

# Interpreting PneumoCaT Results

---

*Written by Georgia Kapatai – 8<sup>th</sup> April 2018*

## Understanding your output files:

Running PneumoCaT will generate an output directory; if not specified otherwise the output path will be inferred from your input directory or fastq path and a `pneumo_capsular_typing` directory will be generated. Within this directory you will find the output files for the mapping-step of the analysis; these are your `*.results.xml` file, a `*.sorted.bam` file and its accompanying index file `*.sorted.bam.bai`. Additionally in this directory you will find the `coverage_summary.txt` file, which can be useful if mapping fails (coverage < 90%), especially if the coverage is close to 90%. This file lists the percent coverage and depth for each serotype. Finally, the `ComponentComplete.txt` file is just a flag to confirm that the analysis was completed.

If the mapping step is enough to determine the serotype, then this will be your output files and the `*.results.xml` file will describe your final serotype prediction. For example, sample PHESPV1910 is a serotype 5 sample which means it can be determined by mapping alone. When viewing the `*.result.xml` file for this, the first line under the `<results>` tag defines the predicted serotype. The `QC_coverage` value is the percent coverage for this serotype, other metrics include the mean and minimum depth; depth corresponds to the number of reads mapping at each position, so minimum 5 means that the lowest number of reads mapped at any position along the capsular sequence is 5. The `*.result.xml` file also returns the second hit with the associated coverage value. In this case, the second hit was serotype 4 with percent coverage of only 33%.

```
<ngs_sample id="PHESPV1910">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <result type="Serotype" value="05">
      <result_data type="QC_coverage" value="99.38"/>
      <result_data type="QC_mean_depth" value="42.1"/>
      <result_data type="QC_minimum_depth" value="5"/>
      <result_data type="QC_meanQ" value="36.4"/>
      <result_data type="second_value" value="04"/>
      <result_data type="QC_coverage_second_value" value="33.39"/>
    </result>
  </results>
</ngs_sample>
```

However, if more than one serotypes with > 90% coverage have been detected then a second directory, called `SNP_based_serotyping`, is created within the `pneumo_capsular_typing` directory that contains the results of the second step, the SNP-based analysis, that uses the CTV database. Here's an example of the mapping `*.results.xml` file for such a case:

```

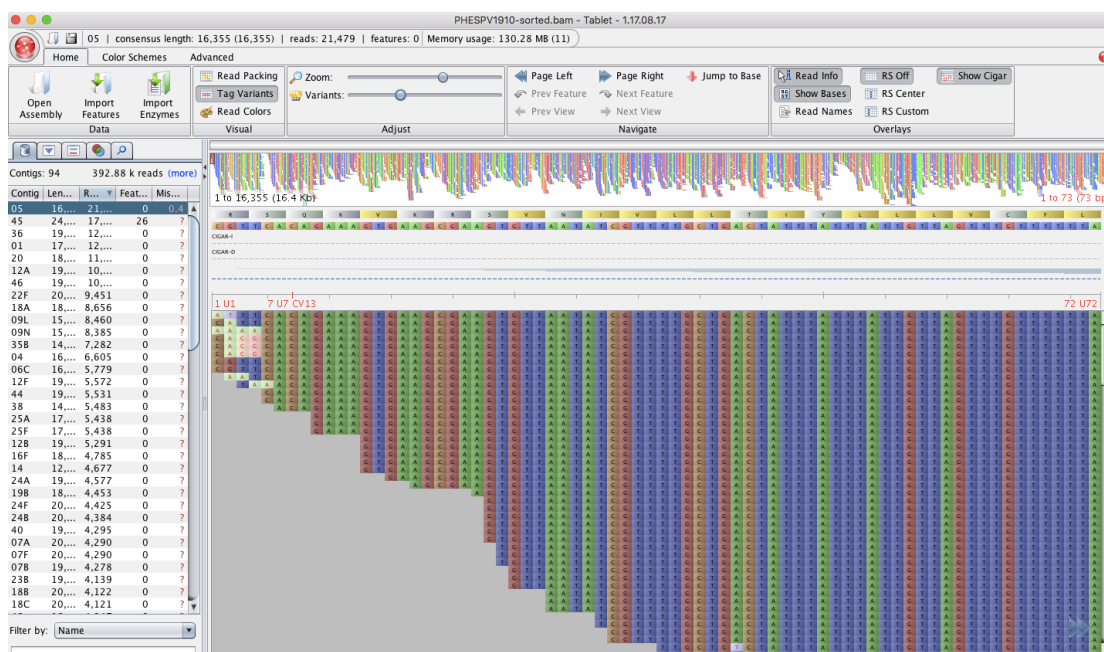
<ngs_sample id="PHESPV0253">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <result type="Serotype" value="40">
      <result_data type="QC_coverage" value="97.80"/>
      <result_data type="QC_mean_depth" value="29.7"/>
      <result_data type="QC_minimum_depth" value="5"/>
      <result_data type="QC_meanQ" value="37.9"/>
      <result_data type="second_value" value="07B"/>
      <result_data type="QC_coverage_second_value" value="97.66"/>
    </result>
  </results>
</ngs_sample>

```

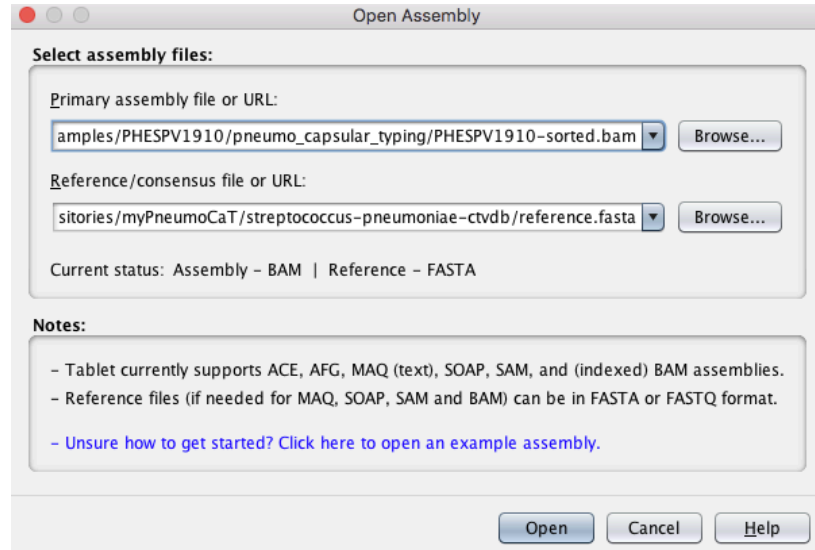
As you can see both first and second hit have percent coverage value >90%. If you look at the coverage\_summary.txt file of this sample you will see that serotypes 40, 7B and 7C all have coverage > 90% so these serotype will be tested in the next step.

Serotype	Coverage	Depth
40	97.7993858751	23.2315487767
07B	97.6569294521	23.1889900747
07C	90.6591316633	19.203608988
24F	58.3706054922	20.2458303691
24B	58.0732047521	20.2377023417

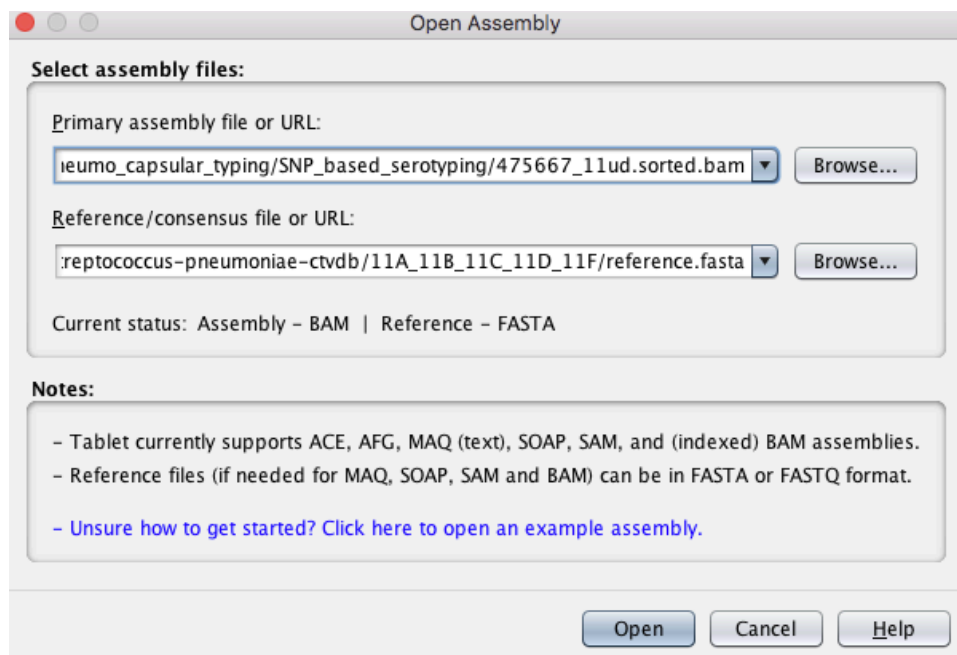
The \*.sorted.bam file is quite large file and if storage is an issue using the -cl flag will remove it when the analysis is complete. However, BAM files are quite useful when troubleshooting; they allow you to visualize the mapping using BAM viewing software like [Tablet](#). At this stage the BAM file allows you to check for possible regions missing or potential recombination events that could results to one gene transferring from one serotype capsular region to another. As the annotation of these reference sequences are available in NCBI you can determine which gene is missing or replaced with another.



To visualize your BAM file you need to open Tablet, then click on the Open Assembly button and navigate to your BAM file (Primary assembly file) and your associated reference file (in this case the reference file with all the capsular regions).



The SNP\_based\_serotyping directory, also has the three main results file; \*.results.xml, \*.sorted.bam and \*.sorted.bam.bai file. In this case to visualize the BAM file you need to select the reference.fasta file within the associated genogroup folder.



If an SNP\_based\_serotyping directory is present then that means that the final serotype will be defined in the \*.results.xml file within this directory. This \*.results.xml is a bit more complicated than the mapping-step \*.results.xml:

```

<ngs_sample id="PHESPV0253">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <!--(START) Serotype Distinction Results (START)-->
    <result type="Serotype_Distinction" value="07B"/>
    <result type="Serotype_Distinction_Total_Hits" value="9/9"/>
    <result type="Serotype_Distinction_Serotypes_Testes" value="07B,07C,40"/>
    <result type="Serotype_Distinction_Gene" value="wcvK">
      <result_data type="Detected" value="Y"/>
      <result_data type="Hits" value="9/9"/>
      <result_data type="SNPs" value="928 AAT:N 309;385 TTT:F 128;706 CAT:H 235;487 ACT:T 162;937 GCA:A
312;46 GAT:D 15;880 CTT:L 293;145 CTT:L 48;946 GGT:G 315"/>
      <result_data type="Allele" value=""/>
      <result_data type="Pseudogene" value=""/>
      <result_data type="pct_Coverage" value="99.9"/>
      <result_data type="Depth(min:max:avg)" value="5:79:45.7"/>
      <result_data type="meanQ" value="38.2"/>
      <result_data type="Length" value="1008"/>
      <result_data type="Coverage_distribution" value="2-1008"/>
      <result_data type="Failure" value="None"/>
    </result>
    <!--(END) Serotype Distinction Results (END)-->
  </results>
</ngs_sample>

```

The first line after the <!--(START) Serotype Distinction Results (START)--> line returns the final predicted serotype, in this case 7B with the next line defining the number of mutations that were tested in order to differentiated the three serotypes (Serotype\_Distinction\_Serotypes\_Testes). For each gene tested in this step a set of metrics is returned, starting with whether is detected or not. Next values include:

- Hits: this is the number of mutations associated with this gene, this is usually 1 except when SNPs are involved as in this case.
- SNPs: this will list the [nt\_position codon:aa aa\_pos] for each position tested.
- Allele: this will return Y/N if presence of an allele is tested
- Pseudogene: this will return Y/N if presence of a pseudogene is tested
- The following are metrics associated with the mapping of the reads to the gene:
  - pct\_Coverage: percent coverage of the length of the gene
  - Depth(min:max\_avg): as mentioned before depth corresponds to the number of reads mapping at each base. A threshold of depth >= 5 is used when determining coverage.
  - meanQ: mean base mapping quality for the gene
  - Length: length of the reference gene sequence
  - Coverage\_distribution: the part of the gene that is covered by reads.
  - Failure: any failure associated with this gene; this could be low coverage, mixed positions etc.

Within this directory there is an additional file, the variant\_summary.yml file, which can help with troubleshooting, when combined with the mutationdb.yml file available in the associated genogroup directory within streptococcus-pneumoniae-ctvdb.

## Troubleshooting examples:

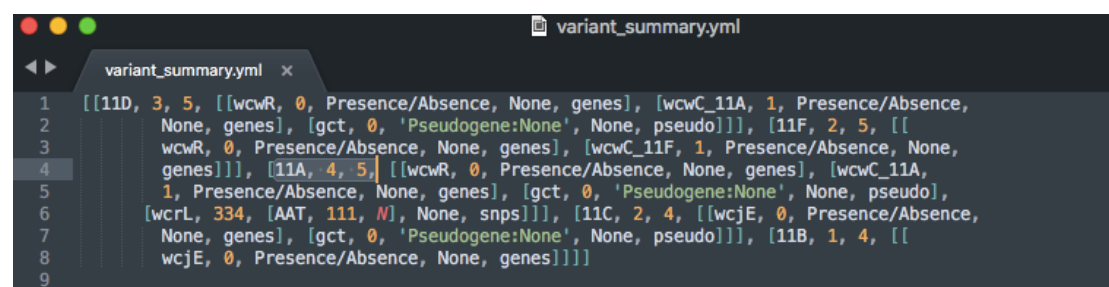
### Example 1:

Let's, for example, look at the variant\_summary.yml file of a failed serogroup 11 isolate and the associated mutationdb.yml file.

The \*.result.xml file within the SNP-based\_analysis directory determines the final serotype prediction, and in this case it returns a 'Serotype undetermined' result due to 4/5 hits:

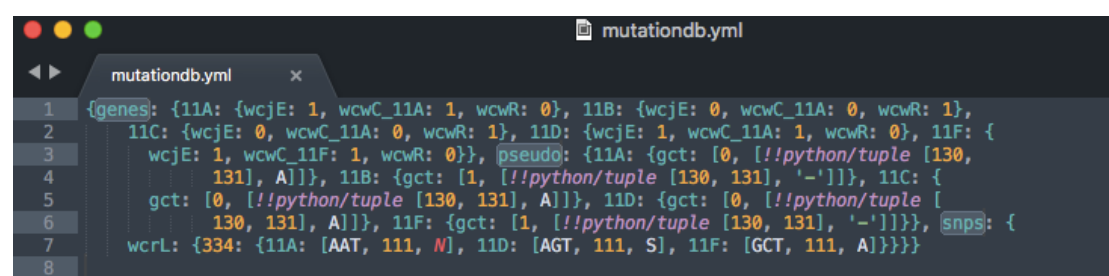
```
<ngs_sample id="475667_11ud">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <!--(START) Serotype Distinction Results (START)-->
    <result type="Serotype_Distinction" value="Serotype undetermined"/>
    <result type="Serotype_Distinction_Total_Hits" value="4/5"/>
    <result type="Serotype_Distinction_Serotypes_Testes" value="11A,11B,11C,11D,
11F"/>
```

The variant\_summary.yml file summarizes the hits for each serotype:



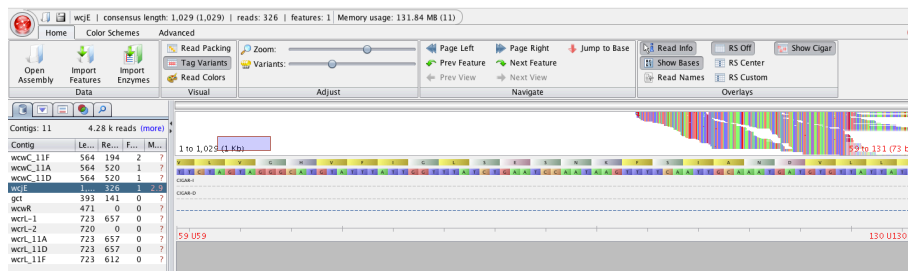
```
1 [[11D, 3, 5, [[wcwR, 0, Presence/Absence, None, genes], [wcwC_11A, 1, Presence/Absence,
2 None, genes], [gct, 0, 'Pseudogene:None', None, pseudo]]], [11F, 2, 5, [[
3 wcwR, 0, Presence/Absence, None, genes], [wcwC_11F, 1, Presence/Absence, None,
4 genes]]], [11A, 4, 5, [[wcwR, 0, Presence/Absence, None, genes], [wcwC_11A,
5 1, Presence/Absence, None, genes], [gct, 0, 'Pseudogene:None', None, pseudo],
6 [wcrL, 334, [AAT, 111, N], None, snps]]], [11C, 2, 4, [[wcjE, 0, Presence/Absence,
7 None, genes], [gct, 0, 'Pseudogene:None', None, pseudo]]], [11B, 1, 4, [[
8 wcjE, 0, Presence/Absence, None, genes]]]]
9
```

The mutationdb.yml file for serogroup 11 defines the expected mutations for each serotype, which for this serogroup includes 'genes' (presence/absence – illustrated with 1/0 respectively), 'pseudo' (pseudogene/coding gene – illustrate with 1/0 with the frameshift mutation defined in this case) and 'SNPs':



```
1 {genes: {11A: {wcjE: 1, wcwC_11A: 1, wcwR: 0}, 11B: {wcjE: 0, wcwC_11A: 0, wcwR: 1},
2 11C: {wcjE: 0, wcwC_11A: 0, wcwR: 1}, 11D: {wcjE: 1, wcwC_11A: 1, wcwR: 0}, 11F: {
3 wcjE: 1, wcwC_11F: 1, wcwR: 0}}, pseudo: {11A: {gct: [0, [!python/tuple [130,
4 131], A]]}, 11B: {gct: [1, [!python/tuple [130, 131], '-']]}, 11C: {
5 gct: [0, [!python/tuple [130, 131], A]]}, 11D: {gct: [0, [!python/tuple [
6 130, 131], A]]}, 11F: {gct: [1, [!python/tuple [130, 131], '-']]}, snps: {
7 wcrL: {334: {11A: [AAT, 111, N], 11D: [AGT, 111, S], 11F: [GCT, 111, A]}}}
8
```

By comparing the expected to the observed mutations we can derive to the cause of the uncertainty for this sample. In this case, it is the absence of the *wcjE* gene. And that can be confirmed by viewing the BAM file:



## Example 2:

In this case we will look into a Failed sample due to low coverage. The \*.result.xml file within the pneumo\_capsular\_typing directory reveals that the top coverage value was 50%:

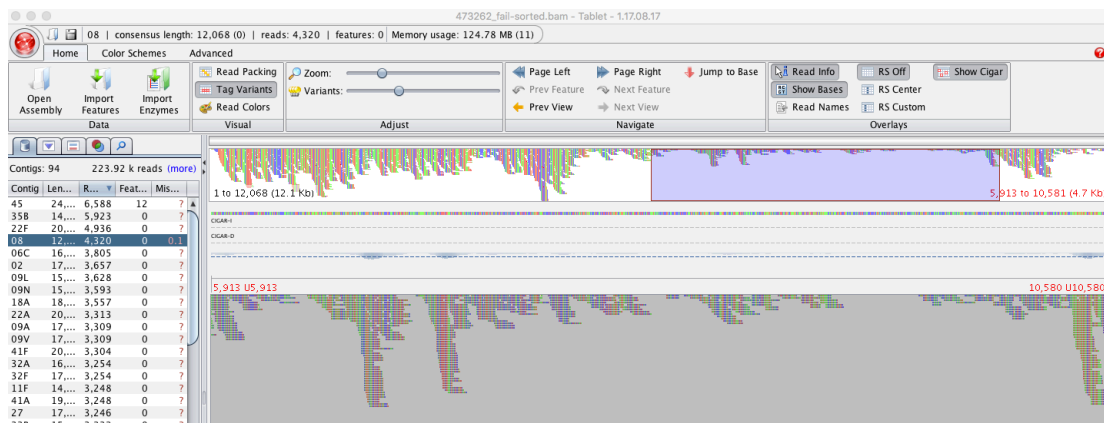
```
<ngs_sample id="473262_fail">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <result type="Serotype" value="Failed">
      <result_data type="QC_coverage" value="50.05"/>
      <result_data type="QC_mean_depth" value="28.3"/>
      <result_data type="QC_minimum_depth" value="5"/>
      <result_data type="QC_meanQ" value="37.6"/>
      <result_data type="second_value" value="Failed"/>
      <result_data type="QC_coverage_second_value" value="37.49"/>
    </result>
  </results>
</ngs_sample>
```

The coverage\_summary.txt file reveals that the top hit was serotype 8:

Serotype	Coverage	Depth
08	50.053856989	16.9892991533
35B	37.4898346435	31.1500086311
06C	34.4397304717	22.067535545
34	32.9946332737	22.6743945353
33F	32.1652755743	22.7999589659
11F	32.0904563198	24.5324362322
06A	31.5063864297	23.2941056496
02	31.4478345332	21.5537136356
06D	30.8661417323	22.2757306226
33B	30.3281299777	24.5855582524

The BAM file visualization with Tablet reveals that the coverage problem is limited to the region between 5,000 and 10,000 bps.





This kind of failure could either be due to dealing with a truly non-typeable isolate or as in this case sequencing service coverage issues. Of course this can only be determined if serological analysis is performed.

### Example 3:

This final example is a case where a result of 15A+ is returned. The '+' is returned in cases where we have priority mutations that define the serotype and other less important mutations that although they seem to be indicative of the serotype population we feel that they could be removed from the CTVdb in the future. Therefore, the '+' is reported to investigate further and monitor the frequency of mismatches to determine if any of those mutations can be removed. These priority mutations are found in 15A and the 7B/7C/40 genogroup. Serotype 15A can actually be called from the mapping step alone as its capsular operon sequence is quite diverse from the other serogroup 15 serotypes. For the 7B/7C/40 genogroup, 9 discriminatory positions were identified in *wcwK* gene, with two of these positions (46 and 385) exhibiting unique codons for each serotype, therefore were defined as priority mutations.

Discriminating positions for genogroup 7B, 7C and 40

Gene	position	07B	07C	40
<i>wcwK</i>	46	[GAT, D, 15]	[GGT, G, 15]	[AAT, N, 15]
<i>wcwK</i>	145	[CTT, L, 48]	[CTT, L, 48]	[TTT, F, 48]
<i>wcwK</i>	385	[TTT, F, 128]	[TGT, C, 128]	[ACT, T, 128]
<i>wcwK</i>	487	[ACT, T, 162]	[GCT, A, 162]	[ACT, T, 162]
<i>wcwK</i>	706	[CAT, H, 235]	[CAT, H, 235]	[TAT, Y, 235]
<i>wcwK</i>	880	[CTT, L, 293]	[CTT, L, 293]	[TTT, F, 293]
<i>wcwK</i>	928	[AAT, N, 309]	[AAT, N, 309]	[AGT, S, 309]
<i>wcwK</i>	937	[GCA, A, 312]	[GAA, E, 312]	[GAA, E, 312]
<i>wcwK</i>	946	[GGT, G, 315]	[GGT, G, 315]	[GAT, D, 315]

For this example, we will investigate a 15A+ sample. As mentioned above, the 15A+ denotes that this sample should be 15A by serology. If you open the \*.results.xml file under the pneumo\_capsular\_typing directory (mapping-based analysis) you will see that the predicted serotype is 15A with 98.85% coverage and the second hit is 15F with only 79.34% identity.

```

<ngs_sample id="422361_15A">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <result type="Serotype" value="15A">
      <result_data type="QC_coverage" value="98.85"/>
      <result_data type="QC_mean_depth" value="25.9"/>
      <result_data type="QC_minimum_depth" value="5"/>
      <result_data type="QC_meanQ" value="36.2"/>
      <result_data type="second_value" value="15F"/>
      <result_data type="QC_coverage_second_value" value="79.34"/>
    </result>
  </results>
</ngs_sample>

```

This is the expected output for a 15A sample. The two capsular loci differ by 4 genes (*glf*, *rlmB*, *rlmD*, *wcjE*) present in 15F but not 15A, therefore a 15A sample can easily be distinguished from 15F but a 15F sample will return coverage >90% for both serotypes, which is why we need to move to the CTVdb-based analysis.

Discriminating mutations for serogroup 15

Gene	variant	15A	15B	15C	15F
<i>wchL</i>	allele	wchL $\alpha$	wchL $\beta$	wchL $\beta$	
<i>wzd</i>	allele	wzd $\alpha$	wzd $\beta$	wzd $\beta$	
<i>glf</i>	detected	N	N	N	Y
<i>rlmB</i>	detected	N	N	N	Y
<i>rlmD</i>	detected	N	N	N	Y
<i>wcjE</i>	detected	N	N	N	Y
<i>wciZ</i>	pseudo		N, [412, 417]	Y, [412, 417]	

While analyzing serogroup 15 we identified the presence of two different alleles for genes *wchL* and *wzd*. To identify the source of the mismatch indicated by '+' we need to look at the \*.results.xml file under the SNP\_based\_serotyping directory:

```

<ngs_sample id="422361_15A">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <!--(START) Serotype Distinction Results (START)-->
    <result type="Serotype_Distinction" value="15A+"/>
    <result type="Serotype_Distinction_Total_Hits" value="5/6"/>
    <result type="Serotype_Distinction_Serotypes_Testes" value="15A,15B,15C,15F"/>
    <result type="Serotype_Distinction_Gene" value="wzd">
      <result_data type="Detected" value="Y"/>
      <result_data type="Hits" value="1/1"/>
      <result_data type="SNPs" value=""/>
      <result_data type="Allele" value="wzd-1"/>
      <result_data type="Pseudogene" value=""/>
      <result_data type="pct_Coverage" value="100.0"/>
      <result_data type="Depth(min:max:avg)" value="6:132:85.3"/>
      <result_data type="meanQ" value="37.1"/>
      <result_data type="Length" value="693"/>
      <result_data type="Coverage_distribution" value="1-693"/>
      <result_data type="Failure" value="None"/>
    </result>
    <result type="Serotype_Distinction_Gene" value="wchL">
      <result_data type="Detected" value="Y"/>
      <result_data type="Hits" value="0/1"/>
      <result_data type="SNPs" value=""/>
      <result_data type="Allele" value="Failed"/>
      <result_data type="Pseudogene" value=""/>
      <result_data type="pct_Coverage" value="99.0"/>
      <result_data type="Depth(min:max:avg)" value="1:125:63.0"/>
      <result_data type="meanQ" value="37.4"/>
      <result_data type="Length" value="1029"/>
      <result_data type="Coverage_distribution" value="1-1019"/>
      <result_data type="Failure" value="Mixed: {'wchL-2': 0.79, 'wchL-1': 0.99}"/>
    </result>
    <result type="Serotype_Distinction_Gene" value="rlmB">

```

As you can see from this file we have 5/6 matches and the culprit is *wchL*:

```

<result_data type="Failure" value="Mixed: {'wchL-2': 0.79, 'wchL-1': 0.99}"/>

```



This line indicates the presence of both alleles. Usually the two alleles are different enough that the reads from one will not map to the second allele. In a normal 15A you expect some coverage of wchL-2 but not enough to pass the 90% threshold. Here's the relevant line extracted from the variant\_summary.yml file of a normal 15A sample and the equivalent for the 15A+ sample:

```
15A - [wchL, wchL-1, {wchL-1: 1.0, wchL-2: 0.8}, None, allele]  
15A+ - [wchL, Failed, Allele,  
        Mixed: {wchL-2: 0.79, wchL-1: 0.99}, allele]
```

Although this is interesting from a research point of view this has no impact on the serotype. However, it will probably lead to the ultimate conclusion that this allele should be removed from the CTVdb if this result occurs too often.

**NOTE:** The above mutation has been removed from the latest release of PneumoCaT – version 1.2