SUPPLEMENTARY MATERIALS

Title
Iterative Sequencing and Variant Screening (ISVS) as a novel pathogenic mutations search strategy - application for *TMPRSS3* mutations screen

Authors:

Urszula Lechowicz[1], Tomasz Gambin[2,3], Agnieszka Pollak[1], Anna Podgorska[1], Piotr Stawinski[1], Andre Franke[4], Britt-Sabina Petersen[4], Malgorzata Firczuk[5], Monika Oldak[1], Henryk Skarzynski[6], Rafal Ploski[7]*

Affiliation:
[1] Department of Genetics, Institute of Physiology and Pathology of Hearing, Warsaw, Poland
[2] Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland
[3] Department of Medical Genetics, Institute of Mother and Child at Warsaw, Warsaw, Poland
[4] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany
[5] Department of Immunology, Center of Biostructure Research, Medical University of Warsaw, Warsaw, Poland
[6] Oto-Rhino-Laryngology Surgery Clinic, Institute of Physiology and Pathology of Hearing, Warsaw, Poland,
[7] Department of Medical Genetics, Centre of Biostructure, Medical University of Warsaw

* Corresponding author: Rafal Ploski, Prof., MD, PhD; Department of Medical Genetics, Centre of Biostructure Research, Medical University of Warsaw, Warsaw, Pawinskiego 3C, 02-106, Poland tel.: + 48 22 572 0606, fax: + 48 22 572 0696, e-mail: rploski@wp.pl

Selection of classification method and model tuning

To select the optimal model for classification of variant pathogenicity we compared the performance of five state-of-the-art supervised machine learning methods, including, decision trees (from R package “CART”), linear discriminant analysis (from R package “MASS”), support vector machines (SVM, from R package “kernlab”), random forest (from R package “RF”) penalized logistic regression (from R package “plr”). For SVM we considered models with linear and radial kernels. The classifiers were evaluated on the raw output of ISVS simulator executed with default parameters and 1000 iterations. Performance comparison of algorithms and model tuning was performed using R package “caret”. This algorithms’ evaluation (see Supplementary Table S2) revealed that SVM, Random Forest and Penalized Logistic Regression significantly outperform two other algorithms (Decision Trees and LDA). In the same time, we found no statistically significant differences among three best performing methods (using comparison test implemented in “caret” R package). Moreover, evaluation results for SVM with radial and linear kernels were comparable.

Since there was no obvious winner we decided to use SVM algorithm for classification of variant pathogenicity. In particular, we selected SVM with linear kernel, because it is not recommended practice to use SVM with other (e.g. radial) kernels when observations from different classes are likely to be linearly separable. Finally, we fine-tuned the SVM model to select optimal value for “C” parameter (optimal model was found for C=1, see Supplementary Table S3).

SUPPLEMENTARY TABLES

Supplementary Table S1. Sequences of primers used in Sanger sequencing

Primers	
Name	**Sequence 5'**
TMPR3e1F	ccgccctctcagagttacag
TMPR3e1R	ttgttttcacctgtcccaca
TMPR3e2F	tgaccaagatgcacctgatg
TMPR3e2R	ccccacagggacagtcagt
TMPR3e3F	ctagagaatgtgccccttgg
TMPR3e3R	taattaaggctgggcagcag
TMPR3e4F	gcactctgaaagagctgttgg
TMPR3e4R	tacagatgggaagggtcagg
TMPR3e5F	cagggatccagagtcactgc
TMPR3e5R	agagcgttaaagcacccaat
TMPR3e6F	ttgccagggtgagtgaactt
TMPR3e6R	tattgggccatactccctca
TMPR3e7F	atctggggcatttttcacag
TMPR3e7R	ctccagcaggtaggggtaca
TMPR3e8F	cccttgcagcacttgtctta
TMPR3e8R	tgatgatgatgggtccacag
TMPR3e9F	ggaccacatcttgcctgataa
TMPR3e9R	aactgatgccaacaccaaca
TMP3e10F	tgctgtgagctgatcgtttt
TMP3e10R	tgactgtgtcccgagcag
TMP3e11F	gcgacacaccagagagcat
TMP3e11R	ttcttctccacgccctgtaa
TMP3e12F	gtcccaactccatagcaagc
TMP3e12R	accaagtcactgctgctgaa
TMP3e13F	agaacagccccacaattcc
TMP3e13R	ctcagagctccaagggtgtc

Supplementary Table S2. Evaluation of performance (accuracy, kappa) of various classification algorithms used for discrimination between pathogenic and nonpathogenic variants

Accuracy						
	Min.	1st qu	Median	Mean	3rd qu	Max.
Decision Tree	0.8902	0.8971	0.9322	0.9174	0.9353	0.9385
LDA	0.9159	0.9204	0.9253	0.9239	0.9274	0.9301
SVM_RADIAL	0.9684	0.9698	0.9725	**0.9721**	0.9732	0.9774
SVM_LINEAR	0.9684	0.9696	0.9725	**0.9721**	0.9736	0.9778
Random Forest	0.968	0.9694	0.9729	**0.9719**	0.9736	0.9767
Penalized Logistic Regression	0.9684	0.9696	0.9729	**0.9722**	0.9741	0.9778
Kappa						
	Min.	1st qu	Median	Mean	3rd qu	Max.
Decision Tree	0.7758	0.79	0.8628	0.832	0.8691	0.8754
LDA	0.8282	0.8376	0.8476	0.8449	0.852	0.8577
SVM_RADIAL	0.9361	0.9389	0.9445	**0.9436**	0.946	0.9544
SVM_LINEAR	0.9361	0.9386	0.9445	**0.9436**	0.9467	0.9551
Random Forest	0.9354	0.9382	0.9452	**0.9433**	0.9467	0.953
Penalized Logistic Regression	0.9361	0.9386	0.9452	**0.9439**	0.9477	0.9551

Supplementary Table S3. Fine-tuning of the SVM model to select optimal value for “C” parameter

C	Accuracy	Kappa
1.00E-03	0.861986	0.7155854
1.00E-02	0.9401231	0.8782235
1.00E-01	**0.9615572**	**0.922119**
1.00E+00	0.9720541	0.9435726
1.00E+01	0.9703162	0.940164

Supplementary Figure S1: Sensitivity of ISVS to input parameters

diseaseCohortSize vs successRate
diseaseFreq vs successRate
diseaseGeneFract vs successRate
patMutCnt vs successRate
successRate
0.0
0.2
0.4
0.6
0.8
1.0
diseaseCohortSize vs patFractDetected
diseaseFreq vs patFractDetected
diseaseGeneFract vs patFractDetected
patMutCnt vs patFractDetected
patFractDetected
diseaseCohortSize vs mutFractDetected
diseaseFreq vs mutFractDetected
diseaseGeneFract vs mutFractDetected
patMutCnt vs mutFractDetected
mutFractDetected
diseaseCohortSize vs avgSteps
diseaseFreq vs avgSteps
diseaseGeneFract vs avgSteps
patMutCnt vs avgSteps
avgSteps
0
1
2
3
4
5
6
diseaseCohortSize vs sequencedSampleNr
diseaseFreq vs sequencedSampleNr
diseaseGeneFract vs sequencedSampleNr
patMutCnt vs sequencedSampleNr
sequencedSampleNr
0
200
400
600
diseaseCohortSize vs screenedVarsNr
diseaseFreq vs screenedVarsNr
diseaseGeneFract vs screenedVarsNr
patMutCnt vs screenedVarsNr
screenedVarsNr
0
5
10
15
20
25
30
diseaseCohortSize
0
1000
2000
3000
4000
5000
diseaseFreq
0.000
0.001
0.002
0.003
0.004
0.005
diseaseGeneFract
0.02
0.04
0.06
0.08
0.10
patMutCnt
5
10
15
20

Supplementary Figure S1: Sensitivity of ISVS to input parameters (continue)

nonPatMutCnt vs successRate
nonPatCumFreq vs successRate
iterNr vs successRate
secondPop vs successRate
nonPatMutCnt vs patFractDetected
nonPatCumFreq vs patFractDetected
iterNr vs patFractDetected
secondPop vs patFractDetected
nonPatMutCnt vs mutFractDetected
nonPatCumFreq vs mutFractDetected
iterNr vs mutFractDetected
secondPop vs mutFractDetected
nonPatMutCnt vs avgSteps
nonPatCumFreq vs avgSteps
iterNr vs avgSteps
secondPop vs avgSteps
nonPatMutCnt vs sequencedSampleNr
nonPatCumFreq vs sequencedSampleNr
iterNr vs sequencedSampleNr
secondPop vs sequencedSampleNr
nonPatMutCnt vs screenedVarsNr
nonPatCumFreq vs screenedVarsNr
iterNr vs screenedVarsNr
secondPop vs screenedVarsNr
successRate
patFractDetected
mutFractDetected
0.0
0.2
0.4
0.6
0.8
1.0
avgSteps
0
1
2
3
4
5
6
sequencedSampleNr
0
200
400
600
screenedVarsNr
0
5
10
15
20
25
30
nonPatMutCnt
0
20
40
60
80
100
nonPatCumFreq
0.05
0.10
0.15
0.20
iterNr
0
2000
4000
6000
8000
10000
secondPop
0.0
0.1
0.2
0.3
0.4
0.5

Abbreviations of X labels: diseaseCohortSize − disease cohort size (default=2000); diseaseFreq − frequency of disease individuals within the population (default=0.001); diseaseGeneFract − the fraction of disease individuals affected by bi−allelic variants in the GENE (default=0.02); patMutCnt − the number of distinct pathogenic mutations (default=10); nonPatMutCnt − the number of distinct non−pathogenic (default=20); nonPatCumFreq − cumulative frequency of non−pathogenic mutations (default=0.05); iterNr − number of repetitions of ISVS experiments (default=10000); secondPop − the ratio of population sizes (default=0, i.e. single population)

Abbreviations of Y labels: successRate − the fraction of ISVS simulations in which all individuals affected by bi−allelic pathogenic mutation were detected; patFractDetected − the average fraction of patients with bi−allelic pathogenic mutation who were properly identified; mutFractDetected − the average fraction of pathogenic variants that were properly identified; avgSteps − the average number of steps in ISVS experiment; sequencedSampleNr − the average number of sequenced samples; screenedVarsNr − the average number of screened variants

Supplementary Figure S2: Sensitivity of variant classification to input parameters

diseaseCohortSize vs SVMsens05
diseaseFreq vs SVMsens05
diseaseGeneFract vs SVMsens05
patMutCnt vs SVMsens05
diseaseCohortSize vs SVMspec05
diseaseFreq vs SVMspec05
diseaseGeneFract vs SVMspec05
patMutCnt vs SVMspec05
diseaseCohortSize vs SVMacc05
diseaseFreq vs SVMacc05
diseaseGeneFract vs SVMacc05
patMutCnt vs SVMacc05
diseaseCohortSize vs SVMsens95
diseaseFreq vs SVMsens95
diseaseGeneFract vs SVMsens95
patMutCnt vs SVMsens95
diseaseCohortSize vs SVMspec95
diseaseFreq vs SVMspec95
diseaseGeneFract vs SVMspec95
patMutCnt vs SVMspec95
diseaseCohortSize vs SVMacc95
diseaseFreq vs SVMacc95
diseaseGeneFract vs SVMacc95
patMutCnt vs SVMacc95
SVMsens05
SVMspec05
SVMacc05
SVMsens95
SVMspec95
SVMacc95
0.0
0.2
0.4
0.6
0.8
1.0
diseaseCohortSize
0
1000
2000
3000
4000
5000
diseaseFreq
0.000
0.001
0.002
0.003
0.004
0.005
diseaseGeneFract
0.02
0.04
0.06
0.08
0.10
patMutCnt
5
10
15
20

Supplementary Figure S2: Sensitivity of variant classification to input parameters (continue)

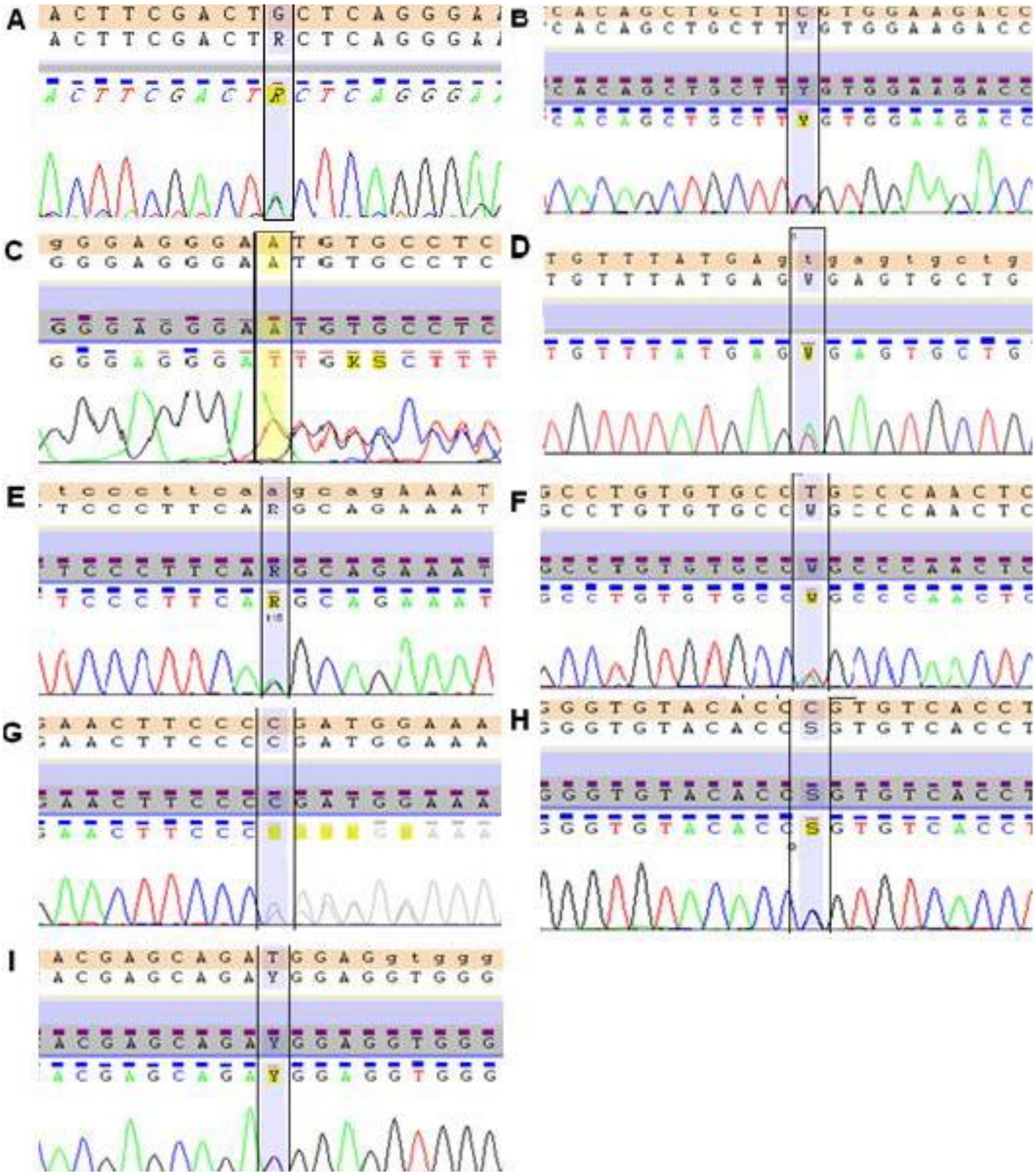
nonPatMutCnt vs SVMsens05
nonPatCumFreq vs SVMsens05
iterNr vs SVMsens05
secondPop vs SVMsens05
nonPatMutCnt vs SVMspec05
nonPatCumFreq vs SVMspec05
iterNr vs SVMspec05
secondPop vs SVMspec05
nonPatMutCnt vs SVMacc05
nonPatCumFreq vs SVMacc05
iterNr vs SVMacc05
secondPop vs SVMacc05
nonPatMutCnt vs SVMsens95
nonPatCumFreq vs SVMsens95
iterNr vs SVMsens95
secondPop vs SVMsens95
nonPatMutCnt vs SVMspec95
nonPatCumFreq vs SVMspec95
iterNr vs SVMspec95
secondPop vs SVMspec95
nonPatMutCnt vs SVMacc95
nonPatCumFreq vs SVMacc95
iterNr vs SVMacc95
secondPop vs SVMacc95
SVMsens05
SVMspec05
SVMacc05
SVMsens95
SVMspec95
SVMacc95
0.0
0.2
0.4
0.6
0.8
1.0
nonPatMutCnt
0
20
40
60
80
100
nonPatCumFreq
0.05
0.10
0.15
0.20
iterNr
0
2000
4000
6000
8000
10000
secondPop
0.0
0.1
0.2
0.3
0.4
0.5

Abbreviations of X labels: diseaseCohortSize − disease cohort size (default=2000); diseaseFreq − frequency of disease individuals within the population (default=0.001); diseaseGeneFract − the fraction of disease individuals affected by bi−allelic variants in the GENE (default=0.02); patMutCnt − the number of distinct pathogenic mutations (default=10); nonPatMutCnt − the number of distinct non−pathogenic (default=20); nonPatCumFreq − cumulative frequency of non−pathogenic mutations (default=0.05); iterNr − number of repetitions of ISVS experiments (default=10000); secondPop − the ratio of population sizes (default=0, i.e. single population)

Abbreviations of Y labels: SVMsens05 − the sensitivity of variant classification with condidence > 50%; SVMspec05 − the specificity of variant classification with condidence > 50%; SVMacc05 − the accuracy of variant classification with condidence > 50%; SVMsens95 − the sensitivity of variant classification with condidence > 95%; SVMspec95 − the specificity of variant classification with condidence > 95%; SVMacc95 − the accuracy of variant classification with condidence > 95%

Supplementary Figure S3: Sensitivity of ISVS and variant classification to the number and cumulative frequency of non−pathogenic mutations

successRate
1.0
0.8
0.6
0.4
0.2
0.0
successRate
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
0.992
0.990
0.988
0.986
0.984
0.982
patFractDetected
1.0
0.8
0.6
0.4
0.2
0.0
patFractDetected
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
0.9980
0.9975
0.9970
0.9965
mutFractDetected
1.0
0.8
0.6
0.4
0.2
0.0
mutFractDetected
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
0.9996
0.9994
0.9992
0.9990
0.9988
0.9986
avgSteps
8
6
4
2
0
avgSteps
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
7
6
5
4
sequencedSampleNr
800
600
400
200
0
sequencedSampleNr
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
700
600
500
400
300
200
100
screenedVarsNr
80
60
40
20
0
screenedVarsNr
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
60
50
40
30
20
10
SVMsens05
1.0
0.8
0.6
0.4
0.2
0.0
SVMsens05
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
1.00
0.95
0.90
0.85
SVMspec05
1.0
0.8
0.6
0.4
0.2
0.0
SVMspec05
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
1.0
0.9
0.8
0.7
0.6
SVMacc05
1.0
0.8
0.6
0.4
0.2
0.0
SVMacc05
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
1.00
0.98
0.96
0.94
0.92
0.90
0.88
SVMsens95
1.0
0.8
0.6
0.4
0.2
0.0
SVMsens95
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
SVMspec95
1.0
0.8
0.6
0.4
0.2
0.0
SVMspec95
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
1.0
0.8
0.6
0.4
0.2
SVMacc95
1.0
0.8
0.6
0.4
0.2
0.0
SVMacc95
20
40
60
80
100
nonPatMutCnt
0.05
0.10
0.15
0.20
nonPatCumFreq
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65



Supplementary figure S4. Sanger sequencing chromatograms showing novel variants identified in the *TMPRSS3* gene: (A) c.218G>A (p.C73Y), (B) c.371C>T (p.S124L), (C) c.579dupA (p.C194Mfs*17), (D) c.782+2T>A, (E) c.953-5A>G (F) c.974 T>A (G) c.999delC (p.D334Mfs*24), (H) c.1306C>G (p.R436G), (I) c.1343 T>C (p.M448T).

A LDLRA domain

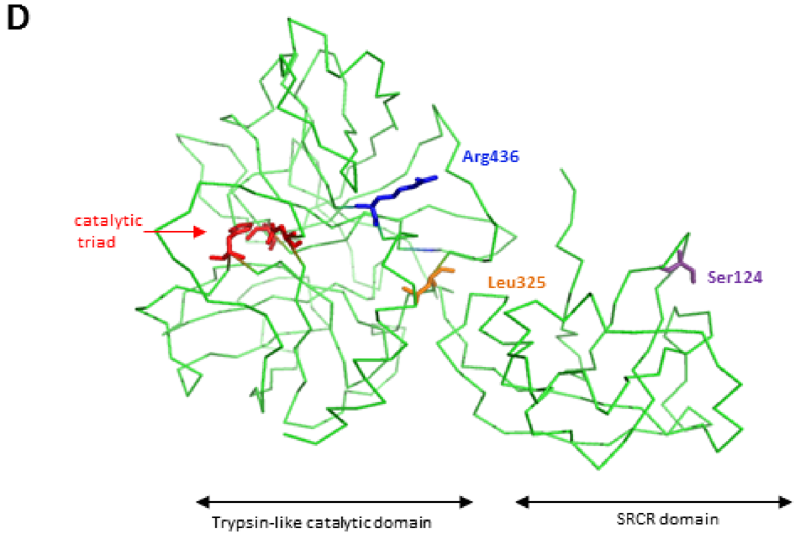
```
                Cys73
72         ...DCSGK-- --YRCR.... SSF.KCIELI A.RC...D.G VSD.CK...D G..EDE---- -YRCVR..  109
consensus  ...pCss s.. . ..F pCs....  s t .   .CI  sbpb.b C...D.G p sD.Cs....D  t..s DE.... . p s C  sp..
```

B SRCR domain

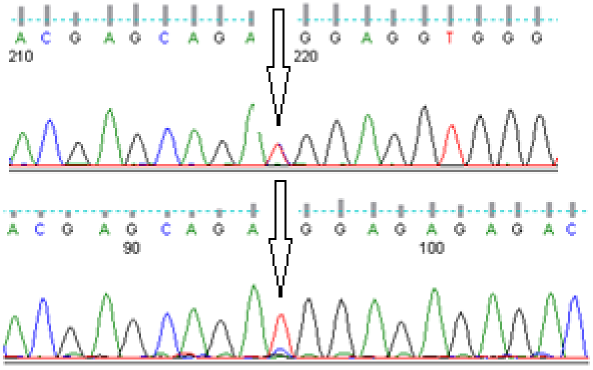
```
                                          Ser124
108        VRVGGQNA.. .----VLQVFTAA---SWKT MC..SDDW.. .KGHYANVAC AQLGFPSYVS SDNL..--RV  158
consensus  IRL s s .... ...sc.........Vb b st... pWGT  VC....D.. W.. .s b p-  As V IC ...pLGa s s t hp s s...........s
```

C Trypsin-like catalytic domain

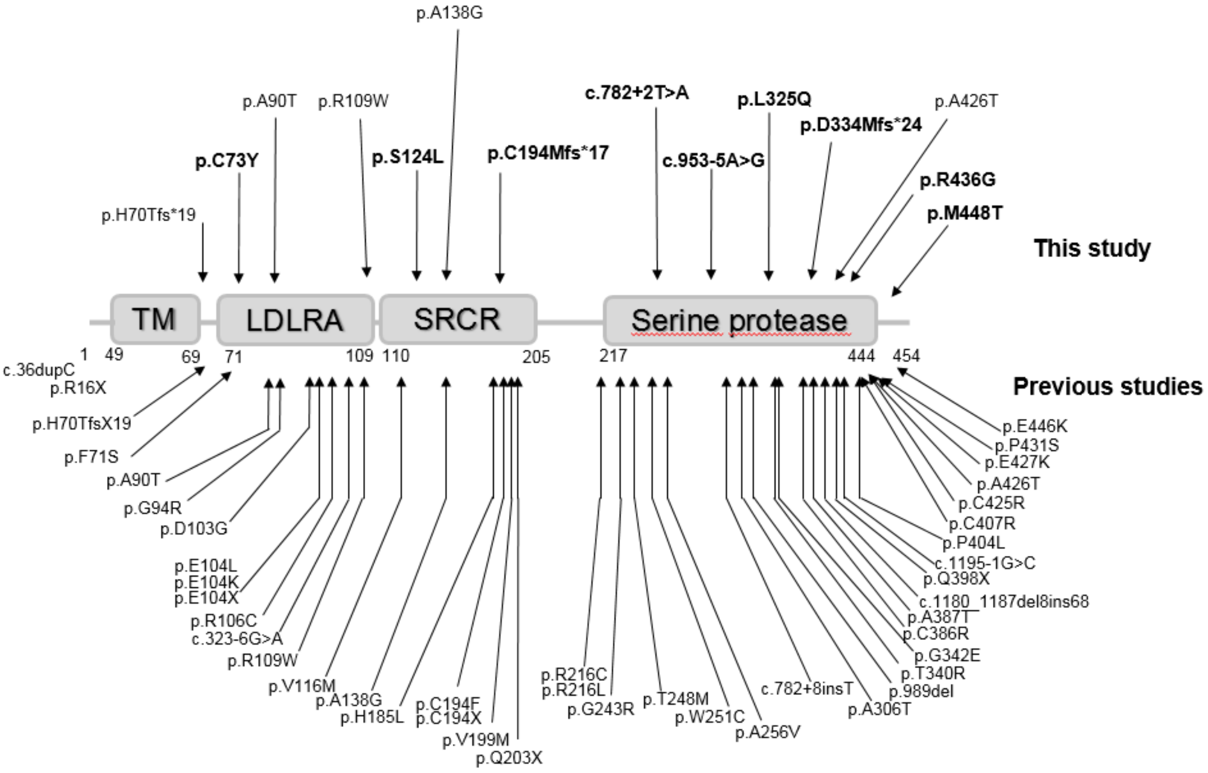
```
                                                                          Leu325
292        YHSKYKPK.. ....RLGNDI A...LMKLAG P... LTFNE- -MIQPVCLPN 327
consensus  . H  spY ss... ......s bs s DI A...L I +L pp  s....I s b ss. ..pI p s I s LPs

                                                                                 Arg436
399        GDSGGPLVCQ ER....RLWK LVGATSFG-- ..-IGCA... .EVNKP.GV. ..YTRVT..S 439
consensus  GDSGGPL  I s p ............h  h   I hGATSaG.......s  C t ............P.GV...Y  *+V  s..p
```



Supplementary figure S5. Conservation of mutated TMPRSS3 variants within LDLRA (**A**), SRCR (**B**), and Trypsin-like catalytic (**C**) domains. The corresponding TMPRSS3 sequences were aligned to SMART domain consensus (60%). Conserved residues are given a grey background. Codes used in a consensus are the following: -; negatively charged (D, E), *; (S, T), l; aliphatic (I, L, V), +; positive (H, K, R), t; tiny (A, G, S), a; aromatic (F, H, W, Y), c; charged (D, E, H, K, R), s; small (A, C, D, G, N, P, S, T, V), p; polar (C,D,E,H,K,N,Q,R,S,T), b; big (E,F,H,I,K,L,M,Q,R,W,Y), h; hydrophobic (A,C,F,G,H,I,L,M,T,V,W,Y). (**D)** –Model of TMPRSS3 comprising two domains, SRCR and catalytic, displayed as ribbons, with mutated residues and active site catalytic triad (colored red) shown in stick representations.



Supplementary figure S6. Chromatograms of g.DNA (A) and c.DNA (B) for carrier of p.M448T mutation in *TMPRSS3* gene.



Supplementary figure S7. Localization of novel TMPRSS3 variants relative to domains of the protein and previously reported mutations. Novel mutations are written in bold.