

Supplementary File S4: Examples of sequences with extended or complex repeat structures, with accession ID and name from NCBI.

Color coding: **brown**: H segment; **blue**: F segment; **green**: S segment; **purple**: conserved Y segment; **pink**: low-scoring Y-segment inferred by position; **red**: conserved K-segment; **orange**: low-scoring K-segment inferred by position.

XP_021833744.1 cold-shock protein CS120-like isoform X1 Prunus avium Y2K10

Seg ID	Length	Start	Sequence
Nterm 0.1	72	1	MAHFGSTPEPTN TDEF GNPV HHSTTGTR KDEF GNPV QHGVA DTGYGTGAGYATHTKPSVVEHHGVPAAFNS
Rpt 1.1	41	73	KDDQYSRDSQTTTGGYGGDGYTGG EKKGI LQV KDKL PGGN
Rpt 2.1	44	114	KDDQYSHDTQTTTGAYGGAGYTGDDTR EKKGI IGQV KDKL PGGQ
Rpt 3.1	44	158	KDDQYCRDSHPTTGTYGGAGCTGGENQ EKKGI IGQV KDKL PGGQ
Rpt 4.1	44	202	KDDQFCRDTHPTTGAYGGAGYTGGEHQ EKKGI IGQV KDKL PGGQ
Rpt 5.1	44	246	KDEQYSHDTHPSTGAFGGGGYTGDDTR EKKGI IGQV KDKL PGGQ
Rpt 6.1	44	290	KDDQYSHDTRKTTGAYGGAGDTGDDTR EKKGI IGQV KDKL PGGQ
Rpt 7.1	44	334	KDDQYSHDTRKTTGAYGGAGDTGDDTR EKKGI IGQV KDKL PGGQ
Rpt 8.1	81	378	KDDQYCHDTHPSTGAFGGGGYTGDDTR EKKGV ADKV REKL PGGQNVHPTTGPYGGAGATGGETRERRGVADKVKEKLPCGP
Rpt 9.1	50	459	KDDQYSHDTHPTNPTSAGHGGVGHTGGEPQLH EKKGL MEIK KDKL PGHNN

ADL59574.1		dehydrin 7		Populus alba x Populus glandulosa		FK14
ID	Length	Start	Sequence			
Nterm	80	1	MAGVNKSHEYETKTKAGGESGAAETRDRLFGFMGKKKEEKQEEVPATGYEENIHRSDNSYPGDGEKKHEHTTVPSNTE			
Rpt 1	41	81	TPLEPEKKKS SYFEQAKDMIP AYKKTEDAPPSPTEAAVHPTE			
Rpt 2	40	122	TPLEPEKKKS SYFEQAKGMIP AYKKTEDGPPSPAETAVHPTE			
Rpt 3	40	162	TPLEPEKKKS SYFEQAKGMIP AYKKTEDAPPSPTEAAVHPTE			
Rpt 4.	41	202	TSLEPEKKKS SYFEQAKERIPT FKKTEDAPSSPAKAAVHHTTE			
Rpt 5	29	243	TPLEPE EKR GGFFD QAKERTPG FKKTEEVs			
Rpt 6	29	272	PRREPE EKR GGFFD QAKERTPG FKKTEEVs			
Rpt 7	29	301	PRREPE EKR GGFFD QAKERTPG FKKTEEVs			
Rpt 8	29	330	PRREPE EKR GGFFD QAKERTPG FKKTEEVs			
Rpt 9	46	359	PRREPE EKR GGFFD QAKERTPG FKKTEEVSPRPAKSAYNEGAFSQTG			
Rpt 10	40	405	TPFEPE EKK GF LDKVKEKVP AHKTEEVPPPPAESAFSHTTE			
Rpt 11	40	445	TPFEPE EKK GLLE KVKEKVP SQKRTEEAPHPPAAAFSHTN			
Rpt 12	42	485	TPFEPE EKR GF NKEKVP THKKTGEFPFPAKPASTEAAVSNTN			
Rpt 13	30	527	TPLEPE EKR GLLD KIKDKMPG HKKTDEVPP			
Cterm	37	557	SEFDSTENVVSHKEEP VKKGMMEKIKDKLPG HRPQI			

AAL83426.1		60 kDa dehydrin-like protein		<i>Cornus sericea</i>	Y36SK2
ID	Length	Start	Sequence		
Nterm	8	1	MAQYGNPT		
Rpt 1	37	9	EGLYKQPTDVYGNPISRTHESGNPVQRTADAQYGNPT		
Rpt 2	37	46	EGVYKQPTDVYGNPISRTHESGNPVQKTAQAQYGNPT		
Rpt 3	37	83	EGVYKQPTDAHGNPISRTHESGNPVQKTAQAQYGNPT		
Rpt 4	37	120	EGVYKQPTDAHGNPVSRTTHESGNPVQKTAQAQYGNPT		
Rpt 5	37	157	EGVDKHPADAHGNPAFRTHESGNPVQKTAQAQYGNPT		
Rpt 6	37	194	EGVYKHPDAHGNPVSRTTHDFGNPVQKTAQAQYGNPT		
Rpt 7	37	231	EGVDKHPADAHGNTAFRTHESGNPVQKTAQAQYGNPT		
Rpt 8	37	268	EGVDKHPDAHKNPISRTHDFGNPVQKTVQAQYGNPT		
Rpt 9	37	305	EGVYKHPDAHRNPISRTHNSGNPIQKTADAQYGNPT		
Rpt 1	37	342	EGVHKHPDAHENPLSRTHESDNPIQKTAGAQNRPNT		
Rpt 11	37	379	EGVHKHPDAHRNPVSRTTHESGNPIQKTADAQYRNPT		
Rpt 12	37	416	EGVHKHPDAHGNPVSRTHEYGNPIRKIADGQGVHPT		
Cterm1	47	453	GGTTEGYITTSSTSTGIGNGGAIGGQQHGGVLQRSGSGSSSEDDGQ		
Cterm2	67	500	GGRRKKKGLTDKKKEKLSGGRKPAQASHPTATTTTTGHDIYKGGQQNQEKKGVMKIKEKLPQHN		

ID	Length	Start	Sequence
Nterm1	57	1	MAEHHHVKPSDESEAAPVGRDDHQGGVESK ERGWFDFLGK KEDKKPQEEVLVSEFEN
Nterm2	58	58	DDHQGGVESK DRGWFDFLGK KEKKPQEEVLVSEFENVSVSEPEPKVDQGGYKDEPKV
Rpt 1	11	116	EGYHKEEPKVD
Rpt 2	16	127	EGYHKEEPKVERPKVD
Rpt 3	21	143	QGYHKEEPKMEGYKDELKVD
Rpt 4	10	164	QGYHEEPPKV
Rpt 5	11	174	EGYHKEEPKID
Rpt 6	10	185	DGYHKEEPKV
Rpt 7	11	195	EGYHKEEPKVD
Rpt 8	11	206	EGYHKEGPKVD
Rpt 9	10	217	QGYNKEEPKV
Rpt 10	21	227	EGYCKEEPVKGYKDELKVD
Rpt 11	10	248	QGGYKEEPKV
Rpt 12	11	258	GGYHKEEPKMD
Rpt 13	16	269	EGYHKEEPKVEGPKVD
Rpt 14	21	285	QGYHKEEPKVEGYKDELKVD
Rpt 15	10	306	QGYHEEPPKV
Rpt 16	11	316	EGYHKEEPKID
Rpt 17	11	327	DGYHKEEPKVD
Rpt 18	16	338	EGYHKEEPKVEGPKVD
Rpt 19	10	354	QGYNKEEPKV
Rpt 20	21	364	EGYCKEEPVKGYKDELKVD
Rpt 21	10	385	QGGYKEEPKV
Rpt 22	11	395	GGYHKEEPKMD
Rpt 23	21	406	EGYHKEEPKVEGYKDELKVD
Rpt 24	10	427	QGYHEEPPKV
Rpt 25	10	437	EGYYKEEPKV
Rpt 26	10	447	EGYHKEEPKV
Rpt 27	11	457	EGYHKEEPKVD
Rpt 28	11	468	EGYHKEEPKMD
Rpt 29	16	479	EGYHKEEPKVEGPKVD
Rpt 30	10	495	QGYHKEESKV
Rpt 31	21	505	EGYCKEEPKVEEYKDELKVD
Rpt 32	10	526	QGYHEEPPKV
Rpt 33	20	536	EGYYKEEPKVASYYKEESKV
Rpt 34	10	556	EGYHKEEPKV
Rpt 35	10	566	AGYYKEEPKV
Cterm1	50	576	DESYKKEEERENKEMKEKKGHTLKEKIAGEKEEEKHEKYKEKHEKYEDTS
Cterm2	49	626	VPVHVEKYEEVAHIPAEPALPE EKKGFLGKIKEKLP GHKKAAEEVHS
Cterm3	50	675	PHPTSTEHASALPPHYEGEASPK EKKGLLGKIKEKIP GYHPKTEEEKEK

KVI07719.1 Apolipoprotein A1/A4/E *Cynara cardunculus* var. *scolymus* K1

ID	Length	Start	Sequence
Nterm1	54	1	MAAIVFYWFCSCSCFAAAISSGLLCFLVLVAAIMADKSV CETVAVVKVEAEEDC
Nterm2	53	55	NEAVVLVEVDRKSGDDCVDGKTKKAELKEKIEEEKEKIGDKIHEAKYKVEEKA
Rpt1	11	108	EELKEKIECDV
Rpt2	18	119	EEAKEKIREKEYEHEYKK
Rpt3	18	137	EEHKEKLKEEVVEAKYKI
Rpt4	11	155	EEFKEKVETKE
Rpt5	11	166	GEIKEKIEEKI
Rpt6	18	177	EEFKEEIEEIVAEAKHKK
Rpt7	11	195	EELKEKIECEI
Rpt8	18	206	EEAKEKIKEKEIEHEYKK
Rpt9	18	224	EERKEKLKG EVDEAKYKI
Rpt10	11	242	EELEKKVECEI
Rpt11	11	253	EEVKEKIEEKI
Rpt12	18	264	EEFKKEIEKKVEEVKYEK
Rpt13	11	282	EELKEKIECEI
Rpt14	18	293	EEAKEKIKEKEYEHEYKK
Rpt15	11	311	EEHKEKLKEEV
Rpt16	18	322	EEAKYKID EVKEKIECKE
Rpt17	11	340	EEVKEKIEEKI
Rpt18	18	351	EEIKEKIEEKVVEAKYEK
Rpt19	11	369	EELKEKIESEI
Rpt20	18	380	EEAKKEIKEREYEHEYKK
Rpt21	11	398	EEHKEKLKEEV
Rpt22	18	409	EEAKYKIEEFKENVEYKV
Rpt23	11	427	EEVKEKIEEKI
Rpt24	18	438	EEFKKEIEEIVAEKGHKK
Rpt25	11	456	EELKEKIKCEI
Rpt26	18	467	EEAKEKIKEKEYEHEYKK
Rpt27	18	485	EEHKEKRKEEVVVKHKI
Rpt28	11	503	EELKEKVECEI
Rpt29	11	514	EEIKEKVEEKI
Rpt30	18	525	KEFKKEVVEKVVEEVKYEK
Rpt31	11	543	EELKEKIECEI
Rpt32	18	554	EAAKEKIKEREYEHEHKK
Rpt33	18	572	EERKEKHKEEVVEAKHKI
Rpt34	11	590	EELKEKVECKE
Rpt35	11	601	EEVKEKIEEKI
Rpt36	11	612	EELKEKIECEE
Rpt37	18	623	EEIKEKIKEKEYEHEHKK
Rpt38	18	641	EERKEKHREEIAEAKHKI
Rpt39	11	659	EELKEKVECKE
Rpt40	18	670	EEVKEKIKELKEKIECEV
Rpt41	18	688	EEVKEKIKEKEIELEYKK
Rpt42	18	706	EEHEEKLKEEIEDVKHKI
Rpt43	11	724	EEIKENIECKV
Rpt44	18	735	EEIEEKIKEKIEEVKENI
Rpt45	11	753	EEKIEEHKEKK
Cterm1	56	764	ELKNEKKELEKEMKELEKIKKHEEERCEVVTLP EPSYEQEEKIVFVEKIEVKA EED
Cterm2	56	820	CVAAPPPPPVAAHSDYTATVEVEHK EKKGIF EKIKQKLH GNHHSKSEEKKEKEHY