Comparison with GenoTyphi: results analysis for the samples that failed BioHansel QC

For the 16 Typhi isolates that failed QC, out of 1,910 datasets analyzed:

1) One dataset was a Paratyphi A strain used in Wong *et al.* (2016) as an outgroup for the phylogenetic tree of the Typhi strains (GenBank ID ERR326600); it failed QC because it has no genotype defined in this genotyping scheme and was missing >5% of the SNP targets.
2) Three datasets were missing >5% of the scheme SNP targets and had mixed signals for both the Wong *et al*. genotypes 4.2.1 and 4.2.2 (GenBank ID ERR204256, ERR213233, and ERR213234; the first 2 were identified as genotype 4.2.1, and the last one as 4.2.2 in the Wong *et al.* paper)
3) One sample produced an unconfident result due to a missing hierarchy level: the sample genotype was 2.3.1 but level 2.3 was missing according to the Wong *et al.* nomenclature (GenBank ID ERR360616). It was identified as genotype 2.3.1 in the Wong *et al.* paper.
4) The other 11 datasets that failed the QC check came out as "mixed sample" when processed by BioHansel as raw (unassembled) Illumina reads. In each case, one of the genotypes identified by BioHansel corresponds to the genotype identified by Wong *et al.*

The one mixed dataset (ERR279178) which failed QC and which gave discordant results between BioHansel and the Wong *et al*. manuscript included three positive genotypes: 4.1.1 with 12X coverage; 2.4.1 (which was reported by Wong *et al*.) with 14X coverage, and 2.2.0 with 43X coverage. The dominant genotype appears to be 2.2.0 in this mixed dataset, based on k-mer coverage values.

Another example of a mixed dataset which failed QC: the WGS dataset # ERR340783 identified as genotype 4.3.1 by Wong *et al*. failed the QC check as it contains sequences from two distinct genotypes (mixed dataset or mixed culture). The BioHansel QC message was: "FAIL: Mixed genotypes found: "2.3.6.1; 2.3.6.3" (corresponding to the original Wong *et al*. genotypes 4.1 and 4.3.1). In the detailed match_results from BioHansel, it can be seen that the k-mer target for 2.3.6.3 (4.3.1) had much more coverage than the k-mer target for 2.3.6.1 (4.1): the negative 2.3.6.1 k-mer had coverage of 63x and the positive had coverage of 23x. Whereas, the negative 2.3.6.3 k-mer had coverage of 20x and the positive had coverage of 69x. Comparing this to the total average coverage of 78.276x, the results are consistent with the dataset being a mixture of two genomes: one sequence from genotype 2.3.6.3 with an approximate genome coverage of ~66x, and the other sequence from genotype 2.3.6.1 with an approximate genome coverage of ~21x. It is likely that the assembly process would have removed evidence of contamination, such that the assembled genome would appear to belong to genotype 2.3.6.3 (e.g. Wong *et al*. genotype 4.3.1), which has the highest genome coverage.

Reference:

Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, Murphy N, Holliman R, Sefton A, Millar M, Dyson ZA, Dougan G, Holt KE, International Typhoid Consortium, 2016. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun. 7:12827. doi: 10.1038/ncomms12827.