FIND

Seq&Treat

# MULTIPLE CONSECUTIVE NUCLEOTIDE VARIANTS

WHY WE NEED TO CARE

# HANDLING DOUBLE AND TRIPLE CODON VARIANTS

- We need to correctly identify double and triple codon variants so that the exact amino acid change can be determined

It's job of the genotyper, which phases appropriately nearby variants:

| | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| Reference | CGA (Arg) | CGA (Arg) | CGA (Gln) |
| Alternative | 100% CGT (Arg) | 50% CGT (Arg)<br>50% CAA (Gln) | 50% CAT (His) |

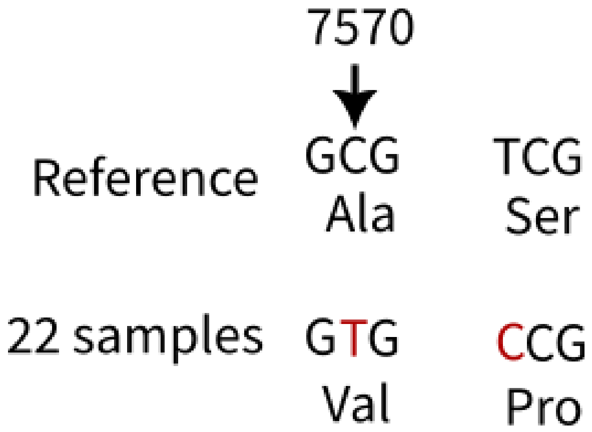
All scenarios lead to different consequence on the protein sequence and need to be accounted for.

THE ISSUE

# THE GENOTYPER DOES NOT KNOW THE CODON LOCATION

- When genotyping, all variants distant less than 2bp away will be grouped together, irrespective of whether they fall on the same codon
- Ideally, we only need to phase variants that fall on the same codon, all other variants could be unphased without consequence
- Real world example on *gyrA* (22 samples in TBKB):

| | | | |
|---|---|---|---|
| NC_00962.3 | 7570 | CGT | TGC |



| | | |
|---|---|---|
| Reference | GCG<br>Ala | TCG<br>Ser |
| 22 samples | GTG<br>Val | CCG<br>Pro |

- If the genotyper unphased the variants, the interpretation would be just the same:

| | | | |
|---|---|---|---|
| NC_00962.3 | 7570 | C | T |
| NC_00962.3 | 7572 | T | C |

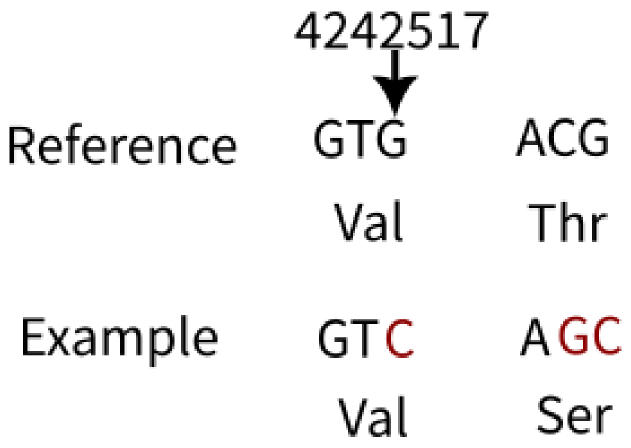
# ETL TO ASSOCIATE ALL RELEVANT FEATURES TO MCNVS

1. We split all MCNV into its constitutive SNV ("atomization")
2. Make sure we don't interpret separately SNVs on the same codon
3. Handle MCNV which constitutive variants each lead to missense but together to synonymous
4. If the MCNV leads to a synonymous on a codon, and each atomic change is also synonymous, we can decompose
5. Handle non CDS related variants (upstream, ribosomal)
6. **Extract missense + nearby synonymous**
    1. The ETL must consider exactly were each constitutive SNV is located on the gene sequence and discard synonymous that fall on a missense associated with the MCNV
    2. If a synonymous is associated with a constitutive SNV and does not fall on any codon where a missense is predicted, then it can be extracted for the MCNV
    3. **Priority is always given to missense (i.e. a missense can exclude a synonymous, the reverse not possible)**

# THE ISSUE

- To determine whether two change occur on the same codon, we use their position on the CDS
- In some cases, the position of missense changes was incorrectly set
- Consequence: some synonymous variants were linked to MCNV although they should have been overseeded by the missense.
- Example (*embC*):

| | | | |
|---|---|---|---|
| NC_00962.3 | 4242517 | GACG | CAGC |



4242517

| | | |
|---|---|---|
| Reference | GTG | ACG |
| | Val | Thr |
| Example | GTC | AGC |
| | Val | Ser |

- Correct features: `embC_c.2655G>C + embC_p.Thr886Ser`
- Incorrect features: `embC_c.2658G>C + embC_p.Thr886Ser`

# OUTPUT FILES NEED CORRECTION

- Current genomic coordinates include 116126 unique entries
  - 462 entries are corrected
  - 894 entries are added
  - Few examples:

| Position | Ref | Alt | Old | New |
|---|---|---|---|---|
| 624 | CGAG | TGCC | dnaA_p.Glu209Ala | dnaA_c.624C>T<br>dnaA_p.Glu209Ala |
| 624 | CGA | TGC | dnaA_p.Glu209Ala | dnaA_c.624C>T<br>dnaA_p.Glu209Ala |
| 903 | CA | GC | dnaA_p.Ile302Leu | dnaA_c.903C>G<br>dnaA_p.Ile302Leu |
| 8579 | CA | TG | gyrA_p.Ile427Val | gyrA_c.1278C>T<br>gyrA_p.Ile427Val |

# EFFECT ON INPUT FEATURES OF THE MODELS

- Overall, synonymous variants are not estimated at all in catalogue v2 :
    - "Silent mutations were assumed to be neutral (“aS”) and were maskedbefore step b (page92)"
- However, there was a "stage 3" implemented to have a look at potential synonymous variants of interest (cf Leonid)
- After propagating the fix, a new database extraction and rerunning R SOLO algorithm:
    - Compared the outputs of database extraction with and without the synonymous fix
    - Compared primary statistics (ie present/presentR/presentS/soloR/soloS)
    - One change out of the 116 variant/drug pairs:

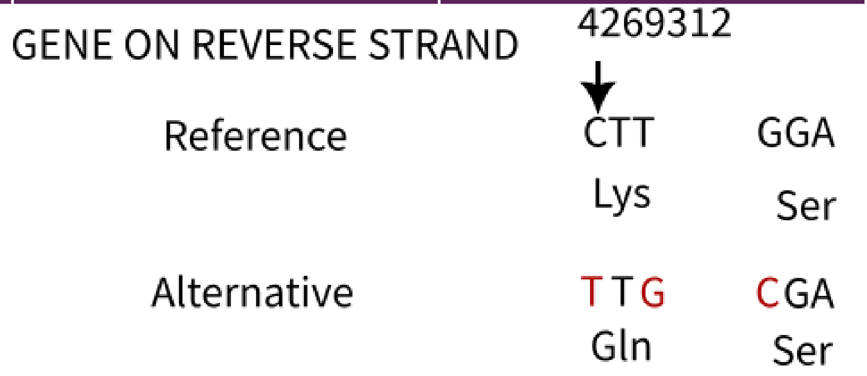
| drug | variant | SOLO_R | SOLO_S before fix | SOLO_S after fix |
|---|---|---|---|---|
| Ethambutol | embB_c.54G>T | 1 | 81 | 80 |

EDGE CASE

# GRADED-VARIANT VS SEEN-VARIANTS

- Initially, only the graded-variant were planned to appear on the genomic coordinates files
- I.e. only variants appearing in the result excel sheet could appear in the second sheet and in the VCF
- However, this led to tricky consequences for MCNV:

| | | | |
|---|---|---|---|
| NC_000962.3 | 4269312 | CTTG | TTGC |



GENE ON REVERSE STRAND

4269312

Reference CTT GGA

Lys Ser

Alternative TTG CGA

Gln Ser

All associated features: `ubiA_c.519C>G + ubiA_p.Lys174Gln`

However:

- `ubiA_p.Lys174Gln` is not present in the catalogue
- `ubiA_c.519C>G` is present in the catalogue

# PROPOSED LOGIC

- Now following rule is applied for MCNVs:
  - If none of the features associated with an MCNV are present in the catalogue, the MCNV will not appear at all in the coordinate file
  - If one of the features associated with the MCNV is present in the catalogue, the MCNV will appear in the coordinate file, and it will be associated with all features it is associated with.
  - Compare:

| Chrosomome | Position | Ref | Alt | Current | Fix 1 | Fix 1 + Fix 2 |
|---|---|---|---|---|---|---|
| NC_000962.3 | 4269312 | CTTG | TTGC | ubiA_c.522G>A | ubiA_c.519C>G | ubiA_c.519C>G<br>ubiA_p.Lys174Gln |