PneumoCaT CTVdb Serotype Determination Reference Document

Software Version 1.21

Dr. Carmen Sheppard
30/09/2019

Introduction

This Document contains summary tables describing the genetic variant information included in PneumoCaT V1.2 available from the Github repository. (https://github.com/phe-bioinformatics/PneumoCaT). These tables are designed to be useful for troubleshooting PneumoCaT outputs and investigating potential serotype variants.

This information is updated from the original tables included in the manuscript (Kapatai et al. 2016) and is designed to be human readable, this information is also included in the yaml files included in the CTVdb in the repository. Usage/interpretation notes are included in this document to further aid users.

The variants described are collated from previously described genetic variants (referenced) and novel discriminatory variants described in the Kapatai et al 2016 publication. These variant positions are constantly under validation by ourselves and from feedback from external users and we encourage users with potentially important variants (i.e more than one example) to get in touch via the GitHub repository or via twitter @PneumoCaT and let us know about them. We fully expect the CTVdb to evolve as further genetic variants are described and their corresponding phenotypic serotypes are determined and validated.

This document consists of a summary table describing all serogroups that are determined by the CTVdb and information on genetic and structural differences. Further reference tables are provided listing the individual genetic sequences used for each serogroup.

Please note: The genetic determinants described in this document are those used in version 1.21 of PneumoCaT as available from Github. Users may create their own custom CTVdb and in this case or in the case of users of previous versions of PneumoCaT these tables may not be a reliable reference.

For further details please refer to the original manuscript – but note the publication refers to PneumoCaT v1.0:

Kapatai, G. et al., 2016. Whole genome sequencing of *Streptococcus pneumoniae* : development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. PeerJ, 4, p.e2477.

Contents

Summary Table

Serogroup	Serotype	Distinguishing genetic features	Functional Effect
6	6A/6B and 6C/6D	A>G 583 in *wciP*	amino acid substitution (Ser195Asn) which results to different rhamnose-ribitol linkages (1 →3 in 6A/6C and 1→in 6B/6D) (Mavroidi et al. 2007; Sheppard et al. 2010; Ko et al. 2013)
	6A/6C and 6B/6D and 6E	*wciNα* in 6A and 6B / *wciNβ*in 6C and 6D	Allele *wciNα* encodes for galactosyl-transferase whereas *wciNβ* is 200 bp shorter and encodes for a glycosyl-transferase - consistent with changes in structure (Park et al. 2007) . *WciNγ* is a chimeric form of *wciNα* (75%) and *wciNβ* (25%).
		wciNγ in 6E	
7 and 40	7A/7F	Frameshift mutation insT 587 in 7A *wcw*D gene	Loss of function of glycosyltransferase leading to loss of side branch for 7A (Mavroidi et al. 2007)
			"Mixed: ['07A','07F']" result corresponds to 7A phenotype
	7B/7C/40	SNPs in *wcwK*	Amino acid changes - *wcwK* encodes for a GT but 7C and 40 structure not known
9	9A/9V	Frameshift mutation delG 722 in 9A *wcjE*	Loss of function of O-acetyltranferase leads to differences in acetylation
	9L/9N	SNPs in genes wchA, wcjA, wcjB and wzy	Amino acid changes - *wcjA* and *wcjB* encode for glycosyltranferases (GT) and changes in these are consistent with presence of glucose in 9N instead of galactose present in residue 3 of the polysaccharide repeat unit of the other three serotypes (Mavroidi et al. 2007)
	9A/9V/9L/9N	Presence of an additional O-acetyltransferase encoded by *wcjD* in 9A-9V	Differences in acetylation
10	10A/10B/10C/10F	10A/10B carries gene *wcrG*, whereas 10C/10F carries genes *wcrH* and *wciG*	*wcrH* encodes for GT and is responsible for side branch linkage
			Galf(1-6)Galp present in 10F but not in 10A; w*crG* encodes for GT and it catalyzes the linkage of Galp(1-6) side branch in 10A (Aanensen et al. 2007)
	10A/10B/10C/10F	10A/10C have *wcrCα* whereas 10B/10F have *wcrCβ*	*wcrCβ* allele is described as *wcrF* and both genes encode for glucosyltransferases and are responsible for the differences observed in the linkage between galactose and ribitol-5-phosphate(Yang et al. 2011)
11	11A/11B/11C/11D/11F	Genes *wcwC* and *wcjE* are present in 11A, 11D and 11F whereas gene *wcwR* is present in 11B/C	*wcwC, wcjE* and *wcwR* are acetyltransferase genes - differences in acetylation
	11A/11B/11C/11D/11F	Frameshift mutation delA 130 in *gct* in	Presence of Gro-1P correlates with an intact *gct* gene in types

		11B and 11F	11A and 11C; *gct* is frameshifted in types 11F and 11B
			Rib-ol is present in the CPS instead of Gro (Mavroidi et al. 2007)
	11A/11D/11F	*wcrL pos 334*: codon AAT (Asn) in 11A; codon ACT (Ser) in 11D and codon GCT (Ala) in 11F	*wcrL* encodes for a GT - donor sugar for WcrL is GlcpNAc in types 11F,
			11B, and 11C but Glcp in type 11A (Mavroidi et al. 2007)
12, 44, 46	12A/B/F/44/46	SNPs in genes *wcxD* and *wcxF*	Both genes encode for GTs present only in this genogroup (Mavroidi et al. 2007) effect on sugar chain unknown (no structure for 12B, 44 and 46)
15	15A/15B/15C/15F	15F has 4 additional genes; *glf, rmlB, rmlD* and *wcjE*	*glf, rmlB* and *rmlD* are involved in sugar biosynthesis; *wcjE* encodes for an acetyltranferase.
	15A/15B/15C	15A *wzd* has 69% identity in the last 300 bps when compared to the 15B/C allele	*wzd* is involved in translocation of mature CPS to the cell surface and thus is responsible for determining the length of the capsule polysaccharide chain
	15B/15C	difference in TA tandem repeat region near position 413 of *wciZ,* leading to frameshift in 15C (Bentley et al. 2006)	*wciZ* encodes for an O-acetyltransferase - differences in acetylation.
			*15B, 15B/C and 15C results can be assigned.
16	16A, 16F	Stage 1 mapping against CPS operon refs only	
17	17A, 17F	Stage 1 mapping against CPS operon refs only	
18	18A/18B/18C/18F	18F has an extra acetyltransferase gene (*wcxM)* and type 18A lacks the acetyltransferase gene *wciX* (Mavroidi et al., 2007)	Differences in acetylation
	18B/18C	G>T 168 in *wciX* leads to early stop codon in 18B (Mavroidi et al., 2007)	*wciX* encodes for an acetyltransferase - difference in acetylation
19	19F	Stage 1 mapping against CPS operon refs only	
	19A/19AF	19AF has 19F wzy	19AF phenotype as 19F despite having overall 19A-like capsular operon sequence
	19B/19C	19B lacks genes *wchU, (HG264*)and *glf*	*wchU* encodes for a putative GT and could be responsible for the additional glucose in the capsular polysaccharide repeat unit of 19C; *glf* encodes for a UDP-galactopyranose mutase whereas *HG264* has no functional product.
22	22A/22F	*wcwA* and *wcwC* share no similarity between 22A and 22F.	*wcwA,* encoding for a putative glycosyl-transferase and *wcwC,* encoding for a putative O-acetyltranferase - structure for 22A unknown

23	23A/23B/23F	distinct *wzy* sequence in all serotypes	wzy encodes for a polymerase and differences in sequence should account for the different polymerization linkages (Mavroidi et al., 2007) - structures for 23A and 23B unknown
	23A/23B/23F	*wchA* is identical in 23B and 23F but distinct in 23A.	*wchA* encodes for a glycosyl-1-phosphatase transferase (Aanensen et al., 2007) - structures for 23A and 23B unknown
25 and 38	25A/25F/38	*wcyV* missing in 38(Mavroidi et al., 2007)	*wcyV,* wcyD *and* wcyC encode for GTs (Aanensen et al. 2007)- no structures available for 25A, 25F or 38
	25A/25F/38	*wcyDα* in serogroup 25 and *wcyDβ* in serotype 38	
	25A/25F/38	SNPs in *wcyC* (Table S4)	
28	28A/28F	SNPs in *wciU* (Table S5)	*wciU* encodes for a GT - no structures available
33 and 37	33A/33F/37	37 carries *tts* - a transferase gene	*tts* is responsible for the polysaccharide capsule synthesis in 37(Waite et al., 2003)
	33A/33F	Frameshift mutation insT 433 in 33F *wcjE* gene	Loss of function of O-acetyltranferase leads to differences in acetylation (Mavroidi et al., 2007)
	33B/33D	*wciNα* in 33B / *wciNβ*in 33C	wciNα encodes for a putative glycosyltranferase whereas wciNα encodes for a putative galactosyltransferase - consistent with differences in structure
	33C	Stage 1 mapping against CPS operon refs only	
35 and 42	35B, 35F	Stage 1 mapping against CPS operon refs only	
	35A/35C/42	SNPs in genes *mnp1, wcrL* and *wzh*	*mnp1* encodes for a putative nucleotidyltranferase (NDP-mannitol pathway), *wcrL*, a GT and *wzh,* a protein-tyrosine phosphatase - consistent with differences in structure
	35A/35C/42	Frameshift mutation insA 248 in 35A *wcrK* (Mavroidi et al., 2007)	wcrK encodes for a GT - consistent with differences in structure
41	41A/41F	Frameshift mutation delG 23 in 41A *wcrX* (Mavroidi et al., 2007)	*wcrX* encodes for a acetyltranferase - differences in acetylation
47	47A, 47F	Stage 1 mapping against CPS operon refs only	

Genetic serotype determinants by serogroup

Serogroup 6

Gene	variant	6A	6B	6C	6D	6A(6E)	6B(6E)
wciN	allele	*wciNα*	*wciNα*	*wciNβ*	*wciNβ*	*wciNγ*	*wciNγ*
wciP	583	[AGT, 194, S]	[AAT, 194, N]	[AGT, 194, S]	[AAT, 194, N]	[AGT, 194, S]	[AAT, 194, N]

Genetic variant 6E is denoted by the predicted serological type with 6E indicated in brackets.

7A, 7F

Gene	variant	7A	7F
wcwD	pseudo	Y, [[587, 588], GT]	N, [[587, 588], G]

In our experience isolates serotyped as 7A often report from PneumoCaT as [mixed: 7A, 7F] rather than 7A. Further investigation required.

Genogroup 7B, 7C and 40

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

Targets in bold red are the main serotype determining positions.

Differences in the other positions are indicated with a "+" at the end of the serotype result and should be checked with serology.

Future updates to the CTVdb after investigations may remove the need for inclusion of these targets in the CTVdb

Serogroup 9

Gene	variant	9A	9V	9N	9L
wcjD	detected	Y	Y	N	N
wcjE	pseudo	Y, [[721, 722], '-']	N, [[721, 722], G]		

9N,9L

Gene	Proposed Function	Position	9N	9L
wchA	UDP-glucosyl-1-phosphate transferase	504	[TCT, 168, S]	[TAT, 168, Y]
wchA	UDP-glucosyl-1-phosphate transferase	879	[TCA, 293, S]	[CCA, 293, P]
wcjA	Glycosyltransferase	414	[TAT, 138, Y]	[CAT, 138, H]
wcjA	Glycosyltransferase	429	[AGT, 143, S]	[GGT, 143, G]
wcjA	Glycosyltransferase	528	[GGT, 176, G]	[GAT, 176, D]
wcjA	Glycosyltransferase	636	[AAT, 212, N]	[GAT, 212, D]
wcjA	Glycosyltransferase	852	[TCA, 284, S]	[GCA, 284, A]
wcjA	Glycosyltransferase	957	[ACT, 319, T]	[ATT, 319, I]
wcjB	Glycosyltransferase	789	[ACC, 263, T]	[GCC, 263, A]
wzy	repeat unit polymerase	846	[AAC, 282, N]	[GAC, 282, D]

Serogroup 10

Gene	variant	10A	10B	10C	10F
wcrG	detected	Y	Y	N	N
wcrH	detected	N	N	Y	Y
wciG	detected	N	N	Y	Y
wcrC	allele	*wcrCα*	*wcrCβ*	*wcrCα*	*wcrCβ*

Serogroup 11

Gene	variant	11A	11B	11C	11D	11F
wcwC	detected	Y	N	N	Y	Y
wcjE	detected	Y	N	N	Y	Y
wcwR	detected	N	Y	Y	N	N
gct	pseudo (130delA)	N	Y	N	N	Y
wcrL	pos 334	AAT			ACT	GCT

Genogroup 12, 44, 46

Genes	Position	12A	12B	12F	44	46
wcil	allele	*wcilα*	*wcilα*	*wcilβ*	*wcilα*	*wcilα*
wcxD	781	[TTG, 260, L]	[TTG, 260, L]	[TTG, 260, L]	[TTG, 260, L]	[ATG, 260, M]
wcxD	793	[TAT, 264, Y]	[TAT, 264, Y]	[TAT, 264, Y]	[TAT, 264, Y]	[CAT, 264, H]
wcxD	805	[TCA, 268, S]	[TCA, 268, S]	[TCA, 268, S]	[TCA, 268, S]	[CCA, 268, P]
wcxD	809	[ATG, 269, M]	[ATG, 269, M]	[ATG, 269, M]	[ATG, 269, M]	[ACT, 269, T]
wcxD	812	[GCA, 270, A]	[GTA, 270, V]	[GTA, 270, V]	[GTA, 270, V]	[GTA, 270, V]
wcxD	845	[GCT, 281, A]	[GCT, 281, A]	[GCT, 281, A]	[GCT, 281, A]	[GTT, 281, V]
wcxF	256	[GCC, 85, A]	[GCC, 85, A]	[GCC, 85, A]	[GCC, 85, A]	[ACC, 85, T]
wcxF	560	[CTT, 186, L]	[CTT, 186, L]	[CTT, 186, L]	[CCT, 186, P]	[CTT, 186, L]
wcxF	703	[ATA, 234, I]	[CTA, 234, L]	[CTA, 234, L]	[CTA, 234, L]	[CTA, 234, L]
wcxF	787	[CTA, 262, L]	[CTA, 262, L]	[CTA, 262, L]	[ATA, 262, I]	[CTA, 262, L]
wcxF	889	[GTT, 296, V]	[GTT, 296, V]	[GTT, 296, V]	[GTT, 296, V]	[ATT, 296, I]
wcxF	916	[GGT, 305, G]	[GGT, 305, G]	[GGT, 305, G]	[GCT, 305, A]	[AGT, 305, S]
wcxF	1120	[GCC, 373, A]	[GCC, 373, A]	[GCC, 373, A]	[ACC, 373, T]	[GCC, 373, A]
wzy	251	[ATT, 83, I]	[ACT, 83, T]	[ACT, 83, T]	[ACT, 83, T]	[ACT, 83, T]

Serogroup 15

Gene	variant	15A	15B	15C	15F
wzd	allele	wzdα	wzdβ	wzdβ	
glf	detected	N	N	N	Y
rmlB	detected	N	N	N	Y
rmlD	detected	N	N	N	Y
wcjE	detected	N	N	N	Y
wciZ	pseudo		N, [412, 417]	Y, [412, 417]	

PneumoCaT can detect mixed populations of 15B and 15C and thus can report 15B, 15C or 15B/C results. We have not tested the sensitivity of mixed serotype detection.

Serogroup 18

Gene	variant	18A	18B	18C	18F
glf	detected	N	Y		
wciX	detected	N	Y	Y	Y
wcxM	detected	N	N	N	Y
wciX	pseudo		Y, [[168, 171], TGA]	N, [[168, 171], GGA]	N, [[168, 171], GGA]

19A, 19AF

Gene	variant	19A	19AF
wzy-1	allele	N	Y
wzy-2	allele	Y	N

19B, 19C

Gene	variant	19B	19C
HG264	detected	N	Y
glf	detected	N	Y
wchU	detected	N	Y

We are aware that some 19C variants may not have HG264. As it has no known function, this determining variant *may* be removed in future updates.

Serogroup 22

Gene	variant	22A	22F
wcwA	detected	Y	N
wcwC	detected	Y	N

Serogroup 23

Gene	variant	23A	23B	23F
wchA	allele	*wchAα*	*wchAβ*	*wchAβ*
wzy	allele	*wzyα*	*wzyγ*	*wzyβ*

Genogroup 25A, 25F and 38

Gene	position	25A	25F	38
wcyD	allele	*wcyDα*	*wcyDα*	*wcyDβ*
wcyV	detected	N	N	Y
wcyC	429	[GCT, 143, A]	[ACT, 143, T]	[GCT, 143, A]
wcyC	465	[ACA, 155, T]	[AAA, 155, K]	[ACA, 155, T]
wcyC	573	[GAA, 191, E]	[GGA, 191, G]	[GAA, 191, E]
wcyC	660	[CGA, 220, R]	[CTA, 220, L]	[CGA, 220, R]
wcyC	891	[ATG, 297, M]	[CTG, 297, L]	[ATG, 297, M]
wcyC	903	[TGG, 301, W]	[CGG, 301, R]	[TGG, 301, W]

Serogroup 28

Gene	position	28A	28F
wciU	118	[TGT, 39, C]	[TAT, 39, Y]
wciU	214	[ATA, 71, I]	[ATG, 71, M]
wciU	253	[ATG, 84, M]	[CTG, 84, L]
wciU	736	[CAA, 245, Q]	[AAG, 245, K]
wciU	739	[CCA, 246, P]	[TCA, 246, S]
wciU	742	[GTT, 247, V]	[GCT, 247, A]
wciU	751	[GTA, 250, V]	[ATA, 250, I]
wciU	769	[AGT, 256, S]	[AAT, 256, N]
wciU	781	[AAA, 260, K]	[AAT, 260, N]
wciU	838	[TCA, 279, S]	[CCA, 279, P]
wciU	880	[GCT, 293, A]	[GTT, 293, V]
wciU	883	[GGA, 294, G]	[GAA, 294, E]
wciU	892	[GAA, 297, E]	[GAT, 297, D]
wciU	895	[CAA, 298, Q]	[CGA, 298, R]
wciU	913	[GTA, 304, V]	[GCA, 304, A]
wciU	985	[ACG, 328, T]	[ATG, 328, M]
wciU	1024	[ATA, 341, I]	[GTT, 341, V]
wciU	1027	[GCA, 342, A]	[CCA, 342, P]
wciU	1036	[TGT, 345, C]	[TGG, 345, W]

Genogroup 33A, 33F, 37

Gene	variant	33A	33F	37
wcjE	pseudo	N, [[433, 434], T]	Y, [[433, 434], TA]	
tts	detected	N	N	Y

Genogroup 33B, 33D

Gene	variant	33B	33D
wciN	allele	*wciNα*	*wciNβ*

Genogroup 35A, 35C and 42

Gene	position	35A	35C	42
wcrK	pseudo	[1, [248, 249], GA]	[0, [248, 249], G]	[0, [248, 249], G]
mnp1	198	[CGC, 66, R]	[TGC, 66, C]	[TGC, 66, C]
mnp1	288	[GTT, 96, V]	[CTT, 96, L]	[CTT, 96, L]
mnp1	456	[GGC,152, G]	[GAC, 152, D]	[GAC, 152, D]
mnp1	540	[GAC, 180, D]	[GCC, 180, A]	[GCC, 180, A]
mnp1	564	[GAT, 188, D]	[AAT, 188, N]	[AAT, 188, N]
mnp1	594	[CAC, 198, H]	[TAT, 198, Y]	[TAT, 198, Y]
wcrL	453	[CAG, 151, Q]	[CAG, 151, Q]	[CGG, 151, R]
wzh	465	[AGG, 155, R]	[AGG, 155, R]	[ATG, 155, M]
wzh	525	[CGT, 175, R]	[CGT, 175, R]	[CCT, 175, P]
wzh	528	[TAT, 176, Y]	[TAT, 176, Y]	[GAT, 176, D]
wzh	567	[CAG, 189, Q]	[CAG, 189, Q]	[CGT, 189, R]

Serogroup 41

Gene	variant	41A	41F
wcrX	pseudo	Y, [[23, 24], '-']	N, [[23, 24], G]

References

Aanensen, D.M. et al., 2007. Predicted Functions and Linkage Specificities of the Products of the Streptococcus pneumoniae Capsular Biosynthetic Loci. *Journal of Bacteriology*, 189(21), pp.7856–7876.

Bentley, S.D. et al., 2006. Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. *PLoS Genetics*, 2(3), p.e31.

Kapatai, G. et al., 2016. Whole genome sequencing of *Streptococcus pneumoniae* : development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ*, 4, p.e2477.

Ko, K.S., Baek, J.Y. & Song, J.-H., 2013. Capsular gene sequences and genotypes of "serotype 6E" Streptococcus pneumoniae isolates. *Journal of clinical microbiology*, 51(10), pp.3395–9.

Mavroidi, A. et al., 2007. Genetic Relatedness of the Streptococcus pneumoniae Capsular Biosynthetic Loci. *Journal of Bacteriology*, 189(21), pp.7841–7855.

Park, I.H. et al., 2007. Discovery of a new capsular serotype (6C) within serogroup 6 of Streptococcus pneumoniae. *Journal of Clinical Microbiology*, 45(4), pp.1225–1233.

Sheppard, C.L. et al., 2010. Streptococcus pneumoniae isolates expressing a capsule with epitopes of both serotypes 6A and 6B. *Clinical and Vaccine Immunology*, 17(11), pp.1820–1822.

Yang, J. et al., 2011. Comparative structural and molecular characterization of Streptococcus pneumoniae capsular polysaccharide serogroup 10. *The Journal of biological chemistry*, 286(41), pp.35813–22.