**A.1. BioHansel Manual**

<https://bio-hansel.readthedocs.io/en/readthedocs/>

**A.1.2. Scheme Input**

Detailed instructions on BioHansel scheme creation can be found online at<https://bio-hansel.readthedocs.io/en/readthedocs/index.html> under the “Subtyping Schemes” section.

A BioHansel SNP scheme is specified as a FASTA file containing at least one record.  Each record specifies a k-mer with a header in the following format:

><position>-<genotype code>

The position is specified as the optional keyword ‘negative’ followed by a positive integer.  The genotype is a numerical

The position is

The format of the k-mers in the fasta file should be the following (for two target SNPs at position 615938 and 3021283 of the reference genome):

>615938-1

CCGGCCTGCTCTCCGA**A**GCACTGACGGATGCCG

>negative615938-1

CCGGCCTGCTCTCCGA**G**GCACTGACGGATGCCG

>3021283-1.1

CCACCTTGGGCTTGCG**A**GTCTACCTCGCGTGGA

>negative3021283-1.1

CCACCTTGGGCTTGCG**G**GTCTACCTCGCGTGGA

**Nomenclature/requirement of the software:** The name of the k-mer should be the position of the SNP site in the reference genome, a dash, and the genotype name.  The name of the alternate base k-mer (the sequence including the “negative” base for the genotype) that is present in all other isolates outside of that genotype, should be preceded by “negative”, as shown above.  I have put in bold the base that differs between the genotype-specific k-mer and the “negative” k-mer above.

**A.1.3. Output**

BioHansel supports three modes of output: simple, standard, and verbose.

Simple Results:  Technician-friendly, includes average k-mer coverage as proxy for genome coverage

Results:  in-depth results for debugging; includes detailed summary of k-mer targets found vs. k-mer targets expected

Detailed Output:  all k-mer matches, useful for genotyping scheme development

A.1.3.1. Equations

Average k-mer coverage = sum of the coverage for all targets found divided by the number of targets found

**A.2. Scheme Development**

Detailed instructions on BioHansel scheme creation can be found online at<https://bio-hansel.readthedocs.io/en/readthedocs/index.html> under the “Subtyping Schemes” section.

A program called “FEHT” is used to do find SNPs that correspond to each group defined in a metadata file (https://github.com/chadlaing/feht). This program is written in Haskell, and can be installed from bioconda.  The program needs two input files: a metadata file (with hierarchical codes), and a SNV table or VCF file.

Each hierarchical code label consists of one or more  integer identifiers (Xn, Xn ≥ 0) separated by dots representing subsequent branches of the hierarchy.  Each level must be given an identifier, but the identifiers need not be contiguous.  For example: the scheme labeled with 1, 1.1, 1.2, 1.5, 1.5.1 would represent a valid set of names for a scheme.  Each level of the hierarchy must be specified. The scheme labeled with 1, 2, 2.1, 1.5.1, 1.5.2, 1.5.3 would not be valid as it is missing a label for genotype 1.5. For example of current BioHansel nomenclatures, see Fig I a-e under the folder “Population\_structure\_backbones” in section 8.1 of the Supplementary materials.

After the SNP positions are identified, the DNA sequence around the SNP site that is conserved across the whole population can be used to form the k-mer target. The current BioHansel scheme have a k-mer lenght of 33 bases. The sequences need to be an exact match for the BioHansel tool to find them, so only one SNP site can be tolerated in the 33 bases k-mers across the whole population.  The genotyping scheme development therefore needs to be done very carefully, and extensively validated across hundreds or thousands of isolates. See current versions of typing schemes in Supplementary materials in section 8.1 in the folder: “BioHansel schemes”.

K-mer target selection should also confirm the k-mer is present in both assembled and raw sequence data at consistent coverage levels. Due to the random nature of Illumina data missing sections of the genome, BioHansel was designed to allow for a portion of k-mers to be missing to assign a type with a default of 5% of the target scheme. New schemes or additions of new targets should be rarely missing from samples to ensure confidence in results.

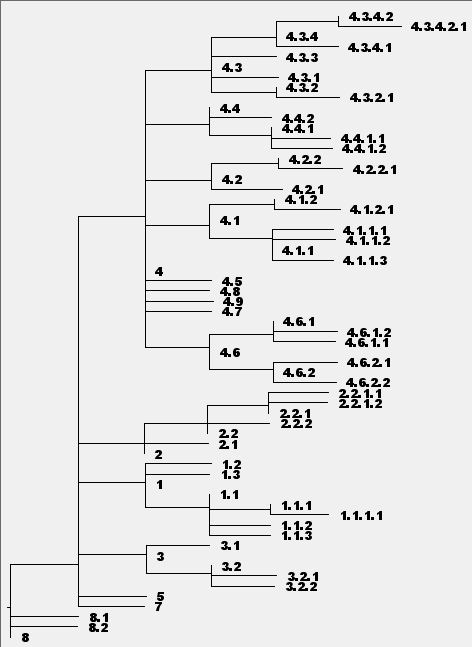
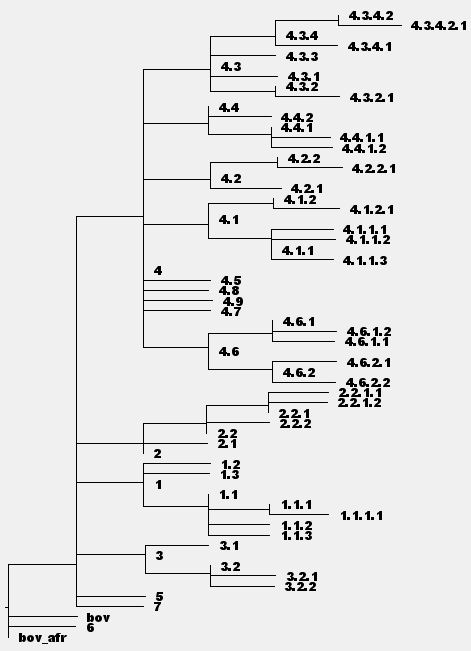
Care will need to be taken with the addition of additional SNP targets per genotype since recombination and intermediate genotypes are possible within a population. Schemes should make efforts to exclude regions with evidence of recombination and use a sufficient set of genomes to capture the routine genetic diversity that is encountered for a given pathogen. Inclusion of multiple SNP targets to define each genotype makes schemes more robust against missing data and recombination events if the designers apply their organism specific knowledge to take into account for its biological complexity.

**A.3 Adaptation of existing SNP-based genotyping schemes for use in BioHansel**

There are 3 main steps for the adaptation of a SNP scheme:

***1-Verify that the nomenclature is strictly hierarchical, and is formatted correctly***

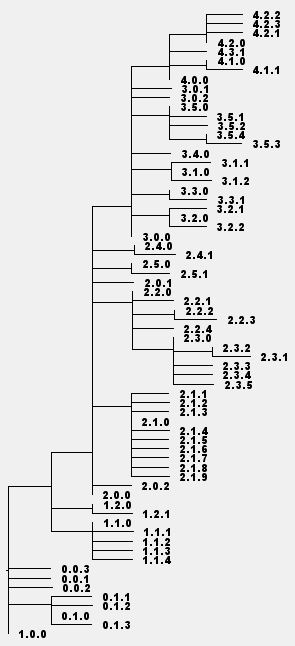
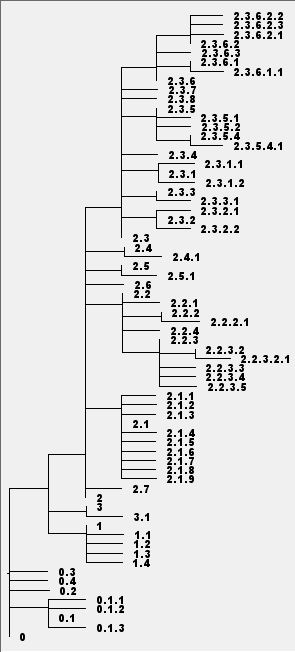
See section A2 above for the correct nomenclature format needed for use in BioHansel. See below the changes made to the genotyping nomenclatures published for *Mycobacterium tuberculosis* (Fig II) and *Salmonella* ser. Typhi (Fig III).



A

B

Fig II. Adaptation of the nomenclature for the *M. tuberculosis* genotyping scheme. Panel A shows the backbone of the population structure based on the SNP positions defined in the genotyping scheme, with the original nomenclature published by Coll et al (2014). Panel B shows the same backbone tree with the nomenclature used in the BioHansel scheme tb\_speciation v1.0.5. Only 3 of the 63 genotypes needed to be renamed for use in BioHansel (red and green circles). See caption for Fig I a-e in section 8.1 of the supplementary materials for additional details on the generation of the population structure backbones.

B

A

Fig III. Adaptation of the nomenclature for the *Salmonella* ser.Typhi genotyping scheme. Panel A shows the backbone of the population structure based on the SNP positions defined in the genotyping scheme, with the original nomenclature published by Wong et al (2016). Panel B shows the same backbone tree with the nomenclature used in the BioHansel scheme Typhi v1.2.0. Most of the genotypes needed to be renamed for use in BioHansel. See caption for Fig I a-e in section 8.1 of the supplementary materials for additional details on the generation of the population structure backbones.

For the *M. tuberculosis* genotyping scheme, the original nomenclature published by Coll et al. was strictly hierarchical, except for the outgroups “bov”, “bov\_afr”, and group 6: only these 3 groups (Fig II) needed to be renamed in order to be compatible with BioHansel. In the case of the *Salmonella* ser. Typhi genotyping scheme, the original nomenclature did not follow a strictly hierarchical pattern, so most genotype needed to be renamed for use in BioHansel (Fig III). BioHansel includes metadata table that will provide to the user the original genotyping nomenclature along with the nomenclature adapted for BioHansel.

***2-Create positive and negative k-mer for the scheme***

Retrieve k-mer sequence surrounding the target SNPs create the BioHansel scheme, including the positive and negative k-mer targets. Create and save the scheme in a fasta file format, using the guidelines described above in sections A1 and A2.

***3-Test the scheme using BioHansel to correct possible nomenclature or k-mer issues***

Run BioHansel on datasets (ideally using both unassembled WGS data and assembled genomes) representing all the genotypes defined in the scheme, and verify that the expected results are obtained, and that at least one representative of each genotype does “PASS” the BioHansel QC. If samples FAIL QC, examine the detailed results from BioHansel to see if the nomenclature is causing problems (in this case all of the samples from the genotype will FAIL QC), or if a k-mer target is not found (in that case, samples may PASS or FAIL QC depending on how many SNPs are used to define each genotype). If a k-mer target is not found, a sequence alignment may be required to see if changes can be made to avoid a problematic genome region (for example, shift the k-mer sequence slighlty upstream or downstream of the target SNP to avoid a region that is not conserved across the whole population), or if degenerate bases should be used as part of the k-mer sequence to account for genetic variability across the population. It is best to try the scheme on as many datasets as possible in order to differentiate between a scheme problem and a low quality WGS dataset or assembly.

**A.4 QC module messages**

A detailed description of the QC module can be found at: <https://bio-hansel.readthedocs.io/en/readthedocs/> in the section “Output: Quality Control”

***Insufficient k-mers found***

A sample will fail QC if there are too many missing k-mers (default is >5% of the number of target SNP positions specified by the scheme).  A target SNP position is considered missing if it is present in the scheme, but the corresponding k-mer(s) are not found in the input data at a sufficient level of coverage.  When processing a FASTA file the default level of coverage required for each k-mer target to be considered present is set to 1x. When processing a FASTQ file the default level of coverage is set to 8x.

A sample will also fail QC if there are too many missing positive k-mers from a single genotype (3 by default).  For example, if a scheme were to associate 5 positive k-mer targets with genotype 1.3.2, but a sample identified as genotype 1.3.2.2 was only found to have 2 of them, it would be reported as missing too many SNP targets for a reliable genotyping call.

***Detection of inconsistent results***

Each hierarchy level present in the scheme nomenclature must be defined by a minimum of one k-mer, and a sample will fail QC if all target k-mers from a higher or lower level in the hierarchy are missing.  For example: If a scheme specifies k-mer targets for 2, 2.1, 2.2, 2.3, 2.3.1, 2.3.2, 2.3.1.1, and 2.3.1.2, but only k-mer targets for 2, 2.3.1, and 2.3.1.2 are found in the sample then even though the sample would be declared genotype 2.3.1.2 it would be reported as unconfident [missing targets for 2.3].  For the same scheme if only k-mer targets for 2, 2.3, and 2.3.1 were found the sample would be declared genotype 2.3.1 and reported as unconfident [missing targets for 2.3.1.1 or 2.3.1.2].

***Detection of possible mixed samples***

A sample will fail QC if there are hierarchical inconsistencies in the targets found.  For example: if a positive target k-mer with the hierarchical rank of 1.2 was found and a positive target k-mer with the hierarchical rank of 1.1.5 was also found, then the sample would fail QC and be reported as a potential mixed sample.

If a scheme includes both positive and negative target k-mers then a sample will also fail QC if both the positive and negative target k-mers for the same SNP position are found.  This is also reported as a potential mixed sample.

***Intermediate genotype***

A warning is issued if the sample passes all of the above quality controls, but there is a mixture of positive and negative k-mer targets present for the final genotype (warning: intermediate genotype).

***Low coverage***

When the input data is in the FASTQ format, a warning is issued if the sample has an average k-mer coverage level below the set QC threshold (20 by default) (warning: low coverage).

**List of QC defaults:**

Min k-mer frequency/coverage default = 8 (--min-kmer-freq)

Max k-mer frequency/coverage default = 1000 (--max-kmer-freq)

QC: Frequency below this coverage are considered low coverage default = 20 (--low-cov-depth-freq)

QC: Min number of kmers missing for Ambiguous Result default = 3 (--min-ambiguous-kmers)

QC: Decimal Proportion of max allowed missing kmers default = 0.05, valid values {0.0 - 1.0} (--max-missing-kmers)

QC: Decimal Proportion of max allowed missing kmers for an intermediate subtype default = 0.05, valid values {0.0 - 1.0} (--max-intermediate-kmers)

QC: Overall kmer coverage below this value will trigger a low coverage warning default = 20 (--low-cov-warning)

Degenerate bases in k-mers:  By default, the tool issues an error to the user if the total number of k-mers to be created from a scheme including one or more degenerate bases exceeds 100,000; this default maximum value can be modified by the user at runtime.

**References:**

Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG, 2014. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 5:4812. doi: 10.1038/ncomms5812.

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