Splice Site Tools

A Comparative Analysis Report

Beth Hellen

Contents

Introduction

Splicing is a process which modifies mRNA after transcription. It allows for introns to be removed and exons joined together to form mature mRNA, ready for translation into protein. The splice site junction, found where an intron meets an exon, contains multiple sequence motifs. These motifs provide signals to allow for correct splicing to occur. The best characterised of these are the acceptor and donor splice site signals. These signals consist of invariant dinucleotides at positions +1, +2, -1 and -2 of the intron and less well conserved nucleotides both within the immediate adjoining exonic sequence and deeper into the intron from the +3 and -3 positions (Seif *et al.,* 1979). The specific splicing of a gene can be easily affected by mutations in the sequence surrounding the splice site junction. This can lead to alternate splicing and thus adversely affect the translated protein (Novoyatleva *et al.*, 2006; Tazi *et al.*, 2009).

In-silico splice site prediction tools can be used to predict the effect of a genetic variant on splicing. A large number of prediction tools are currently available, either as standalone programs or as part of the Alamut (http://www.interactive-biosoftware.com/alamut/doc/1.5/index.html) or Human Splicing Finder (Desmet, 2009) interfaces. Some small analyses of these algorithms have been carried out, but no large scale analyses (Hartmann *et al.*, 2008; Holler *et al.*, 2009; Houdayer *et al.*, 2008). Although the UV guidelines (Bell *et al.*, 2007) provided by the CMGS (http://www.cmgs.org/) suggest several splice site prediction algorithms, the performance of these algorithms have not been formally assessed and may give divergent results. This analysis aims to provide an assessment of the performance of these algorithms in the prediction of splicing-related variant pathogenicity. It will also assess the scope of the splice-site prediction tools to ensure that they can be used in the most appropriate way. The analysis will allow scientists to use splice site prediction tools in the prediction of pathogenesis with more confidence.

In this analysis, six of the most common donor and acceptor prediction algorithms have been assessed for their ability to predict the pathogenicity of splice site variants. The algorithms chosen were those suggested by the UV guidelines, plus MaxEntScan, which are used as part of the Alamut and HSF splicing interfaces. The six algorithms were: GeneSplicer (Pertea *et al.*, 2001), Human Splicing Finder (HSF) (Desmet et al., 2009), MaxEntScan (Yeo & Burge, 2004), NetGene2 (Brunak *et al.*, 1991), NNSplice (Reese *et al.*, 1997) and SSFL, an algorithm based on Alex Dong Li's Splice Site Finder (no longer available). In each algorithm the splice signal given by the wild type sequence is compared to the splice site signal given by a mutated sequence supplied by the user.

Methods

Six algorithms were assessed for their ability to predict disruption to normal splicing patterns, caused by genetic variants. SSFL, MaxEntScan, NNSplice and GeneSplicer were accessed through the Alamut interface. HSF and a second implementation of MaxEntScan were accessed through the HSF interface. Netgene2 was implemented using a stand alone web interface. The majority of these methods were chosen because they had been recommended by the UV guidelines; MaxEntScan was included because it is used in both the HSF and Alamut splicing interfaces. A set of 265 pathogenic variants and 15 non-pathogenic variants from a total of 180 genes (see figure 1 and appendix 1) were retrieved from the literature. These variants were used to assess the splice site prediction algorithms using their default settings and recommended lengths of sequence. Sensitivity (equation 1), specificity (equation 2) and accuracy (equation 3) were calculated, as were the standard errors for each of the statistics. For the purposes of this analysis a true positive was defined as a pathogenic variant correctly classified as pathogenic and a true negative was a non-pathogenic variant correctly classified as non-pathogenic. A change in splice site signal of ≥10% was considered to predict a pathogenic effect.

$$\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}} \quad (1)$$

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}} \quad (2)$$

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{numbers of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}} \quad (3)$$

A second set of sensitivity, specificity and accuracy calculations were made for those variants which did not fall into the invariant di-nucleotide positions at -1, -2, +1, +2. The dataset consisted of 110 pathogenic variants and 15 non-pathogenic variants. The variants occurred in 83 different genes. This analysis will allow the algorithms to be assessed on their performance with the more difficult splice site variants.

The UV guidelines for splice site analysis recommend the use of three prediction algorithms to give a consensus prediction. Combinations of three high performing algorithms were compared to determine whether the accuracy was improved. The criteria required to categorise a variant as pathogenic or non-pathogenic was that at least two of the algorithms must agree on the prediction. The accuracy scores were calculated and compared to those given by the single algorithms.

To test the range of predictions made by the algorithms at each intronic position near the splice site junction, an in-silico analysis was performed. Thirteen acceptor and donor splice site junctions from BRCA1 and BRCA2 were analysed. Only junctions where the wild type splice site signal was found by all four of the highest performing algorithms were used. The wild type base at each position from +1 to +10 or -1 to -10 was artificially mutated in-silico to each of the remaining 3 nucleotides and the proportional change in splice site signal given by each algorithm was recorded. The mean change in splice site prediction (equation 4) at each position was plotted for each algorithm. The mean change in splice site signal strength is described in equation 4, where SS_M is the mutated splice site signal, SS_W is the wild type splice site signal and N is the number of examples analysed.

$$MeanChangeSS = \frac{\Sigma(SS_M/SS_W)}{N} \quad (4)$$

Results

Pathogenic and non-pathogenic splice site related variants retrieved from the literature were found at a range of positions relative to the splice site junction (Figure 1). The majority of splice site related pathogenic mutations used in this analysis were found within intronic positions between 1 and 10 nucleotides from the splice site junction. However, >40 of the variants were found in positions within the exon, and pathogenic mutations were also found at >100bp from the splice site junction. Only 15 non-pathogenic variants were found and they mainly occurred at positions further from the splice site junction. The small number of non-pathogenic variants arises from the problem of non-reporting of negative results. This is likely to increase the error associated with the specificity scores.

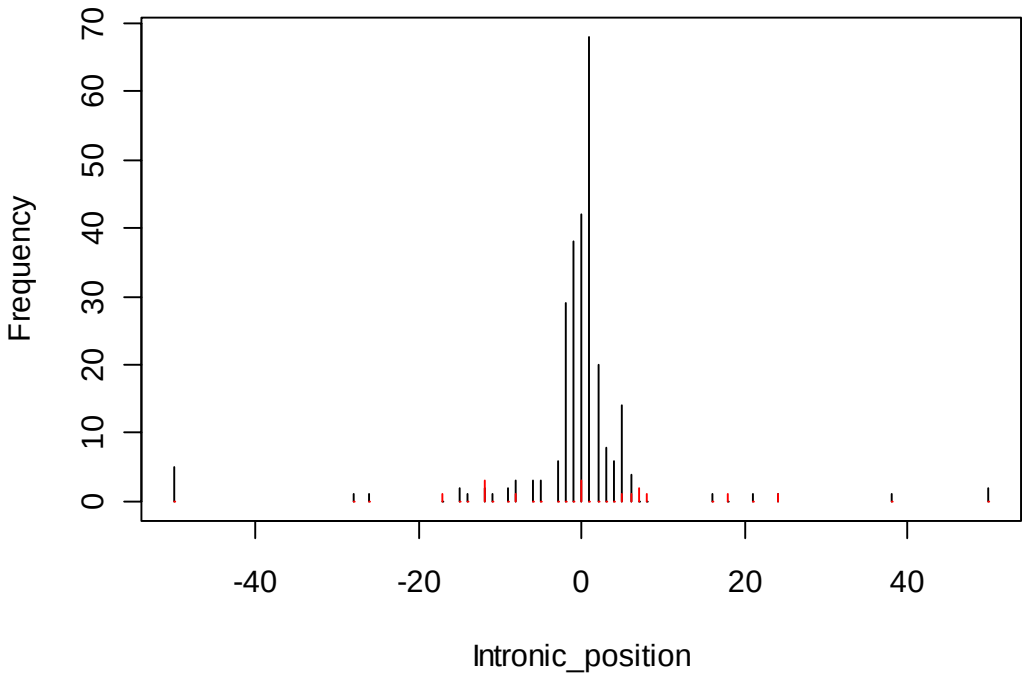


Figure 1 Chart showing the position of variants retrieved from the literature. Variants in exonic positions are shown at 0, variants >50bp from the splice site junction are binned and represented as a single frequency at 50bp from the splice site. Black lines represent the frequency of pathogenic variants and red lines represent the frequency of non-pathogenic variants.

The sensitivity, specificity and accuracy scores showed that the four highest performing algorithms were NNSplice, MaxEntScan, GeneSplicer and SSFL (Figure 2). These algorithms achieved between 80 and 92% accuracy and sensitivity. The specificity scores (between 73 and 93%) were less reliable due to the smaller number of variants tested. These four algorithms are those implemented through the Alamut interface. It is possible that the ease of interpretation of the results, when using the Alamut interface, has influenced this result. With the HSF interface it was more difficult to determine the predicted difference in splice site signal.

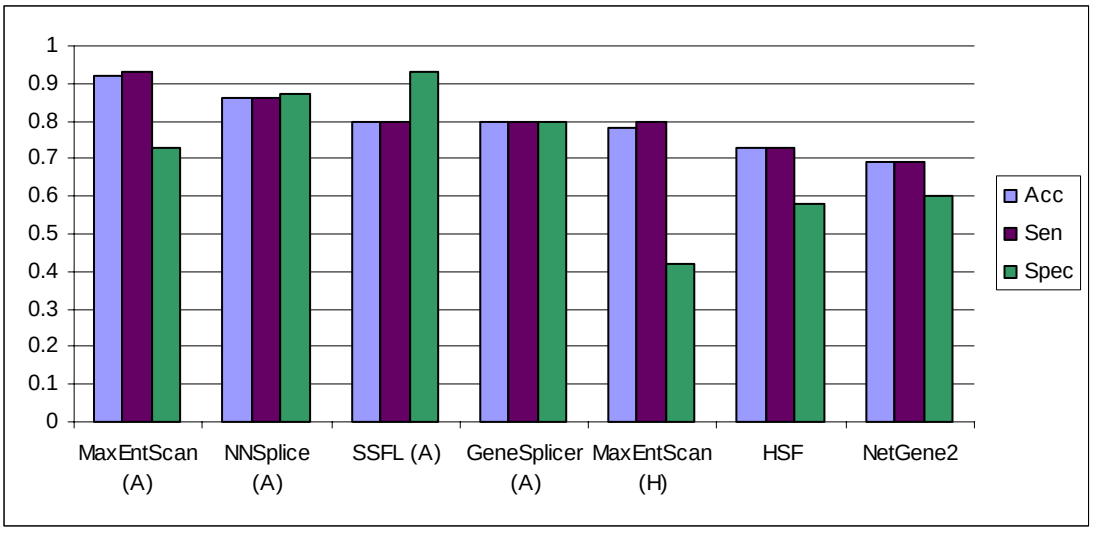


Figure 2 Accuracy, Sensitivity and Specificity values for each of the splice site prediction algorithms tested. Sensitivity measures the ability to predict pathogenic variants (TP) and specificity measures the ability to predict non-pathogenic variants (TN).

The removal of variants occurring at +1, +2, -1 and -2 positions reduced the performance of the algorithms, as was expected (Figure 3). However, two algorithms (MaxEntScan & NNSplice) still achieved an accuracy score of >80%. Therefore it can be seen that these algorithms perform reasonably well, even with variants where it is more difficult to predict the splicing effect.

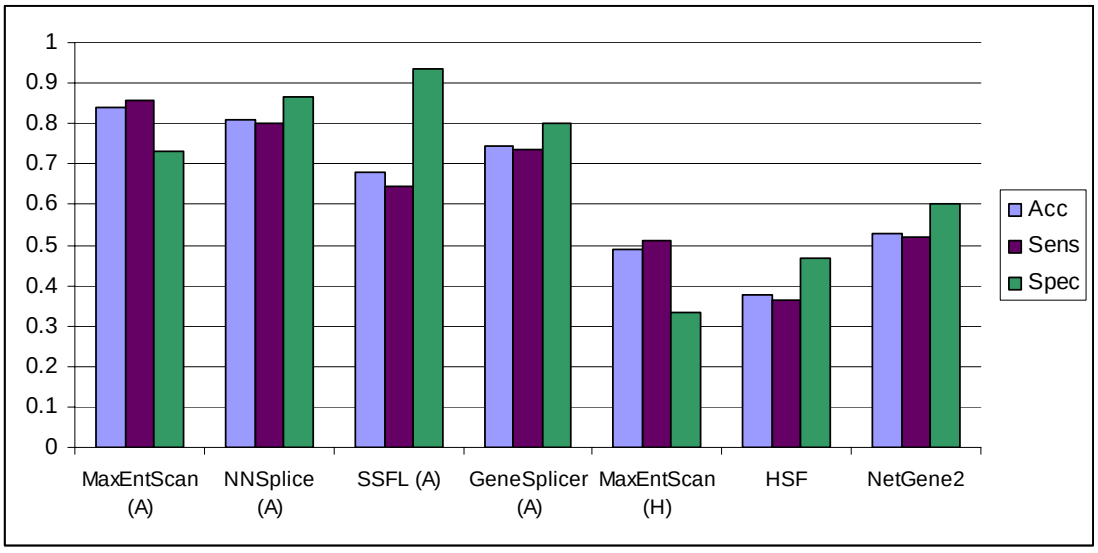
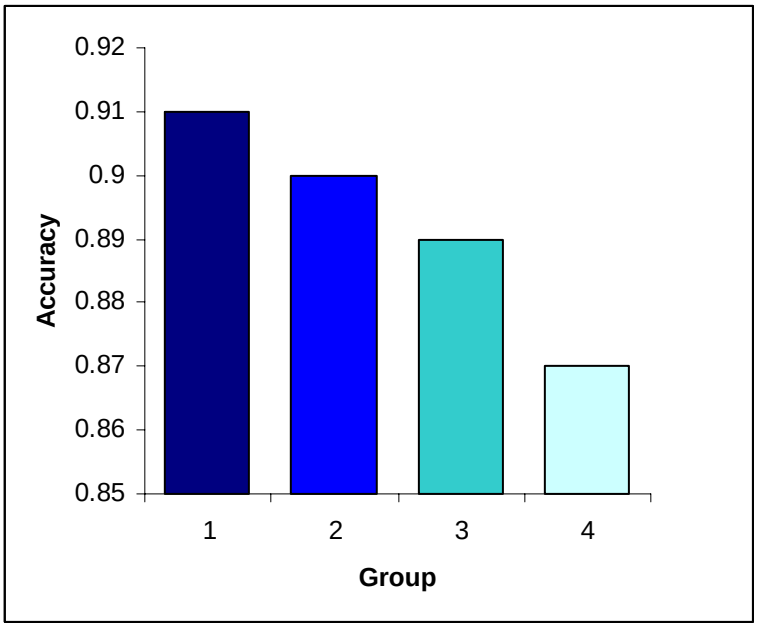


Figure 3 Accuracy, Sensitivity and Specificity values for each of the splice site prediction algorithms tested. Only variants which did not occur at one of the +1, +2,-1 or -2 positions were analysed.

The accuracy given by the consensus prediction of splice site signals was found to be between 86% and 92% for all combinations (Figure 4). The highest accuracy obtained through a consensus method was comparable to that given by MaxEntScan when implemented through Alamut. None of the consensus methods achieved an accuracy that was significantly higher than the individual algorithms.



	SSFL	MES	NNSPlice	Genesplicer
Group 1	X	X		X
Group 2	X	X	X	
Group 3		X	X	X
Group 4	X		X	X

Figure 4 The chart shows the accuracy obtained by combining results from three algorithms and using the consensus to predict pathogenicity of variants. The accompanying table describes the combinations of programs used in each consensus group.

Genetic variants which occur in the invariant dinucleotides at -1, -2, +1 and +2 were predicted to always disrupt splice site signalling (Figure 5). This would be assumed by most users and so no further information is gained by using the splice site prediction tools at these positions. The algorithms were shown to be the most useful for the prediction of both pathogenic and non-pathogenic splice site variants when applied to positions between +3 and +7 and -3 to at least -10 (Figure 5). At positions further from the splice site junction, no disruption in splice site signal was seen. The scope of these tools can therefore be defined as the prediction of the disruption of splice sites within these regions. The effect of variants on splice sites further than this cannot be predicted by any of the algorithms. The tools are, however, able to predict new splice sites at other positions. This could occur if the variant caused the sequence surrounding the new splice site to become a closer match to the statistical models used by the tools.

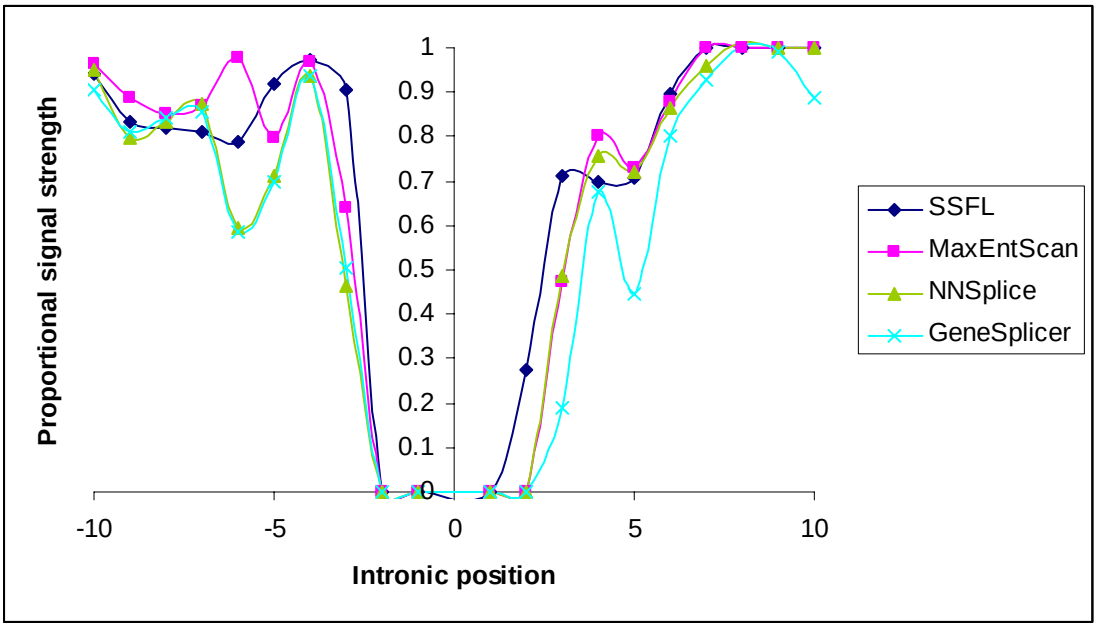


Figure 5 Graphs showing the proportional signal strength change on known splice sites when a mutation was introduced at positions in the intron between -1 and -10 or between +1 and +10. A score of 1 indicates that no disruption in the splice site signal was observed, a score of 0 indicates that the signal was completely destroyed. Lines between points have been added to ease interpretation although the data is discrete.

Conclusions

The four algorithms used in Alamut were shown to have a high degree of accuracy and users can be confident in the safe interpretation of these results as part of the assessment of a variant. It should still be noted that the algorithms alone are not sufficient evidence for a clinical decision. These algorithms, with the exception of SSFL, can be used as standalone web tools as well as via the Alamut interface. However, the results obtained through alternative implementations may differ, as shown by the MaxEntScan results obtained through Alamut and HSF.

The range of splice site signal strength predictions given by the algorithms is determined by the position of the variant. At +1, +2, -1 or -2 the algorithms always predict a large change in splice site signal, as would be predicted by experts. Variations in the wild type sequence further than +7 or -10 from the splice site junction do not cause any reduction in the wild type splice site signal predicted by the algorithms. Variants found between these two regions show a range of splice site reduction predicted by the algorithms and it is in this range that the algorithms are likely to be the most useful. This mirrors the reduction in occurrence of pathogenic variants found in the literature at these positions. The algorithms are still useful for prediction of splice site signals related to variants further into the intron, however it is only new splice sites which can be detected, not the reduction in wild type splice sites.

Although the use of three different algorithms is suggested in the UV guidelines, the accuracy was not improved by using a consensus method, therefore there does not seem to be a need for this step. However, as the Alamut interface performs all four analyses simultaneously, it is easy to compare predictions without a formal consensus method. The Alamut interface also contains methods to predict splicing enhancer or silencer motifs (ESE, ESS etc.) and branch point motifs. These methods have not been assessed and as the mechanisms by which these motifs regulate splicing are less clearly understood, the methods should be only be used with caution.

References

Bell, J., Bodmer, D., Sistermans, E., Ramsden, S. (2007) Practice guidelines for the interpretation and reporting of unclassified variants in clinical molecular genetics. Available: http://cmgs.org/BPGs/pdfs current bpgs/UV GUIDELINES ratified.pdf

Brunak, S., Engelbrecht, J., Knudsen, S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, 220:49-65.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., Béroud, C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.*, 37(9):e67.

Hartmann, L., Theiss, S., Niederacher, D., Schaal, H. (2008) Diagnosis of pathogenic splicing mutations: does bioinformatics cover all bases? *Front Biosci.*, 13:3252-72.

Holla, Ø. L., Nakken, S., Mattingsdal, M., Ranheim, T., Berge, K.E., Defesche, J.C., Leren, T.P. (2009) Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: Comparison of wet-lab and bioinformatics analyses. *Mol. Genet. Metab.,* 96(4):245-252.

Houdayer, C., Dehainault, C., Mattler, C., Michaux, D., Caux-Moncoutier, V., Pagès-Berhouet, S., d'Enghien, C.D., Laugé, A., Castera, L., Cauthier-Villars, M., Stoppa-Lyonnet, D. (2008) Evalutation of in silico splice tools for decision-making in molecular diagnosis. *Hum. Mutat.*, 29(7): 975-82.

Novoyatleva, T., Tang, Y., Rafalska, I., Stamm, S. (2006) Pre-mRNA missplicing as a cause of human disease. *Prog. Mol. Subcell. Biol.*, 44:27-46.

Pertea, M., Lin, X., Salzberg, S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, 29(5):1185-90.

Reese, M.G., Eeckman, F.H., Kulp, D., Haussler, D. (1997) Improved splice site detection in Genie. *J. Comp. Biol.,* 4(3):311-23.

Seif, I., Khoury, G., Dhar, R. (1979) BKV splice sequences based on analysis of preferred donor and acceptor sites. *Nucleic Acids Res.,* 6(10):3387-98.

Tazi, J., Bakkour, N., Stamm, S. (2009) Alternative splicing and disease. *Biochim. Biophys. Acta.*, 1792(1):14-26.

Yeo, G. and Burge, C.B. (2004) Maximum entropy modelling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.,* 11(2-3):377-394.

Appendix 1 Variants found in literature

Table 1 The number of pathogenic and non-pathogenic variants found for each gene in the literature search.

Gene	*# Pathogenic Variants*	*# Non-Pathogenic Variants*
AAAS	1	
ABCA1	1	
ABCA4	2	
ACADVL	1	
ACAT1	1	
ACOX1	1	
AIP	1	
AIRE	1	
ALDOB	1	
ALS2	2	
APC	1	
APOA5	1	
APOB	2	
ARSA	1	
ARSB	2	
ATM	3	
ATP2C1	3	
ATP7B	2	
BRCA1	11	3
BRCA2	17	
BTK	5	
CASR	1	
CDH23	1	
CERKL	1	
CETP	1	
CHM	1	
CHRNA1	1	
COG1	1	
COG7	1	
COL1A1	2	
COL4A3	1	
COL7A1	1	
COL8A2	0	2
CRYBA1	1	
CTSK	1	
CYBA	1	
CYBB	5	
CYP11A1	1	
DDC	1	
DFNA5	1	
DGUOK	1	
DMD	2	
DOK7	1	
DSPP	1	1
EDA	1	
EFNB1	1	
ERCC3	1	
ERCC8	1	
F11	2	
F13A1	1	
F5	2	
FAS	1	
FBN1	1	
FECH	2	
FGB	1	
FGFR1	2	

Gene	*# Pathogenic Variants*	*# Non-Pathogenic Variants*
KLK8	1	
KRIT1	1	
KRT1	1	
L1CAM	0	2
LDLR	4	2
LHB	1	
LMNA	2	
LPIN2	1	
MANBA	1	
MAPT	1	
MCOLN1	1	
MECP2	1	
MEN1	1	
MERTK	1	
MFSD8	2	
MIP	1	
MPV17	1	
MPZ	1	
MSH2	1	
MSX1	1	
MTM1	1	
MYBPC3	1	
MYO15A	2	
MYO7A	1	
NF1	1	
NPC1	1	
NR2E3	1	
OTC	1	
PAH	1	
PAK3	1	
PCCA	3	
PCCB	2	
PDHA1	1	
PHEX	3	
PHYH	1	
PITX2	2	
PKHD1	1	
PMM2	3	
PMS2	1	
POMGNT1	1	
POU1F1	1	
PPOX	1	
PROP1	1	
PRPF31	2	
PTEN	1	
PYGM	6	
RAPSN	1	
RB1	3	1
REEP1	1	
RHO	1	
RS1	3	
RSPO1	1	
SBDS	1	
SETX	1	
SLC12A3	1	
SLC25A20	2	

Table 1 Continued…

Gene	*# Pathogenic Variants*	*# Non-Pathogenic Variants*
FTSJ1	1	
GAMT	1	
GBA	2	
GBE1	1	
GDAP1	1	
GHR	1	
GHRHR	1	
GLB1	2	1
GLRX5	1	
GNPTAB	1	
GNPTG	1	
GNS	1	
GRN	2	2
HBB	1	
HEXB	3	
HMGCL	2	
IDS	11	
IGHMBP2	1	
IKBKAP	1	
ITPA	2	
IVD	1	
KCNH2	2	
KCNQ1	1	
KIF5A	1	

Gene	*# Pathogenic Variants*	*# Non-Pathogenic Variants*
SLC26A4	5	
SLC40A1	1	
SLC4A11	2	
SMARCB1	1	
SPAST	1	
SPG11	1	
SPINK1	1	
SPR	1	
STK11	1	1
TCIRG1	1	
TFR2	1	
TG	8	
TGM1	1	
TMC1	1	
TMEM67	1	
TNFRSF1A	2	
TRAPPC2	1	
TREM2	1	
UPB1	2	
VCAN	1	
VPS33B	1	
WT1	2	
XK	1	
ZMPSTE24	1	