Supplementary Material

SeqTailor: a user-friendly webserver for the extraction of DNA or protein sequences from next-generation sequencing data

Peng Zhang, Bertrand Boisson, Peter D. Stenson, David N. Cooper, Jean-Laurent Casanova, Laurent Abel, and Yuval Itan

Table of Contents

	-
Case Study	9
Case Study 1: DNA sequence extraction for a variant in MSH2	9
Case Study 2: DNA sequence extraction for a variant in BRCA21	1
Case Study 3: DNA sequence extraction for a variant in IL2RG1	3
Case Study 4: Protein sequence extraction for a variant in BRAF1	5
Case Study 5: Protein sequence extraction for a variant in GJB21	8
References	0

Supplementary Tables and Figures

CUD UCCUDI 1 CTCIII	CUP USCUP12 1 CTC4	
CHR_HSCHRI_I_CIGII	CHR_HSCHR13_1_C1G4	CHR_HSCHR18_2_C1G2
CHR_HSCHR1_1_C1G3	CHR_HSCHR13_1_C1G5	CHR_HSCHR18_2_C1G2_1
CHR_HSCHR1_1_CTG31	CHR_HSCHR13_1_CTG6	CHR_HSCHR18_3_CTG2_1
CHR_HSCHR1_1_CTG32_1	CHR_HSCHR13_1_CTG7	CHR_HSCHR18_4_CTG1_1
CHR_HSCHR1_2_CTG3	CHR_HSCHR13_1_CTG8	CHR_HSCHR18_5_CTG1_1
CHR_HSCHR1_2_CTG31	CHR_HSCHR14_1_CTG1	CHR_HSCHR18_ALT2_CTG2_1
CHR_HSCHR1_2_CTG32_1	CHR_HSCHR14_2_CTG1	CHR_HSCHR18_ALT21_CTG2_1
CHR_HSCHR1_3_CTG3	CHR_HSCHR14_3_CTG1	CHR_HSCHR19_1_CTG2
CHR HSCHR1 3 CTG31	CHR HSCHR14 7 CTG1	CHR HSCHR19 1 CTG3 1
CHR HSCHR1 3 CTG32 1	CHR HSCHR14 8 CTG1	CHR HSCHR19 2 CTG2
CHR HSCHR1 4 CTG3	CHR HSCHR15 1 CTG1	CHR HSCHR19 2 CTG3 1
CHR HSCHR1 4 CTG31	CHR HSCHR15 1 CTG3	CHR HSCHR19 3 CTG2
CHR HSCHR1 4 CTG32 1	CHR HSCHR15 1 CTG8	CHR HSCHR19 3 CTG3 1
CHR HSCHR1 5 CTG3	CHR_HSCHR15_2_CTG3	CHR HSCHR19 4 CTG2
CHR_HSCHR1_5_CTG32_1	CHR_HSCHR15_2_CTG8	CHR HSCHR19 4 CTG3 1
CHP_HSCHP1_6_CTG3	CHP_HSCHP15_3_CTG3	CHP_HSCHP10_5_CTG2
CHP_HSCHP1_8_CTG3	CHP_HSCHP15_3_CTG9	CHR_HSCHR10/J_010 5217 AP_CTG2_1
		CHR_HSCHRI9KIK_0010-521/-AB_CTO5_1
CHR_HSCHRI_9_CIG3	CHR_HSCHR15_4_CIG8	CHR_H5CHR19KIR_0019-4050-A_C1G5_1
CHR_HSCHRI_ALI2_I_CIG32_I	CHR_HSCHR15_5_CIG8	CHR_HSCHR19KIR_0019-4656-B_C1G3_1
CHR_HSCHRI0_I_CIGI	CHR_HSCHR15_6_CIG8	CHR_HSCHR19KIR_502960008-1_CTG3_1
CHR_HSCHR10_1_C1G2	CHR_HSCHR16_1_CIG1	CHR_HSCHR19KIR_502960008-2_C1G3_1
CHR_HSCHR10_1_CTG3	CHR_HSCHR16_1_CTG3_1	CHR_HSCHR19KIR_7191059-1_CTG3_1
CHR_HSCHR10_1_CTG4	CHR_HSCHR16_2_CTG3_1	CHR_HSCHR19KIR_7191059-2_CTG3_1
CHR_HSCHR10_1_CTG6	CHR_HSCHR16_3_CTG1	CHR_HSCHR19KIR_ABC08_A1_HAP_CTG3_1
CHR_HSCHR11_1_CTG1_1	CHR_HSCHR16_3_CTG3_1	CHR_HSCHR19KIR_ABC08_AB_HAP_C_P_CTG3_1
CHR_HSCHR11_1_CTG1_2	CHR_HSCHR16_4_CTG1	CHR_HSCHR19KIR_ABC08_AB_HAP_T_P_CTG3_1
CHR_HSCHR11_1_CTG2	CHR_HSCHR16_4_CTG3_1	CHR_HSCHR19KIR_CA01-TA01_1_CTG3_1
CHR_HSCHR11_1_CTG3	CHR_HSCHR16_5_CTG1	CHR_HSCHR19KIR_CA01-TA01_2_CTG3_1
CHR_HSCHR11_1_CTG3_1	CHR_HSCHR16_5_CTG3_1	CHR_HSCHR19KIR_CA01-TB01_CTG3_1
CHR_HSCHR11_1_CTG5	CHR_HSCHR16_CTG2	CHR_HSCHR19KIR_CA01-TB04_CTG3_1
CHR_HSCHR11_1_CTG6	CHR_HSCHR17_1_CTG1	CHR_HSCHR19KIR_CA04_CTG3_1
CHR_HSCHR11_1_CTG7	CHR_HSCHR17_1_CTG2	CHR_HSCHR19KIR_FH05_A_HAP_CTG3_1
CHR_HSCHR11_1_CTG8	CHR_HSCHR17_1_CTG4	CHR_HSCHR19KIR_FH05_B_HAP_CTG3_1
CHR_HSCHR11_2_CTG1	CHR_HSCHR17_1_CTG5	CHR_HSCHR19KIR_FH06_A_HAP_CTG3_1
CHR HSCHR11 2 CTG1 1	CHR HSCHR17 1 CTG9	CHR HSCHR19KIR FH06 BA1 HAP CTG3 1
CHR HSCHR11 2 CTG8	CHR HSCHR17 10 CTG4	CHR HSCHR19KIR FH08 A HAP CTG3 1
CHR HSCHR11 3 CTG1	CHR HSCHR17 11 CTG4	CHR HSCHR19KIR FH08 BAX HAP CTG3 1
CHR HSCHR12 1 CTG1	CHR HSCHR17 12 CTG4	CHR HSCHR19KIR FH13 A HAP CTG3 1
CHR HSCHR12 1 CTG2	CHR_HSCHR17_2_CTG1	CHR HSCHR19KIR FH13 BA2 HAP CTG3 1
CHR_HSCHR12_1_CTG2_1	CHR_HSCHR17_2_CTG2	CHR HSCHR19KIR FH15 A HAP CTG3 1
CHR_HSCHR12_2_CTG1	CHR_HSCHR17_2_CTG4	CHR HSCHRIOKIR FHIS B HAP CTG3 1
CHR_HSCHR12_2_CTG2	CHR_HSCHR17_2_CTG5	CHR_HSCHR19KIR_G085_A_HAP_CTG3_1
CHR_HSCHR12_2_CTG2_1	CHR_HSCHR17_2_CTG1	CHR_HSCHRI9KIK_0085_A_HAI_CTG5_1
		CHR_HSCHRI9KIK_0085_BAI_HAP_CT05_1
CHR_HSCHR12_3_CTG2_1	CHR_HSCHR17_3_CTG2	CHR_HSCHRI9KIR_C248_A_HAP_C105_1
CHR_HSCHR12_5_C1G2_1	CHR_HSCHR17_5_CTG4	CHR_HSCHRI9KIR_G248_BA2_HAP_CIG5_1
CHR_HSCHR12_4_CTG2	CHR_HSCHR1/_4_CIG4	CHR_HSCHRI9KIR_GRC212_AB_HAP_CIG3_1
CHR_HSCHR12_4_C1G2_1	CHR_HSCHR1/_5_CIG4	CHR_HSCHR19KIR_GRC212_BA1_HAP_C1G3_1
CHR_HSCHR12_5_C1G2	CHR_HSCHR17_6_CTG4	CHR_HSCHR19KIR_HG2393_C1G3_1
CHR_HSCHR12_5_CTG2_1	CHR_HSCHR17_7_CTG4	CHR_HSCHR19KIR_HG2394_CTG3_1
CHR_HSCHR12_6_CTG2_1	CHR_HSCHR17_8_CTG4	CHR_HSCHR19KIR_HG2396_CTG3_1
CHR_HSCHR12_7_CTG2_1	CHR_HSCHR17_9_CTG4	CHR_HSCHR19KIR_LUCE_A_HAP_CTG3_1
CHR_HSCHR12_8_CTG2_1	CHR_HSCHR18_1_CTG1	CHR_HSCHR19KIR_LUCE_BDEL_HAP_CTG3_1
CHR_HSCHR12_9_CTG2_1	CHR_HSCHR18_1_CTG1_1	CHR_HSCHR19KIR_RP5_B_HAP_CTG3_1
CHR_HSCHR13_1_CTG1	CHR_HSCHR18_1_CTG2	CHR_HSCHR19KIR_RSH_A_HAP_CTG3_1
CHR_HSCHR13_1_CTG2	CHR_HSCHR18_1_CTG2_1	CHR_HSCHR19KIR_RSH_BA2_HAP_CTG3_1

CHR_HSCHR19KIR_T7526_BDEL_HAP_CTG3_1	CHR_HSCHR3_1_CTG3	CHR_HSCHR6_1_CTG4
CHR_HSCHR19LRC_COX1_CTG3_1	CHR_HSCHR3_2_CTG2_1	CHR_HSCHR6_1_CTG5
CHR_HSCHR19LRC_COX2_CTG3_1	CHR_HSCHR3_2_CTG3	CHR_HSCHR6_1_CTG6
CHR HSCHR19LRC LRC I CTG3 1	CHR HSCHR3 3 CTG1	CHR HSCHR6 1 CTG7
CHR HSCHR19LRC LRC J CTG3 1	CHR HSCHR3 3 CTG2 1	CHR HSCHR6 1 CTG8
CHR HSCHR19LRC LRC S CTG3 1	CHR HSCHR3 3 CTG3	CHR HSCHR6 1 CTG9
CHR HSCHR19LRC LRC T CTG3 1	CHR HSCHR3 4 CTG1	CHR HSCHR6 8 CTG1
CHR HSCHR19LRC PGF1 CTG3 1	CHR HSCHR3 4 CTG2 1	CHR HSCHR6 MHC APD CTG1
CHR HSCHR19LRC PGF2 CTG3 1	CHR HSCHR3 4 CTG3	CHR HSCHR6 MHC COX CTG1
CHR HSCHR2 1 CTG1	CHR HSCHR3 5 CTG2 1	CHR HSCHR6 MHC DBB CTG1
CHR HSCHR2 1 CTG15	CHR HSCHR3 5 CTG3	CHR HSCHR6 MHC MANN CTG1
CHR HSCHR2 1 CTG5	CHR HSCHR3 6 CTG2 1	CHR HSCHR6 MHC MCF CTG1
CHR_HSCHR2_1_CTG7	CHR_HSCHR3_6_CTG3	CHR_HSCHR6_MHC_QBL_CTG1
CHR HSCHR2 1 CTG7 2	CHR HSCHR3 7 CTG2 1	CHR HSCHR6 MHC SSTO CTG1
CHR_HSCHR2_2_CTG1	CHR_HSCHR3_7_CTG3	CHR_HSCHR7_1_CTG1
CHR HSCHR2 2 CTG15	CHR HSCHR3 8 CTG2 1	CHR HSCHR7 1 CTG4 4
CHR_HSCHR2_2_CTG7	CHR_HSCHR3_8_CTG3	CHR_HSCHR7_1_CTG6
CHR_HSCHR2_2_CTG7_2	CHR_HSCHR3_9_CTG2_1	CHR_HSCHR7_1_CTG7
CHR_HSCHR2_3_CTG1	CHR_HSCHR3_9_CTG3	CHR_HSCHR7_2_CTG1
CHR_HSCHR2_3_CTG15	CHR_HSCHR4_1_CTG12	CHR_HSCHR7_2_CTG4_4
CHR_HSCHR2_3_CTG7_2	CHR_HSCHR4_1_CTG4	CHR_HSCHR7_2_CTG6
CHR_HSCHR2_4_CTG1	CHR_HSCHR4_1_CTG6	CHR_HSCHR7_2_CTG7
CHR_HSCHR2_4_CTG7_2	CHR_HSCHR4_1_CTG8_1	CHR_HSCHR7_3_CTG1
CHR_HSCHR2_5_CTG7_2	CHR_HSCHR4_1_CTG9	CHR_HSCHR7_3_CTG4_4
CHR_HSCHR2_6_CTG7_2	CHR_HSCHR4_11_CTG12	CHR_HSCHR7_3_CTG6
CHR_HSCHR2_7_CTG7_2	CHR_HSCHR4_12_CTG12	CHR_HSCHR8_1_CTG1
CHR_HSCHR2_8_CTG7_2	CHR_HSCHR4_2_CTG12	CHR_HSCHR8_1_CTG6
CHR_HSCHR20_1_CTG1	CHR_HSCHR4_2_CTG4	CHR_HSCHR8_1_CTG7
CHR_HSCHR20_1_CTG2	CHR_HSCHR4_3_CTG12	CHR_HSCHR8_2_CTG1
CHR_HSCHR20_1_CTG3	CHR_HSCHR4_4_CTG12	CHR_HSCHR8_2_CTG7
CHR_HSCHR20_1_CTG4	CHR_HSCHR4_5_CTG12	CHR_HSCHR8_3_CTG1
CHR_HSCHR21_1_CTG1_1	CHR_HSCHR4_6_CTG12	CHR_HSCHR8_3_CTG7
CHR_HSCHR21_2_CTG1_1	CHR_HSCHR4_7_CTG12	CHR_HSCHR8_4_CTG1
CHR_HSCHR21_3_CTG1_1	CHR_HSCHR4_8_CTG12	CHR_HSCHR8_4_CTG7
CHR_HSCHR21_4_CTG1_1	CHR_HSCHR4_9_CTG12	CHR_HSCHR8_5_CTG1
CHR_HSCHR21_5_CTG2	CHR_HSCHR5_1_CTG1	CHR_HSCHR8_5_CTG7
CHR_HSCHR21_6_CTG1_1	CHR_HSCHR5_1_CTG1_1	CHR_HSCHR8_6_CTG1
CHR_HSCHR21_8_CTG1_1	CHR_HSCHR5_1_CTG5	CHR_HSCHR8_6_CTG7
CHR_HSCHR22_1_CTG1	CHR_HSCHR5_2_CTG1	CHR_HSCHR8_7_CTG1
CHR_HSCHR22_1_CTG2	CHR_HSCHR5_2_CTG1_1	CHR_HSCHR8_7_CTG7
CHR_HSCHR22_1_CTG3	CHR_HSCHR5_2_CTG5	CHR_HSCHR8_8_CTG1
CHR_HSCHR22_1_CTG4	CHR_HSCHR5_3_CTG1	CHR_HSCHR8_9_CTG1
CHR_HSCHR22_1_CTG5	CHR_HSCHR5_3_CTG1_1	CHR_HSCHR9_1_CTG1
CHR_HSCHR22_1_CTG6	CHR_HSCHR5_3_CTG5	CHR_HSCHR9_1_CTG2
CHR_HSCHR22_1_CTG7	CHR_HSCHR5_4_CTG1	CHR_HSCHR9_1_CTG3
CHR_HSCHR22_2_CTG1	CHR_HSCHR5_4_CTG1_1	CHR_HSCHR9_1_CTG4
CHR_HSCHR22_3_CTG1	CHR_HSCHR5_5_CTG1	CHR_HSCHR9_1_CTG5
CHR_HSCHR22_4_CTG1	CHR_HSCHR5_6_CTG1	CHR_HSCHR9_1_CTG6
CHR_HSCHR22_5_CTG1	CHR_HSCHR5_7_CTG1	CHR_HSCHR9_1_CTG7
CHR_HSCHR22_6_CTG1	CHR_HSCHR5_8_CTG1	CHR_HSCHRX_1_CTG3
CHR_HSCHR22_7_CTG1	CHR_HSCHR5_9_CTG1	CHR_HSCHRX_2_CTG12
CHR_HSCHR22_8_CTG1	CHR_HSCHR6_1_CTG10	CHR_HSCHRX_2_CTG3
CHR_HSCHR3_1_CTG1	CHR_HSCHR6_1_CTG2	CHR_HSCHRX_3_CTG7
CHR_HSCHR3_1_CTG2_1	CHR_HSCHR6_1_CTG3	

Table S1: The 329 alternate loci and scaffolds in Human GRCh38 assembly supported in theSeqTailor webserver.

protein-coding
IG genes
TR genes
non-stop decay
nonsense mediated decay
polymorphic pseudogene
processed pseudogene
processed transcript
pseudogene
transcribed processed pseudogene
transcribed unitary pseudogene
transcribed unprocessed pseudogene
translated processed pseudogene
unitary pseudogene
unprocessed pseudogene

Table S2: The 15 transcript biotypes that belonging to the categories of protein coding and pseudogenes, that have been adopted in the data collection in the SeqTailor webserver.

Comments	Actions
Genetic Variants in VCF Files (CHROM, POS,	ID, REF, ALT)
Normal	Normal
Duplicated	Skipped
Field mistake	Skipped
Empty REF allele	Skipped
Unmatched REF allele to the reference genome	Skipped
Identical REF and ALT allele	Skipped
Negative window size	Skipped
Too long window size	Reduced window size to 5,000bp
Genetic Ranges in BED Files (CHROM, STAR	Γ, END)
Normal	Normal
Duplicated	Skipped
Field mistake	Skipped
Negative START	Set Start to 0
Negative END	Skipped
Negative START and END	Skipped
START is greater than END	Swap START with END
Too long genomic range	Trim at 10,000bp from START

Table S3: The exception handling in the SeqTailor webserver for VCF files and BED files.

Organism	Standard Genetic Codes	Mitochondrial Genetic Codes
Human	Table 1	Table 2
Chimpanzee	Table 1	Table 2
Mouse	Table 1	Table 2
Rat	Table 1	Table 2
Cow	Table 1	Table 2
Chicken	Table 1	Table 2
Lizard	Table 1	Table 2
Zebrafish	Table 1	Table 2
Fruitfly	Table 1	Table 5
Arabidopsis	Table 1	Table 1
Rice	Table 1	Table 1

Table S4: The standard and mitochondrial genetic codes tables for the supported 11 organisms in the SeqTailor webserver, according to the NCBI Taxonomy Database.

Genetic Code Table 1									
Codon	Amino A	Acid		Codon Amino Aci					
TTT	Phe	F		ATT	Ile	Ι			
ТСТ	Ser	S		ACT	Thr	Т			
TAT	Tyr	Y		AAT	Asn	N			
TGT	Cys	С		AGT	Ser	S			
TTC	Phe	F		ATC	Ile	Ι			
TCC	Ser	S		ACC	Thr	Т			
TAC	Tyr	Y		AAC	Asn	Ν			
TGC	Cys	С		AGC	Ser	S			
TTA	Leu	L		ATA	Ile	Ι			
TCA	Ser	S		ACA	Thr	Т			
TAA	Ter	*		AAA	Lys	Κ			
TGA	Ter	*		AGA	Arg	R			
TTG	Leu	L		ATG	Met	М			
TCG	Ser	S		ACG	Thr	Т			
TAG	Ter	*		AAG	Lys	K			
TGG	Trp	W		AGG	Arg	R			
СТТ	Leu	L		GTT	Val	V			
ССТ	Pro	Р		GCT	Ala	Α			
CAT	His	Н		GAT	Asp	D			
CGT	Arg	R		GGT	Gly	G			
CTC	Leu	L		GTC	Val	V			
CCC	Pro	Р		GCC	Ala	Α			
CAC	His	Н		GAC	Asp	D			
CGC	Arg	R		GGC	Gly	G			
СТА	Leu	L		GTA	Val	V			
CCA	Pro	Р		GCA	Ala	Α			
CAA	Gln	Q		GAA	Glu	Е			
CGA	Arg	R		GGA	Gly	G			
CTG	Leu	L		GTG	Val	V			
CCG	Pro	Р		GCG	Ala	Α			
CAG	Gln	Q		GAG	Glu	Е			
CGG	Arg	R		GGG	Gly	G			

 Table S5: The genetic codes table 1.

Genetic Code Table 2									
Codon	Amino .	Acid		Codon Amino Ac					
TTT	Phe	F		ATT	Ile	Ι			
TCT	Ser	S		ACT	Thr	Т			
TAT	Tyr	Y		AAT	Asn	Ν			
TGT	Cys	С		AGT	Ser	S			
TTC	Phe	F		ATC	Ile	Ι			
TCC	Ser	S		ACC	Thr	Т			
TAC	Tyr	Y		AAC	Asn	Ν			
TGC	Cys	С		AGC	Ser	S			
TTA	Leu	L		ATA	Met	М			
TCA	Ser	S		ACA	Thr	Т			
TAA	Ter	*		AAA	Lys	Κ			
TGA	Trp	W		AGA	Ter	*			
TTG	Leu	L		ATG	Met	М			
TCG	Ser	S		ACG	Thr	Т			
TAG	Ter	*		AAG	Lys	Κ			
TGG	Trp	W		AGG	Ter	*			
CTT	Leu	L		GTT	Val	V			
CCT	Pro	Р		GCT	Ala	Α			
CAT	His	Η		GAT	Asp	D			
CGT	Arg	R		GGT	Gly	G			
CTC	Leu	L		GTC	Val	V			
CCC	Pro	Р		GCC	Ala	Α			
CAC	His	Н		GAC	Asp	D			
CGC	Arg	R		GGC	Gly	G			
CTA	Leu	L		GTA	Val	V			
CCA	Pro	Р		GCA	Ala	А			
CAA	Gln	Q		GAA	Glu	Е			
CGA	Arg	R		GGA	Gly	G			
CTG	Leu	L		GTG	Val	V			
CCG	Pro	Р		GCG	Ala	А			
CAG	Gln	Q		GAG	Glu	Е			
CGG	Arg	R		GGG	Gly	G			

 Table S6: The genetic codes table 2.

Genetic Code Table 5									
Codon	Amino	Acid		Codon Amino A					
TTT	Phe	F		ATT	Ile	Ι			
TCT	Ser	S		ACT	Thr	Т			
TAT	Tyr	Y		AAT	Asn	Ν			
TGT	Cys	С		AGT	Ser	S			
TTC	Phe	F		ATC	Ile	Ι			
TCC	Ser	S		ACC	Thr	Т			
TAC	Tyr	Y		AAC	Asn	Ν			
TGC	Cys	С		AGC	Ser	S			
TTA	Leu	L		ATA	Met	М			
TCA	Ser	S		ACA	Thr	Т			
TAA	Ter	*		AAA	Lys	K			
TGA	Trp	W		AGA	Ser	S			
TTG	Leu	L		ATG	Met	М			
TCG	Ser	S		ACG	Thr	Т			
TAG	Ter	*		AAG	Lys	K			
TGG	Trp	W		AGG	Ser	S			
CTT	Leu	L		GTT	Val	V			
CCT	Pro	Р		GCT	Ala	Α			
CAT	His	Н		GAT	Asp	D			
CGT	Arg	R		GGT	Gly	G			
CTC	Leu	L		GTC	Val	V			
CCC	Pro	Р		GCC	Ala	Α			
CAC	His	Н		GAC	Asp	D			
CGC	Arg	R		GGC	Gly	G			
CTA	Leu	L		GTA	Val	V			
CCA	Pro	Р		GCA	Ala	Α			
CAA	Gln	Q		GAA	Glu	Е			
CGA	Arg	R		GGA	Gly	G			
CTG	Leu	L		GTG	Val	V			
CCG	Pro	Р		GCG	Ala	Α			
CAG	Gln	Q		GAG	Glu	Е			
CGG	Arg	R		GGG	Gly	G			

Table S7: The genetic codes table 5.



DNA Sequence Extraction from BED

Figure S1: Runtime performance in extracting DNA reference sequences from varying sizes of input BED data, by SeqTailor-standalone, BEDTools, and SAMtools, in a script-based manner.

Case Study

SeqTailor aims to makes it efficient to further investigate the genomic variant data and renders sequence-based software more accessible. To demonstrate the practical power of SeqTailor to bridge the gap between genomic variations and sequence-based tools for analyses and predictions, we exhibited a case study on pathogenic genetic variants with different effects and different clinical consequences identified in five human genes (*MSH2*, *BRAF*, *GJB2*, *BRCA2*, and *IL2RG*) by the HGMD professional database (1) and the ClinVar database (2).

Case Study 1: DNA sequence extraction for a variant in MSH2

A single nucleotide variant (ClinVar: SCV000107433.2, dbSNP: rs587779138), changes T to G on Chr2:47635062 forward strand of GRCh37 assembly, is a deep intronic variant of *MSH2* gene that has been shown to cause Lynch syndrome through the creation of a new splice donor site with pseudoexon activation (3).

NM_000251.2(MSH2):c.212-478T>G

Allele ID:	96369
Variant type:	single nucleotide variant
Cytogenetic location:	2p21
Genomic location:	 Chr2: 47407923 (on Assembly GRCh38) Chr2: 47635062 (on Assembly GRCh37)
HGVS:	 NG_007110.2:g.9800T>G NM_000251.2:c.212-478T>G NC_000002.12:g.47407923T>G (GRCh38) LRG_218t1:c.212-478T>G NC_000002.11:g.47635062T>G (GRCh37) NM_000251.1:c.212-478T>G LRG_218:g.9800T>G

p25.3	p24	p23	p22	p21	p16	p15	p14	p13	p12	p11.2	p11.1	q11.	2 q12	q13	q14.1	q14.3	q21.2	q22	q23	q24.2	q31	q32.1	q32.3	q33	q34	q35	q36
		_		_				_																		_	
												<u> </u>									<u> </u>						
	<u>Re</u>	gion 🗸	MSH2	\diamond	NM_000	251.2		\$				▶ •	•••	••	••	•••			•								
			Gene		Transcri	pt							Exons	: clicl	k an e	xon abo	ove to z	oom ir	n, mouse ov	er to see	details						4
6	0	NG 000		4	20							- 3	Z.										SD T			Treat	
9	0	NC_000	002.11 •		4					а, н н	6 🗖	2	<u> </u>										× 10		*	ITACK	s •
<u> </u>		47	,635,040				47,	635,05	0				47,	rs5	877	79138			47,635,6	70			47,635	,080			
G	A A	TT	TTG	A A	TTA	A	T (i C	TG	A A	A 1	ΤG	iA	БΤ	T.	A A (GG	ст	TTG	GG	GG/	АСТ	GTI	G	GG	A A	ιT
C 1	ГΤ	A A	ААС	ΤT	AAT	T	A (: G .	A C	ТΤ	T.	A C	т	A	Α	TT	сс	GΑ	AAC	сс	c c 1	FGA		۱C	сс	TT	A
Gene	es,	NCBI H	lomo sa	pien	s Annot	atic	n F	elea	se 10)5				6	14												
\rightarrow	_					-	-		\rightarrow		_	_	>	-	-		_	_	\rightarrow			>		\rightarrow			_
Clir	nVar	r Short	. Varia	tion	s based	l on	dbs	NP B	uild	150	(Ho	mo	sapi	ən(

With SeqTailor, the DNA sequence (+/-100 bp) around this variant (chr2-47635062-T-G) was rapidly extracted. In the output sequence, this variant is located at position 101.



The output sequences can be directly used as the input for sequence-based splicing prediction tools (e.g. NetGene2 (4)), to evaluate the impact of this variant on splicing. A number of tools are available for this purpose, and NetGene2 was selected here for demonstration purposes only. When provided with the reference sequence and alternative sequence, NetGene2 gave the splicing predictions for both sequences, as shown below. In this example, NetGene2 identified donor splice site at position 101 on the alternative sequence, but no donor splice site in the reference sequence.

Prediction on the reference sequence
Donor splice sites, direct strand
No donor site predictions above threshold.
Prediction on the alternative sequence
Donor splice sites, direct strand
pos 5'->3' phase strand confidence 5' exon intron 3' 101 1 + 0.83 CTGAAATGAG^GTAAGGCTTT

Case Study 2: DNA sequence extraction for a variant in BRCA2

An indel (ClinVar: SCV000637244.1, dbSNP: rs587781422), deletes GG and inserts TA on Chr13:32954282–32954283 forward strand of the GRCh37 assembly, is a splicing variant of the *BRCA2* gene that has been associated with hereditary breast-ovarian cancer (5). This 2-bp variant sits exactly at the splicing site, spanning from the last nucleotide of exon 24 of *BRCA2* gene to the first nucleotide of the following intron.

Allele ID: 150706			
Variant type:	Indel		
Cytogenetic location: 13q13.1			
Genomic location:	 Chr13: 32380145 - 32380146 (on Assembly GRCh38) Chr13: 32954282 - 32954283 (on Assembly GRCh37) 		
HGVS: NG_012772.3:g.69666_69667delGGinsTA NM_000059.3:c.9256_9256+1delGGinsTA NC_000013.11:g.32380145_32380146delGGinsTA (GRCh38) LRG_293t1:c.9256_9256+1delGGinsTA NC_000013.10:g.32954282_32954283delGGinsTA (GRCh37) LRG_293:g.69666_69667delGGinsTA			
p13 p12 p11.2 p11.1 q11 q12.1	q12.2 q13 q14.1 q14.2 q21.1 q21.2 q21.3 q22 q31	q32 q33	
Region V BRCA2 NM_000059.3 Image: Constraint of the second			
🦕 😓 NC_000013.10 - 🗘 🖒 Q ⊂		🔆 Tools 🗸 🔹 Tracks 🗸	
32,954,260	32,954,270 32, rs587781422 32,954,290	32,954,300	
ТТСТССТТТСТСТ	T G T G A A A A A A A G G T A A T G C A C A A T A T A G T T	A A T T T T T	
AACAGCAAAGACA	A C A C T T T T T T T T G T C C A T T A C G T G T T A T A T C A A	ттаааааа	
Genes, NCBI Homo sapiens Annota	tion Release 105		
	an alkown puild 150 (Wang series and	2	

NM_000059.3(BRCA2):c.9256_9256+1delGGinsTA

By submitting this variant (chr13-32954282-GG-TA) and choosing to annotate the nearest splice site, SeqTailor extracted the ref./alt. DNA sequences, and provided the distance from the variant to the nearest splice site (+ve distance: downstream, -ve distance: upstream, and 0: exactly at the splice site), as well as the belonging gene symbol, transcript ID, exon number, and donor/acceptor site information. Please note that, in SeqTailor, the nearest splice site refers to the first nucleotide of the nearest exon (as acceptor site) or the last nucleotide of the nearest exon (as donor site).

🔀 Reference Ge	nome: Hur	nan [Hor	no sapiens] ((GRCh37/hg19)	
∓ Coordinate:	💿 1-based	O 0-	based		
≓ Strand:	\bigcirc both	o fo	orward	 reverse 	
For Genomic V	ariants in VCF				
🛠 Window Size:	(in bp)				
💿 uniform	(+/-): 25	bp			
 different 	(+):	bp	(-):	bp	
兆 Nearest Splic	e Site Annotati	on:	\bigcirc no	\bigcirc canonical	💿 all
🗮 Neighbor Var	iants Within Wi	ndow:	o no	\bigcirc yes	
🜔 Output Sequ	ence: 💿 re	ef & alt	\bigcirc ref	\bigcirc alt	
두 Genomic Var	iants: (no more t	han 10,00	0 genomic vc	ariants)	
🖹 provide the	first 5 columns o	f the gen	omic variants	s in VCF format. (<i>che</i>	eck sample VCF)
chr13 329	54282 .		GG TA		
>13_32954282_GG_TA	+ ref <mark>Nearest</mark>	Splice:(CACAATA);BRCA2;ENS	T00000380152;exc	n_24;donor_site
>13_32954282_GG_TA	+ alt Nearest	Splice:(;BRCA2;ENS	_ T00000380152;exc	n_24;donor_site
TGTCGTTTCTGTTGTGAA	AAAAACA <mark>TA</mark> TAATG	CACAATA	AGTTAATTTT	"T	

The output sequences can be rapidly used for splicing prediction (e.g. NNSPLICE (6)). Again, the tool used here were selected for demonstration purposes only. NNSPLICE identified a donor splice site in the reference sequence, but not in the alternative sequence.

F	Reference Sequence:					
	Donor site predictions for 129.85.163.169.19962.0 :					
	Start	End	Score	Exon	Intron	
	20	34	0.95	aaaaca	aggtaatgca	
A	Alternative Sequence:					
	Donor	site pr	edictions	for 129	.85.163.169.20244.0 :	
	Start	End	Score	Exon	Intron	

Case Study 3: DNA sequence extraction for a variant in IL2RG

A single nucleotide variant (ClinVar: SCV000637244.1, dbSNP: rs886039387), changes A to G on ChrX:70330553 reverse strand of the GRCh37 assembly, is an intronic variant of the *IL2RG* gene that has been reported in association with X-linked severe combined immunodeficiency (7). This variant does not directly change the encoded amino acid sequence, but the experimental studies have shown that this intronic mutation causes aberrant splicing in the mRNA as shown by RT-PCR on B-cell line of an individual with this variant (7).

Allele ID:	260341	260341			
Variant type: single nucleotide variant					
Cytogenetic location:	Xq13.1				
Genomic location:	 ChrX: 71110703 (on a ChrX: 70330553 (on 	Assembly GRCh38) Assembly GRCh37)			
HGVS: NG_009088.1:g.5851A>G NM_000206.2:c.270-15A>G NC_000023.11:g.71110703T>C (GRCh38) LRG_150t1:c.270-15A>G NC_000023.10:g.70330553T>C (GRCh37) LRG_150:g.5851A>G					
p22.3 p22.2 p22.1 p21.3 p21.2 p21.1	p11.4 p11.3 p11.23 p11.21 q12	q13 q21.1 q21.2 d	12].3 q22.1 q22.3 q23 q24	q25 q26	q27
					-
Region IL2RG NM_0002 Gene Transcrip	206.2 🗘 🕅 🕸 🖣	Exons: click an exon above to zoo	om in, mouse over to see details		
匀 🗦 NC_000023.10 - 🗘 🖒 Q 🛛		<u></u>		🔀 Tools 🗸 🛙	🕈 Tracks 🗸
70,330,530	70,330,540	70,338 rs886039387	70,330,560	70,330,570	
GAGTTCTTGTACC	TAGAGGAGAA	AGGTTGGAAG	GAAGAGGAAC	AGTGGG	GCCA
С Т С А А G А А С А Т G G	ATCTCCTCTT	TCCAACCTTC	сттстсстт 🤉	TCACCC	ссст
Genes, NCBI Homo sapiens Annota	ation Release 105				
٠ ٩	<		<	← →	
:linVar Short Variations based on dbSNP Build 150 (Homo sapieh]					

NM_000206.2(IL2RG):c.270-15A>G

As the *IL2RG* gene is on the reverse strand, its VCF format is converted to (chrX-70330553-T-C). SeqTailor extracted its ref./alt. DNA sequences, and informed its nearest splice site (acceptor site) is located at 15bp downstream from the position of the variant. The sequence colored in orange represent the exons.

	🔀 Reference Ger	Human [Ho	mo sapiens]	(GRCh37/hg19)		
		o 1-bas	ed 🛛 🔿 0	-based		
	≓ Strand:	\bigcirc both	\bigcirc f	orward	 reverse 	
	<u>For Genomic Va</u>	ariants in \	/CF			
	🗶 Window Size:	(in bp)				
	o uniform	(+/-): 50	bp			
	\bigcirc different	(+):	bp	(-):	bp	
	♣ Nearest Splice	e Site Anno	tation:	\bigcirc no	\bigcirc canonical	💿 all
	🗮 Neighbor Vari	ants Withir	n Window:	o no	\bigcirc yes	
	Output Seque	ence:	ref & alt	\bigcirc ref	\bigcirc alt	
	두 Genomic Vari	ants: (no mo	pre than 10,0	00 genomic v	ariants)	
	🗎 provide the	first 5 colum	ns of the ger	iomic variant	s in VCF format. (<i>ch</i>	eck sample VCF)
	chrX 7033	80553	•	т С		
>X N TC	_70330553_T_C - r earestSplice:+15; TGGATATCTGCAGTACC	cef IL2RG;ENSI CCAGATTGGCC	:0000037420 :CCACTGTTCC	2;exon_3;a	cceptor_site AACCTTTCTCCTCTAG	GTACAAGAACTCGGA
	ATGATAAAGTCCAGAAG	576				
N	earestSplice:+15;	IL2RG;ENSI	0000037420	2;exon_3;a	cceptor_site	
TC ¹	TGGATATCTGCAGTACO <mark>ATGATAAAGTCCAGAA</mark> G	CAGATTGGCC TG	CCACTGTTCC	TCTTCCTTCC	AGCCTTTCTCCTCTAG	GTACAAGAACTCGGA
					1560	
					тэрр	

The output ref./alt. DNA sequences were directly applied to Human Splicing Finder (8) for splicing analysis. As shown below, Human Splicing Finder predicted a new acceptor site might be created at the position of this variant, by giving splicing score 92.5 for the mutant versus the splicing score 63.55 for the wild-type.

	HSF Matrices Sequence Position	cDNA Position	Splice site type	Motif	New splice site	Wild Type	Mutant If c exor	ryptic site use, I length variation	Variation (%)
	✓ HSF Matrices								
	1 Ttetggatat etgeagtace eagattg <mark>gee ceaetgttee tetteettee ageetttete etetaggtae aagaae</mark> tegg ataatgataa agteeagaag 101 tg								
	Mutant sequence								
	Total sequence	length: 102 nucle	otides						
1	101 tg	Caglace Caga	LUGGCC CCaCLG	Jilee letteettee	adcolloc	CLCLAGGLAC	aagaactegg	alaalyalaa	agreeayaay
	4 mb above bab above								

Case Study 4: Protein sequence extraction for a variant in BRAF

A single nucleotide variant (ClinVar: SCV000616361.3, dbSNP: rs121913355), changes G to A on Chr7: 140481402 reverse strand of the GRCh37 assembly, is a missense variant of *BRAF* that replaces the glycine (G) residue in position 469 with a glutamic acid (E), and has been associated with cardio-facio-cutaneous syndrome (9).

Allele ID:	29013						
Variant type:	single nucleotide variant						
Cytogenetic location:	7q34						
Genomic location:	 Chr7: 140781602 (on Assembly GRCh38) Chr7: 140481402 (on Assembly GRCh37) 						
Other names:	• p.G469E:GGA>GAA						
Protein change:	G469E						
HGVS: p22 p21 p15.3 p15.1 p14	 NG_007873.3:g.148163G>A NM_004333.5:c.1406G>A NP_004324.2:p.Gly469Glu NC_000007.14:g.140781602C>T (GRCh38) LRG_299t1:c.1406G>A NC_000007.13:g.140481402C>T (GRCh37) NG_007873.2:g.148163G>A NM_004333.4:c.1406G>A P15056:p.Gly469Glu LRG_299p1:p.Gly469Glu LRG_299p1:p.Gly469Glu LRG_2999:g.148163G>A 						
Region ∨ BRAF NM_0043 Gene Transcript	Exons: click an exon above to zoom in, mouse over to see details						
🦕 😓 NC_000007.13 - 🗘 🖒 🔍 🤇	🕕 🔍 🝈 🗮 🛬						
140,481,380	140,481,390 144 rs121913355 140,481,410 140,481,420 14						
CCATGCCACTTTCC	С Т Т Б Т А Б А С Т Б Т Т С С А А Т Б А Т С С А Б А Т С С А А Т Т С Т Т Б Т С С						
GGTACGGTGAAAGG	G A A C A T C T G A C A A G G T T T A C T A G G T C T A G G T T A A G A A A C A G G						
Genes, NCBI Homo sapiens Annota	.tion Release 105						
ClinVar Short Variations based	on dbSNP Build 150 (Homo sapien.						

As the *BRAF* gene is located on the reverse strand, the HGVS nomenclature of this variant becomes (g.140481402C>T), thus its VCF format becomes (chr7-140481402-C-T). SeqTailor was then used to annotate the variant by the built-in SnpEff (10), followed by extracting the protein sequence (+/- 25 aa) around this missense variant of *BRAF*.

Z Reference Genome: Human [Homo sapiens] (GRCh37/hg19)
X Window Size: (in aa)
 entire amino acid sequence
uniform (+/-): 25 aa
different (+): aa (-): aa
\equiv Protein Sequence Annotation: \bigcirc canonical \bigcirc all
Output Sequence: o ref & alt o ref alt
Variants: (no more than 10,000 genomic variants)
provide the first 5 columns of the genomic variants in VCF format. (<i>check sample VCF</i>)
chr7 140481402 . C T
>7_140481402_C_T BRAF ENST00000288602 missense_variant p.Gly469Glu ref RDSSDDWEIPDGQITVGQRIGSGSFGTVYKGKWHGDVAVKMLNVTAPTPQQ
>7_140481402_C_T BRAF ENST00000288602 missense_variant p.Gly469Glu alt RDSSDDWEIPDGOITVGORIGSGSFETVYKGKWHGDVAVKMLNVTAPTPOO

Using the ref./alt. protein sequences, protein family or domain prediction tools (e.g. Pfam (11)) can be used to determine if the variant will lead to a loss of functionally important protein domains. In this case, *BRAF* is a protein kinase transducing mitogenic signals from cell membrane to nucleus, and its kinase domain plays a key role in its function. A Pfam search identified the protein kinase domain in the ref. protein sequence, but not in the alt. protein sequence, suggesting the missense variant may damages the conserved kinase domain, thereby impairing *BRAF* protein function.

Pfam	Search of refe	erence protein seque	nce		
	Sequence search results Show the detailed description of this results page. We found 1 Pfam-A match to your search sequence (all significant)				
	Show the search options and sequence that you submitted. <u>Return</u> to the search form to look for Pfam domains on a new sequence.				
	Significant Pfam-A Matches Show or hide all alignments.				
	Family	Descript	ion	Entry type	Clan
	Pkinase Tyr	Protein tyrosine kinase		Domain	<u>CL0016</u>
Pfam	Search of alte	ernative protein sequ	ence		
Sequence search results Show the detailed description of this results page. We did not find any Pfam-A matches to your search sequence					
	<u>Show</u> the search <u>Return</u> to the sea	h options and sequence th arch form to look for Pfan	nat you submitted. n domains on a new	sequence.	

Furthermore, the protein sequence and the altered amino acid can be submitted to PolyPhen-2 (12), to predict the functional effect of the missense variant. In this case, PolyPhen-2 assigned a score of 0.999 to this variant, implying a 'probably damaging' effect.



Case Study 5: Protein sequence extraction for a variant in GJB2

A single nucleotide variant (ClinVar: SCV000840535.3, dbSNP: rs80338942), deletes T on Chr13:20763554 revsere strand of the GRCh37 assembly, is a frameshift variant of the *GJB2* gene that changes the amino acid from Leu to Arg at position 56, and has been found to cause Nonsyndromic hearing loss and deafness (13).

Allele ID:	32049			
Variant type:	Deletion			
Cytogenetic location: 13q12.1				
Genomic location:	 Chr13: 20189415 (on Assembly GRCh38) Chr13: 20763554 (on Assembly GRCh37) 			
Other names:	 NM_004004.5(GJB2):c.167delT(p.Leu56Argfs) NM_004004.5(GJB2):c.167delT 			
HGVS:	 NG_008358.1:g.8561delT NM_004004.5:c.167delT NP_003995.2:p.Leu56Argfs NC_000013.11:g.20189415delA (GRCh38) NC_000013.10:g.20763554delA (GRCh37) NM_004004.5:c.167del NC_000013.10:g.20763554del (GRCh37) 			
p13 p12 p11.2 p11.1 q11 q12.1	q12.2 q12.3 q13 q14.1 q14.2 q14.3 q21.1 q21.2 q21.3 q22 q31 q32 q33 q3			
Kegion ✓ GJB2 ⓒ NM_004004.5 Gene MM_004004.5 Transcript				
20,763,540	Image: Weight of the second			
A C G T T C T T G C A T G C A A G A A C G T Segmental Duplications, Eichler Lab Warning: No track data found in this range 1000 Genomes Phase 3 Strict Accessibi	G C C T G G C T G C A G G G T G T T G C A G A C A A A G T C G G C C G G A C C G A C G T C C C A C A A C G T C T G T T T C A G C C G lity Mask Image: Contract of the contract			

NM_004004.5(GJB2):c.167delT (p.Leu56Argfs)

In this example, the reference protein sequence was the entire protein sequence of transcript ENST00000382844 of gene *GJB2*. The variant led to a frameshift occurring at position 56 by changing L to R, thus the following amino acids in the protein sequence will be translated differently. SeqTailor extracts and alters the reference CDS sequence, followed by re-translating the altered CDS sequence to the alternative protein sequence. In this example, the new protein sequence will terminate at 24 amino acids downstream from the frameshifted amino acid. Once the stop codon is encountered, SeqTailor gives a '*' symbol to inform the sequence termination.

🔀 Reference Genome:	Human [Homo sapiens] (GRCh37/hg19)
🛠 Window Size: (in aa)	
💿 entire amino acid s	equence
○ uniform (+/-): 2	5 aa
⊖ different (+):	аа (-): аа
🖹 Protein Sequence Anno	otation: 💿 canonical 🔷 all
Output Sequence:	o ref&alt 🛛 ref 🔷 alt
Variants: (no more than 1	0,000 genomic variants)
🗎 provide the first 5 colu	mns of the genomic variants in VCF format. (<i>check sample VCF</i>)
chr13 20763554	. AG G
>13_20763554_AG_G GJB2 ENST MDWGTLQTILGGVNKHSTSIGKIWLTV LIFVSTPALLVAMHVAYRRHEKKRKFI FSMQRLVKCNAWPCPNTVDCFVSRPTE >13_20763554_AG_G GJB2 ENST MDWGTLQTILGGVNKHSTSIGKIWLTV *	00000382844 frameshift_variant p.Leu56fs ref LFIFRIMILVVAAKEVWGDEQADFVCNTLQPGCKNVCYDHYFPISHIRLWALQ KGEIKSEFKDIEEIKTQKVRIEGSLWWTYTSSIFFRVIFEAAFMYVFYVMYDG KTVFTVFMIAVSGICILLNVTELCYLLIRYCSGKSKKPV 00000382844 frameshift_variant p.Leu56fs alt LFIFRIMILVVAAKEVWGDEQADFVCNTRSQAARTCATITTSPSPTSGYGPCS

References

- 1. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D. and Cooper, D.N. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*, **136**, 665-677.
- 2. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. and Maglott, D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, **42**, D980-985.
- 3. Clendenning, M., Buchanan, D.D., Walsh, M.D., Nagler, B., Rosty, C., Thompson, B., Spurdle, A.B., Hopper, J.L., Jenkins, M.A. and Young, J.P. (2011) Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer*, **10**, 297-301.
- 4. Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol*, **220**, 49-65.
- Acedo, A., Sanz, D.J., Duran, M., Infante, M., Perez-Cabornero, L., Miner, C. and Velasco, E.A. (2012) Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. *Breast Cancer Res*, 14, R87.
- 6. Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997) Improved splice site detection in Genie. *J Comput Biol*, **4**, 311-323.
- 7. Kumaki, S., Ishii, N., Minegishi, M., Ohashi, Y., Hakozaki, I., Nonoyama, S., Imai, K., Morio, T., Tsuge, I., Sakiyama, Y. *et al.* (2000) Characterization of the gammac chain among 27 unrelated Japanese patients with X-linked severe combined immunodeficiency (X-SCID). *Hum Genet*, **107**, 406-408.
- Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M. and Beroud, C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*, 37, e67.
- 9. Niihori, T., Aoki, Y., Narumi, Y., Neri, G., Cave, H., Verloes, A., Okamoto, N., Hennekam, R.C., Gillessen-Kaesbach, G., Wieczorek, D. *et al.* (2006) Germline KRAS and BRAF mutations in cardio-facio-cutaneous syndrome. *Nat Genet*, **38**, 294-296.
- 10. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80-92.
- Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, 44, D279-285.
- 12. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248-249.
- Amorini, M., Romeo, P., Bruno, R., Galletti, F., Di Bella, C., Longo, P., Briuglia, S., Salpietro, C. and Rigoli, L. (2015) Prevalence of Deafness-Associated Connexin-26 (GJB2) and Connexin-30 (GJB6) Pathogenic Alleles in a Large Patient Cohort from Eastern Sicily. *Ann Hum Genet*, **79**, 341-349.