Guidelines for Inclusion of Strains in the Microbial Reference Genome Collection & Determination of Finishing Level

The HMP plans to sequence to high-quality draft level, or collect from publicly available sources, a total of 1000 reference genomes isolated from human body sites. Approximately 15% of these will be taken to improved levels of finishing. Here we explain the biological justification for strain selection & identification of strains to be taken to improved finishing levels.

1. Phylogeny and uniqueness of the species.
It is anticipated that the finishing or improvement of the genomes of species that represent novel lineages will enable broad representation of as many lineages as possible, regardless of other criteria, and will provide improved scaffolding for the metagenomic data that are being produced. These genomes will also provide valuable information to groups beyond those involved in metagenomics studies.

2. Established clinical significance.
From the initial work within HMP body site-specific working groups, as well as from external sources and literature on individual strains, we have knowledge regarding relevance to health or disease states. We believe that any strain that has an **established** clinical significance to some health or disease condition should be included in the subset proposed to receive some level of improvement.

3. Abundance (dominance) in a body site.
Similarly, some strains have accompanying information on abundance and relative abundance in the various body sites. We believe that any strains that have **established** information on abundance in a body site should be included in the subset proposed to receive some level of improvement. Additional reasoning for these isolates includes:
(a) more predominant organisms will contribute the largest number of shotgun reads and thus should be sequenced to aid in identifying these reads;
(b) more prevalent organisms will most likely have a bigger impact on metabolic capabilities of the community and thus one would want to know their metabolic pathways. This can only be obtained by complete genome sequences or finished genomes.

4. Identical species found in different body sites.
For obvious reasons, duplicate species present an interesting data set that might have different metabolic capabilities dependent on the body sites where they are found. As an example, The Microbial Reference Genome Project Catalog currently includes isolates of *Gardnerella vaginalis* collected from vagina as well as skin.

5. Opportunity to explore pan-genomes.
Again, isolates that have already been closed by other genome sequencing efforts outside of the HMP may be from other environmental niches, and by having additional closed isolates we can obtain more information on the associated pan-genomes. For example, we are all aware of the extra Megabase of DNA obtained when the genome of *E. coli* O157 was compared to *E. coli* K12 as the finished reference genome.

6. Poor quality draft assembly that needs some improvement.
In situations where a genome did not assemble well, we may propose some level of manual improvement to yield a better assembly

7. Other.
In situations where there is some valid criteria other than those justifications listed above.