ISSN: 2378-3648

Sokolov et al. J Genet Genome Res 2019, 6:044
DOI: 10.23937/2378-3648/1410044
Volume 6 | Issue 1
Open Access

Journal of
Genetics and Genome Research

RESEARCH ARTICLE

SVMOneClAS: Pipeline for Efficient Splicing Events Calling

Sokolov A[1*], Mazur A[1], Zhigalova N[1], Prokhortchouk A[1], Gruzdeva N[2] and Prokhortchouk E[1,2]

[1]*Center "Bioengineering", Russia, 117312, Moscow, Prospekt 60-Letiya Oktyabrya, 7/1., Russia*

[2]*National Research Center "Kurchatov Institute", Russia, 123182, Moscow, pl. Kurchatova, 1. Russia*

***Corresponding author:** *Alexey Sokolov, Center "Bioengineering", Russia, 117312, Moscow, Prospekt 60-Letiya Oktyabrya, 7/1., Russia, Tel: 79031332150*

Abstract

Splicing is a part of mRNA maturation process when exons of pre-mRNA transcript are joined in multiple ways. In case of alternative splicing, some exons may be excluded from the final mRNA, and the resulting ensemble of mRNAs creates different protein isoforms, allowing multiple proteins to be coded by a single gene. To detect new events of alternative splicing in human blood cells, we used RNA-seq technology. A pipeline based on Python "Scikit-Learn" library was developed for efficient splicing events calling. We found 8728 potential candidates, 20 of which were selected for RT-PCR and 8 were finally confirmed as novel events in 7 genes (MPPE1, CTDP1, ENOSF1, SEH1L, TXNL4A, C18orf1, and ME2). SVMOneClAS are freely available from the https://github.com/alsokolov-dev/SVMOneClAS

Keywords

Alternative splicing, Machine learning, Support vector machine, RNA-seq

Introduction

Alternative splicing occurs generally in eukaryotes and is considered to be a key factor in protein functional complexity and diversity [1]. It has been observed that ~95% of the multiexon genes in humans are subject to alternative splicing events [2]. Among different modes of alternative splicing, exon skipping appears to be the most common mode characterized by exon inclusion or exclusion depending on mRNA production conditions [1]. At the same time alternative 5' or 3' splice sites and intron retention modes are demonstrated in cancer cells [3]. Several protein products can result from one gene by the use of alternative mRNA splicing, and some of these splice isoforms can lose or gain a function. Therefore, apart from the protein diversity, alternative splicing is a cause of abnormal splicing variants resulting in functional consequences. This way, aberrant mRNA transcripts are implicated in a large number of human genetic disorders. Furthermore, hundreds of genes are found to be abnormally spliced in cancer cells, which is functionally important for cancer development [4-6].

Recent discoveries and technological advances have given an opportunity to deeply investigate alternative splicing events providing new insight into physiological and biochemical network in different systems. Such studies not only reveal mechanisms of cellular processes regulation but also clear the way to the development of new therapeutic strategies to disorders treatment. Pathogenesis of a number of neurodegenerative, cancer, blood, inflammatory, and hemolytic disorders was found to be associated with alternative splicing [7,8]. For example, study of stress responses mechanisms in brain and blood described stress-induced changes associated with alternative splicing patterns of acetylcholine pre-mRNA and functional acetylcholine properties [9]. Another research in the field of thrombosis and hemostasis discovered genes producing alternatively spliced protein isoforms that can break some coagulation cascades and affect blood clotting [10].

Current studies of alternative splicing and its mechanisms showed an important role of such events in the cellular processes; however, characterization of new alternative splicing events as well as their tissue specificity is still in high demand. Here a combination of PCR, Sanger sequencing, and RNA-seq data analysis was successfully applied to find new events of alternative splicing.

Citation: Sokolov A, Mazur A, Zhigalova N, Prokhortchouk A, Gruzdeva N, et al. (2019) SVMOneClAS: Pipeline for Efficient Splicing Events Calling. J Genet Genome Res 5:044. doi.org/10.23937/2378-3648/1410044

Accepted: June 03, 2019: **Published:** June 05, 2019

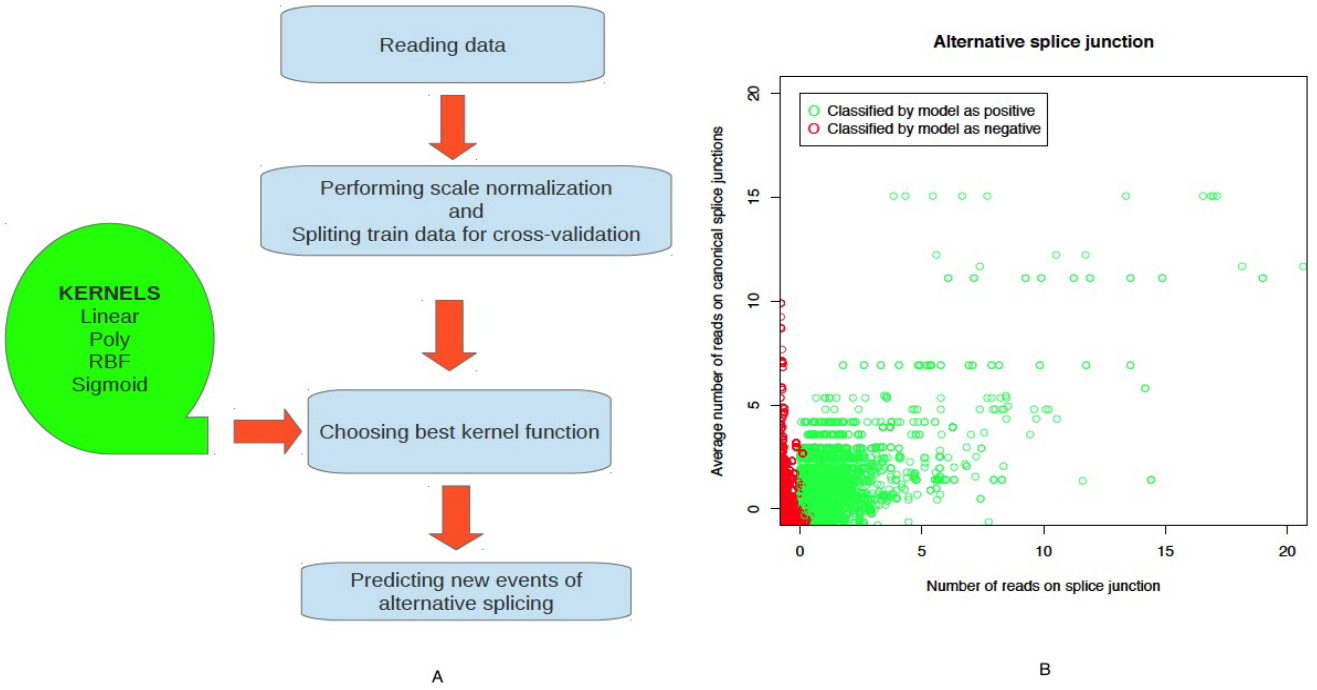


Figure 1: A) Pipeline outline; B) New events of alternative splicing.

Implementation

In order to identify new events of alternative splicing, we used a «One Class Support Vector Machine» algorithm [11-15]. This algorithm is often used for novelty detection, when one is able to decide whether a new observation (in our case a new alternative splicing event) belongs to the same distribution compared to existing observations (canonical events of alternative splicing).

The pipeline workflow was as follows: A data set of n canonical splicing events described by p features was considered; new observations (new events of alternative splicing) to that data set were added. Then the question was: do these events come from the same distribution or more precisely are these events so similar to the canonical that we cannot distinguish it? This self-taught algorithm identified a rough, close frontier, delimiting the contour of the canonical splicing events distribution, plotted in embedding p-dimensional space. Then, if new events of alternative splicing lay within this delimited subspace, they are considered as potential new events of alternative splicing. On the other hand, if they lay outside, we can say that they are abnormal with a given confidence in our assessment.

The pipeline outline is shown on Figure 1A. 4 types of kernels were used to best fit the data («linear», «poly», «rbf» and «sigmoid»). Data set was divided on train and test sets for purposes of cross-validation procedure. The training set is used to train the classifier with two parameters: Gamma (kernel coefficient for “rbf” and “poly” types of kernels functions) and degree (degree of kernel function). The test set is used to evaluate the accuracy of each classifier and then kernel function and parameters gamma and degree that gave smallest value of error on cross validation step was used for subsequent analysis.

Results

Whole blood RNA isolated from an individual N, whose genome was sequenced previously [16,17], was used to prepare a library according Illumina protocols. Illumina GAIIx RNA-seq reads generated were mapped to reference human genome and splice junctions (Table S1, Table S2, Figure 2 and Figure S1) (we used RefSeq database of known exons to construct all possible exon junctions) of human genome using Bowtie2 [18,19].

We used 4 parameters for each splicing event: number of reads at this splice junction, average number of reads at all canonical splice junctions of this transcript, gene length, and a number of exons. Compared with papers that use Support Vector Machines (SVM) to find new events of alternative splicing [20-23], we didn’t use sequence information, which makes our algorithm more appropriate for searching for different alternative splicing events between various tissues or conditions of the same organism. We found 8728 (28% of all possible events) potential candidates (Figure 1B), 20 of which were selected for subsequent analysis Table S3 (none of those events were found by cufflinks).

To confirm new events of alternative splicing we performed RT-PCR with specific primers, PCR products were fractionated in agarose gel, amplicons of expected size were eluted and sequenced (Sanger).

Only 15 events out of 20 tested produced fragments of expected size and 11 were proved by sequencing to be alternative splicing (Figure S2, Figure S3, Figure S4,

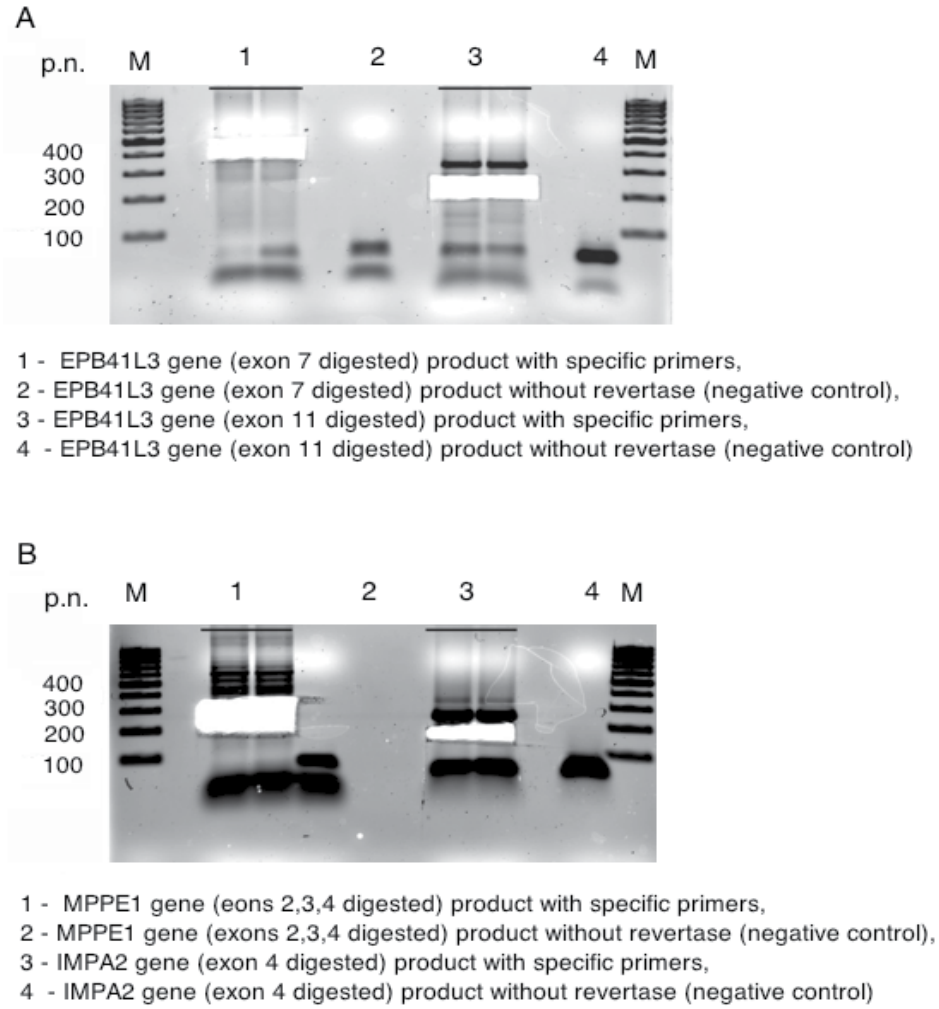


Figure 2: Finding and confirming new events of alternative splicing.
A, B) Electrophoresis of PCR products for genes EPB41L3 and MPPE1, IMPA2 resp.

Figure S5, Figure S6, and Figure S7). Next, we translated DNA sequences of the amplicons and ran corresponding ORF against UniProt database. This procedure filtered out additional 3 events. The rest 8 events (referred as "8 events") were localized in 7 genes, i.e. MPPE1, CTDP1, ENOSF1, SEH1L, TXNL4A, C18orf1, ME2 and most of them are already described in HAVANA database. Then we isolated RNA from whole blood of 10 healthy individuals of Slavic origin and performed RT-PCR with specific for "8 events" primer sets. Among 10 tested patients, 8 carried full set of "8 events" and the rest 2 shared 6 out of "8 events" (Figure S7 and Figure S8).

All new events represented an exon skipping mode of alternative splicing. When exon skipped, protein may miss some specific sites or domains after translation. In our case proteins may miss metal-binding sites (as for genes MPPE1, ENOSF1, ME2), active sites (as for genes ENOSF1, ME2), topological domain (as for gene C18orf1), or some particular sequences, for example, repeat 55-96 (as for SEH1L gene).

Acknowledgements

The authors thank K.G. Skryabin, P. V. Mazin and M.V. Kovalchuk for the support and careful attention to the project. This study was supported by the Ministry of Education and Science of Russia (project № 14.740.11.0759) and by the Scientific Research Agreement №248 with the Institute of Biomedical Chemistry RAMS (a part of the Human Proteome Project, Russia).

References

1. Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72: 291-336.
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40: 1413.
3. Kim E, Goren A, Ast G (2008) Insights into the connection between cancer and alternative splicing. Trends Genet 24: 7-10.
4. Fackenthal JD, Godley LA (2008) Aberrant RNA splicing and its functional consequences in cancer cells. Dis Model Mech 1: 37-42.
5. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6: 386-398.
6. Skotheim RI, Nees M (2007) Alternative splicing in cancer:

noise, functional, or systematic? Int J Biochem Cell Biol 39: 1432-1449.

7. Benz EJ Jr., Huang SC (1997) Role of tissue specific alternative pre-mRNA splicing in the differentiation of the erythrocyte membrane. Trans Am Clin Climatol Assoc 108: 78-95.
8. Ladd AN (2013) CUG-BP, Elav-like family (CELF)-mediated alternative splicing regulation in the brain during health and disease. Mol Cell Neurosci 56: 456-464.
9. Pick M, Flores-Flores C, Soreq H (2004) From brain to blood: alternative splicing evidence for the cholinergic basis of Mammalian stress responses. Ann N Y Acad Sci 1018: 85-98.
10. Bogdanov VY (2006) Blood coagulation and alternative pre-mRNA splicing: an overview. Curr Mol Med 6: 859-869.
11. Lohoff FW, Ferraro TN, Brodkin ES, Weller AE, Bloch PJ (2010) Association between polymorphisms in the metallophosphoesterase (MPPE1) gene and bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 153B: 830-836.
12. Greenberg DA, Cayanis E, Strug L, Marathe S, Durner M, et al. (2005) Malic enzyme 2 may underlie susceptibility to adolescent-onset idiopathic generalized epilepsy. Am J Hum Genet 76: 139-146.
13. Kalaydjieva L (2006) Congenital cataracts-facial dysmorphism-neuropathy. Orphanet J Rare Dis 1: 32.
14. Zhang Y, Lindblom T, Chang A, Sudol M, Sluder AE, et al. (2000) Evidence that dim1 associates with proteins involved in pre-mRNA splicing, and delineation of residues essential for dim1 interactions with hnRNP F and Npw38/PQBP-1. Gene 257: 33-43.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: Machine Learning in Python. JMLR 12: 2825-2830.
16. Chekanov NN, Boulygina ES, Beletskiy AV, Prokhortchouk EB, Skryabin KG, et al. (2010) Individual Genome of the Russian Male: SNP Calling and a de novo Assembly of Unmapped Reads. Acta Naturae 2: 122-126.
17. Skryabin KG, Prokhortchouk EB, Mazur AM, Boulygina ES, Tsygankova SV, et al. (2009) Combining two technologies for full genome sequencing of human. Acta Naturae 1: 102-107.
18. Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.
19. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: 25.
20. De Bona F, Ossowski S, Schneeberger K, Rätsch G (2008) Optimal spliced alignments of short sequence reads. BMC Bioinformatics 24: 174-180.
21. Jean G, Kahles A, Sreedharan V, Bona F, Rätsch G (2010) RNA-Seq Read Alignments with PALMapper, Curr Protocols Bioinformatics.
22. Rätsch G, Sonnenburg S (2004) Accurate splice site detection for Caenorhabditis elegans. In: Schölkopf B, Tsuda K, Vert JP, Kernel Methods in Computational Biology. Edited by MIT Press.
23. Rätsch G, Sonnenburg S, SchÄolkopf B (2005) RASE: recognition of alternatively spliced exons in C. elegans, Bioinformatics.

Table S1: Reads mapped on genome.

Unmapped, reads	Mapped, reads	Unique, reads	Non-unique, reads	Exons
33,796,715	25,139,655	22,967,443	2,172,212	15,546,383

Table S2: Reads mapped on splice junctions.

Unmapped, reads	Mapped, reads	Unique, reads	Non-unique, reads
53,557,143	5,379,227	2,924,752	2,454,475

Table S3: None of those events were found by cufflinks.

Event number	Gene symbol	Exons in junction	Skipped exons	Gel electroforesis	Sanger sequencing	Confirming in SwissProt database	Functional sites in skipped exons
1	EPB41L3	6,8	7	+	+	+	spectrin-actin binding domain
2	EPB41L3	10,12	11	+	+	+	spectrin-actin binding domain
3	MPPE1	1,5	2,3,4	+	+	-	3 metal-binding sites
4	IMPA2	3,5	4	+	-	-	no
5	ATP5A1	8,1	9	+	-	-	no
6	CTDP1	10,12	11	+	+	-	no
7	TYMS	1,4	2,3	-	-	-	no
8	ENOSF1	3,6 or 2,5	4,5 or 3,4	+	-	-	no
9	ENOSF1	3,7 or 2,5 or 2,6	4,5,6 or 3,4 or 3,4,5	+	+	-	active and metal-binding sites
10	ENOSF1	4,6 or 3,5	5,4	+	+	-	no
11	NDC80	6,8	7	-	-	-	no
12	KIAA0802	7,9	8	+	+	+	no
13	SEH1L	2,4	3	+	+	-	repeat 55-96
14	ATP5A1	4,6	5	-	-	-	
15	ATP5A1	10,12	11	+	-	-	no
16	TXNL4A	1,3	2	+	+	-	no
17	TYMS	1,3	2	-	-	-	no
18	ENOSF1	8,11	9,10	-	-	-	no
19	C18orf1	3,5	4	+	+	-	topological domain
20	ME2	13,15	14	+	+	-	active and metal-binding sites

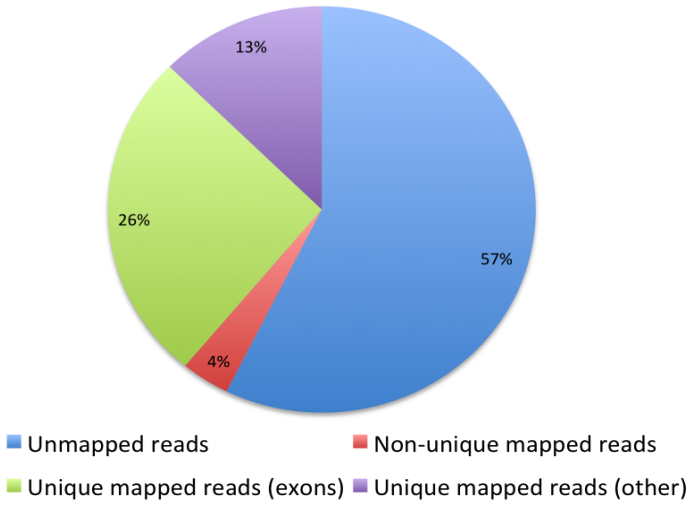


Figure S1: Results of mapping reads on genome.

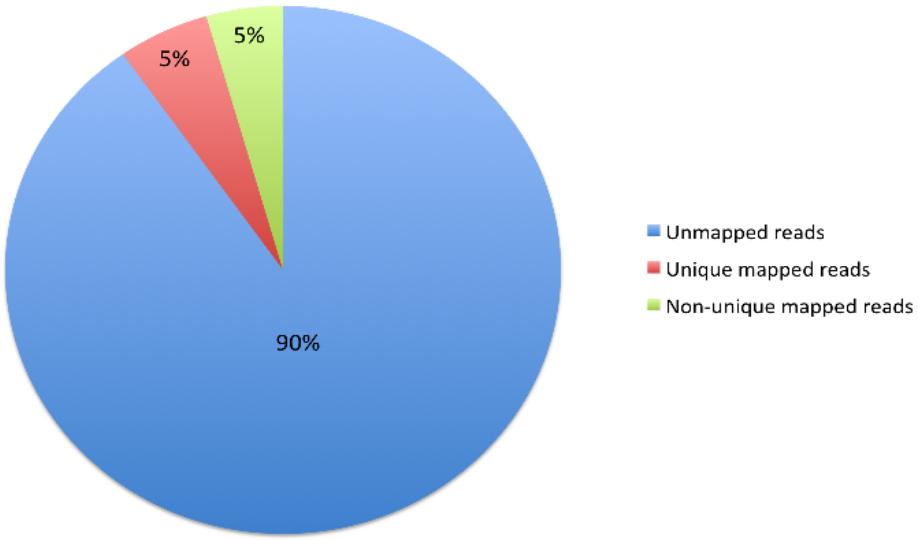
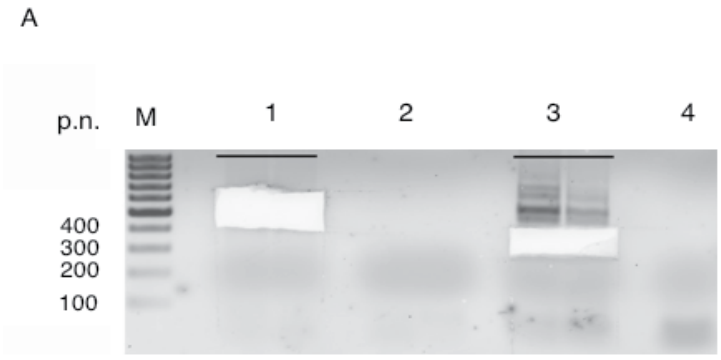
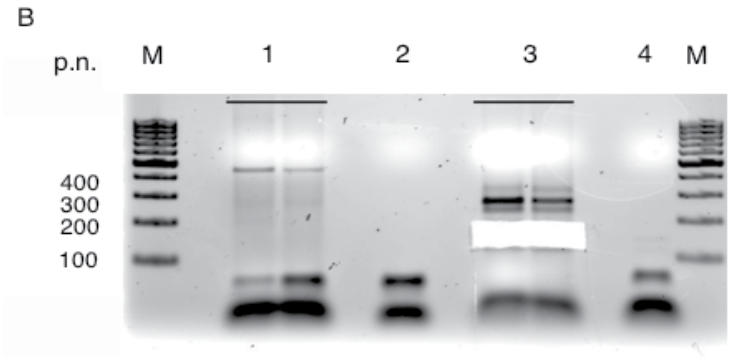


Figure S2: Results of mapping reads on splice junction.



1 - ATP5A1 gene (exon 5 digested) product with specific primers,
2 - ATP5A1 gene (exon 5 digested) product without revertase (negative control),
3 - CTDP1 gene (exon 11 digested) product with specific primers,
4 - CTDP1 gene (exon 11 digested) product without revertase (negative control)



1 - TYMS gene (exons 2,3 digested) product with specific primers,
2 - TYMS gene (exons 2,3 digested) product without revertase (negative control),
3 - ENOSF1 gene (exon 5 digested) product with specific primers,
4 - ENOSF1 gene (exon 5 digested) product without revertase (negative control)

Figure S3: A, B Electrophoresis of PCR products for genes ATP5A1, CTDP, TYMS and ENOSF1 resp.

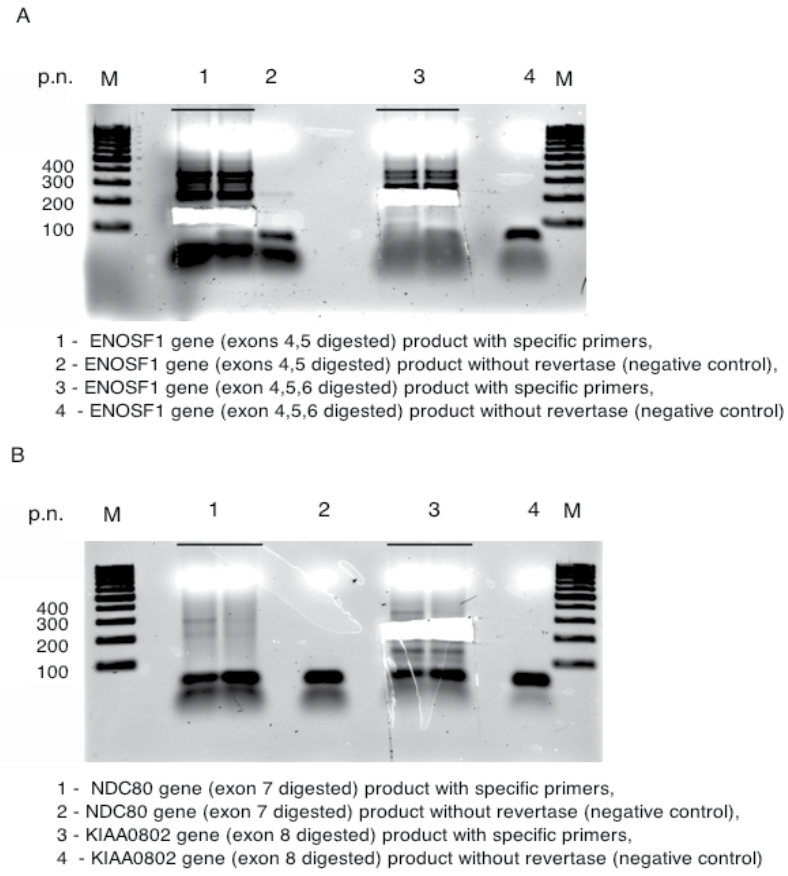


Figure S4: A, B Electrophoresis of PCR products for genes ENOSF1, NDC80, and KIAA0802 resp.

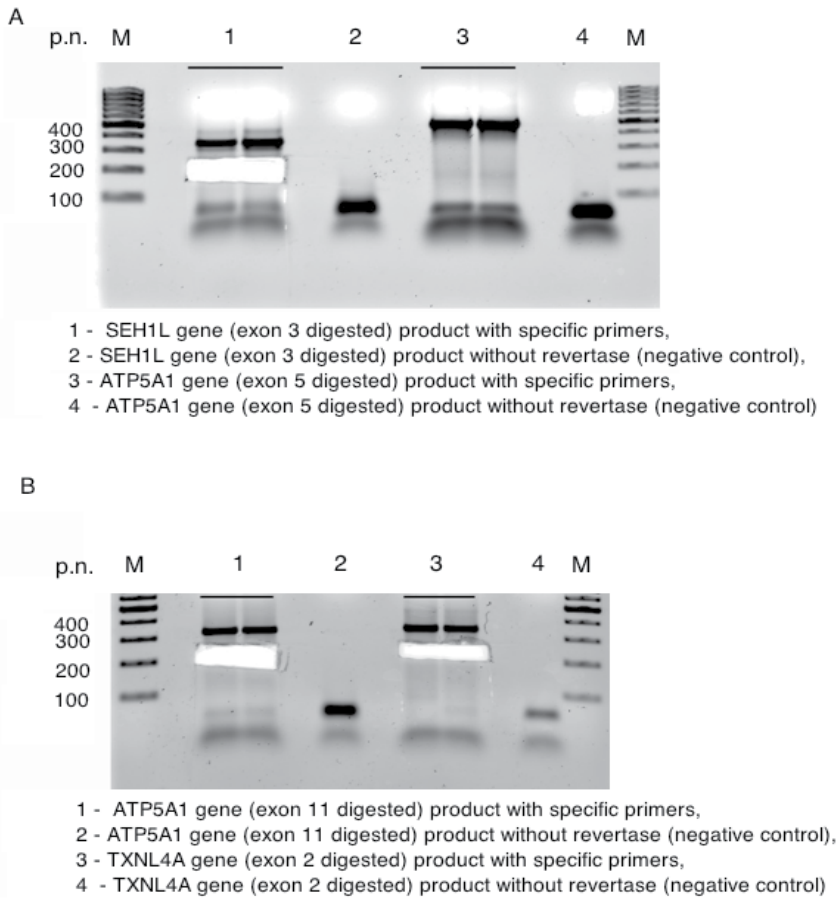


Figure S5: A, B Electrophoresis of PCR products for genes SEH1L, ATP5A1, TXNL4A resp.

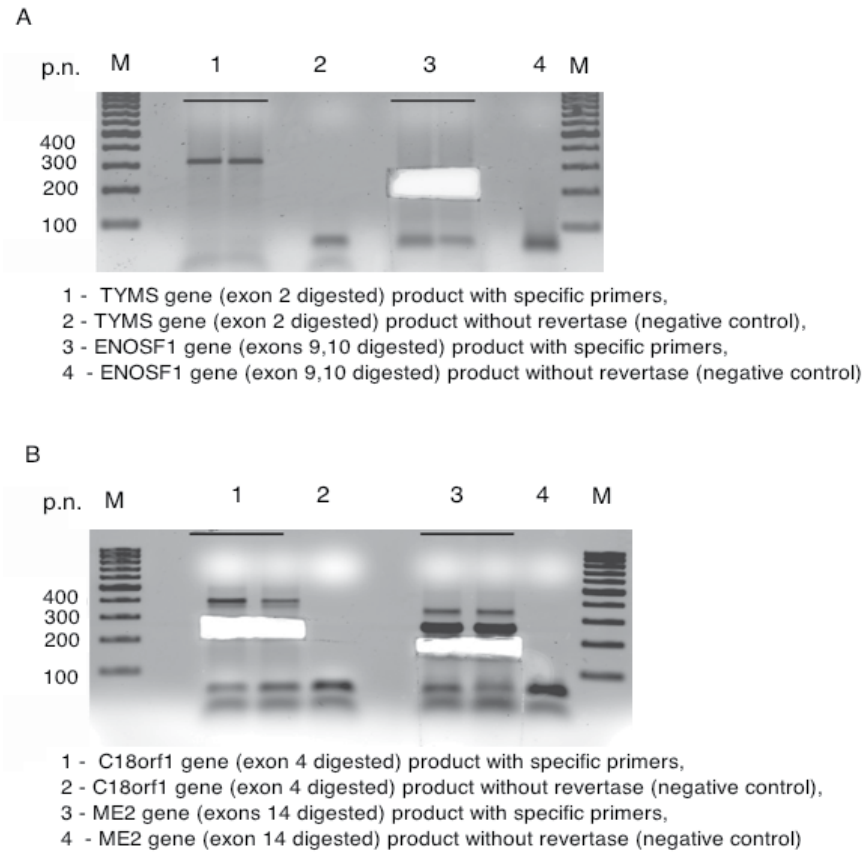


Figure S6: A, B Electrophoresis of PCR products for genes TYMS, ENOSF1, C18orf1 and ME2 resp.

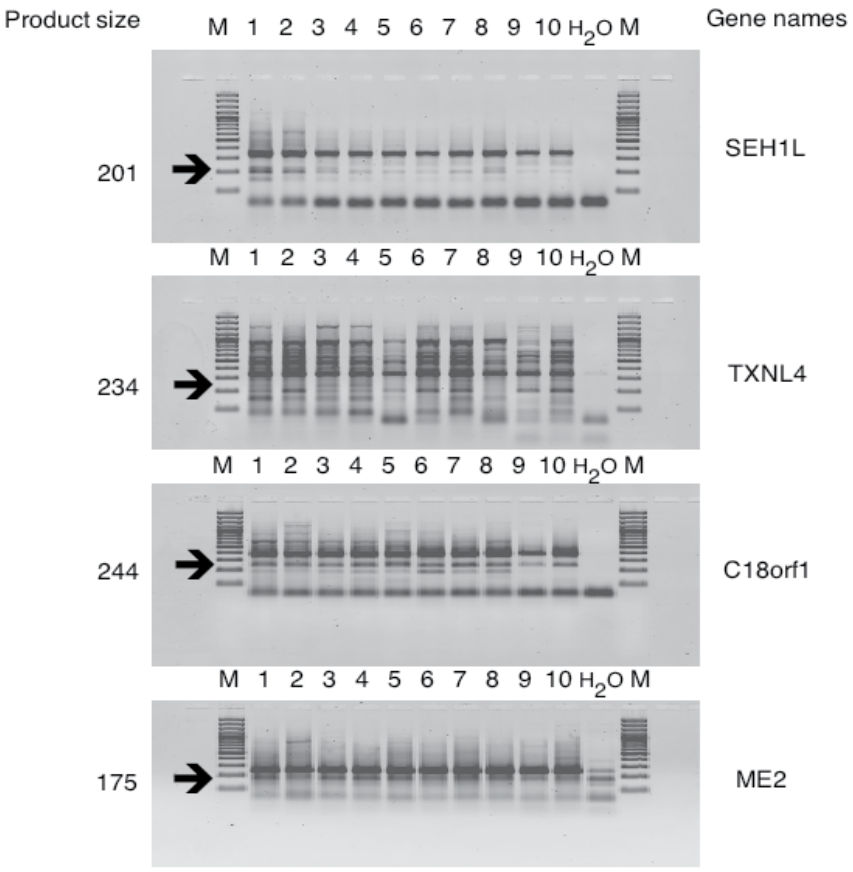


Figure S7: Confirming of new splice events by RT-PCR with specific primers on 10 healthy blood donors 1,2,3,4,5,6,7,8,9,10 - samples numbers; M - Marker; H_2O - Negative controls; arrows indicate specific product size.

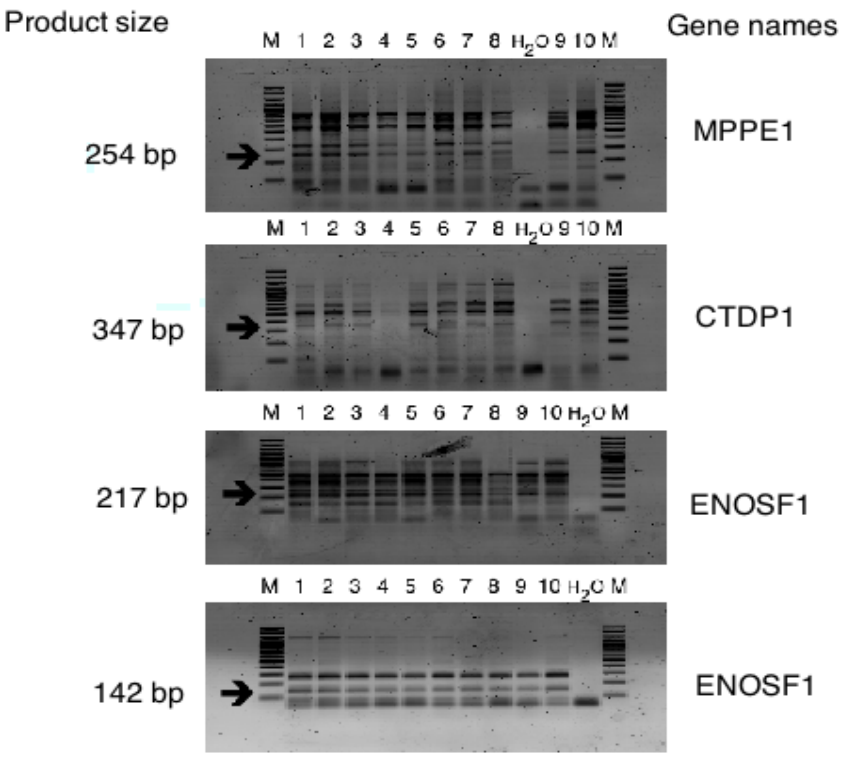


Figure S8: Confirming of new splice events by RT-PCR with specific primers on 10 healthy blood donors 1,2,3,4,5,6,7,8,9,10 - samples numbers; M - Marker; H_2O - Negative controls; arrows indicate specific product size.