## Strains working group: Proposal for selecting, finishing, or "improving" strains:

The Genome Centers are proposing to sequence and/or collect a total of 1000 reference genomes, and to upgrade 15% of these draft genome sequences to further levels of finishing or improvement. This document sets out a biological justification for choosing strains so that a clear rationale for those strains to be included or upgraded can be evaluated and approved by the NIH. These are:

**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 with the sub-working groups, as well as from other sources and literature on the individual strains, we do have knowledge on 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 include:

(a) the more predominant organisms will contribute the largest number of shotgun reads and thus should be sequenced to aid in identifying these reads;

(b) the 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. Duplicate species but found in different body sites.**
For obvious reasons, duplicate species present an interesting data set that might have different metabolic capabilities dependent on which body sites they are found. For example on the strain Master list we currently have isolates of *Gardnerella vaginalis* that have been 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 criteria other than those justifications listed above.