Instruction of use for incorporation of the mutations catalogue version 2 results into bioinformatic pipelines

Initial considerations

Aiming for the most accurate and simplest incorporation of the graded-variants into any existing bioinformatic pipeline, we have chosen the nucleotidic changes using the genome sequence as reference for unit of matching (hereafter termed genomic-variant), for the following reasons:

- All laboratory experiments and bioinformatic pipelines rely on DNA sequencing to this date to identify drug resistance markers
- Graded-variant can represent changes on either protein or nucleotide sequences. Choosing the most universal coordinate system (i.e. genomic-variant) leads to better consistency.
- The Variant Calling Format (VCF) standard is already well established and most likely to be used in the primary steps of any bioinformatic analysis

All our analyses use as **genomic reference NC_000962.3**, thus the genomic-variants we are circulating refer to that sequence only. If you are using any other genomic references in your bioinformatic analyses, you will first have to **translate our genomic-variants into your genomic reference** of use.

Excel file

In the excel file, two sheets are provided. The second sheet, "Genomic_coordinates" provides the mapping between the graded-variant ("*variant*" column from the grading sheet) and all associated genomic-variants. The columns that uniquely define a single genomic-variant are identical to the fields used in the VCF specifications:

- *chromosome* ([CHROM] in VCF)
- *position* ([POS] in VCF)
- *reference_nucleotide* ([REF] in VCF)
- *alternative_nucleotide* ([ALT] in VCF)

To perform the matching inside your bioinformatic pipelines please perform the following steps in your preferred coding language:

1. read the second excel sheet using the first row as header.
2. using **all last four (*chromosome, position, reference_nucleotide, alternative_nucleotide*) columns**, perform an **exact match** on your own list of variants, which must include the same set of information. Alternatively, as the *chromosome* column is constant in our case, you can use only the three last columns.
3. this step will allow you to link your own list of mutations to the graded-variant naming convention we have used for the purposes of the analyses in the mutations catalogue version 2.
4. read the first excel sheet ("Catalogue_master_file") using row number 3 as header.
5. perform an exact match between the graded-variant value you have incorporated at step 3, and the "*variant*" column from the excel sheet that you have read at step 4
6. this final step will allow you to perform the final link between your own list of variants and the v2 grading for each drug

VCF file

In addition to the excel sheet, we are also providing a VCF file providing the same data. This file is purely coordinate base, and includes an "INFO "graded_variant" tag which will include all possible graded-variant linked to a genomic-variant. **The VCF will not include any GENOTYPE column. Similarly to the excel data, exact matching must be performed on all four CHROM, POS, REF and ALT columns.**

However, as each genomic-variant will only occur once in our VCF file, thus all graded-variant associated with it **have been concatenated and separated with an ampersand ("&") in the INFO "graded-variant" tag**. These values **must be split on subsequent operations before performing an exact match merge with the final grading data**.

Important additional information about the matching protocol

- **Each graded-variant can be linked to more than one genomic-variant** (for instance, different genomic-variants can lead to identical missense)
- **Each genomic-variant can be linked to more than one graded-variant** (for instance, multiple consecutive genomic-variants are split into each of their atomic, constitutive changes)

- All our genomic-variants inserted into our database are normalized using *bcftools norm.* This is step is particularly relevant for insertion and deletion genomic-variant consistent representation, thus matching for these genomic-variant will only be guaranteed if your own list of variants is properly normalized as well.
- We do not provide genomic-variants for deletion graded-variants (for instance, for the graded-variant “pncA_deletion”)
- We do not provide genomic-variants for LoF graded-variant that are never classified as group 1 or 2 for any drug or that are not subject to an epistatis rule
- For every gene of our candidate gene list, we have included all theoretical single, multiple, constant-length genomic-variants in our coordinate data. Following that, every possible genomic-variant leading to a missense or non-sense graded-variant will appear in the coordinate data, even if we did not observe that genomic-variant in our database of samples
- Contrarily to constant-length changes, we cannot ensure that all varying-length graded-variants are associated with all possible genomic-variant theoretically leading to them (because an infinite number of nucleotide changes could potentially be involved). We can only ensure they are linked to all genomic-variants we have observed. It is thus possible that some rare genomic-variants leading to in-frame or frameshift insertions and deletions graded-variants, escape the automatic matching proposed and described here. However, we have provided all genomic-variants that occur in our full database of samples, which is a more extensive set of samples included in the grading analysis. Flagging of these variants will require to be implemented in an additional step.