The Development of pCoMiC: Predicted Classification of Mutations in CFTR

Aishani Tewari - 20796813

Connor Clark-Baba - 20830237

Faizaan Adtani - 20826317

BIOL 469

December, 08, 2023

**1.0 Introduction**

Cystic fibrosis transmembrane conductance regulator (CFTR) is responsible for the transportation of chloride through membranes created by the CFTR protein, which is located on the surface of the lungs6. It normally functions to maintain the balance of salt and water ions as it transports through membranes and it is part of a class of membrane proteins that couple the hydrolysis of ATP called ATP-binding cassette (ABC) transporter superfamily1. CFTR is the only member of this superfamily that transports chloride ions1,3. Along with this, it is the only ligand-gated channel that consumes its ligand (ATP) during the gating cycle, which is the opening and closing of the CFTR channel3.

The structure of CFTR is made up of 1,480 amino acids and consists of 5 domains1. This includes two membrane spanning domains (MSDs), two nucleotide-binding domains (NBD1 and NBD2), and a regulatory (R) domain1,3. MSDs are responsible for anchoring the protein within the membrane by their transmembrane helices1. NBDs are involved in binding and hydrolyzing ATP, which is crucial for a proper functioning CFTR1. Following this, the phosphorylation of the R domain by protein kinase A (PKA) leads to conformational change that allows for chloride ions to flow through1.

Cystic Fibrosis (CF) is a genetic disease occuring by the inheritance of two mutated CFTR genes, one from each parent1,10. This genetic anomaly results in absence of functional CFTR protein or the presence of dysfunctional variant2. Individuals inheriting one mutated gene are carriers of CF but will not develop the disease10.

**1.1 Five Classes of Mutations**

There are 5 classes of CFTR mutations that occur in the CFTR gene. Class I is a protein production mutation which interferes with the production of the CFTR protein by nonsense and splice mutations2. Nonsense mutations lead to an early stop codon being introduced in the nucleotide sequence, resulting in an unstable protein being produced which gets rapidly degraded by cells2,8. Splicing mutations disrupt the cell’s ability to read protein-building instructions, causing errors in determining the start and end points2. The oversight of amino acids interferes with proper protein synthesis2,8. Class II is a protein processing mutation which is the most common form of mutation found in patients2. Normally, the CFTR protein folds into a specific 3-dimensional (3-D) shape to function properly but with a Class 2 mutation, the 3-D shape cannot be formed due to alterations in the amino acid sequence2,8. Class III is called a gating mutation where the gate that allows the chloride ions is always closed, eliminating all forms of chloride transport through the membrane2. Class IV is a conduction mutation and disrupts the structure and properties of the CFTR channel making it harder for chloride ions to flow despite the 3-D shape being correct8. Lastly, Class V is called an insufficient protein mutation where a normal CFTR protein is made but the amount is lower than normal8. This results in limited production, dysfunction, or rapid degradation of CFTR proteins2,8.

Overall, each class of mutation of CFTR can be present in patients which results in CF and require different forms of subsequent treatment. In Canada, over 4,300 individuals are affected yearly by CF and 1 in 25 are carriers for a mutated CFTR gene12.

**1.2 Symptoms and Diagnosis**

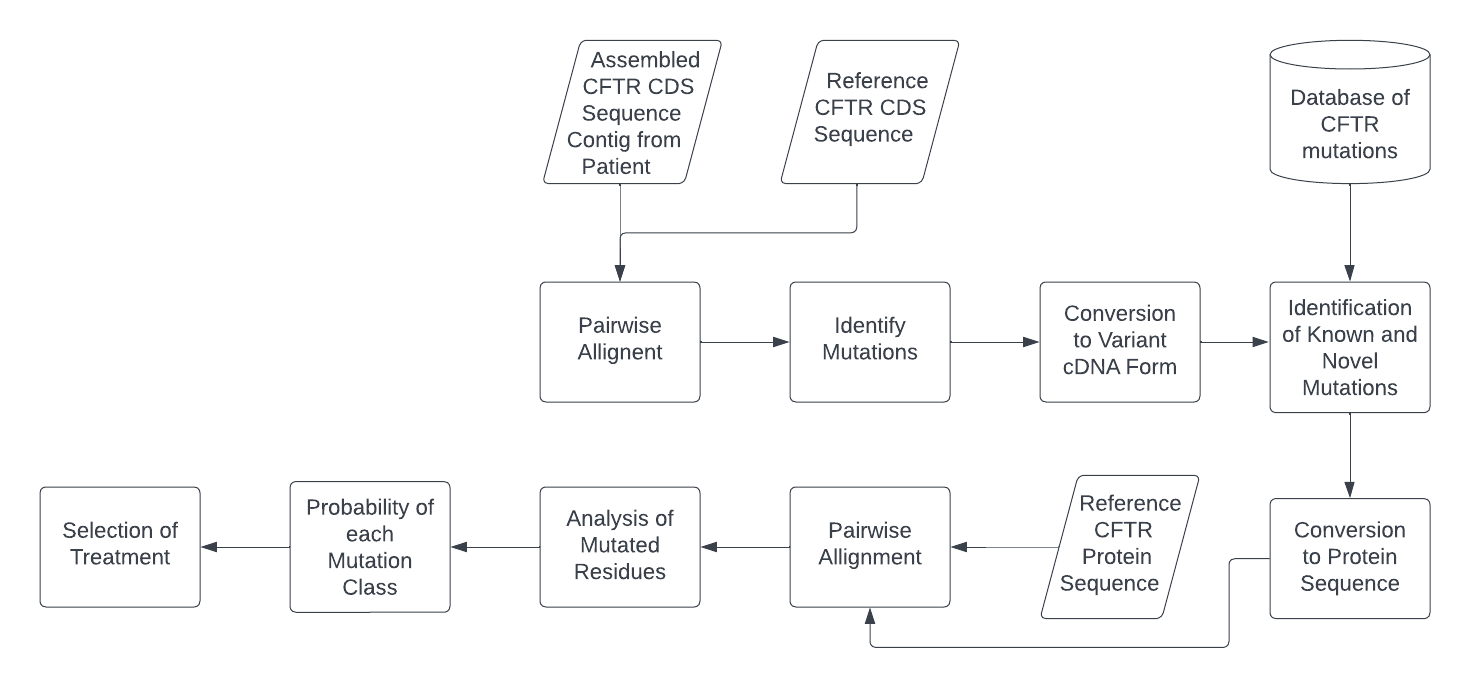
CF predominantly affects the lungs, digestive system and some major organs10. The mutated CFTR gene causes a thick and sticky mucus to accumulate in the lungs and elevated sweat salt levels10. This excess mucus leads to lung blockages and damages causing infections9,10. CF diagnosis involves genetic testing and looking at chloride levels by a sweat test; any abnormalities reflect mutations in CFTR9.

**2.0 Objectives**

The objective of this report is to document the development of a pipeline to analyze and classify mutations in patient CFTR coding sequences (CDS) for efficient characterization and treatment selection. Traditional classification methods involve extensive clinical and experimental trials to determine the corresponding phenotype. The pipeline introduced in this report, predicted Classification of Mutations in CFTR (pCoMiC), aims to limit the need for substantial *in vivo* or *in vitro* testing for phenotypic classification of mutations in CFTR, greatly reducing time and cost alongside rapidly extending databases of known mutants.

**3.0 Methods**

**3.1 Pipeline Workflow**



**Figure 1:** *Predicted Classification of Mutations in CFTR (pCoMiC) workflow.*

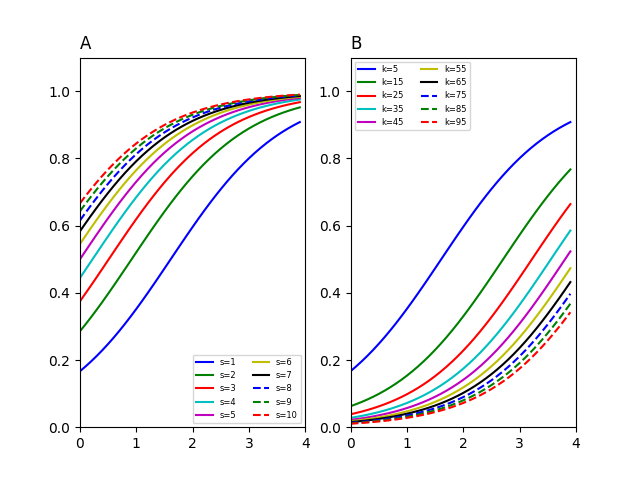
An assembled CFTR CDS contig was obtained from patient data and aligned to a reference CDS. Single base-pair mutations were identified and converted to variant cDNA form through a cleaning function. The identified mutations were compared with a database of known CFTR mutations. Both patient and reference CDS’ were converted to amino acid sequences (AAseq) and aligned. The mutated residues were analyzed through a custom probability calculation function to predict the likelihood of the mutation to fall into each of the five classes. The output probabilities can then be used to determine appropriate selection of treatment (Fig.1 ).

**3.2 Probability Calculation Function**

The custom probability function considers hydropathy, relative charge and hydrogen bonding capability (HBC) of mutated residues to determine the probabilities for each class. Hydropathy scores were obtained from the Kyte-Doolittle scale4,7. Relative charges were derived from the isoelectric point (pI) of amino acids at the average pH of the human body (7.4). HBCs were derived from visual analysis of each amino acid, varying from 0 to 3.

Average hydropathy and charge of the surrounding region (3 residues up- and downstream of the mutation including the mutated residue) were calculated for both patient and reference AAseqs. Deletions in either sequence were considered. The absolute value of the percent change between hydropathy and charge were then determined representing the hydropathy score (HyS) and relative charge score (CS), respectively. The difference in hydrogen bonding between the reference and mutated residues was calculated (ΔHBC = HBCref - HBCmut).

[1]



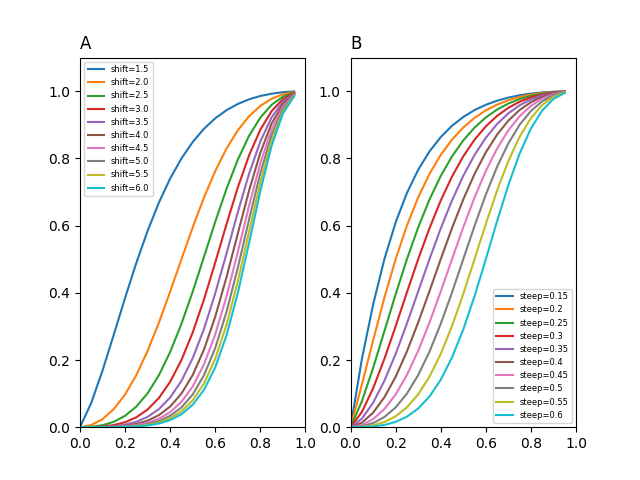
**Figure 2**: *Effects of varying “s” and “k” in sigmoidal normalizing function. Shifting the curve upwards results in smaller initial values having a higher relative increase in their normalized scores. Increasing the steepness makes higher initial values much more significant than smaller when normalized. A. Parameter “k*” *was held at 5, increasing “s” shifts curve left though decreases steepness. B. Parameter “s” was held at 1, increasing “k” shifts curve right and increases steepness.*

This number was then normalized through a custom sigmoidal function (1) and was taken as the HBC score (HbS). If ΔHBC was ≤ 0, the value was not normalized and HbS was considered 0. In equation 1, *Score* is the original ΔHBC score and *s* and *k* both determine the steepness and x-axis shift of the sigmoidal curve. The effects of varying *s* and *k* are shown (Fig. 2). It was determined that values of 1 and 5 for *s* and *k*, respectively, were optimal. HyS and CS were considered in mutation scoring for classes II-V. HbS was considered for classes II, IV and V.

**3.2.1 Calibrating Characteristic Weights and Parameters of Probability Function**

Probabilities of each mutation class were calculated through 5 separate functions. HyS, CS and HbS should range from 0 to 1 and were normalized prior to probability calculations if outside of this range (1) with appropriate parameter values determined through testing.

[2]



**Figure 3**: *Effects of varying “shift” and “steep” on probability score sigmoidal function. Relative changes in x-axis shifting and steepness are similar to sigmoidal normalizing function (Fig. 2). A. Parameter “steep” was held at 0.45, increasing “shift” shifts curve right with little effect on steepness. B. Parameter “shift” was held at 2, increasing “steep” increases steepness of curve with shifting effects.*

A second sigmoidal function (2) was implemented to determine probability scores for each class of mutation. Here, *Score* denotes the combined weighted scores from all considered characteristics for a specific mutation class, and, generally, *shift* and *steep* control the x-axis shift and steepness of the sigmoidal curve, respectively (Fig. 3). Weighing of each characteristic varied between each mutation class and optimal weights were determined through iterative adjustments and analyses (see Section 4.1 for all parameter testing and calibration).

**3.3 Testing pCoMiC**

To test pCoMiC, 54 characterized mutations in the CDS were selected from the CFTR database to be classified. Of the known mutants, 23 fell into class I, 16 class II, 11 class III, 6 class IV, 0 class V, 4 benign and 3 unknown. Importantly, mutations can and often will fall into more than one class hence the sum of all classes exceeding total mutations. Mutations were predicted to be in a class if pCoMiC output a probability of ≥ 60% for said class. The predicted (P) classifications were compared with expected (E) to determine the efficacy of pCoMiC. Classification was deemed correct if E and P output shared any classes due to the overlapping nature of the mutations (e.g. E: II, III P: III and E: III, P: II, III are both deemed correct).

**3.4 Data and Code Availability**

The reference CFTR CDS sequence was obtained from NCBI GenBank. Patient data was limited due to confidentiality constraints. As such, mutated sequences were derived from the reference sequence through a custom function. An extensive database of known mutations was obtained from Clinical and Functional Translation of CFTR (CFTR2). Classifications of each known mutation were derived from information in the CFTR2 database and various submissions in ClinVar Miner. The reference CFTR protein sequence was derived from the reference CDS sequence through a custom function. Pairwise alignments were performed using the PairwiseAlignment function in the BioPython package. All analyses were performed in Python (v3.10.4) and source code is available upon request.

**4.0 Results**

**4.1 Calibration of Initial Characteristic Scores and Weights**

The three amino acid characteristics selected, hydropathy, relative charge, and HBC, play large roles in determining protein conformation and thus which class a mutation will fall into. In addition, the position of the mutation in the CDS greatly affects the phenotype of the resultant protein15. As such, optimal parameter values and weighted scores differed between the 5 scoring functions. It was determined that HySs and CSs needed to be normalized prior to use in scoring functions in a similar fashion to HbS (1). HySs and CSs were often weighted too high (in some cases with initial scores of greater than 3 for HyS), greatly increasing the predicted probabilities for classes II-V.

Values for *k* (1) were iteratively adjusted for each normalization step until the optimal value outputting the maximum percentage of correct classifications was determined (see Appendix, Table 1). Values for *s* (1) were held constant at 5 to avoid excessive testing. Initial HySs varied the most and were subjected to a piecewise function for normalization steps to penalize extremely high values to a larger degree. Initial CSs rarely exceeded 2 and thus was normalized consistently for all initial values exceeding 1.

**4.2 Parameter Variation Between Scoring Functions and Probability Calculation Theory**

HySs and CSs were accounted for in all scoring functions but class I since they play a vital role in determining protein conformation (class II), forming and stabilizing ATP-binding pockets and the channel pore (classes III and IV) and, to a lesser extent, shaping the regulatory domain of CFTR (class V)1,3,5,15. HbS was not accounted for in the class III scoring function unless the residue which was mutated was initially one that was capable of forming hydrogen bonds with ATP to stabilize and facilitate its binding (e.g. S, T, Y, etc.).

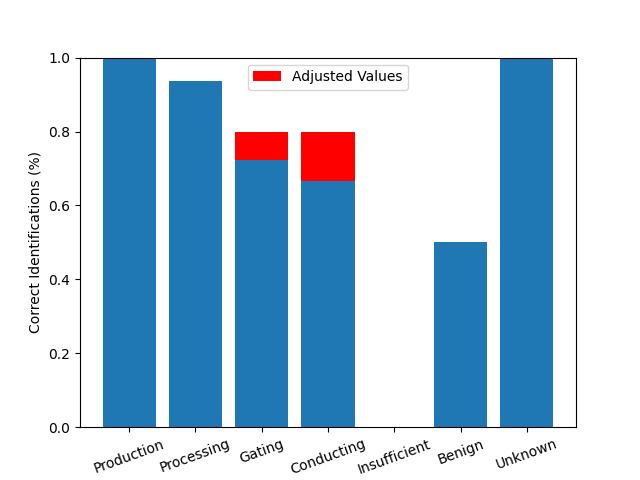
As mentioned, the position of the mutation in the CFTR CDS has a significant impact on the resultant phenotype. NBD1 and 2 are crucial regions for binding ATP and interact heavily with the regulatory domain to maintain correct conformation5. Therefore, it was determined that increasing the weights of all three scores for mutations within NBD1 or 2 resulted in significant improvements for correctly identifying class II and III mutations. A *steep* value (2) of 0.25 for mutations within these domains and 0.5 outside was found to be optimal resulting in 93.8% of class II mutations being correctly identified though only 58.3% of class III mutations. To improve the latter, an additional constraint was added to increase weighted scores if the initial residue was integral in facilitating binding of ATP, similar to as described above relating to HbS. For these residues, a *shift* value (2) of 1.65 and 2 for all other residues was determined optimal. In this case, the *shift* value was altered to avoid excessive changes in *steep* and inflating scores. Correct identification of class III mutations subsequently increased to 72.3%. It should be noted that with only 11 characterized class III mutations being used, this number is lower than expected and characterization of only 4 additional class III mutations saw an increase to 80.0%.

As for class IV mutations, transmembrane (TM) regions 1 and 6 of MSD1 and TM12 of MSD2 are important domains which have been shown to compromise the interior of the channel1,6,15. Mutations within these regions may affect the conformation of the channel and the resultant efficacy of chloride specific transport. Thus, *steep* was determined to produce optimal results at 0.25 within these regions and 0.55 outside. Additionally, mutations resulting in a less positively charged channel interior will affect conductance of the negatively charged chloride ions6. In such events, decreasing *steep* by 0.1 and *shift* by 0.5 from their initial values was shown to attain peak performance. Similarly to class III mutations, only 6 class IV mutations were characterized yielding a relatively low 66.7% correct identification rate which increased to 80.0% with 4 additional class IV mutations analyzed.

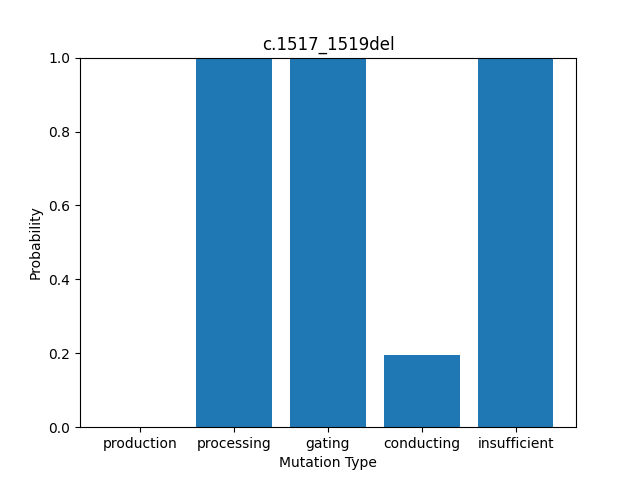
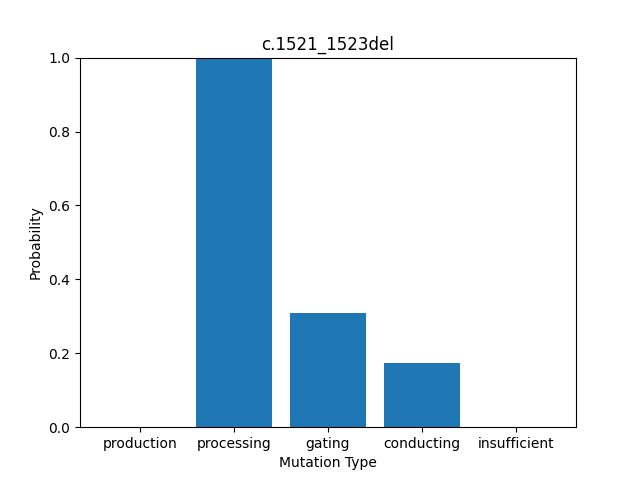
Class I mutations did not require any parameter calibration as pCoMiC will simply identify if a premature stop codon has been introduced during pairwise alignment and check if the length of the protein sequences differs significantly to assign a score of 0 or 1. Class V mutations are considerably more difficult to identify as they primarily involve mutations within non-coding regions of the CFTR gene which are not currently analyzed through this pipeline (see Section 5.0). Currently, pCoMiC checks if the mutation occurs within the regulatory domain to assign a score for class V mutations, though the results have not been tested and any output should not be considered for subsequent treatment selection at the moment.

**4.3 Efficacy of pCoMiC for Known Mutations**

Of the 54 known mutations, pCoMiC correctly classified 49 mutants producing a correct identification rate of 90.7%. The 3 unknown mutations were Pro67Leu, Leu1077Pro and Ala559Thr and were predicted as II and III, II and III, and II, respectively. Pro67Leu has been identified in patients also containing the Phe508del mutation, indicating the mutation produces a recessive allele and is likely associated with class II symptoms2,14. Leu1077Pro has loosely been associated with class III symptoms and Ala559Thr is predicted to disrupt protein conformation meaning it is most likely class II2,14. Thus, all 3 unknowns were deemed correctly characterized.

**Figure 4**: *Correct identification rates of the 54 known mutations for each mutation type. Red bars depict adjusted values for class III and IV after 4 additional mutations for each class were identified. No class V mutations were included.*

As for other classes, class I was correctly classified at a rate of 100.0%, II at 93.8%, III at 72.3% (80.0% when adjusted), IV at 66.7% (80.0% when adjusted) and benign at 50.0% (Fig. 4). No class V mutations were tested (see Sections 4.2 and 5.0).

****

B

A

**Figure 5**: *Sample representation of bar plot output from pCoMiC. Mutation type plotted on x-axis and corresponding predicted probability on y-axis. Figure titles importantly denote the variant cDNA form, not protein. A. Calculated probabilities for F508del mutation. Correctly categorized as class II, processing mutation. B. Calculated probabilities for I507del. Though correctly categorized as class II, class III and V are drastically inflated from extremely high HpS.*

Many mutations were classified effectively and reliably such as F508del, the most common mutation in CF patients2 (Fig. 5A). However, some continued to show skewed results such as I507del which output an initial HpS of 10.3 being normalized to 0.999 and drastically inflating its weight in scoring functions for classes II, III and V (Fig. 5B). No false negatives were reported as a result of heavily penalizing malignant mutations wrongly classified as benign during parameter calibration (given a score of -2 during correct identification procedures). Unfortunately, the increased stringency resulted in a 50% false positive rate (Fig. 4). Bar graphs and csv output for all tested mutations are shown (see Appendix, Figs. 6-9, Tables 1 & 2).

**5.0 Discussion and Conclusions**

Currently, pCoMiC excels at identifying mutations which follow typical conventions though is limited to the analysis of the CDS of the CFTR gene. This limits its ability to analyze alternative splicing events in non-coding regions which result in recognition factors potentially being spliced out and insufficient levels of protein being produced, indicative of a class V mutation15. As such, class V mutations cannot be correctly characterized and will require a more in-depth analysis than investigating solely mutated residues.

Most importantly, pCoMiC does not analyze the interactions between the various regions of the CFTR protein and instead compares the effects of mutations on the average amino acid characteristics of surrounding regions. This greatly reduces the likelihood of correctly identifying certain class II, III and IV mutations. In the analyses performed, scores for these mutations are weighted considerably higher if the mutation resides in an important domain (i.e. NBD1 and 2, TM1, 6 and 12, etc.). However, mutations outside of these domains, which may interact with said regions and influence protein conformation, are weighted at baseline resulting in incorrect classification. Such mutations compromise 3 of the 5 incorrect classifications. As a result, the optimal parameters and weights determined were of lower stringency than expected leading to smaller differences in average amino acid characteristics producing higher scores than anticipated. This is reflected in the remaining two incorrect classifications which were expected to be benign though both classed as II and III (see Appendix and Supplementary Data, Table 2).

In future development, adapting pCoMiC to accept a Gen-Bank-like file which details the positions of and interactions between notable domains will allow for a more complete analysis and is predicted to increase the correct classification rate to upwards of 95% with a false positive rate of < 10%. This also may enable pCoMiC to be tailored to classify mutations in any gene, though additional information regarding mutant classes will be required as input and scoring functions will need to be generalized.

In summary, pCoMiC demonstrates the ability to correctly characterize 90.7% of CFTR mutations in their respective classes, enabling appropriate treatment selection. The software is currently being adapted for command-line usage making it extremely user-friendly and capable of efficiently generating output in less than a minute. Though in the initial stages of development, pCoMiC displays effective use in efficient mutation classification and expansion of existing CFTR databases, alluding to a promising future for *in silico* mutant classification of the CFTR CDS.

# 

# **6.0 References**

[1]: Corradi, V., Vergani, P., & Tieleman, D. P. (2015). Cystic Fibrosis Transmembrane Conductance Regulator (CFTR). *Journal of Biological Chemistry*, *290*(38), 22891–22906. https://doi.org/10.1074/jbc.m115.665125

[2]: Cystic Fibrosis Foundation. (2022). *Types of CFTR Mutations | Cystic Fibrosis Foundation*. www.cff.org; Cystic Fibrosis Foundation. https://www.cff.org/research-clinical-trials/types-cftr-mutations

[3]: Hwang, T. C., & Kirk, K. L. (2013). The CFTR ion channel: gating, regulation, and anion permeation. *Cold Spring Harbor perspectives in medicine, 3*(1), a009498. https://doi.org/10.1101/cshperspect.a009498

[4]: Kyte, J. & Doolittle, R. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*(1), 105-132. 10.1016/0022-2836(82)90515-0

[5]: Lewis, H., et al. (2004). Structure of nucleotide-binding domain 1 of the cystic fibrosis transmembrane conductance regulator. *The EMBO Journal, 23*(2), 282-293. 10.1038/sj.emboj.7600040

[6]: Linsdell, P. (2005). Mechanism of chloride permeation in the cystic fibrosis transmembrane conductance regulator chloride channel. *Experimental Physiology*, *91*(1), 123–129. https://doi.org/10.1113/expphysiol.2005.031757

[7]: Maheshwari, R., & Dhathathreyan, A. (2004). Investigation of surface properties of amino acids: polarity scale for amino acids as a means to predict surface exposed residues in films of proteins. *Journal of Colloid and Interface Science*, *277*(1), 79–83. [https://doi.org/10.1016/j.jcis.2004.04.02](https://doi.org/10.1016/j.jcis.2004.04.023)3

[8]: Marisa Wexler, M. (2023, October 10). *Types of CFTR mutations*. Cystic Fibrosis News Today. https://cysticfibrosisnewstoday.com/types-of-cftr-mutations/

[9]: Mayo Clinic. (2021, November 23). *Cystic Fibrosis - Symptoms and Causes*. Mayo Clinic; Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/cystic-fibrosis/symptoms-causes/syc-20353700

[10]: National Heart, Lung, and Blood Institute. (2022, March 24). *Cystic Fibrosis - Causes*. Www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/cystic-fibrosis/causes#:~:text=The%20CFTR%20protein%20has%20also

[11]: Sharma, N., & Cutting, G. R. (2020). The genetics and genomics of cystic fibrosis. *Journal of Cystic Fibrosis*, *19*, S5–S9. https://doi.org/10.1016/j.jcf.2019.11.003

[12]: The Canadian Cystic Fibrosis Registry. (2021). *2021 annual data report*. https://www.cysticfibrosis.ca/uploads/2021-Annual-Data-Report-WEB-AODA.pdf

[13]: The Clinical and Functional Translation of CFTR. (2023, April 7). *CFTR2 Variant List History*. Cftr2.org. https://cftr2.org/mutations\_history

[14]: University of Utah Health. (2023). *ClinVar Miner*. https://clinvarminer.genetics.utah.edu/

[15]: Veit, G., et al. (2016). From CFTR biology toward combinatorial pharmacotherapy: expanded classification of cystic fibrosis mutations. *Molecular Biology of the Cell, 27*(3), 424-433. 10.1091/mbc.E14-04-0935

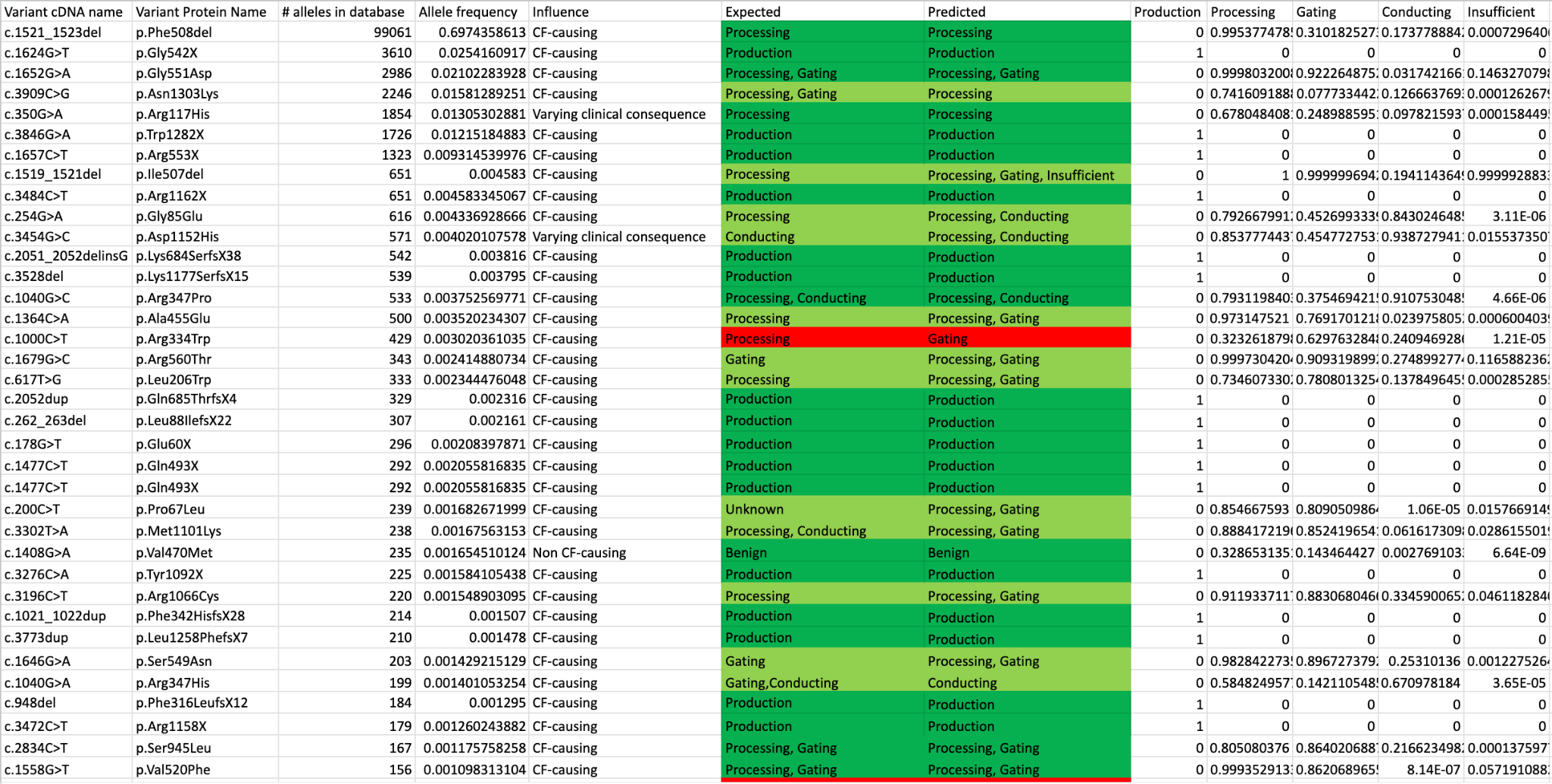
[16]: Zhou, Z., Wang, X., Liu, H.-Y., Zou, X., Li, M., & Hwang, T.-C. (2006). The Two ATP Binding Sites of Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Play Distinct Roles in Gating Kinetics and Energetics. *The Journal of General Physiology*, *128*(4), 413–422. <https://doi.org/10.1085/jgp.200609622>

**Appendix**

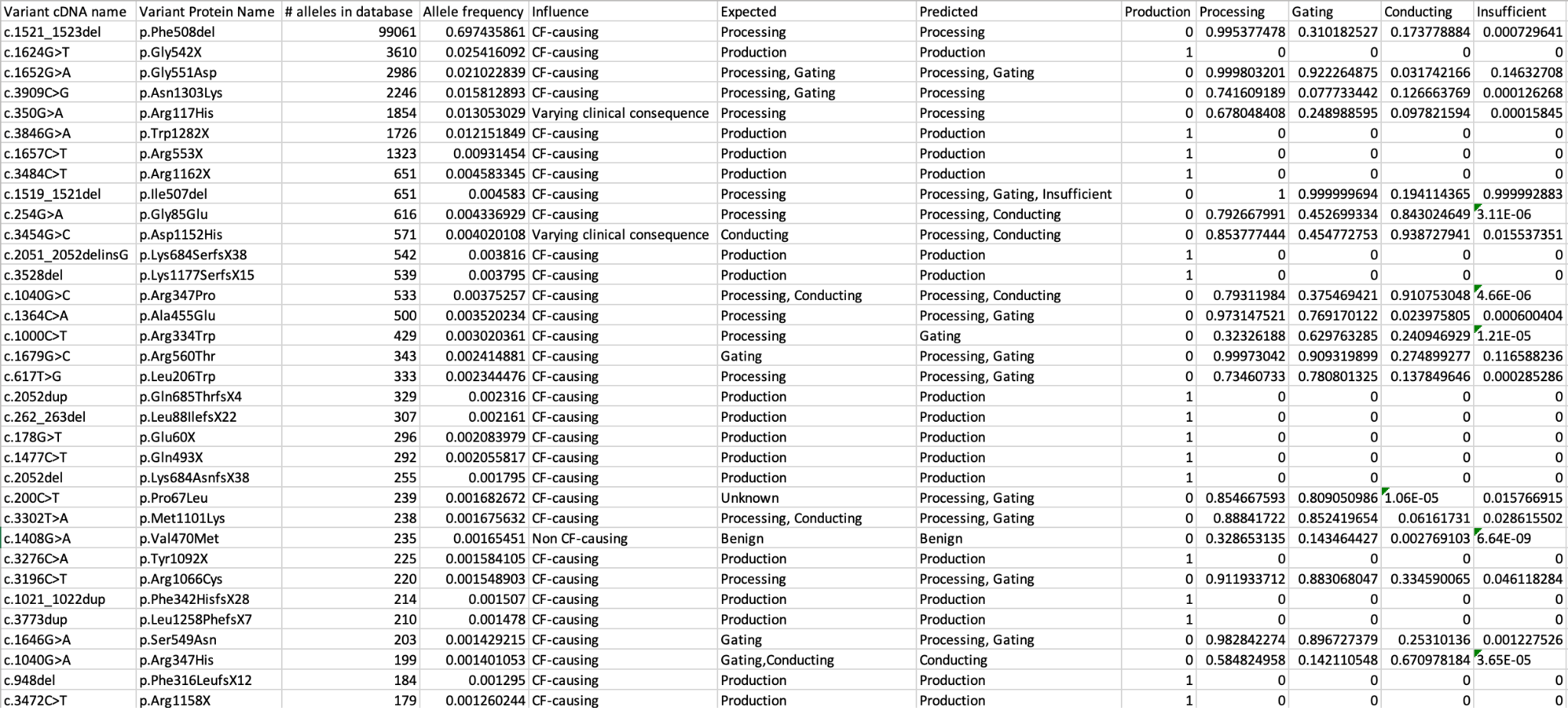
**Table 1**: *Iterative adjustments of k in sigmoidal normalizing functions for varying values of HyS and CS and corresponding percentage of correct classifications. Optimal parameter values determined are highlighted in green.*

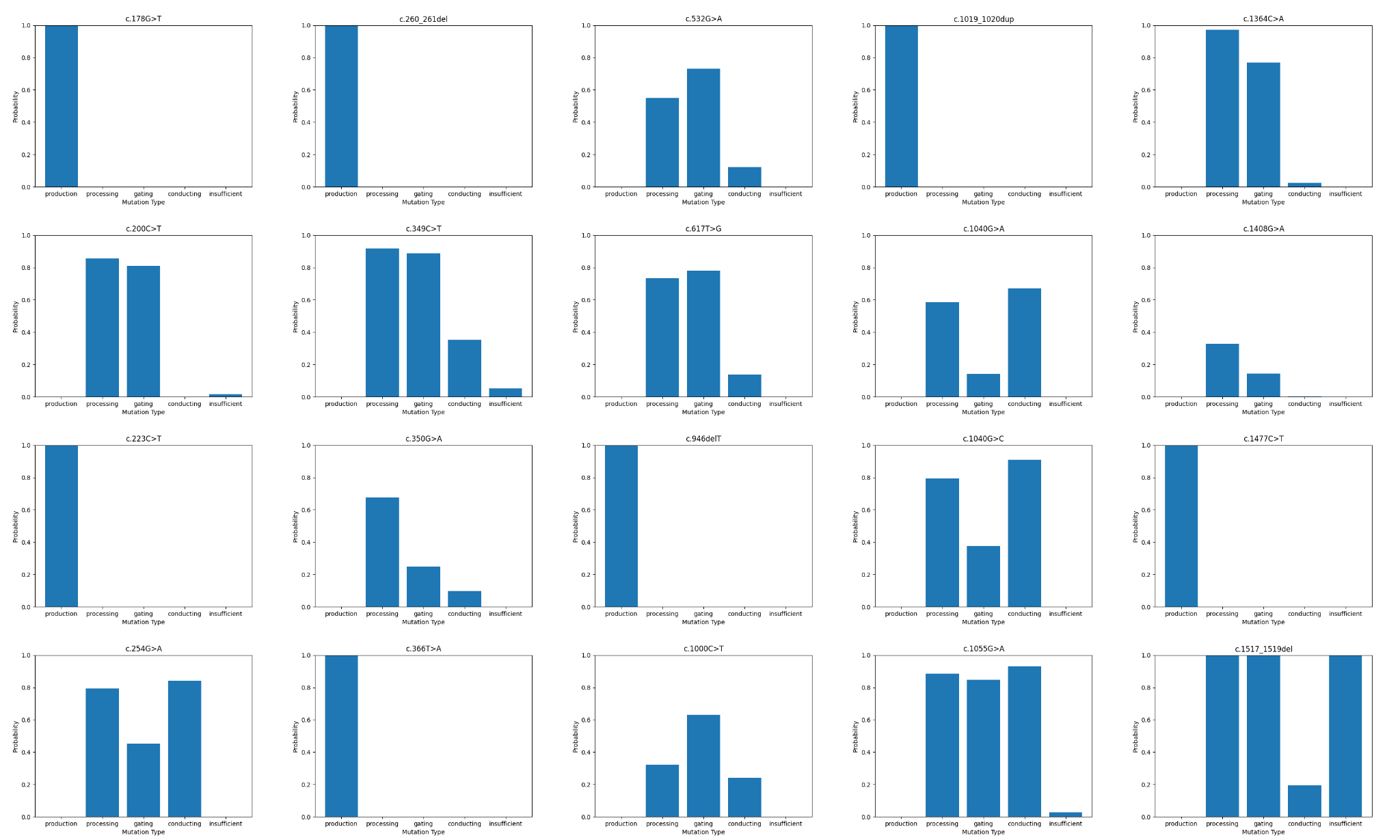
|  |  |  |
| --- | --- | --- |
| Score Type and Value | *k* | Correct Classifications (%) |
| HyS > 3 | 40 | 42.6 |
| 55 | 51.9 |
| 70 | 72.2 |
| 85 | 90.7 |
| 100 | 22.2 |
| 2 < HyS < 3 | 25 | 64.8 |
| 40 | 74.1 |
| 55 | 88.9 |
| 70 | 77.8 |
| 85 | 44.4 |
| 1 < HyS < 2 | 1 | 9.3 |
| 15 | 92.6 |
| 30 | 70.3 |
| 45 | 63.0 |
| CS > 1 | 1 | 33.3 |
| 15 | 87.0 |
| 30 | 74.1 |
| 45 | 27.8 |

**Table 2:** *Screenshot depicting formatted output of pCoMiC highlighting matches in expected and predicted classifications. Dark green cells indicate exact matches, light green cells partial matches and red cells incorrect classifications.*

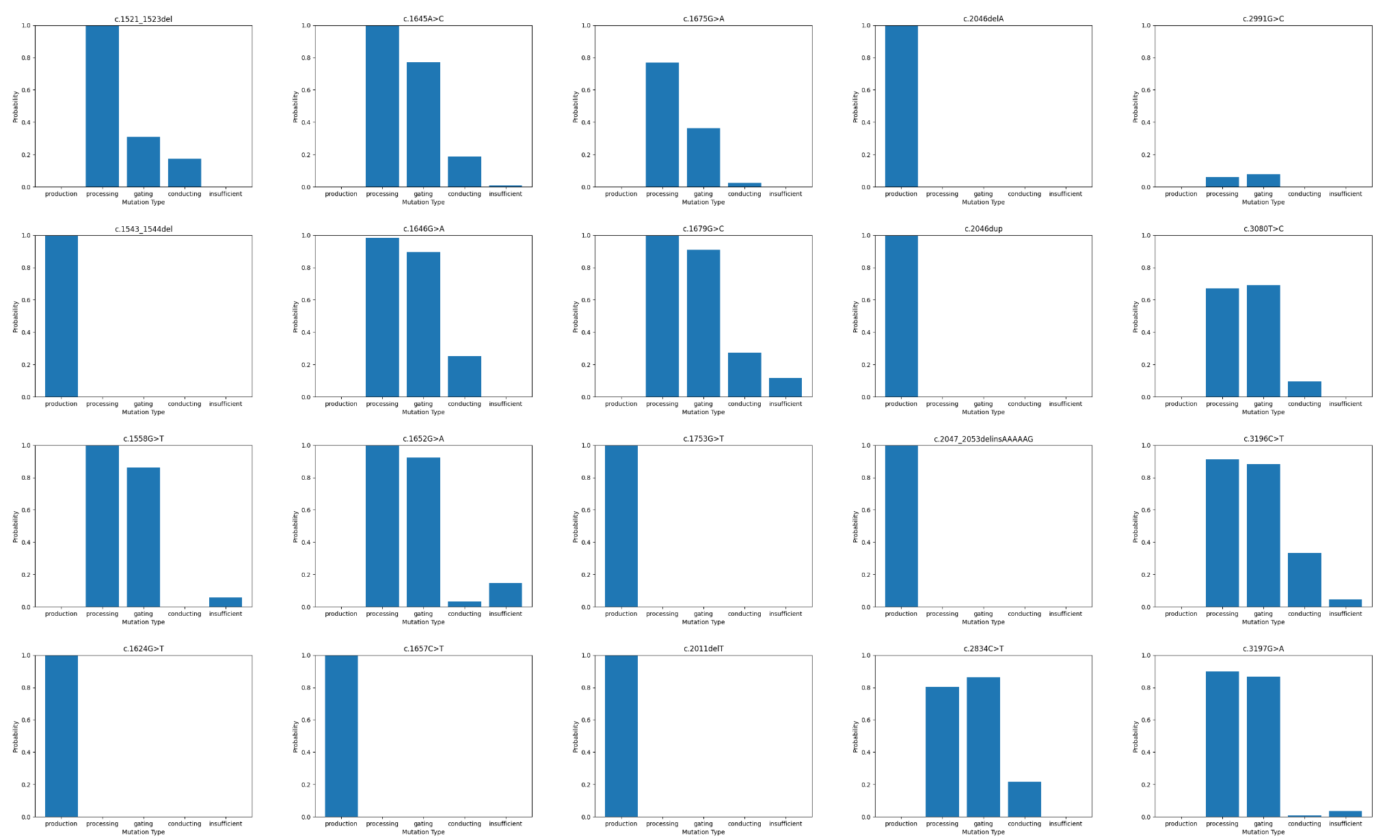
****

**Table 3:** *Screenshot depicting raw .csv output of pCoMiC including information obtained from CFTR2 database for known mutations (variant cDNA name to Expected) alongside calculated information (Predicted to Insufficient).*

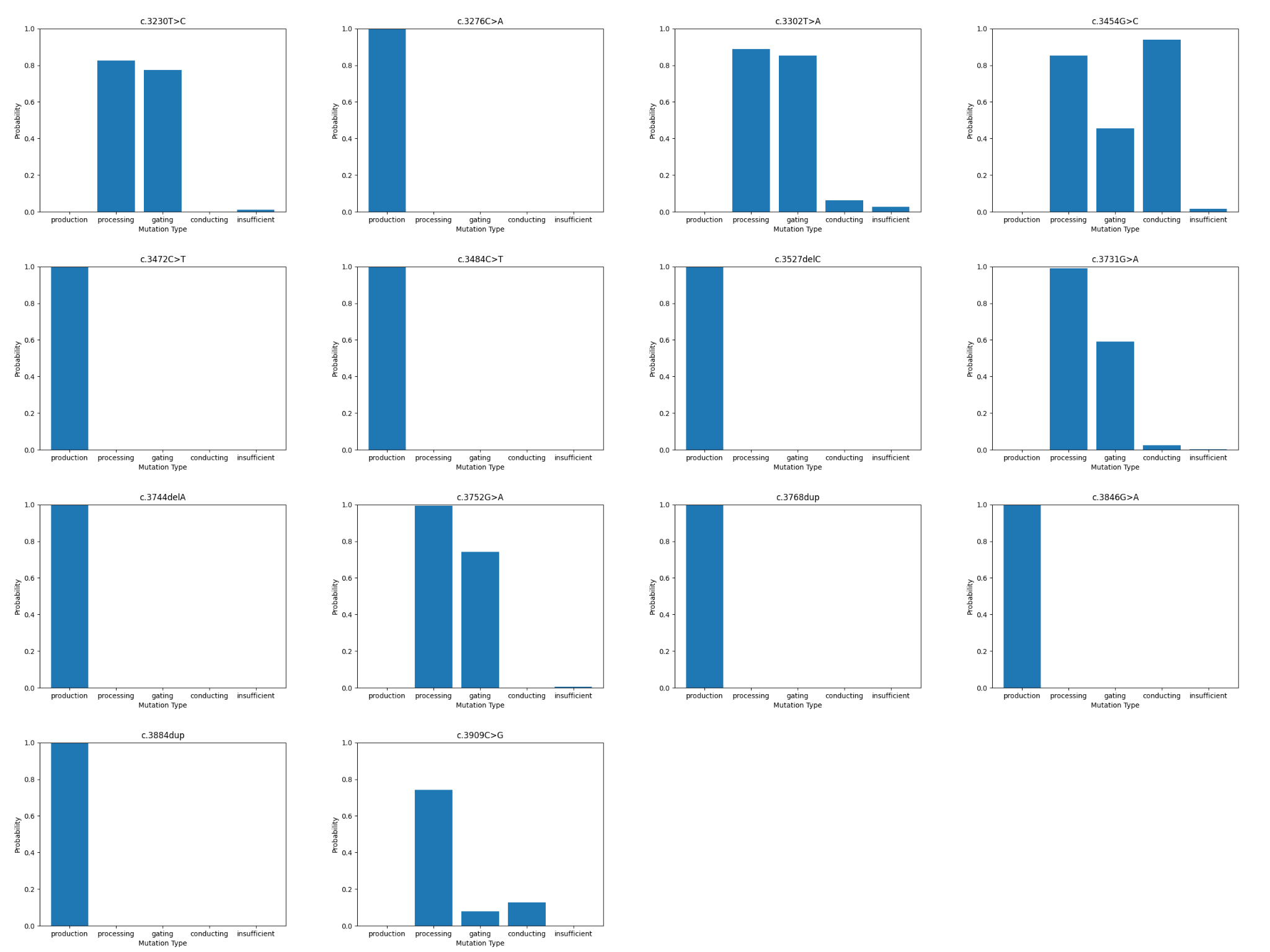




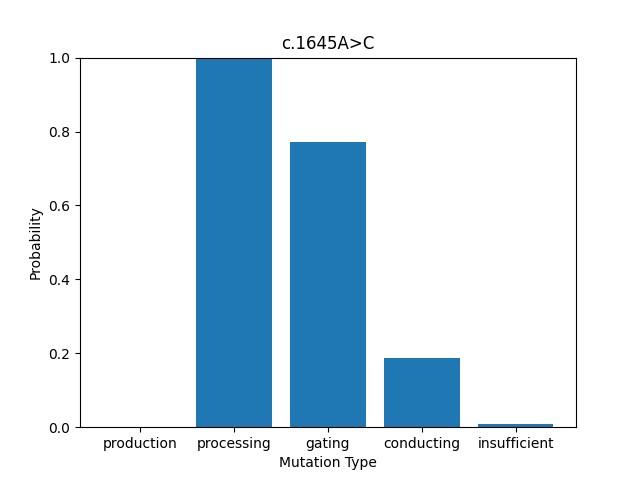
**Figure 6:** *Merged image of the first 20 graphical outputs of pCoMiC testing depicting probability scores of the five mutation classifications in variant cDNA form.*



**Figure 7:** *Merged image of the next 20 to 40 graphical outputs of pCoMiC testing depicting probability scores of the five mutation classifications in variant cDNA form.*



**Figure 8:** *Merged image of the first 14 graphical outputs of pCoMiC testing depicting probability scores of the five mutation classifications in variant cDNA form.*



**Figure 9:** *Enlarged graph depicting calculated probabilities among the five mutant classifications for the mutation of variant cDNA form c.1645A>C (variant protein form of p.Ser549Arg).*