Contamination Testing:

The 13 most abundant subtypes of each of the four *Salmonella enterica* subsp. *enterica* serovar schemes available in BioHansel (Enteritidis, Heidelberg, Typhi, Typhimurium), as well as the 13 most abundant subtypes of the *Mycobacterium tuberculosis* scheme, were used to test the detection of contamination by the BioHansel QC module. To get these most abundant subtypes for each pathogen, BioHansel was run on directories of Unicycler assemblies that contained all the publicly available sequencing data for each of the pathogens. Then, a small python script was run on the final BioHansel output which just tallied up the 13 most abundant subtypes and put it into a *.txt file.

One draft assembly or closed genome was chosen for each of these most abundant subtypes (the accessions are listed in Tables S3 to S7) and was used to create synthetic raw reads using the create_reads.sh bash script. This script makes use of Art Illumina to generate synthetic next generation sequencing (NGS) Illumina reads based on the MiSeq v3 with inputs for a seed (42), read length (250), and fragment size (800). The create_reads.sh script looks at the folder it is in for any *.fasta files and uses those to create reads of the same seed and size. This process generates 26 synthetic paired end reads (13 pairs) that will be used for the contamination testing

Contamination thresholds were set as differing fractions of one BioHansel subtypes reads contaminated by another subtypes reads to see how well BioHansel could detect mixed datasets. The fractions used were: 0/60, 1/59, 3/57, 6/54, 8/52, 10/50, 20/40, 30/30. The subtypes were combined in a matrix all against all such that in the end, 156 contaminated datasets were generated per contamination fraction. An example of a smaller 10/50 matrix would look as such:

		Contaminant Subtype (%reads = 20)				
Sample Subtype (%reads = 80)		1	1.1	1.2	1.3	2
	1	x	1 + 1.1	1 + 1.2	1 + 1.3	1 + 2
	1.1	1.1 + 1	x	1.1 + 1.2	1.1 + 1.3	1.1 + 2
	1.2	1.2 + 1	1.2 + 1.1	x	1.2 + 1.3	1.2 + 2
	1.3	1.3 + 1	1.3 + 1.1	1.3 + 1.2	x	1.3 + 2
	2	2 + 1	2 + 1.1	2 + 1.2	2 + 1.3	x

Here, the first number represents the subtype of which the sample would contain 80% of its sequence data from and the second number would be the contaminant where the sample would contain 20% of its sequence data (for the 10/50 contamination). If this matrix was scaled up to have 13 subtypes on each of the axes, then there would be 169 (13 *13) boxes of which 156 (13 *12) were the contaminated reads.

This matrix was setup using the contaminate_data.sh script which takes into account the genome size and the read length to determine how many reads from the subtype and the contaminant should be used for each contaminated dataset. These contaminated reads were made using seqtk which takes random reads based on a seed (42 was used here) to and outputs them to the newly

created contaminated read file. These new reads are then sorted based on their contamination level into different directories.

Finally, BioHansel was run using the -D (directory) command on the different contamination level directories to see if the contamination was detected with the subtype flagged as a fail. Bash commands were used to check how many of the 156 data points failed the contamination test for each subtype at each contamination level.