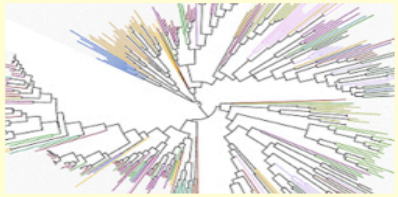
The HMP DACC is the central repository for all HMP data, providing specialized data management and analysis infrastructure to facilitate discoveries about the microbiome. The HMP-DACC web site will provide web-based query and visualization tools, comprehensive computational analysis of HMP data, quality control measures, and links to well-documented Standard Operating Procedures. The DACC is also strongly committed to outreach and training. Please visit the above tabs for more information on each of these topics.

Focus Areas of the HMP

Reference Genomes

Approximately 600 microbes from cultured and uncultured bacteria, plus several non-bacterial microbes will be sequenced. Combined with existing and other planned efforts, the total reference collection should reach 1000 genomes. These sequences will provide a benchmark against which further sequence data can be compared.



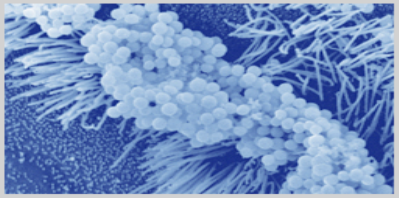


16S RNA sequencing

16s RNA sequencing will be used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome at each site.

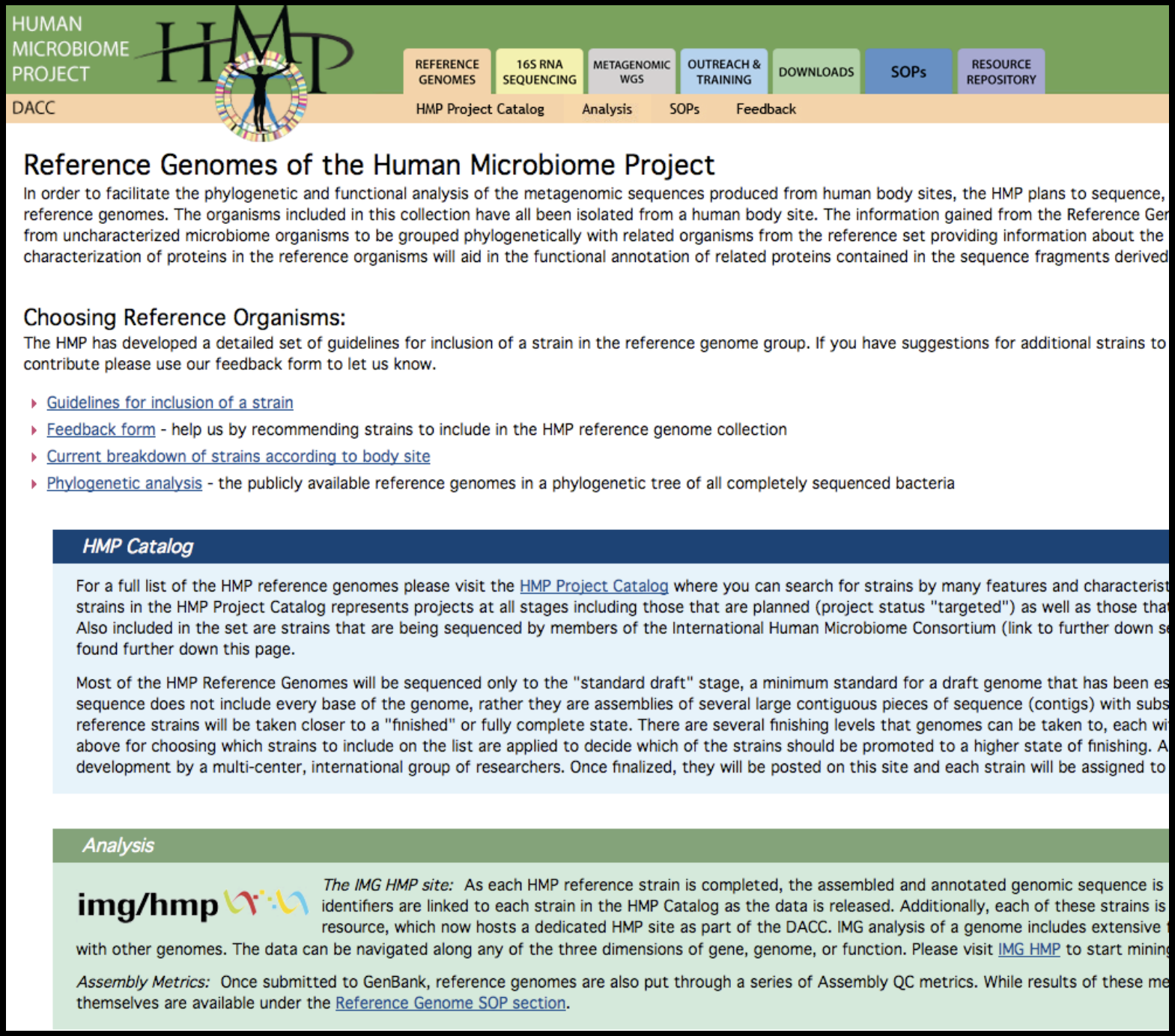
Metagenomic WGS

Expanding on the 16S data, Whole Genome Shotgun (WGS) sequencing will be performed on samples taken from human subjects. Coupled with the other data generated during the project, this will provide insights into the genes and pathways present in the human microbiome.



Outreach & Training

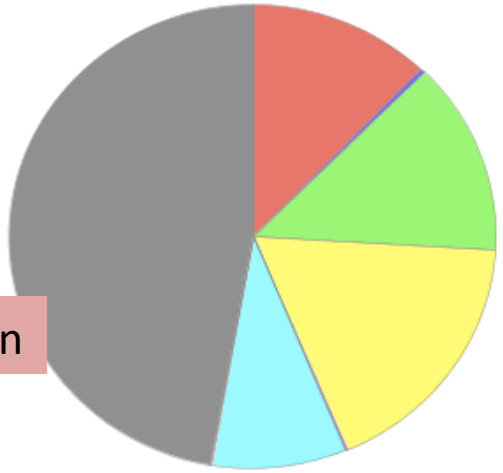
We encourage and welcome community feedback. As well, we offer a number of workshops and training opportunities.


HUMAN MICROBIOME PROJECT
HMP
DACC
REFERENCE GENOMES
16S RNA SEQUENCING
METAGENOMIC WGS
OUTREACH & TRAINING
DOWNLOADS
SOPs
RESOURCE REPOSITORY
HMP Project Catalog
Analysis
SOPs
Feedback
Reference Genomes of the Human Microbiome Project
In order to facilitate the phylogenetic and functional analysis of the metagenomic sequences produced from human body sites, the HMP plans to sequence,
reference genomes. The organisms included in this collection have all been isolated from a human body site. The information gained from the Reference Ge
from uncharacterized microbiome organisms to be grouped phylogenetically with related organisms from the reference set providing information about the
characterization of proteins in the reference organisms will aid in the functional annotation of related proteins contained in the sequence fragments derived
Choosing Reference Organisms:
The HMP has developed a detailed set of guidelines for inclusion of a strain in the reference genome group. If you have suggestions for additional strains to
contribute please use our feedback form to let us know.
Guidelines for inclusion of a strain
Feedback form - help us by recommending strains to include in the HMP reference genome collection
Current breakdown of strains according to body site
Phylogenetic analysis - the publicly available reference genomes in a phylogenetic tree of all completely sequenced bacteria
HMP Catalog
For a full list of the HMP reference genomes please visit the HMP Project Catalog where you can search for strains by many features and characterist
strains in the HMP Project Catalog represents projects at all stages including those that are planned (project status "targeted") as well as those tha
Also included in the set are strains that are being sequenced by members of the International Human Microbiome Consortium (link to further down s
found further down this page.
Most of the HMP Reference Genomes will be sequenced only to the "standard draft" stage, a minimum standard for a draft genome that has been es
sequence does not include every base of the genome, rather they are assemblies of several large contiguous pieces of sequence (contigs) with subs
reference strains will be taken closer to a "finished" or fully complete state. There are several finishing levels that genomes can be taken to, each wi
above for choosing which strains to include on the list are applied to decide which of the strains should be promoted to a higher state of finishing. A
development by a multi-center, international group of researchers. Once finalized, they will be posted on this site and each strain will be assigned to
Analysis
img/hmp
The IMG HMP site: As each HMP reference strain is completed, the assembled and annotated genomic sequence is
identifiers are linked to each strain in the HMP Catalog as the data is released. Additionally, each of these strains is
resource, which now hosts a dedicated HMP site as part of the DACC. IMG analysis of a genome includes extensive
with other genomes. The data can be navigated along any of the three dimensions of gene, genome, or function. Please visit IMG HMP to start minin
Assembly Metrics: Once submitted to GenBank, reference genomes are also put through a series of Assembly QC metrics. While results of these me
themselves are available under the Reference Genome SOP section.

HMP Project Catalog

www.hmpdacc.org

Body Sample Site	Count	Percentage
Urogenital tract	75	12%
Blood	2	0%
Oral	83	13%
Skin	109	18%
Nasopharyngeal	1	0%
Airways	55	9%
Eye	1	0%
Gastrointestinal tract	291	47%



Body site breakdown

HUMAN MICROBIOME PROJECT

DACC

REFERENCE GENOMES | 16S RNA SEQUENCING | METAGENOMIC WGS | OUTREAC TRAINI

HMP Home | HMP Project Catalog | HMP Statistics

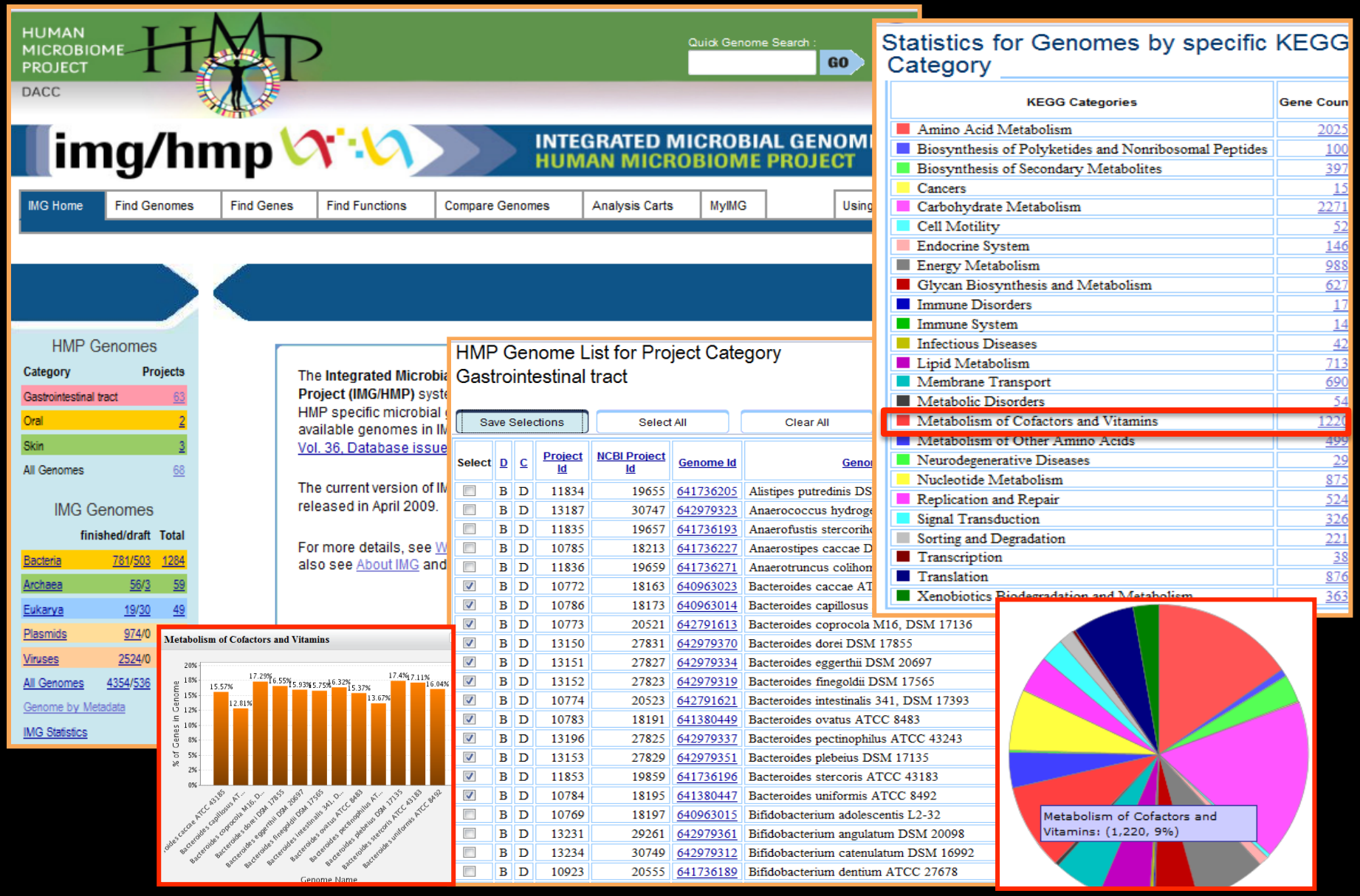
Human Microbiome Projects Catalog

Scrollable Table View | Search Projects | Export to Excel

HMP ID	Genome Species Strain	NCBI Project ID	Genbank ID	Gene Count	Body Site	Project Status	Sequencing Center	Funding Source	Strain Repository ID	Cross Reference ID
0001	Abiotrophia defectiva ATCC 49176	33011	ACIN00000000	3,346	Oral	incomplete	Washington Univ, USA	NIH	ATCC 49176	
0002	Abiotrophia para-adiacens HMP2				Airways	targeted	USA			
0003	Abiotrophia sp. HMP3				Airways	targeted	USA			
0004	Achromobacter piechaudii ATCC 43553				Airways	incomplete	BCM-HGSC, USA		ATCC 43553	
0005	Achromobacter xylosoxidans C54				Airways	incomplete	Broad Institute, USA	NIH		
0006	Achromobacter xylosoxidans HMP6				Airways	targeted	USA			
0007	Achromobacter xylosoxidans HMP7				Airways	targeted	USA			
0008	Acidaminococcus sp. D21_V1	34117	ACGB00000000	2,055	Gastrointestinal tract	incomplete	Broad Institute, USA	NIH		
0009	Acidovorax sp.				Skin	targeted	USA			
0010	Acinetobacter baumannii ATCC 19606	38509		4,506	Urogenital tract	incomplete	Broad Institute, USA	NIH	ATCC 19606	
0011	Acinetobacter baumannii HMP11				Skin	targeted	USA			
0012	Acinetobacter calcoaceticus RUH2202	38337			Skin	incomplete	Broad Institute, USA	NIH		
0013	Acinetobacter sp. SH024				Skin	incomplete	Broad Institute, USA	NIH		
0014	Acinetobacter genomosp. 13TU RUH2624	38511			Skin	incomplete	Broad Institute, USA	NIH		
0015	Acinetobacter haemolyticus ATCC 19194				Airways	incomplete	BCM-HGSC, USA		ATCC 19194	
0016	Acinetobacter johnsonii SH046	38339			Skin	incomplete	Broad Institute, USA	NIH		
0017	Acinetobacter lwoffii SH145	38343		2,796	Skin	incomplete	Broad Institute, USA	NIH		
0018	Acinetobacter radioresistens SH164, DSM 20098	38345			Gastrointestinal tract	incomplete	Broad Institute, USA	NIH	DSM 20098	
0019	Acinetobacter radioresistens SK82	34081			Skin	incomplete	J. Craig Venter Institute, USA	NIH		
0020	Acinetobacter sp. 6013113	33017			Skin	incomplete	Washington Univ, USA	NIH		
0021	Acinetobacter sp. 6013150	33071			Skin	incomplete	Washington Univ, USA	NIH		
0022	Acinetobacter sp. 6014059	33073			Skin	incomplete	Washington Univ, USA	NIH		

Complete List of targeted and/or sequenced strains Search and Download

Genome Analysis at IMG



Do you have strains?

The HMP is actively seeking strains that have been isolated from a human body site. If you have any you are willing to share please let me know or fill out the feedback form on the DACC web site. There are several strains listed on the HMP Catalog as “targeted”. We want these but may not have identified a source for them. We’d love to know if you have one of these you are willing to share.

Focus 2/3:
Microbiome sampling and sequencing, 16S RNA and WGS

- Jumpstart Phase Goal:
  - Recruit 250 healthy volunteers for sampling at 5 body sites
    - Age 18-40
    - Total maximum of 18 sites (with subsites included)
    - 220 have been screened
    - 123 have been sampled
- Reference Data Set Generation Goal:
  - sequence the samples

Sampling

- Two sampling centers:
  - Baylor College of Medicine
  - Washington University
- Consistency
  - Make sure sampling at the same exact sites in the same way
  - 18 sites precisely defined/described
- Protocols
  - sampling
  - DNA extraction
  - Much more

Sequencing

16S ribosomal subunit sequencing

Whole metagenome shotgun sequencing

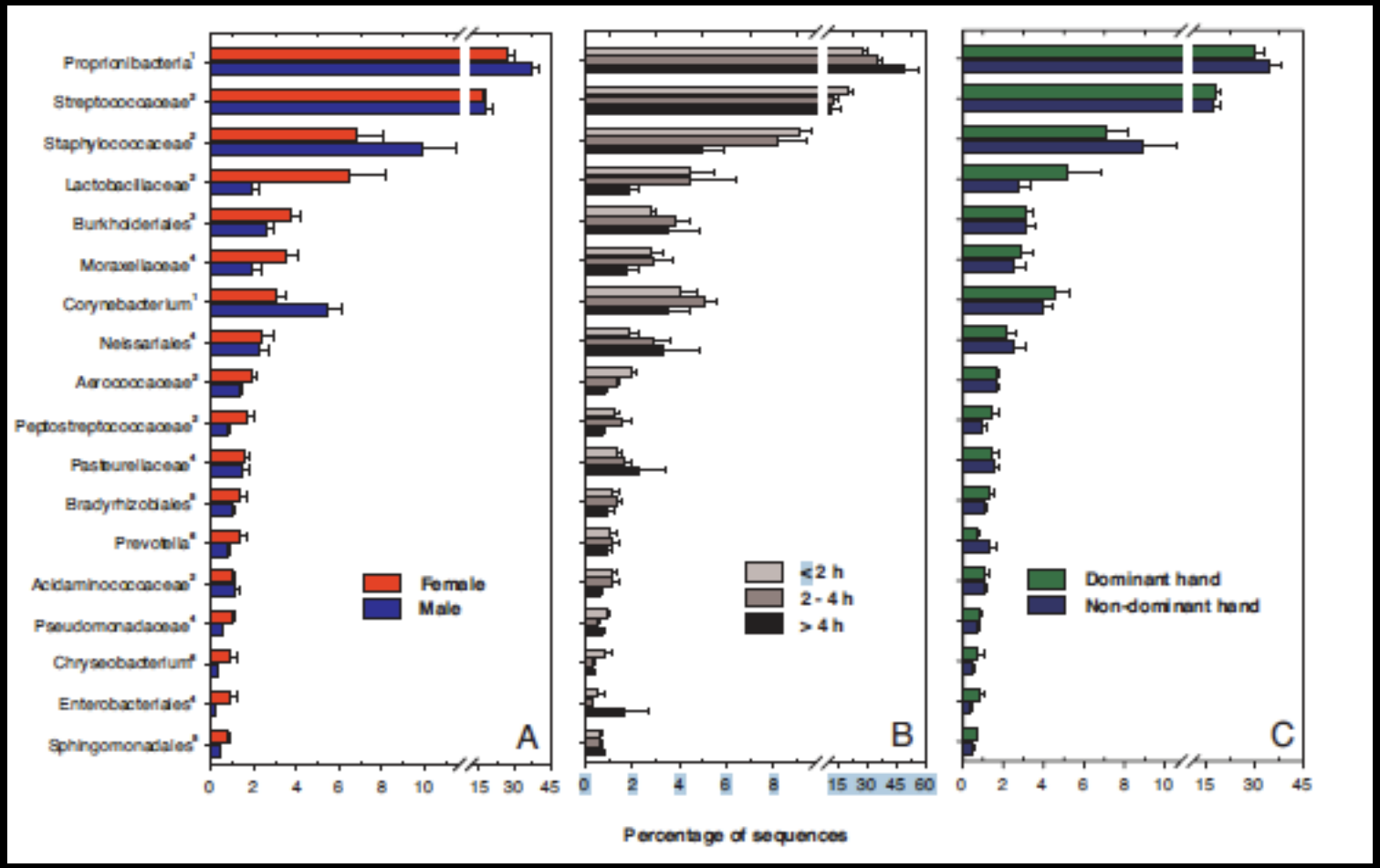
Microbiome Sequencing and Analysis: 16S Ribosomal RNA analysis

- 16S rRNA gene sequencing will be used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome.

- Establish standards across centers
  - Same sample, different centers
    - Initially lots of variability
    - After standardization much more consistent across centers
  - Positive Controls
    - Mock community sent to each center
    - More standardization

- The hope is that the wider metagenomic community can make use of these standards which will then allow comparisons across data sets produced by different groups.

Abundance profiles



Rob Knight's group, Proc Natl Acad Sci U S A. 2008 November 18; 105(46): 17994–17999.

Microbiome Sequencing and Analysis: WGS metagenome sequencing

To try to get at the functions that are encoded by the microbiomes. Look for metabolic capability, see what pathways/systems are present in the community that may not be present in isolated species. Look for interaction with host pathways/systems.

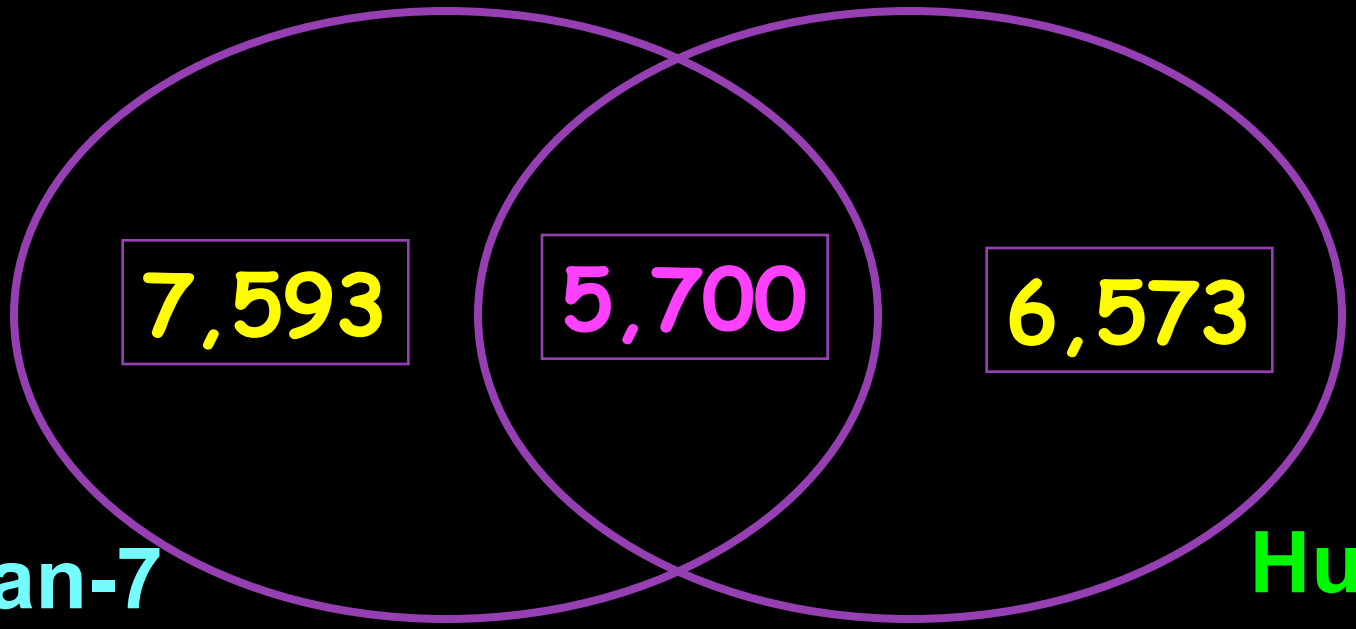
HMP WGS Metagenomic Sequencing

Has only just begun…..

Metagenomic Analysis of the Human Distal Gut Microbiome

Steven R. Gill,[1*‡] Mihai Pop,[1†] Robert T. DeBoy,[1] Paul B. Eckburg,[2,3,4]
Peter J. Turnbaugh,[5] Buck S. Samuel,[5] Jeffrey I. Gordon,[5] David A. Relman,[2,3,4]
Claire M. Fraser-Liggett,[1,6] Karen E. Nelson[1]

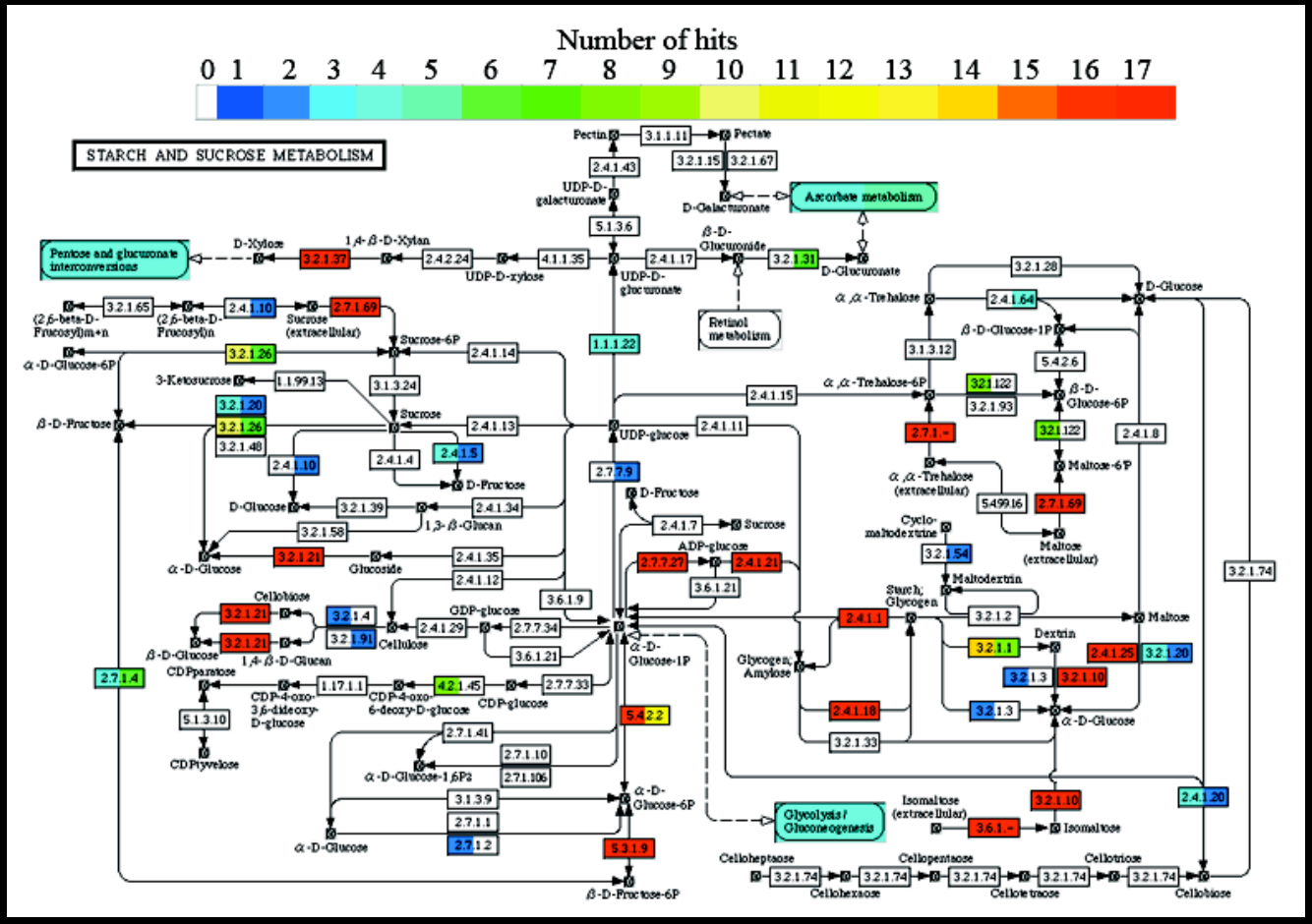
Amount of overlap shared between 19,866 unique blastx database matches for the two random libraries



Glycan metabolism

The plant polysaccharides we consume are rich in xylan-, pectin- and arabinose-containing carbohydrate structures. The human genome lacks most of the enzymes required for degrading these glycans.

At least twenty six different glycoside hydrolase families are encoded in the microbiome, many of which are not present in the human glycobiome.



Metagenomic Analysis of the Human Distal Gut Microbiome

Steven R. Gill,[1*‡] Mihai Pop,[1†] Robert T. DeBoy,[1] Paul B. Eckburg,[2,3,4] Peter J. Turnbaugh,[5] Buck S. Samuel,[5] Jeffrey I. Gordon,[5] David A. Relman,[2,3,4] Claire M. Fraser-Liggett,[1,6] Karen E. Nelson[1]

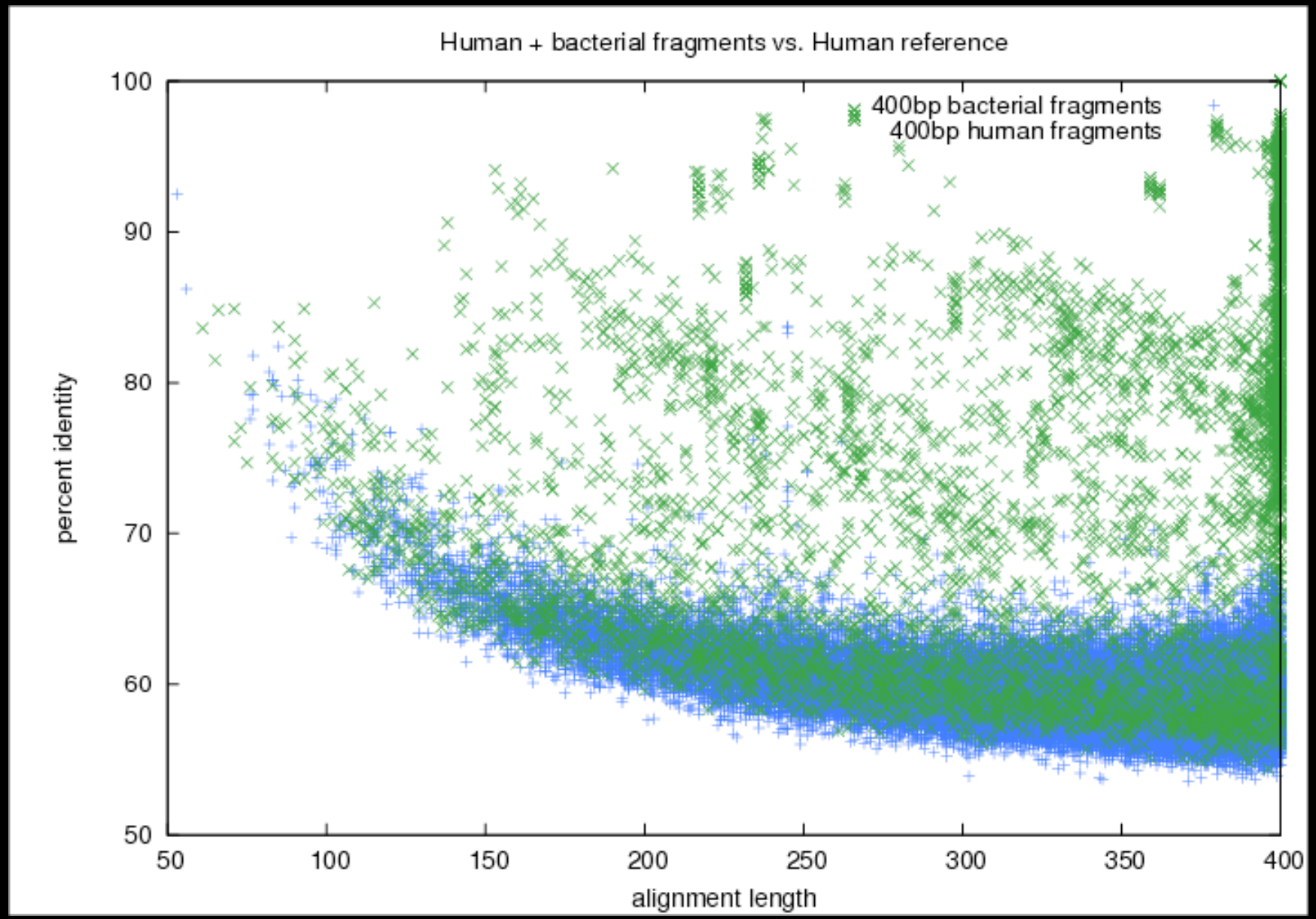
Metadata – data about the data

- About the strains
  - body site
  - repository id
  - much more
- About the human subjects
  - gender
  - age
  - dietary habits
  - smoker/non-smoker
  - and much, much more

Human Contamination

- not only microbial DNA in the samples
- possible identification of subjects based on their DNA sequence mixed in with the microbiome sequence, possible 18S sequence mixed in with 16S
- Filtering
  - Take out human

    OR
  - Only include what you are confident is bacterial
  - Both are problematic
- Currently a serious concern of NIH
  - dbGaP – database of genotypes and phenotypes, NCBI
    - Open OR Closed access options


Human + bacterial fragments vs. Human reference
400bp bacterial fragments
400bp human fragments
percent identity
alignment length
100
90
80
70
60
50
50
100
150
200
250
300
350
400

Initiative 2:
Reference Data Set Generation

Sequencing really ramps up now:

- 400 more Reference Genomes
- Sequence the samples from the 250 volunteers

- Plus moving into new territory – eukaryotes, viruses, unculturables

Initiative 3: Demonstration Projects

The demonstration projects are UH2/UH3 grants where the UH2 pilot phase lasts one year during which awardees have the opportunity to show the correlations they hypothesize between microbiomes and disease

A subset of the funded UH2 projects will be chosen to move on to the UH3 ramp-up phase and receive another 3 years of funding to further research on their topic.

15 UH2's have been funded, the following slides show the focus of a sampling of these projects

Skin disorders: Acne

- Huiying Li and Robert Modlin
  - The microbiome of the microcomedome is less complex then many other human environments and is thus more tractable for in-depth metagenomic sampling
  - They will examine the role of *P. acne*, examine its diversity in normal and disease states, and examine non-*P. acne* strains for their correlation with disease

Skin disorders: Psoriasis

- Martin Blaser
  - Inflammatory disease of the skin, highly prevalent, unknown cause
  - Examine the microbiome in skin with disease and normal skin in both patients and healthy individuals
  - Attempt to determine what role the microbiome may play in response to treatment

Febrile Illness

- Gregory Storch
  - Fevers of unknown origin
  - This project will examine the viral microbiome and determine if correlations can be found with febrile illness

vaginosis

- Gregory Buck
- Jacques Ravel

  - Vaginosis is a disease caused by a disruption in the normal vaginal biota, it can lead to increased risk for pregnancy loss, preterm birth, and STD infection
  - These projects will examine correlations between the microbiome, disease, and host characteristics/ behavior

NEC

- Phillip Tarr
- Necrotizing enterocolitis
  - Disorder that effects 10% of premature infants
  - Many lines of evidence that the microbiome plays a key role in this illness
  - Infants will be tracked and sampled every day from birth through age 35 days (as virtually all cases occur by age 35 days)
  - Thus samples can be compared pre- and post-event to determine correlations and possible causation

And several others ….

Supporting Initiatives: Research into Ethical, Legal, and Social Implications (R01s)

- Patient perspectives
- Federal regulations
- Policy analysis
- Privacy issues

Suporting Initiatives continued

Development of new tools for the computational analysis of HMP data (R01s)

- Annotation methods
- Phylogenetic analysis methods
- Metagenome assembly

Development of new technologies needed for studying the HM (R01s)

- Single cell sequencing
- Cell sorting
- Representation of rare community members
- Unculturable species

HMP management

- Working groups
  - lots of conference calls – and I mean lots
- Face-to-face meetings
- HMP Research Network
- Steering Committee

Huge Volume of Data

- Reference Genomes

+

- 16S RNA analysis of metagenomic samples from Ref. Data set and Demo projects

+

- Metagenomic sequence from Ref. Data set and Demo projects

= An unprecendented amount of data!

What is the DACC doing?

- Collecting and distributing all this data
- Helping with standardization
- Offering centralized common pipelines for Center and Demo project use
- Providing a repository of protocols and SOPs
- Providing access to sequence, annotation, and analysis data
- Providing a comprehensive web resource
  - www.hmpdacc.org

HUMAN MICROBIOME PROJECT

DACC

REFERENCE GENOMES | 16S RNA SEQUENCING | METAGENOMIC WGS | OUTREACH & TRAINING | DOWNLOADS | SOPs | RESOURCE REPOSITORY

www.hmpdacc.org

Welcome

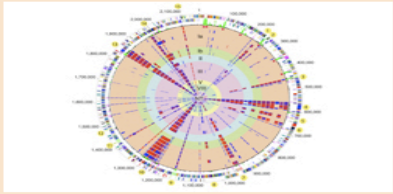
Welcome to the Data Analysis and Coordination Center (DACC) for the Human Microbiome Project (HMP). The HMP was launched by the National Institutes of Health Roadmap for Medical Research and is designed to fuel research into the multitude of microbes that live in the various environments of the human body. A major goal of the HMP is to look for correlations between changes in the microbiome and human health. More information about the project can be found on the NIH Roadmap website at http://nihroadmap.nih.gov/hmp.

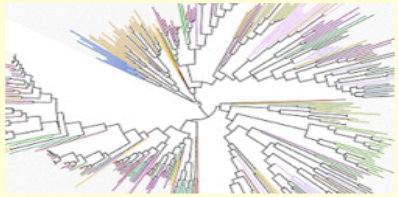
The HMP DACC is the central repository for all HMP data, providing specialized data management and analysis infrastructure to facilitate discoveries about the microbiome. The HMP-DACC web site will provide web-based query and visualization tools, comprehensive computational analysis of HMP data, quality control measures, and links to well-documented Standard Operating Procedures. The DACC is also strongly committed to outreach and training. Please visit the above tabs for more information on each of these topics.

Focus Areas of the HMP

Reference Genomes

Approximately 600 microbes from cultured and uncultured bacteria, plus several non-bacterial microbes will be sequenced. Combined with existing and other planned efforts, the total reference collection should reach 1000 genomes. These sequences will provide a benchmark against which further sequence data can be compared.



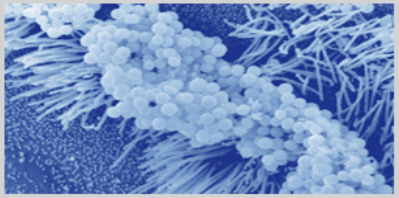


16S RNA sequencing

16s RNA sequencing will be used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome at each site.

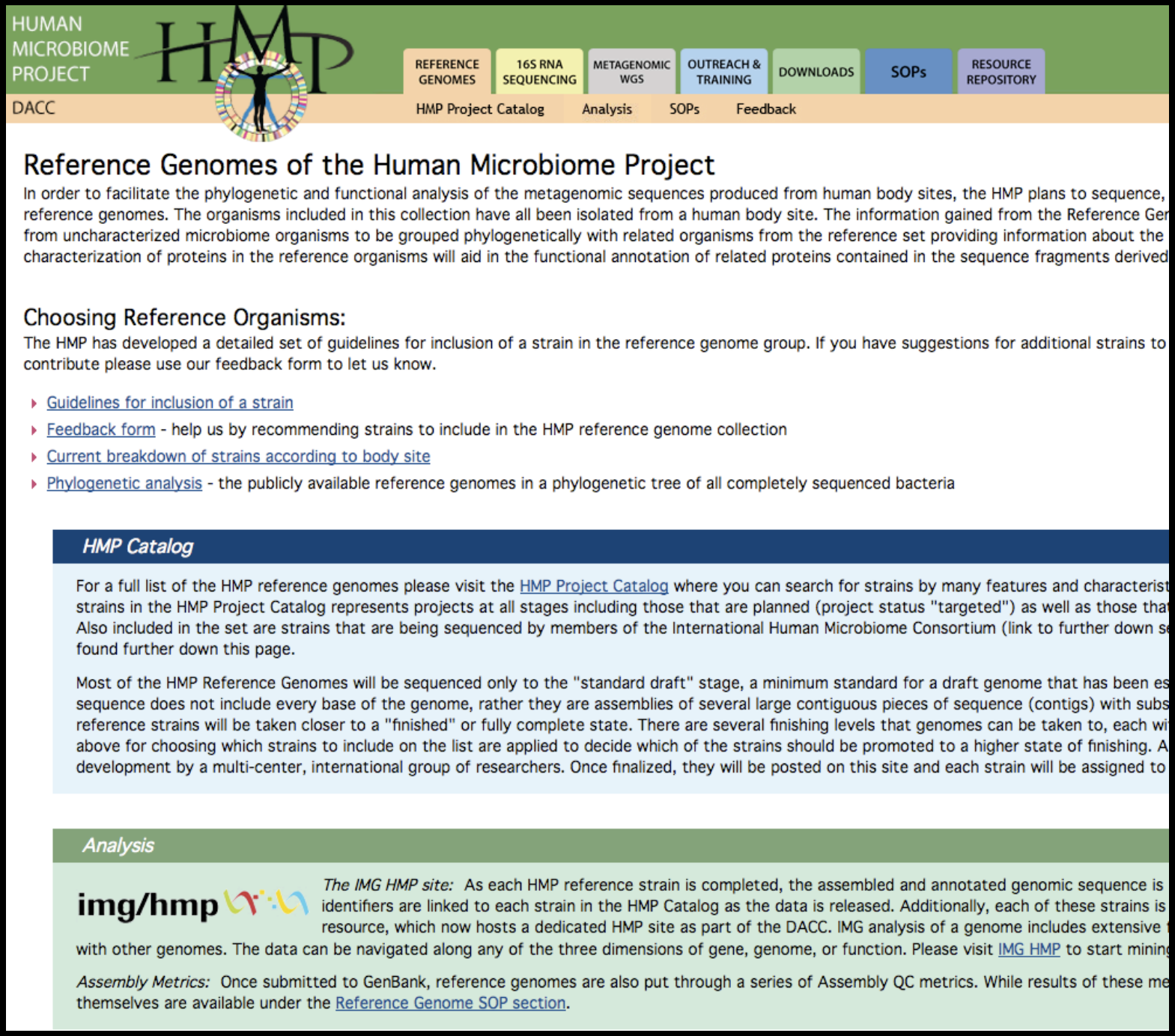
Metagenomic WGS

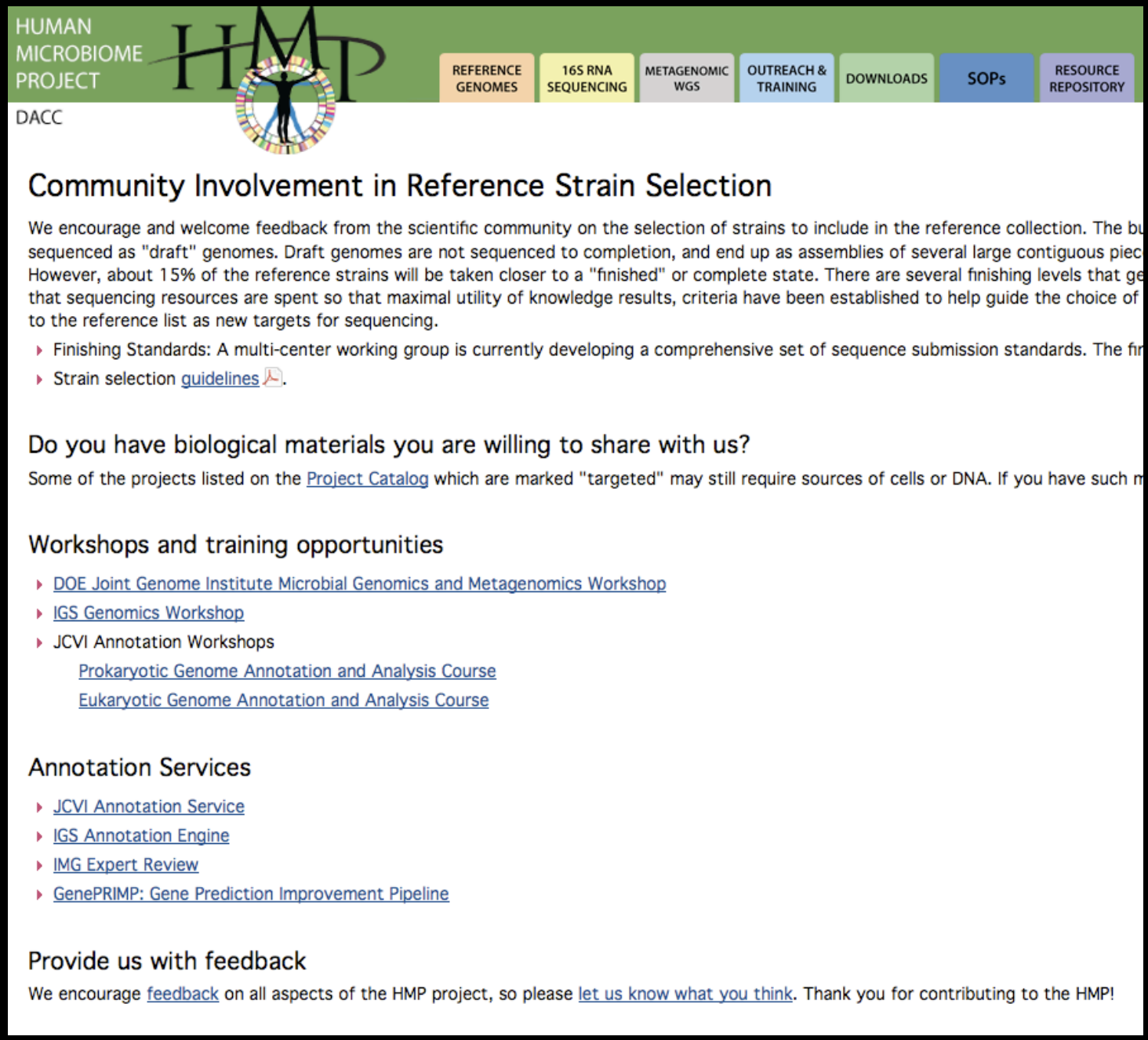
Expanding on the 16S data, Whole Genome Shotgun (WGS) sequencing will be performed on samples taken from human subjects. Coupled with the other data generated during the project, this will provide insights into the genes and pathways present in the human microbiome.



Outreach & Training

We encourage and welcome community feedback. As well, we offer a number of workshops and training opportunities.


HUMAN MICROBIOME PROJECT
HMP
DACC
REFERENCE GENOMES
16S RNA SEQUENCING
METAGENOMIC WGS
OUTREACH & TRAINING
DOWNLOADS
SOPs
RESOURCE REPOSITORY
HMP Project Catalog
Analysis
SOPs
Feedback
Reference Genomes of the Human Microbiome Project
In order to facilitate the phylogenetic and functional analysis of the metagenomic sequences produced from human body sites, the HMP plans to sequence,
reference genomes. The organisms included in this collection have all been isolated from a human body site. The information gained from the Reference Ge
from uncharacterized microbiome organisms to be grouped phylogenetically with related organisms from the reference set providing information about the
characterization of proteins in the reference organisms will aid in the functional annotation of related proteins contained in the sequence fragments derived
Choosing Reference Organisms:
The HMP has developed a detailed set of guidelines for inclusion of a strain in the reference genome group. If you have suggestions for additional strains to
contribute please use our feedback form to let us know.
Guidelines for inclusion of a strain
Feedback form - help us by recommending strains to include in the HMP reference genome collection
Current breakdown of strains according to body site
Phylogenetic analysis - the publicly available reference genomes in a phylogenetic tree of all completely sequenced bacteria
HMP Catalog
For a full list of the HMP reference genomes please visit the HMP Project Catalog where you can search for strains by many features and characterist
strains in the HMP Project Catalog represents projects at all stages including those that are planned (project status "targeted") as well as those tha
Also included in the set are strains that are being sequenced by members of the International Human Microbiome Consortium (link to further down s
found further down this page.
Most of the HMP Reference Genomes will be sequenced only to the "standard draft" stage, a minimum standard for a draft genome that has been es
sequence does not include every base of the genome, rather they are assemblies of several large contiguous pieces of sequence (contigs) with subs
reference strains will be taken closer to a "finished" or fully complete state. There are several finishing levels that genomes can be taken to, each wi
above for choosing which strains to include on the list are applied to decide which of the strains should be promoted to a higher state of finishing. A
development by a multi-center, international group of researchers. Once finalized, they will be posted on this site and each strain will be assigned to
Analysis
img/hmp
The IMG HMP site: As each HMP reference strain is completed, the assembled and annotated genomic sequence is
identifiers are linked to each strain in the HMP Catalog as the data is released. Additionally, each of these strains is
resource, which now hosts a dedicated HMP site as part of the DACC. IMG analysis of a genome includes extensive
with other genomes. The data can be navigated along any of the three dimensions of gene, genome, or function. Please visit IMG HMP to start minin
Assembly Metrics: Once submitted to GenBank, reference genomes are also put through a series of Assembly QC metrics. While results of these me
themselves are available under the Reference Genome SOP section.


HUMAN
MICROBIOME
PROJECT
DACC
REFERENCE GENOMES
16S RNA SEQUENCING
METAGENOMIC WGS
OUTREACH & TRAINING
DOWNLOADS
SOPs
RESOURCE REPOSITORY
Community Involvement in Reference Strain Selection
We encourage and welcome feedback from the scientific community on the selection of strains to include in the reference collection. The bu
sequenced as "draft" genomes. Draft genomes are not sequenced to completion, and end up as assemblies of several large contiguous piec
However, about 15% of the reference strains will be taken closer to a "finished" or complete state. There are several finishing levels that ge
that sequencing resources are spent so that maximal utility of knowledge results, criteria have been established to help guide the choice of
to the reference list as new targets for sequencing.
Finishing Standards: A multi-center working group is currently developing a comprehensive set of sequence submission standards. The fir
Strain selection guidelines.
Do you have biological materials you are willing to share with us?
Some of the projects listed on the Project Catalog which are marked "targeted" may still require sources of cells or DNA. If you have such m
Workshops and training opportunities
DOE Joint Genome Institute Microbial Genomics and Metagenomics Workshop
IGS Genomics Workshop
JCVI Annotation Workshops
Prokaryotic Genome Annotation and Analysis Course
Eukaryotic Genome Annotation and Analysis Course
Annotation Services
JCVI Annotation Service
IGS Annotation Engine
IMG Expert Review
GenePRIMP: Gene Prediction Improvement Pipeline
Provide us with feedback
We encourage feedback on all aspects of the HMP project, so please let us know what you think. Thank you for contributing to the HMP!

DACC Training Workshops: 2009

- IGS Genomics Workshop
  - July 28-30
  - September 15-17
  - November 10-12

    http://www.igs.umaryland.edu/workshop

- DOE JGI Microbial Genomics and Metagenomics Workshop
  - September 14-18

    http://www.jgi.doe.gov/meetings/mgm/

challenges and future directions

- Moving on to other members of the communities
  - Eukaryotes
  - viruses
- How to get representation of the rare members of the communities
  - These are likely the most interesting and we know very little about them
- Unculturables needed as reference strains
- Transcriptomics – what is being expressed

Related IGS activities

- GSCID
  - Genome Sequencing Center for Infectious Disease
  - Prokaryotic and eukaryotic pathogens and vectors of disease
  - White paper submission process
  - http://gscid.igs.umaryland.edu/
- Annotation Engine
  - Free annotation service for prokaryotic genome sequences
  - http://ae.igs.umaryland.edu/cgi/index.cgi

acknowledgements

- Owen White (PI), Jennifer Wortman (Project Manager)
- Heather Huot, Jonathan Crabtree, Mark Mazaitis, Victor Felix, Joshua Orvis, Anup Mahurkar
- Nikos Kyrpides, Victor Markowitz, Amy Chen, Todd DeSantis, Rob Knight, Gary Andersen and all the people who work with them on this project
- The Sequencing Centers

- NIH and the program staff on this project