ADDITIONAL SUPPLEMENTARY INFORMATION

Position dependent termination and widespread obligatory frameshifting in *Euplotes* translation.

Alexei V. Lobanov[1]*, Stephen M. Heaphy[2]*, Anton A. Turanov[1], Maxim V. Gerashchenko[1], Sandra Pucciarelli[3], Raghul R. Devaraj[3], Fang Xie[4], Vladislav A. Petyuk[4], Richard D. Smith[4], Lawrence A. Klobutcher[5], John F. Atkins[2], Cristina Miceli[3], Dolph L. Hatfield[6], Pavel V. Baranov[2#], Vadim N. Gladyshev[1#]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[2]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

[3]School of Biosciences and Biotechnology, University of Camerino, Camerino, MC, Italy

[4]Pacific Northwest National Laboratory, Richland, WA 99352, USA

[5]Department of Molecular Biology and Biophysics, University of Connecticut Health Center, Farmington, CT 06032, USA

[6]Molecular Biology of Selenium Section, Mouse Cancer Genetics Program, Center for Cancer Research, National Institutes of Health, Bethesda, MD 20892, USA

* These authors contributed equally to this work

[#] Corresponding authors: Vadim N. Gladyshev (vgladyshev@rics.bwh.harvard.edu) and Pavel V. Baranov (p.baranov@ucc.ie).

Table of content

Supplementary Table S1. *E. crassus* and *E. focardii* genome assemblies. 3

Supplementary Note 1. *E. crassus* proteins with recoded and frameshift sites identified by mass spectrometry analyses. 4

Supplementary Note 2. Executable analysis document supporting proteomics component. 8

Supplementary Note 3. Representative profiles of ribosome density mapped to E. crasus transcripts and supporting BLAST hits alignments. 27

Supplementary Note 4. IGV screenshots of ribo-seq reads alignments in the vicinity of selected frameshifting sites 37

Supplementary Table S1. *E. crassus* and *E. focardii* genome assemblies.

Species	Assembler	Assembly size, kbp	Number of contigs	Number of nanochromosomes*
Euplotes focardii	ABYSS	91,569	363,689	7,199
	NEWBLER	**94,015**	**109,492**	**12,922**
	SOAP	200,640	1,144,956	4
	SSAKE	118,465	374,877	8,879
	VELVET	114,730	301,971	4,996
Euplotes crassus	CELERA	19,350	12,326	247
	NEWBLER	**59,563**	**56,588**	**14,194**
	PCAP	64,474	70,328	8,097

* Contigs containing both telomeric caps were designated as nanochromosomes. The assemblies shown in bold were used for further analyses.

Supplementary Note 1. *E. crassus* proteins with recoded and frameshift sites identified by mass spectrometry analyses.

a. Five out of nine selenoproteins (encoded by genes with UGA codon reassigned to code for selenocysteine) were detected by whole lysate high-throughput MS/MS analysis. Selenocysteine is shown in red. Sequences of the identified peptides are highlighted in yellow.

>eTR1
MDYSDTPQEESTHSYDYDLFVIGGGSGGLACAKVAQEAGAKVAVADFVKPTPKGTKWKVGGTCVNVGCIPKKLMHYSALLGNSYHDQVE
SGWEHEKPSHDWGKMITNVNNHIRGINFGYKADMRKRGIKFHEKFASFVDPHTVQLVDKKGKTEMITSNYFVIATGGRPLYPDIPGAKE
HAITSDDIFWMKDNPGKTLVVGASYVALEcAGFLHHFGNEVSVCVRSIFLRGFDQDMAQKIAKDMELSGINFIRDSVPTKIEKDEETGK
LTcFLTVGGEETTVEVDTVLFAIGRYAVTADLNLGNAGLIAEKNGKFITDKYQKTNVDNIYAIGDVLHGKLELTPTAIQAGRLLADRLF
AGGTTTMDFYDVPTTIFTPLEYGcVGYSEEDAREEYGDFIKVYHTYFQPLEWNFAKSIYKERNcYVKIIVNTADNDRVIGFHILCPNAG
EITQGIAIAIKVGVTKPQLDNCVGIHPTIAEEMTNLHIDKADNPDPIKSDCUS

>eP22
MESSDDKVGCVQSILVVLEGLNDDSSIITGLEILIKLIKNILKSPHEEKFRNIKKTNKAISTKLLSLSGIEDLILALGYKDDNDEFYVF
DIDKYSDLYKLKRAIQEFHDEKRKKYMTPEELEKFEILQEQKRKFYEDNKKKAKARKDLENGMKFDREEKNQEEIKSSKANHLNFGANV
VKFQPPAPASRUG

>eSelW2
MDSTTKGHIVVNYCGGUGYLPKARYVQEAVENRFPGDFSFDLKADVGKTGRLEVTVFVGDDTEGKLVHSKDKGQGFVKDSNVDSVLDSI
AALLE

>eGPx1
MGAALCFKKRKEKLETTVESLFEISAEDIDGQEHLLADLAKDKKCIMVVNVASKUGLTKTHYTQMVKIHNKYKDKGFEIFAFPCNQFLS
QEPGSNEDIKKFAREKYGAEFQLFSKIDVNGPNTHEVFRFCRRHSPLYDDETDTIQNIPWNFAKFLIDEEGNVVNYYSPKSNPDVCVPM
IEEMLGL

>eGPx2
MGQVFFKSKKEKLATTVKSLFEISAKDIDGQTHLLADLAEGRKCTMVVNVASKUGLTKTHYKQMVKIHNKYRDHGFEIFAFPCNQFMSQ
EPGTHEQIKKFAQEKYGAEFPLFSKVDVNGPDTHEVFKFCRRHSPLYDAEKDVVQNIPWNFAKFLIDEKGQVVEYYTPKQNPDLCVPKI
EEMLGL

b. Sequences of proteins predicted to contain frameshifting. Sites of frameshifting are shown with an exclamation point highlighted in red. Sequences of the identified peptides are highlighted in yellow.

>comp7880_c0_seq1 AAATAA
MDNIPDYLVLRLNGTSFLDRREEILSIEYNSIFTFAEFKMEACRTLRVKYDPKCRLFDKEGIETFEDDLNMLKSHDVLYLASRGEDFDY
SSVLNDYDRKDVLGEGGFGKVYHAVNRETGEDVAIKFMDISHYLTHADQIEEIYREADALQKLNHSHIISLHKAFVQRKEVILIMEYAG
GGELKDRVEEMKDMDEIYARFIFQQICSAMSYCHNRGLIHRDLKLENVLFKEKGGMIIKIVDFGIAGVCKPQEKEKTDSGTLSYMPPEV
LSGEKLEAGPGIDIWALGVMLYTMIYGKLPFYGDTEDEIINCIIKKKPSFKDKKTISKELKDLLIKVLNKDSDKRLSMFDLQNHKWMEM
QDEEILKSIEESKLEQEQEEEKK!NEEDELIAFDKLNIKDDKKSSKAGSDYNSLSAHSSNPSGGRKKKMRGSSPRSGMNGTGKKKKKT
IKKKAT

>comp8353_c0_seq1 AAATAA
KKAHAVFNGGVTALCFSKKRTTLYTAGGDGTFLVWPVGAKPNPNQSVEPADFFSSPDLSNIPQIDNETDDSVKYYKELLQEQFLDSEIP
RKKEFKAEISKELEDIRSHIVDLVEDNRKAQEIEQLERHEFVIDVKKKQSMEQDGWQQRQLITKQAKRKQLEYECLKERVKSATWSTME
THSTACISLDSDLLLYNYGIRLRTPLEKKTLNRVLHFRRMELRQQITGMETRANKILDQSLFSNHDETYIMNRIRGVQHYEVDEEAPII
TQAKSATKRKNKIAQQEASKSTADGALKKGKRKPKVDLNQFRLGANKPKLLDEDDFDDKLKDRAQNRDESNIAELRWKIVTKKKELEDL
KKDIHILGSWDLLYEPSDLYTDGRKKMQIEMLQDVVFALKQEYNKEFERFQRFKDDQIFAIQERSNRITEILEDLKREEELFHPRTHPL
ETPASILEVKPEEVTVQKYLTAEERAVEDEKHRIEQERLKALEGDNVGQRGIKNMLGGTYELKKNKGIMEETLEREEWMSKPIEDMTED
EKLKFKEFQQREKELQEEKDKKRKAWDQELKKNRIEIEEICFKFEDELKKIHKKRLYYDMRVYEQELYIIRLTLMLHENKEIKQEAALI
AEKKDKLEQELVESKNTINNFQRMYEDFDSEFKASSAIAEQEKGLRDLFPNAPFRQILAFVRNGGKAANRARGFRGQTEENTLEQEALA

LLCKLDPYIIVDENAVKAKFEQENVKEEYSYDRDKIVNLTPGEFDTLVQERENRNKIDKERKGMEQEIANLSGHKEFCEINANDLEEAY
EDIKASHTEIESRMEKLKYNFEAVVYMLQGQVEVAQAPVATDYKDAILVNTGVIEDENKKVVQEGNTNVK!KLEEITKFKRKLNHETWK
NDKLKLEIKDLLERAIDVQLYKVTKDTQEIIKGNHRTKDEDEKKRLEDQINNLQENAGARIEVINKKKKKLRKEINEKRKENNELETRA
RDLQTNVDQRNLIIDLRSKGPSGGGDDRLQDPIKRFKEVATVRKYKEIVDQQKEEIEFLEDELERFRSRTFPSFANMHARQDYAD

>comp7194_c0_seq1 AAATAG
MDTDLIDQIQKNMDQDPQLKDLFEPSSEENSQNDHGFTQGSKKFYSQDSAMAPPKVSRSKERELAFLK!RAQEIGLEPYNEYHGKKKTM
IKKQTPGKDLIFSNIRHKESTSQQRRDHASRDEQQLSNKSLALKTGTNKEKIQKHKLHQSQRTVGKKTFECKEAKPSNGVAQKRVEVIE
ISSKTSSSGNYSTPIESCPPIENCPSVESCPPVDHKTHEVVSLDSSNSDDKNIDDQPLNPKQKKRDKKKEQVDAKNARDFDSNPPKKSR
MVSSTPTVNSEVMDKYLQQTPPDQIEIVEFDSRPKWPTENCSNFQDQLKLMASQRKRNKADTLGSNMFQEMKSTLSNIEASMDSPQGPI
QKEYIHLKESIYPFWQSTFHHLEWNDDSDANKIKSREELKERMLDSLQYFSGHKLYRYADPADVKRGLLQNYPFVEDSDSKTEVKLLEG
KFHMLKIIVDGITKKIRVSIIKKDFT

>comp2566_c0_seq1 AAATAA
MWGRKKKAEPKKSETKKRAPAKKATGTRKTKPPAKAKFSKLESKEEVKEEIKHDSETIFKVYATDQEMGRPIIGSEGIQNLASDLKLDI
ASSAELIVFMWHCECEEYGQISKSEFQKGCDKLGVKDFSHFK!KSVPKKLSATLAMQDTPKEFRPFYKFAFTFHRTDGKNVPVETCQVL
FGLIFSDKYPILKTFFKFLAEKEVTHLTLDQWDSTYDLIRENPENLDNYDEYAAWPTLMDDFYQWYGENK

>comp6054_c0_seq1 AAATAA
MEDEKNESIQNFKAMAECDDDGIAFQYLDSNNWDLAQAYDQYQNTHQFNQSNPTSTPAPTSFPGGADVDMSPGADAEESAFDDIPDIPN
IGIPQMDQPSPVPQESNAPGSGLGNITSQFSNFASSIQSNLQNLTGGMFSGVMGGGMMGTDMNTQSNFSNRNLTAAQEFLFQFRKKNGM
HVILPKFVNNTFEEIGQESKRLRRPVFFYLHNDKGDSCNIVDQSVIGEEMTRMLLNKYICVGVNVNTEEGRKLLTALEIPKAPFIGITY
IDENGTLQNIGSRSGDEINVMALSEMDEAASGVFNAIFDGDTTDLTFHIEDSNLQLLETEEFKAEISAQMGNRTFDGYNEEPPRRRGPE
IDPTTGFPVGMTPQQIQDKILKDQQRQEYAEIDEKNKVLIEERKK!KQEENNLKRKEELEKKAKIEKLEEEKKEMAEIVRSNLPEEPSE
GTPDTITIQFRFPDGNHKQVRRFYKTDKVQLLYDYITSFGNENGFEAAHTHFSIIQNFPKKFFEDMNKTLEEEGLSNCTLMIKEHSHVE

>comp7882_c0_seq1 AAATAA
MNPKGSKREKRVKGSKNKSAKEAFLKKIDAAMKVYDYVDETKDVKGKSERLNAINELQNLLQDQKSVSQLIIPNLDSCMQMIEKNIFRC
LPNIKKSNLAFSETGIDQEEETDPAWPHVQGVYEFFLQLIMNDSIEVKLLKGYVTPEFVSRFLELFDSEEAVERDYLKNILHKLYAKLV
PRRKMIRKAINETFYQLIHEGHKFNGASELLDILASIISGFAVPLREEHVIFFNNIIIRLHKVQTCSEFFEQLLRCSMLFLTKDKSLAI
SLLKGLLKYWPFANCVKETLFLTELQEVLEIVDDDKIGDLVIPLFRRIVKCIGGTHLQVADRAMCFFENDYFLTKLRIYKDVTFPMLVP
VIVELSENHWHKILQESLVALKVILKEIDSAAFDEAQQISK!KDHRRFIVKPNVEKRTELDAKWERLNTTLKSTSAGFTPPDVPFKTSE
LISNYNGLYRKIYDKEKFIND

>comp7341_c0_seq1 AACTAA
MNNMIDCNHSPSLSDESVGADGDKECFSGERLEGSNNLPEEGITSISPDSLVNKASTLLRALTLFTSFSEFDTQNSPPPKTSKLFNDFS
LKLNNMKTCLSQAKTSDSEEKLPALKEEIESIKASIGEELALTPLGQALLWNLIEGDN!KDSSMKIFDELDHRCQDIANLHEVIAQKDA
EIQTLSKQIRKLAKFRRTSSLVSEETEDGDSASQSDSGSMTQLSRSSSLLNKLNLNTKSRLTLCLNKVRDLMLVKELKEPLQKINTLSF
NPGLLALEDAKEFLKNCFPLEVASFHFNKDSLLRNDLEKFLDVLLRTNEYVTDEIVLSNFVIDQDSLVKILSNFKNKEVVSFNSCKMSL
SNPPEFGDSLDGATLKHLYLNFCGDKSHGDWASNPAHFENLINGLSHSPDLKASLKDIWMEGSGLKKDKARDILDTFGFHSTKIWILYG

>comp5116_c0_seq1 AATTAA
MLPKPTNNMEKIKQQEYYKYKIRKVFECIAKESNHNNDITN!KNEIAYILRYFSQFPSEAQVTDYVIQKIEDDEPNDFIKLTKFEPYLL
DVIENNEFEPSPPEHLLAAFRVLDTDKGRIPIDVLNNLLTTEGIPFRKEEMDSFQEFALDKSQKFVYYEDYVAKLVEENDKHVEEFLKE
YPTFKPPINQ

>comp7670_c0_seq1 AAATAG, AATTAA
MDEETIEKYCQALDLSVSPDNDLRKQAEAFIIEGMETPGFIAAMLHISSNPDLNRDRKIDITQAAAIQFKNIVETHWKYKDDEYAKEMR
EDGYKVIIIPDETKTYVKENILTAYINVHSEKVAKQFDFIVRCITKHDFPDKWPDLANKVKDYIESDDLYGSEMFVGLYTLKSICKRYE
YEFDAKREPLNEIADILFPRLEAITTCVEGDNSDQGSRLKNLIGHCFYISNQISLCKRYLDPSMLDFIVKFNTSALEAEIDNSLTQPTE
SIEEIDHRAESFQWKLKMTAMNFLFRIFQKFSNPQYVNETMKPIAEHCINNYAEGIINLANTLIAKAK!SIYIDRQVLSYCFKVVSTSI
NQTSYREMIKPLIPEILTSHCVPAMLLTEKDTEDFEADPVEFIRKARDPNPNIYTARNSVLEMIRNVTQHKSNQDKGALPDFLESFFGF
LLENLSECIKQDAPDFRIKDALLLCLGQIAPTLLMYDQFHDQLNQVLTGAVFQDLTSENELVKYRALWVYGQCSRVPMEDDHRLEVGKL
LFQLMNDENTAVKITASTSLYKNLRNNSMKEAFKSELASILEAYLGLMDTIDNEELIAGLEEVVSLYEDCIGPYAIELCSKIVENFN!K
ITGKEQEEEEYGTMGMATSGLVVTIRSIINSCKGDPETLLKLEPVIFPVVVRSLSADGCEYLDEAMDCITAILNFTQSATERMWALFPH
LIKIIVGGPEDEEGGYAFDYFTSMEDYFRSLIKYGHDGMLTKKIGNDPVMILLIKGIIKILQLVKEGDVNTNAYICIVIVETLLLEFPG
KLDQLLPTFIKILCTELSNKEITKEFRLHALTLVAHCFIYNCTLTLGALTDLKVLVPVCQNFFSYLKKFSEVEHLRGLIYGITALLRMD

EMPDVIKGSIQKIIESLIDLMRKYTRERILELRSKFEDKRNRWDEGTDEYNNLDAPFQKLSEWMEEYKDDAYSEDDDDDDNFEEDDYLW
SRSDSCYYKSCLEDKEAPLFFKETLEDFRENKEEVYRGIIELIPEDSQKLLEMIMERCEYMQSLQS

>comp5528_c0_seq1 ATATAG
MSSQEILANSITNTVDKEKSAQEEQDDEVIIDDQNPLLEDDLQI!DEPEQKVNTDEPDQRNQEDEASENEQNLSDFINNTEFSYQSSST
TQNLKNLLIQSTIGLALKLKPKLGKMLITNSDGGCCDDIRKSLDLNKQLLGEDVADLISVKQITWDESLQVSGTRYDYICITGSHFSQE
FVQILEKIVPTVLSVVDDQERVFLLGPTSEEDISQFEDNVGDTIFESKKEEIDVPTKNSNSLGQFGDPDNHYSSENKDLIDEGIFDDDQ
VLQREGHLGLPSLEDDNEDDYFEPIG

>comp8412_c0_seq1 ATATAA
MSKNTKSKQEIDSTSKKLSRKERKNLEYIQYAKERKEYQKWEKEEADNLGFEGEESAPPVQNTSTAAGEKKKKGKKEKKPTDKVTTSPE
EQKKAYSETVAKQDALIAELQVKLDQREKRMKDLKKTQEQKLSKLKDQTMKLTKERDDAKASLSKVEDKCNGRLGDYQELIESVNRENL
DNEKLI!NDLTNDKANLKDIVFDLMFEKQKEGNSATEHPVVIPDNLQEEFNRKSQTKSSQNNPQLKTLLEFANEQIKRLQKEIKEARNH
LPKPELGEIDEATEEKED

>comp6034_c0_seq2_ATATAA
MSKNTKSKKQVTSNAKKGGNKKGKKAEPVQPPKEKKELAEWDLEDMPNFGFEPKKIAPTASKGPAVSGDKKKKGKKEKKTVEDTLISIE
EAKRANPEEIARQETLITELKSQLEQKDQAIADLEKDQKEQFKQLTEQAQQLTEERDETRAALAVAEGQCNQKLDDFKQTVDRVNRNFL
ENEKLI!SELTSEKSNLKDIVFDLMFEKQKEGDSAPEGEEEITDEITDEIHGEFDRNTRRQAPQDNSQVKTLLDFANEQIKRLQTELKE
VRNQIPESALQELDQIDETLKETED

>comp2483_c0_seq1 ATATAA, AAATAA
MEKFGDLRASRHRVKHSKMKSKNHRQYEQEEVKHARSRPDRFDPPQIDEESKYSAGIDEAIQLVEITEKGVCKINPIAMNIVKGIKTKV
GIISVVGPYRTGKSFLLNRLLGQQDGFEIGPTVQSCTRGIWIWGKPVKVSEDMHVILMDTEGLGSCNRTMNIDIKIFTLSVLLSSMFVY
NCLNAIDENALEVLSLVVNLAKYISNQKKNDSMDVYNQANYSPYFMWVVRDFSLQMMPSEELEKAGHDPATYWDKLENQEAAAKEYLEK
SLEAIDLGTINEENKRTVTKKNEIRKAIKNFFHQREATCLFRPINEEEKLRIVNKIPYEDLRKPFRKQVEHLINKIYYNVKPKSINGQT
LTGKMFAQMLEEYTSSMNNNGMPEINTAWDRVMDTEIKRVLQESTTKINYQLQEVVIDKMPMPLKQLISIERNVRKSALKLLYDPNIKN
APKDKLSRLQDKFIENLDEIFEGIFNENEIISKRQAKDLLPRMYQKIKAMINKGEFETIHDFSDIYGKMAISYFDNTNEPENYKI!IQN
FQINTVFEDLDEIMQTQVQRHESQNQEYETKLETKDHQIEHLNEQLKKEKTKNKDREQELRSKNMNIRSNLEEEIQMIK!NQISNKDQQ
FESLGTMIKEGWNNSEQVLKEIKAKEIENIKATIALESQKKIKELETTHQQEQLHYKKEMKKALEKTLAGLKMSYEDEIRLLKKNVKDQ
DKKITLLKKMCVRKDTQIQMLEEKAQSNEKQDKLQKEHRDILFELARAFKEGTTPGKDSTTNESATPY

>comp6951_c0_seq1 GAGTAA
MYSTKFRRVMTMAPLLLANPALALCEEPSTADRIRGNYENKIRFFAAPE!KIFETFSNIREEDGQVYMSYQDFFHSLTPYNFVASKDDD
DDDDDEENKDKEKEKEPGYFDKFTPEIMTIVDANQDKKIDFNEYIFFITLLQLPEGEVMRIIEKVNPEERKINKAQFAKYLTKLRKCTA
LGLKQMSKSFMPDGRKISTDEDHISKTILLHLFNDKEYITIEDFCELKSKLKHALLHYEFYQFDVDEDETISAESFAKSLLSCLNYTQA
SKYSRRIHSLKLEGRVSFKEYVAFHNLIEKADIIKMKISTYRFLSLGMFRDLCDDFAKLDPYCNQNKVSISDTQIATFFKVLDEDENGA
LEYDEVVDILEGKKNIGLGKEDKFKREMMEKIDRYIKKFQKYVGWT

>comp3853_c0_seq1 GTATAA
MSEENKEEVKGTTHTDEDQYHHGFGNHFESEAIEGALPKHRNNPQQCKFGLYAEQISGTPFTYPRAKMQRSWLYRIMPTVAHPPYKALK
DYNNLWIANFARDDDEEVFTTPQQMRWTPIDLPSEEITFVQGIQTV!TGAGDPSMKAGINMGVYTCNTSMKNEAFFSSDGDIMIVPQLG
KLSIMTEFGHIEAESWEVVVIPRGIKFAVEVNEDCRGYYCELYDGHLQIPDLGPIGTNGSANPRDFAIPKAKYFDETNEFRVIQKYLGK
FFEYTIPHNIFDIVAWHGNYYPYKYDCHHFNTMGSISYDHPDPSVFTVLTCQTPDHGQAALDFAIFPPRWLSMEDTFRPPYFHRNTMNE
FMGNVAGQYDAKEEGFSPGAVSLHSCMSAHGPEAEVVEKASTCELKPQKVGEGCLAFMFETCYTMKVTKSFMHDLEGATDSYSVNSSKA
VDESYHDCWKGMKRLFDPNDPDAGYKKYLSEHKN

>comp5973_c0_seq1 TTATAG
MSHLKNFQFSSVQITEIDTYIEHLYSENMDLKLKGCISILYLCFSAENMEEMIEHESLLPAVSRILRDDYKKSLDLSLYLLNVFYAYSH
FTEFHPLLIENQIGDTCVKIIEYEIKRYKARVNEYTKTAQLVKQTQQTPSADTDLKELQNNFRKEEKRLSVTIKKQEKVLFVTFHILLN
LAEDLKIERKMKKRRIVPLLVSMLERNNPDLLYIVLSFLKKLSVFGSNKDDMLELDIMKKLNRFIPCQNALLTQTALRLLFNLSFDNEI
RERVNAIGMIPKLVELLKVAQYRSILLRILYHLSSDDKIKATFAYTSCIPLVYQLVIHFPDAIIGKELIALAINLTTNKTNAALISQDD
QLEALIERAFKYNDVLLFRVVRNIAQFGPVTNIDIYEKYMDKIIELTKQCGDNTDLQIELIGTLVYINIEKWDTVLSQGDFLDFIHNNL
VSDYSEDDLVLETIMLIGTMCRSEKCAEAIAGSYIIGMLHELLGAKQEDDEMVQQILYTYHRLLYYRVTREIMLEQTQIVNVILELLND
KNPNIRKLVNSTLDLVQLHDEIWKQEIKTKKFEMHNEVYL!GLMEEYEAQAEALDEEALYDYYAQDPEALAALENGEFGEDDQWLDQND
LAQRIWNGEMDPDQMMDPNQMMDPNQ

>comp4582_c0_seq1 TTATAA
LTVDSFILLADKKNCITLFSTFQDLISKIARKKHIFALNKNNEFAPNPMIGFIQNLCDKIYTITTDKEGKTAKEFLQDFEYCHNEEAEI
DIPKPKIMKKMPPGIKKRLLADYNAKVEAAKKEVSKRAKNKVVISSSKTLLKEAFDLQDYHSFEYLFQYVEDKGIGYAELLHCLRNESN
RKIFVLILDYVLTTLPEEEFELIDVTNTTTQEISLRELFPDIDLKEFCLALYDSKNVPGEIPLKSKYLYTKLEVYTKKYSTLDEKNRTK
FLTTSLLVTGNTNPNEGYECLITKILDIISLYNIEIPIYYEGEVQEALRSNFKKLIFIRNYELVLKLQEIVKKNLRGLYEKL!SQEHLS
YISRLVQNSDAELGLEKDLLDSNEDVGVINSRQAVKDTLVFCLEQITLFDSYNQINFNDAPEKIIHMVKGFIHLGGFVNINVGFDSLDN
KEVNEGDVSEIKRLVTVTEQYKEAKHKFSQRLLLTEFFSTFQKYNDLSYELIDVDSPLRWIIDCPGQESPNFFEYCIEQGNIDLAMKLI
ESCDISEVISMFPLQERTISNLLNSPHIFEFLKKMSKEEEKIKKLIDRTNILEIPIMKLENSLSIDFDDEEDGNSGKPKYTQDQLTLYY
FCYLKDSLVPKVGKEIPYFNLLFFNQGQFKYSLDELVKVLPLDKIKELNSLGYQIGKIISQKPVNTKDLIEI

>comp7073_c0_seq1 AAATAA, AAATAA, TCCTAA
KVSKESESLSQNKNKPIKRRKITEDDKVEHLLSNSNSNSQQQVNQVKPREEIKQPPQDPHKDQNMDADMARLESLDIPKVGRQPDKHKD
HPMETDHDQKPQADANQQARDPEKPVRVPEDMRIPPSQPHVNPHLTEAPLRDAPSSQPLRAPQASPIHEIETAKKGKHVAPEVIRPDND
VDMSKNMFENKSDRPKMQQERAQVVTTPQFTEQVPQRKDKAVVHKSISEIKKENDAPNHRDRKGRANDLSAK!KLKFIDKYSTSQGRKE
LGNMIRRISGPQVQGIVRLMRQFHVGNKEGKEFKFSLNTLTPAQCARVGMLIEGISDPGSASSTGRAQTGKPGSHGERSSSAVGSQDAS
GAGRVSERERE IERRKRAEEEARYKERRK!KEHEMKLQERKKDELRRKEQEQRKEENRKFKEQQELLRRQEHERQQESDPHGPSYESKS
PVPPTTSEQQEAARVKAHQEQLEQKRLEEEKRKQAEAERERIEQERRRAEEDKRRKAEELRKSEEQERQRELAKLRLEEERRKKKEAEE
QRRREEERKRLELIKQKEEERRRQENSSPSKKSS!KACEEERRKREQEQLRKREEERRRQQELEQKKKEEEQRIREEQERRKREMEELR
IREEEQRRRQQEEEDRKRQELQRKLKEDEERRLKAEQERKQREEEQRRIREEQERQRREEEQRRLREEQERKRKQEEERKRKEEEERKR
KEEEEERKRKREQEELRLREEQERKRREEEERRRIEEERRRMEEERKRKEEEERKR

Supplementary Note 2. Executable Analysis Document Supporting Proteomics Component.

1 Introduction

The vignette describes and reproduces all the steps that aimed to confirm frameshifts in the *Euplotes crassus* proteome. The global 8M urea soluble proteome was digested using conventional trypsin protocol and alternatively with Glu-C protease under high pH (7.5) conditions. The latter restricts specificity of Glu-C cleavages to C-terminal of glutamic acid (E). The peptides resulting from trypsin digest were fractionated using two different approaches: with strong cation exchange (SCX) and high pH reverse phase (HPRP) chromatographies. The peptides from Glu-C digest were fractionated using HPRP only.

The datasets were deposited to PRIDE and available by this link http://dx.doi.org/10.6019/PXD004333. Summary of the datasets shown in the table below:

Dataset Prefix	Digestion Enzyme	Fractionation Chromatography Type
Euplotes_1_SCX	trypsin	SCX
Euplotes_1_HPRP_1	trypsin	HPRP
Euplotes_1_HPRP_2	Glu-C (pH 7.5)	HPRP

Preprocessing of the `raw` files prior MS/MS searches was done in two steps. First, the raw files were processed with DeconMSn to correct for wrong assignments of monoisotopic peaks. The parameters are as follows:

```
DeconMSN.exe -I35 -G1 -F1 -L6810 -B200 -T5000 -M3 -XCDTA
```

At the second step the peak files were processed with DtaRefinery to perform post-acquisition recalibaration of parent ion mass-to-charge ratios. The peak lists (concatenated dta files in this case) were searched using MS-GF+ tool against 6-frame translated *Euplotes Crassus* genome concatenated with tentatively frameshifted sequences and common contaminants. The 6-frame translated FASTA file, `DtaRefinery` and `MS-GF+` parameter files are available in `extdata` folder of the `EuplotesCrassus.proteome` package.

For example:

```
fpath <- system.file("extdata",
                     "MSGFDB_GluC_StatCysAlk_10ppmParTol.txt",
                     package="EuplotesCrassus.proteome")
```

```
cat(readLines(fpath, n=12), sep = '\n')
## #Parent mass tolerance
## #  Examples: 2.5Da or 30ppm
## #  Use comma to set asymmetric values, for example "0.5Da,2.5Da" will set 0.5Da to the left (expMass<t
## PMTolerance=10ppm
##
## #Max Number of Modifications per peptide
## # If this value is large, the search will be slow
## NumMods=3
##
## #Modifications (see below for examples)
## StaticMod=C2H3N1O1,    C,  fix, any,       Carbamidomethyl       # Fixed Carbamidomethyl C (alkylation
```

2 Post MS/MS Search Analysis Steps

2.1 Prerequisites

2.1.1 Dowloading Datasets

To download the datasets we will take advantage of rpx R package. Note, this step may take awhile (10-30 min) depending on the speed of the internet connection. However, if they are downloaded the script will use the available datasets instead of downloading them again.

```
library(rpx)
id <- "PXD004333"
px <- PXDataset(id)
repoFiles <- pxfiles(px)
mzids <- grep('*msgfplus.mzid.gz', repoFiles, value=T)
system.time(pxget(px, mzids))
##    user  system elapsed
##   0.295   0.012   3.000
```

2.1.2 Reading Frameshift Marks

The FASTA files containing 595 sequences with frameshifts availabe as a part of this package and available as `system.file("extdata", "Euplotes_Crassus_frameshifts.fasta", package="EuplotesCrassus.proteome")`. There is an additional FASTA file with frameshift locations marked with exclamation mark `!`.

```
library(Biostrings)
fasta_clean <- readAAStringSet(
    system.file("extdata",
                "Euplotes_Crassus_frameshifts.fasta",
                package="EuplotesCrassus.proteome"),
    format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
fasta_marks <- readAAStringSet(
    system.file("extdata",
                "Euplotes_Crassus_frameshifts_with_mark.fasta",
                package="EuplotesCrassus.proteome"),
    format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
length(fasta_clean)
```

```
## [1] 595
```

2.2 Processing of MS/MS Search Results

2.2.1 Trypsin Digest Fractionated by SCX

For processing of MS/MS identification we will use MSnID R package. First step is to read the LC-MS/MS datasets corresponding to 25 SCX fractions.

```
library(MSnID)
trypscx <- grep('Euplotes_1_SCX_.*msgfplus.mzid.gz', repoFiles, value=T)
trypscxPrj <- MSnID()
system.time(trypscxPrj <- read_mzIDs(trypscxPrj, trypscx, backend = 'mzR'))
##    user  system elapsed
##   4.829   0.214   5.106
```

Assess the peptide termini for their corresponding cleavage patterns. We will lleave peptides that resuted only from proper trypsin cleavave events. That is we won't allow peptide resulting from irregular clevages.

```
trypscxPrj <- assess_termini(trypscxPrj, validCleavagePattern="[KR]\\.[^P]")
trypscxPrj <- apply_filter(trypscxPrj, "numIrregCleavages == 0")
```

Note, that for this project we are interested only in peptides covering the sites of the frameshifting events. So if a peptide identification can be explained by a regular protein sequence we are not interested in pursuing this identification. The protein/accession names of normal (non-frameshifted) sequences starts with `Contig` or `Contaminant`. If the FASTA entry sequence is a results of the frameshift event if starts with `comp`. Therefore in the code below we retain only peptide-to-spectrum matches that can appear only due to frameshifted sequences.

```
#' Rule on how to split the names.
#' Contig + Contaminants - main piece
#' comp - sequences with frameshifts
trypscxPrj.main <- apply_filter(trypscxPrj, "!grepl('comp', accession)")
trypscxPrj.fmsh <- apply_filter(trypscxPrj, "grepl('comp', accession)")
#' if peptide matches to the main piece we don't care about it
trypscxPrj.fmsh <- apply_filter(trypscxPrj.fmsh,
                                "!(peptide %in% peptides(trypscxPrj.main))")
show(trypscxPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  25
## #PSMs: 442 at 58 % FDR
## #peptides: 348 at 67 % FDR
## #accessions: 291 at 66 % FDR
```

Setting-up and optimizing filtering options for MS/MS identifications. Since the number of peptides mapping frameshifted sequences is rather low we will loosed up the FDR of the identification up to 5%, however, then follow-up with manual spectra validation.

```
trypscxPrj.fmsh$mme.ppm <- abs(mass_measurement_error(trypscxPrj.fmsh))
trypscxPrj.fmsh$score <- -log10(trypscxPrj.fmsh$`MS.GF.SpecEValue`)
trypscxPrj.fmsh <- apply_filter(trypscxPrj.fmsh, "mme.ppm < 10")

filtr <- MSnIDFilter(trypscxPrj.fmsh)
filtr$mme.ppm <- list(comparison="<", threshold=5.0)
filtr$score <- list(comparison=">", threshold=8.0)
```

```
#' pre-optimization with brute-force approach
filtr.grid <- optimize_filter(filtr, trypscxPrj.fmsh, fdr.max=0.05,
                              method="Grid", level="peptide", n.iter=20000)
evaluate_filter(trypscxPrj.fmsh, filtr.grid)
```

```
##                  fdr   n
## PSM       0.02970297 104
## peptide   0.03703704  56
## accession 0.04166667  50
```

```
#' fine tune with optimization using simulated annealing technique
filtr.sann <- optimize_filter(filtr.grid, trypscxPrj.fmsh, fdr.max=0.05,
                              method="SANN", level="peptide", n.iter=20000)
evaluate_filter(trypscxPrj.fmsh, filtr.sann)
```

```
##                  fdr   n
## PSM       0.02941176 105
## peptide   0.03636364  57
## accession 0.04081633  51
```

```
trypscxPrj.fmsh <- apply_filter(trypscxPrj.fmsh, filtr.sann)
show(trypscxPrj.fmsh)
```

```
## MSnID object
## Working directory: "."
## #Spectrum Files:  18
## #PSMs: 105 at 2.9 % FDR
## #peptides: 57 at 3.6 % FDR
## #accessions: 51 at 4.1 % FDR
```

Finally we will extract only those peptides that exactly span the frameshift sites. That is their sequences should be present/identifiable in normal FASTA file, however missing in the file with frameshifts masked with the exclamation mark !.

```
#' extract only those that map frameshift sites
library(dplyr)
pepSeq <- unique(trypscxPrj.fmsh$pepSeq)
pepSeqMapped_to_clean <- pepSeq %>%
    sapply(grep, x=fasta_clean) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqMapped_to_with_marks <- pepSeq %>%
    sapply(grep, x=fasta_marks) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqFmsh_trypscx <- setdiff(pepSeqMapped_to_clean, pepSeqMapped_to_with_marks)
print(pepSeqFmsh_trypscx)
```

```
## [1] "SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK" "WTPIDLPSEEITFVQGIQTVTGAGDPSMK"
## [3] "ESNHNNDITNKNEIAYILR"             "KKKQEENNLKR"
```

Reporting extra information on the peptide sequences spanning frameshift sites: dataset, scan, charge, score, and mass measurement error.

```
meta_tryp_scx <- trypscxPrj.fmsh %>%
    apply_filter('pepSeq %in% pepSeqFmsh_trypscx') %>%
    psms %>%
```

```
    select(spectrumFile,MS.GF.SpecEValue,mme.ppm,spectrumID,chargeState,peptide) %>%
    rename(SpecEValue = MS.GF.SpecEValue, charge = chargeState, `MME (ppm)`=mme.ppm) %>%
    mutate(spectrumFile = sub('_msgfplus.mzid.gz','',spectrumFile))
library(xtable)
print(xtable(meta_tryp_scx, display = c('d','s','e','f','s','d','s')),
      include.rownames=FALSE,
      comment = FALSE,
      size='scriptsize',
      floating = F)
```

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.53e-21	0.08	index=8908	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.07e-20	1.10	index=8896	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	7.29e-19	1.10	index=8897	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	2.17e-15	0.94	index=8895	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_18_13Nov09_Falcon_09-09-15	9.27e-17	0.11	index=5912	2	K.ESNHNNDITNKNEIAYILR.Y
Euplotes_1_SCX_20_13Nov09_Falcon_09-09-15	2.23e-11	0.70	index=10317	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_22_13Nov09_Falcon_09-09-15	4.36e-10	3.76	index=9720	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_23_13Nov09_Falcon_09-09-15	2.47e-09	1.64	index=9440	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_24_13Nov09_Falcon_09-09-15	3.42e-10	8.85	index=2127	3	R.KKKQEENNLKR.K

2.2.2 Trypsin Digest Fractionated by HPRP

All the processing steps are conceptually the same as in the section above.

```
tryphprp <- grep('Euplotes_1_HPRP_1_.*msgfplus.mzid.gz', repoFiles, value=T)
tryphprpPrj <- MSnID()
system.time(tryphprpPrj <- read_mzIDs(tryphprpPrj, tryphprp, backend = 'mzR'))
##    user  system elapsed
##   2.716   0.175   2.945
```

```
tryphprpPrj <- assess_termini(tryphprpPrj, validCleavagePattern="[KR]\\.[^P]")
tryphprpPrj <- apply_filter(tryphprpPrj, "numIrregCleavages == 0")
```

```
tryphprpPrj.main <- apply_filter(tryphprpPrj, "!grepl('comp', accession)")
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj, "grepl('comp', accession)")
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj.fmsh,
                                 "!(peptide %in% peptides(tryphprpPrj.main))")
show(tryphprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  24
## #PSMs: 511 at 49 % FDR
## #peptides: 399 at 62 % FDR
## #accessions: 293 at 78 % FDR
```

```
tryphprpPrj.fmsh$mme.ppm <- abs(mass_measurement_error(tryphprpPrj.fmsh))
tryphprpPrj.fmsh$score <- -log10(tryphprpPrj.fmsh$`MS.GF.SpecEValue`)
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj.fmsh, "mme.ppm < 10")

filtr <- MSnIDFilter(tryphprpPrj.fmsh)
filtr$mme.ppm <- list(comparison="<", threshold=5.0)
filtr$score <- list(comparison=">", threshold=8.0)
filtr.grid <- optimize_filter(filtr, tryphprpPrj.fmsh, fdr.max=0.05,
                              method="Grid", level="peptide", n.iter=20000)
evaluate_filter(tryphprpPrj.fmsh, filtr.grid)
##                  fdr   n
## PSM       0.02631579 195
## peptide   0.04504505 116
## accession 0.07142857  75
```

```
filtr.sann <- optimize_filter(filtr.grid, tryphprpPrj.fmsh, fdr.max=0.05,
                              method="SANN", level="peptide", n.iter=20000)
evaluate_filter(tryphprpPrj.fmsh, filtr.sann)
##                  fdr   n
## PSM       0.02604167 197
## peptide   0.04504505 116
## accession 0.07142857  75
```

```
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj.fmsh, filtr.sann)
show(tryphprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  23
## #PSMs: 197 at 2.6 % FDR
## #peptides: 116 at 4.5 % FDR
```

```
## #accessions: 75 at 7.1 % FDR
```

```
library(dplyr)
pepSeq <- unique(tryphprpPrj.fmsh$pepSeq)
pepSeqMapped_to_clean <- pepSeq %>%
    sapply(grep, x=fasta_clean) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqMapped_to_with_marks <- pepSeq %>%
    sapply(grep, x=fasta_marks) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqFmsh_tryphprp <- setdiff(pepSeqMapped_to_clean, pepSeqMapped_to_with_marks)
print(pepSeqFmsh_tryphprp)
```

```
##  [1] "FFAAPEK"                       "ELAFLKRAQEIGLEPYNEYHGKKK"
##  [3] "VVQEGNTNVKK"                   "WTPIDLPSEEITFVQGIQTVTGAGDPSMK"
##  [5] "IIQNFQINTVFEDLDEIMQTQVQR"      "KSSKACEEERRKR"
##  [7] "LINDLTNDK"                     "LISELTSEK"
##  [9] "IVENFNK"                       "LSQEHLSYISR"
## [11] "LINDLTNDKANLK"
```

```
meta_tryp_hprp <- tryphprpPrj.fmsh %>%
    apply_filter('pepSeq %in% pepSeqFmsh_tryphprp') %>%
    psms %>%
    select(spectrumFile,MS.GF.SpecEValue,mme.ppm,spectrumID,chargeState,peptide) %>%
    rename(SpecEValue = MS.GF.SpecEValue, charge = chargeState, `MME (ppm)`=mme.ppm) %>%
    mutate(spectrumFile = sub('_msgfplus.mzid.gz','',spectrumFile))
library(xtable)
print(xtable(meta_tryp_hprp, display = c('d','s','e','f','s','d','s')),
      include.rownames=FALSE,
      comment = FALSE,
      size='scriptsize',
      floating = F)
```

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	7.58e-11	0.08	index=3031	1	R.FFAAPEK.I
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	2.44e-09	0.00	index=3046	2	R.FFAAPEK.I
Euplotes_1_HPRP_1_05_17Nov09_Falcon_09-09-14	1.46e-09	5.31	index=8245	3	R.ELAFLKRAQEIGLEPYNEYHGKKK.T
Euplotes_1_HPRP_1_06_17Nov09_Falcon_09-09-14	5.54e-10	2.21	index=759	2	K.VVQEGNTNVKK.L
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	5.93e-22	2.11	index=8644	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	2.18e-21	0.78	index=8638	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	3.05e-21	2.11	index=8646	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	4.19e-16	0.82	index=8639	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.19e-21	0.70	index=8806	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.20e-21	1.57	index=8812	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	5.49e-20	1.64	index=8802	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	4.33e-15	1.53	index=8810	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	4.51e-21	0.33	index=10684	2	K.IIQNFQINTVFEDLDEIMQTQVQR.H
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	1.36e-11	1.25	index=10678	3	K.IIQNFQINTVFEDLDEIMQTQVQR.H
Euplotes_1_HPRP_1_18_17Nov09_Falcon_09-09-15	5.08e-09	2.64	index=13785	2	K.KSSKACEEERRKR.E
Euplotes_1_HPRP_1_20_17Nov09_Falcon_09-09-15	1.91e-11	0.00	index=3425	1	K.LINDLTNDK.A
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	6.65e-11	1.67	index=3600	2	K.LISELTSEK.S
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	2.55e-10	0.78	index=3602	1	K.LISELTSEK.S
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	1.89e-09	0.49	index=2595	2	K.IVENFNK.I
Euplotes_1_HPRP_1_23_17Nov09_Falcon_09-09-15	3.01e-13	1.01	index=2200	2	K.LSQEHLSYISR.L
Euplotes_1_HPRP_1_24_17Nov09_Falcon_09-09-15	2.45e-16	1.41	index=2709	2	K.LINDLTNDKANLK.D

2.2.3 Glu-C Digest Fractionated by HPRP

All the processing steps are conceptually the same as in the section above. The only substantial diffence is the specification of the enzyme digestion rule.

```
gluchprp <- grep('Euplotes_1_HPRP_2_.*msgfplus.mzid.gz', repoFiles, value=T)
gluchprpPrj <- MSnID()
system.time(gluchprpPrj <- read_mzIDs(gluchprpPrj, gluchprp, backend = 'mzR'))
##    user  system elapsed
##   2.780   0.190   3.027
```

```
gluchprpPrj <- assess_termini(gluchprpPrj, validCleavagePattern="E\\.[^P]")
gluchprpPrj <- apply_filter(gluchprpPrj, "numIrregCleavages == 0")
```

```
gluchprpPrj.main <- apply_filter(gluchprpPrj, "!grepl('comp', accession)")
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj, "grepl('comp', accession)")
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj.fmsh,
                                 "!(peptide %in% peptides(gluchprpPrj.main))")
show(gluchprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  24
## #PSMs: 555 at 67 % FDR
## #peptides: 440 at 80 % FDR
## #accessions: 297 at 89 % FDR
```

```
gluchprpPrj.fmsh$mme.ppm <- abs(mass_measurement_error(gluchprpPrj.fmsh))
gluchprpPrj.fmsh$score <- -log10(gluchprpPrj.fmsh$`MS.GF.SpecEValue`)
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj.fmsh, "mme.ppm < 10")

filtr <- MSnIDFilter(gluchprpPrj.fmsh)
filtr$mme.ppm <- list(comparison="<", threshold=5.0)
filtr$score <- list(comparison=">", threshold=8.0)
filtr.grid <- optimize_filter(filtr, gluchprpPrj.fmsh, fdr.max=0.05,
                              method="Grid", level="peptide", n.iter=20000)
evaluate_filter(gluchprpPrj.fmsh, filtr.grid)
##                   fdr  n
## PSM       0.02222222 46
## peptide   0.03448276 30
## accession 0.05000000 21
```

```
filtr.sann <- optimize_filter(filtr.grid, gluchprpPrj.fmsh, fdr.max=0.05,
                              method="SANN", level="peptide", n.iter=20000)
evaluate_filter(gluchprpPrj.fmsh, filtr.sann)
##                   fdr  n
## PSM       0.02222222 46
## peptide   0.03448276 30
## accession 0.05000000 21
```

```
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj.fmsh, filtr.sann)
show(gluchprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  18
## #PSMs: 46 at 2.2 % FDR
```

```
## #peptides: 30 at 3.4 % FDR
## #accessions: 21 at 5 % FDR
```

```
library(dplyr)
pepSeq <- unique(gluchprpPrj.fmsh$pepSeq)
pepSeqMapped_to_clean <- pepSeq %>%
    sapply(grep, x=fasta_clean) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqMapped_to_with_marks <- pepSeq %>%
    sapply(grep, x=fasta_marks) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqFmsh_gluchprp <- setdiff(pepSeqMapped_to_clean, pepSeqMapped_to_with_marks)
print(pepSeqFmsh_gluchprp)
```

```
## [1] "NFNKITGKEQEEEE"                    "SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE"
## [3] "NLDNEKLINDLTNDKANLKDIVFDLMFE"      "NKIRFFAAPEKIFE"
## [5] "MQDEEILKSIEESKLEQEQEEEKKNE"        "VYLGLMEEYE"
```

```
meta_gluc_hprp <- gluchprpPrj.fmsh %>%
    apply_filter('pepSeq %in% pepSeqFmsh_gluchprp') %>%
    psms %>%
    select(spectrumFile,MS.GF.SpecEValue,mme.ppm,spectrumID,chargeState,peptide) %>%
    rename(SpecEValue = MS.GF.SpecEValue, charge = chargeState, `MME (ppm)`=mme.ppm) %>%
    mutate(spectrumFile = sub('_msgfplus.mzid.gz','',spectrumFile))
library(xtable)
print(xtable(meta_gluc_hprp, display = c('d','s','e','f','s','d','s')),
      include.rownames=FALSE,
      comment = FALSE,
      size='scriptsize',
      floating = F)
```

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide
Euplotes_1_HPRP_2_06_22Nov09_Falcon_09-09-15	6.80e-07	2.95	index=13369	2	E.NFNKITGKEQEEEE.Y
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.78e-17	0.19	index=9982	3	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.33e-07	0.57	index=9974	4	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.74e-16	0.44	index=10771	3	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.03e-07	1.11	index=10770	4	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	2.09e-09	0.43	index=3933	3	E.NKIRFFAAPEKIFE.T
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	1.62e-07	0.07	index=3930	2	E.NKIRFFAAPEKIFE.T
Euplotes_1_HPRP_2_15_17Nov09_Falcon_09-09-17	2.83e-07	1.61	index=1758	2	E.MQDEEILKSIEESKLEQEQEEEKKNE.E
Euplotes_1_HPRP_2_21_22Nov09_Falcon_09-09-17	2.17e-07	0.10	index=6671	1	E.VYLGLMEEYE.A
Euplotes_1_HPRP_2_22_22Nov09_Falcon_09-09-17	2.12e-08	0.88	index=6753	1	E.VYLGLMEEYE.A

2.3 Compendium of Peptides Covering Frameshift Locations

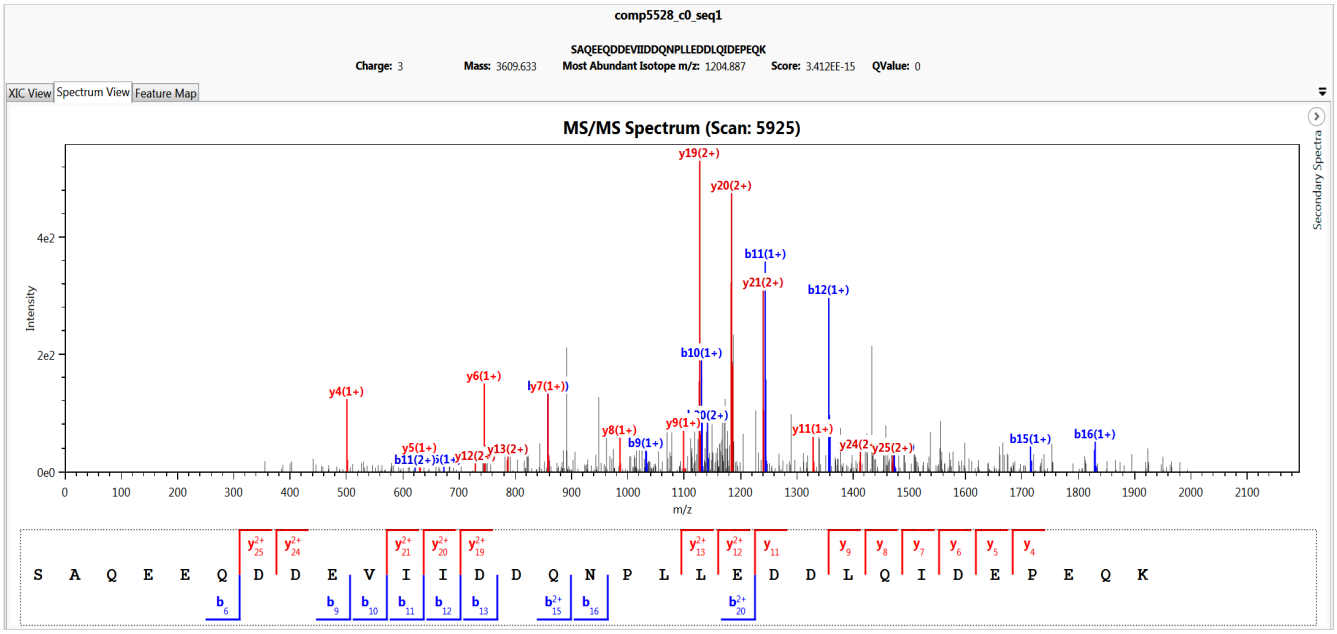
Final set of peptides and corresponding references to LC-MS/MS datasets and spectra. Overall, **4**, **11**, and **6** unique peptide sequences spanning the frameshift sites were identified in `trypsin/SCX`, `trypsin/HPRP`, and 'Glu-C/HPRP' experiments, respectively.

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide	experiment
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V	trypsin/SCX
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.53e-21	0.08	index=8908	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.07e-20	1.10	index=8896	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	7.29e-19	1.10	index=8897	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	2.17e-15	0.94	index=8895	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_18_13Nov09_Falcon_09-09-15	9.27e-17	0.11	index=5912	2	K.ESNHNNDITNKNEIAYILR.Y	trypsin/SCX
Euplotes_1_SCX_20_13Nov09_Falcon_09-09-15	2.23e-11	0.70	index=10317	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_22_13Nov09_Falcon_09-09-15	4.36e-10	3.76	index=9720	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_23_13Nov09_Falcon_09-09-15	2.47e-09	1.64	index=9440	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_24_13Nov09_Falcon_09-09-15	3.42e-10	8.85	index=2127	3	R.KKKQEENNLKR.K	trypsin/SCX
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	7.58e-11	0.08	index=3031	1	R.FFAAPEK.I	trypsin/HPRP
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	2.44e-09	0.00	index=3046	2	R.FFAAPEK.I	trypsin/HPRP
Euplotes_1_HPRP_1_05_17Nov09_Falcon_09-09-14	1.46e-09	5.31	index=8245	3	R.ELAFLKRAQEIGLEPYNEYHGKKK.T	trypsin/HPRP
Euplotes_1_HPRP_1_06_17Nov09_Falcon_09-09-14	5.54e-10	2.21	index=759	2	K.VVQEGNTNVKK.L	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	5.93e-22	2.11	index=8644	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	2.18e-21	0.78	index=8638	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	3.05e-21	2.11	index=8646	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	4.19e-16	0.82	index=8639	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.19e-21	0.70	index=8806	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.20e-21	1.57	index=8812	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	5.49e-20	1.64	index=8802	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	4.33e-15	1.53	index=8810	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	4.51e-21	0.33	index=10684	2	K.IIQNFQINTVFEDLDEIMQTQVQR.H	trypsin/HPRP
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	1.36e-11	1.25	index=10678	3	K.IIQNFQINTVFEDLDEIMQTQVQR.H	trypsin/HPRP
Euplotes_1_HPRP_1_18_17Nov09_Falcon_09-09-15	5.08e-09	2.64	index=13785	2	K.KSSKACEEERRKR.E	trypsin/HPRP
Euplotes_1_HPRP_1_20_17Nov09_Falcon_09-09-15	1.91e-11	0.00	index=3425	1	K.LINDLTNDK.A	trypsin/HPRP
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	6.65e-11	1.67	index=3600	2	K.LISELTSEK.S	trypsin/HPRP
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	2.55e-10	0.78	index=3602	1	K.LISELTSEK.S	trypsin/HPRP
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	1.89e-09	0.49	index=2595	2	K.IVENFNK.I	trypsin/HPRP
Euplotes_1_HPRP_1_23_17Nov09_Falcon_09-09-15	3.01e-13	1.01	index=2200	2	K.LSQEHLSYISR.L	trypsin/HPRP
Euplotes_1_HPRP_1_24_17Nov09_Falcon_09-09-15	2.45e-16	1.41	index=2709	2	K.LINDLTNDKANLK.D	trypsin/HPRP
Euplotes_1_HPRP_2_06_22Nov09_Falcon_09-09-15	6.80e-07	2.95	index=13369	2	E.NFNKITGKEQEEEE.Y	Glu-C/HPRP
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.78e-17	0.19	index=9982	3	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.33e-07	0.57	index=9974	4	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.74e-16	0.44	index=10771	3	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.03e-07	1.11	index=10770	4	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	2.09e-09	0.43	index=3933	3	E.NKIRFFAAPEKIFE.T	Glu-C/HPRP
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	1.62e-07	0.07	index=3930	2	E.NKIRFFAAPEKIFE.T	Glu-C/HPRP
Euplotes_1_HPRP_2_15_17Nov09_Falcon_09-09-17	2.83e-07	1.61	index=1758	2	E.MQDEEILKSIEESKLEQEQEEEKKNE.E	Glu-C/HPRP
Euplotes_1_HPRP_2_21_22Nov09_Falcon_09-09-17	2.17e-07	0.10	index=6671	1	E.VYLGLMEEYE.A	Glu-C/HPRP
Euplotes_1_HPRP_2_22_22Nov09_Falcon_09-09-17	2.12e-08	0.88	index=6753	1	E.VYLGLMEEYE.A	Glu-C/HPRP

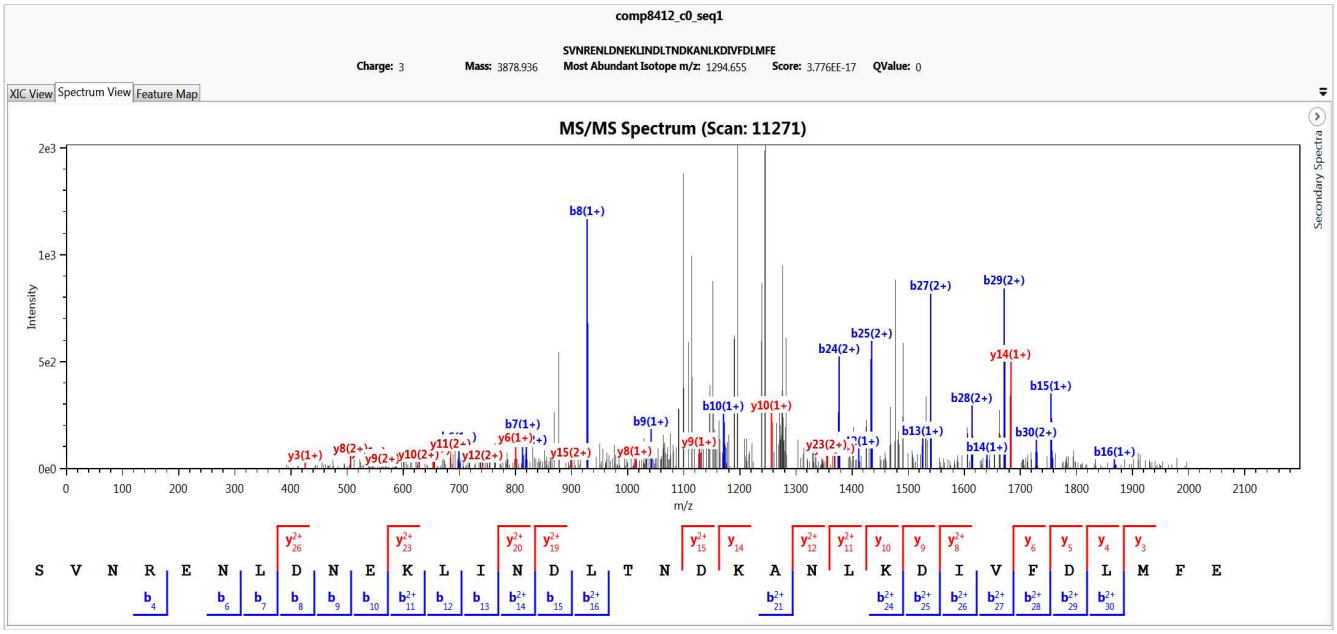
3 Manual Validation

Manual valiation was perfomed by LCMSSpectator. The spectra that have passed the consensus opinion of 5 independed experts are shown below. Necessary raw and mzIdenML files to reproduce the analysis are available at http://dx.doi.org/10.6019/PXD004333. Note, the `MS/MS scan number` is not the same identifier as `spectrumID` in the table above.

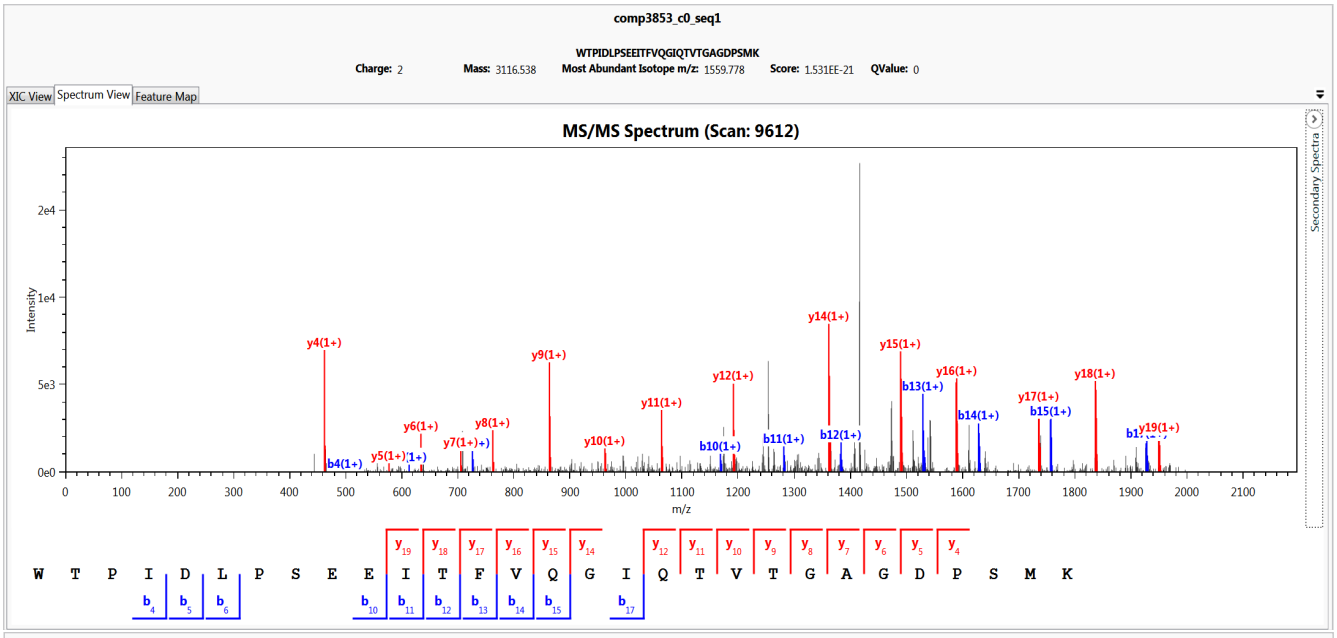
SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK

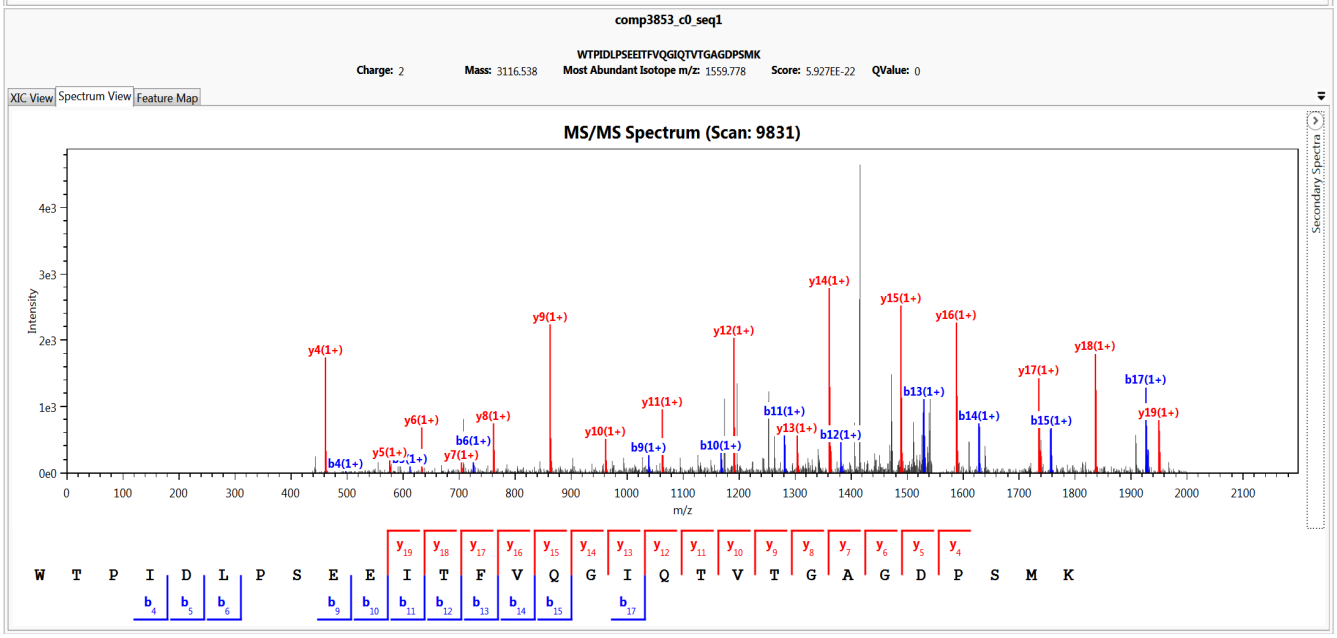


SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE

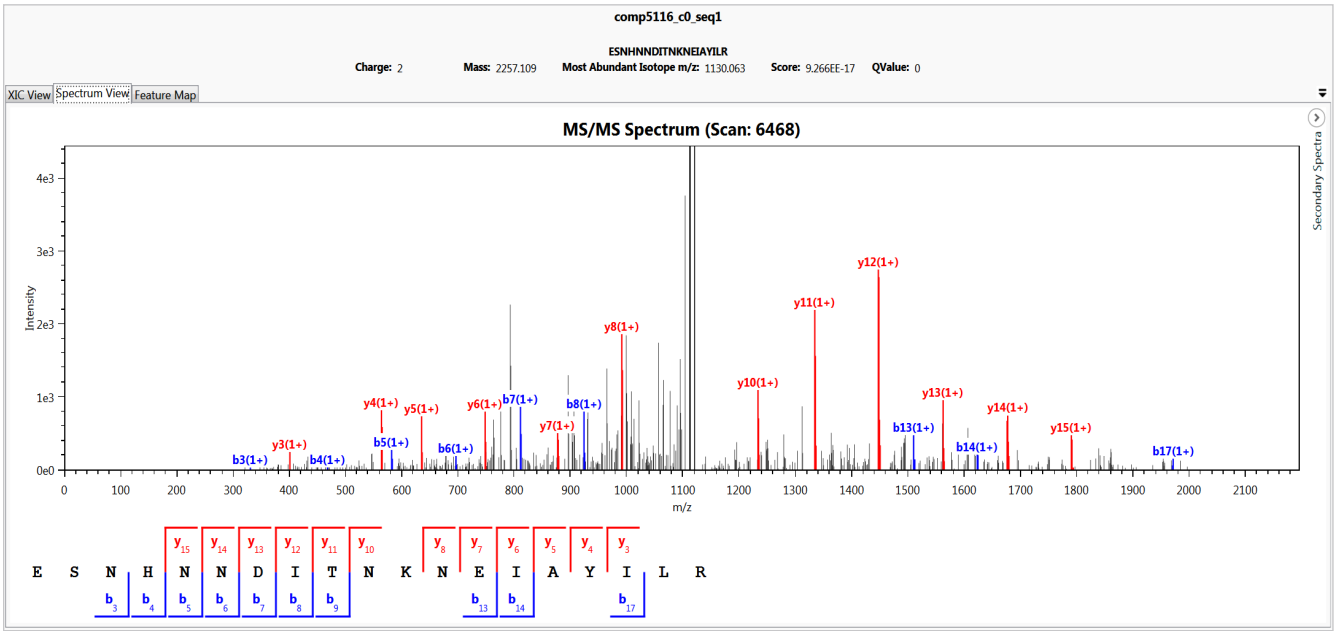


WTPIDLPSEEITFVQGIQTVTGAGDPSMK

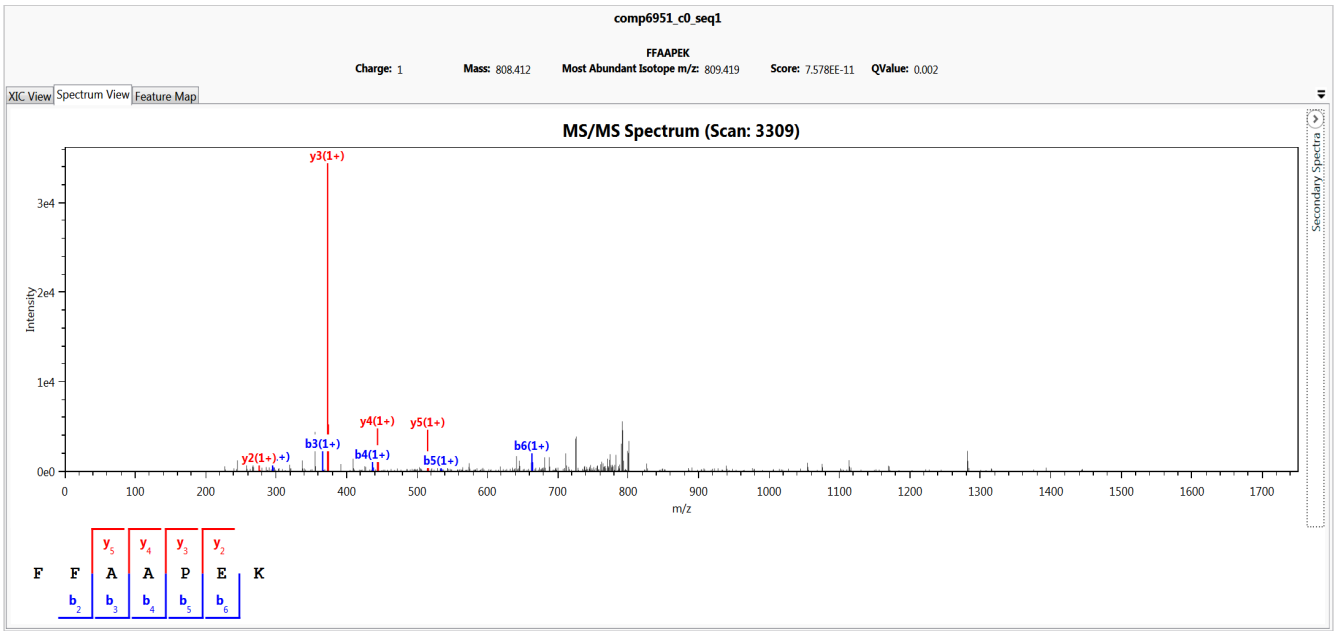




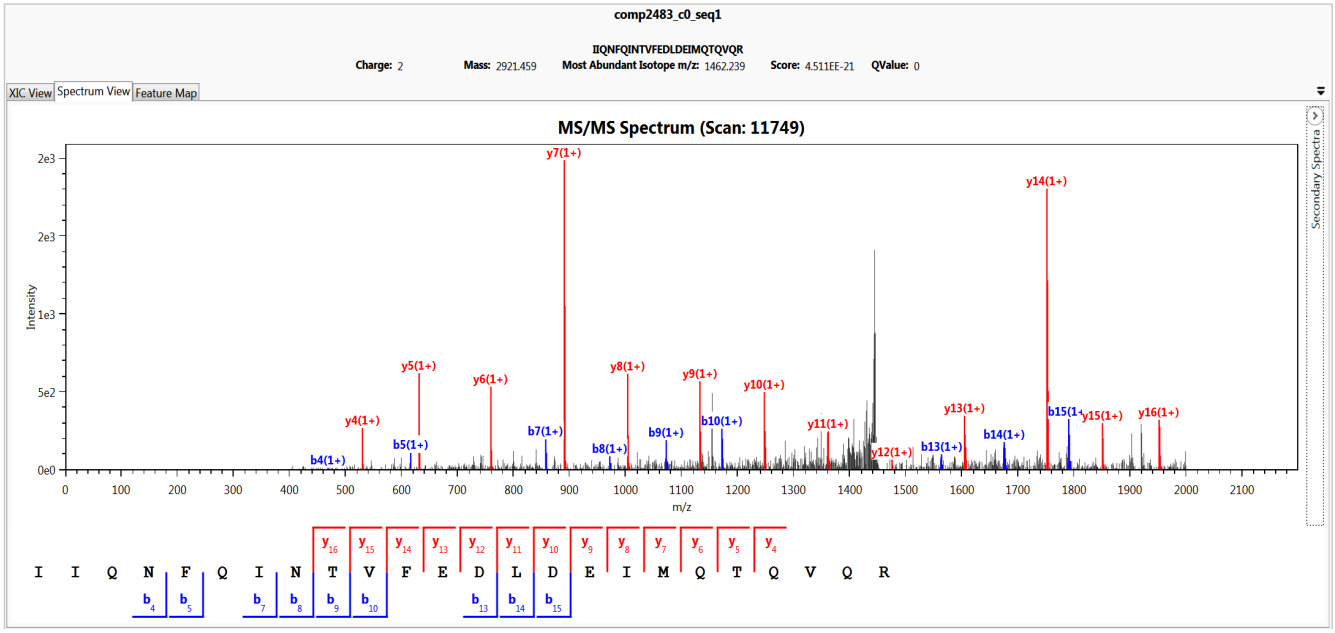
ESNHNNDITNKNEIAYILR



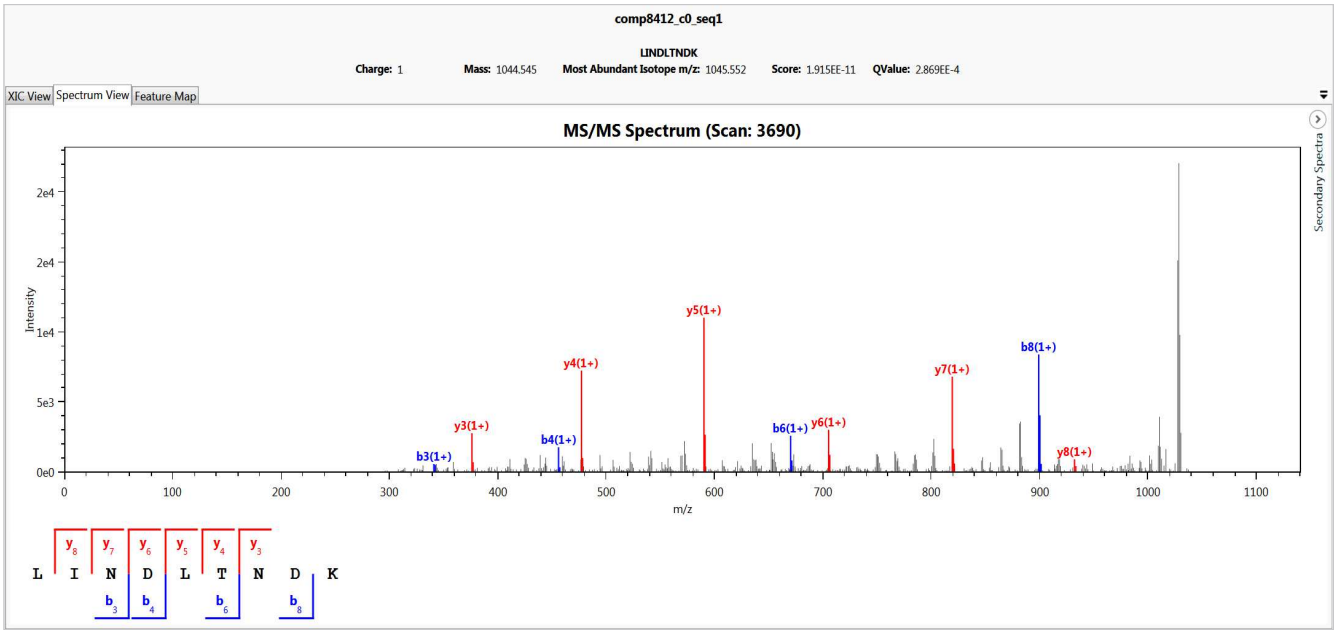
FFAAPEK



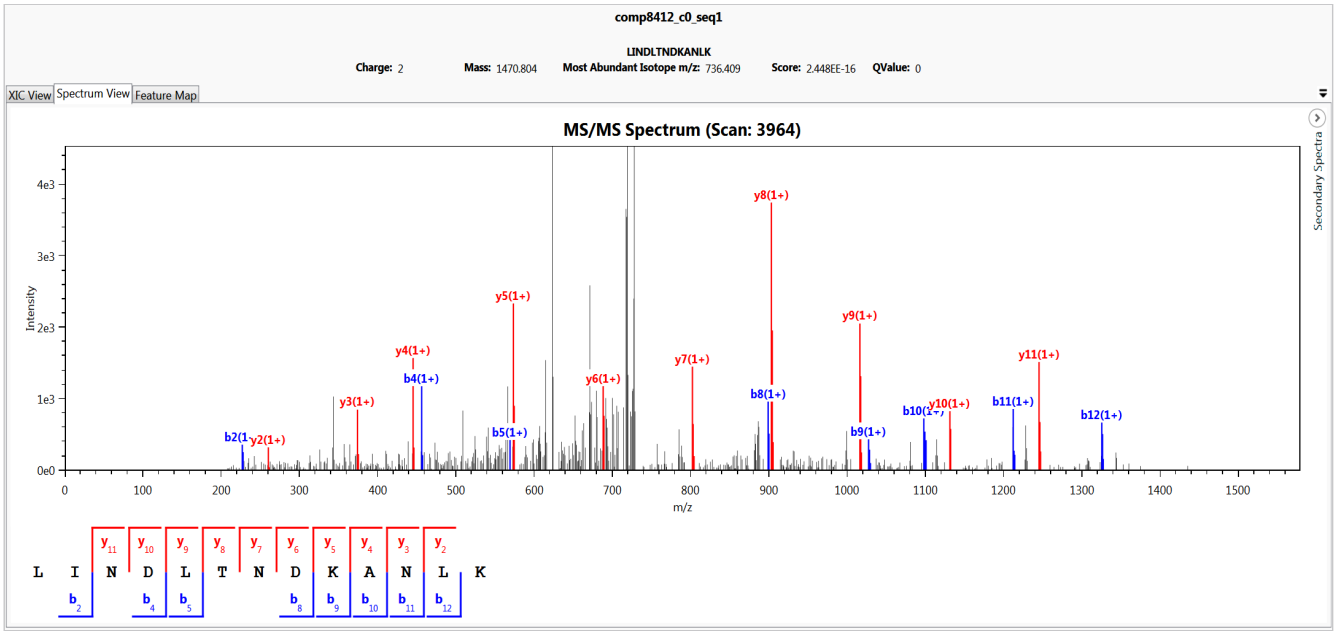
IIQNFQINTVFEDLDEIMQTQVQR



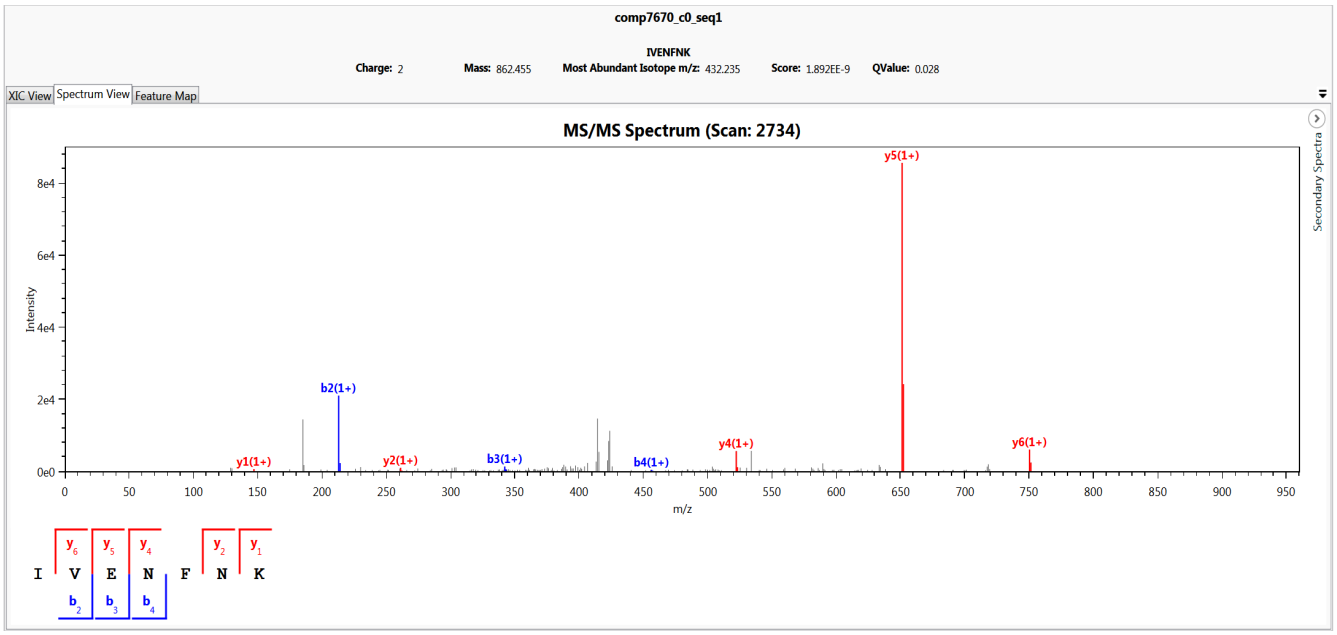
LINDLTNDK



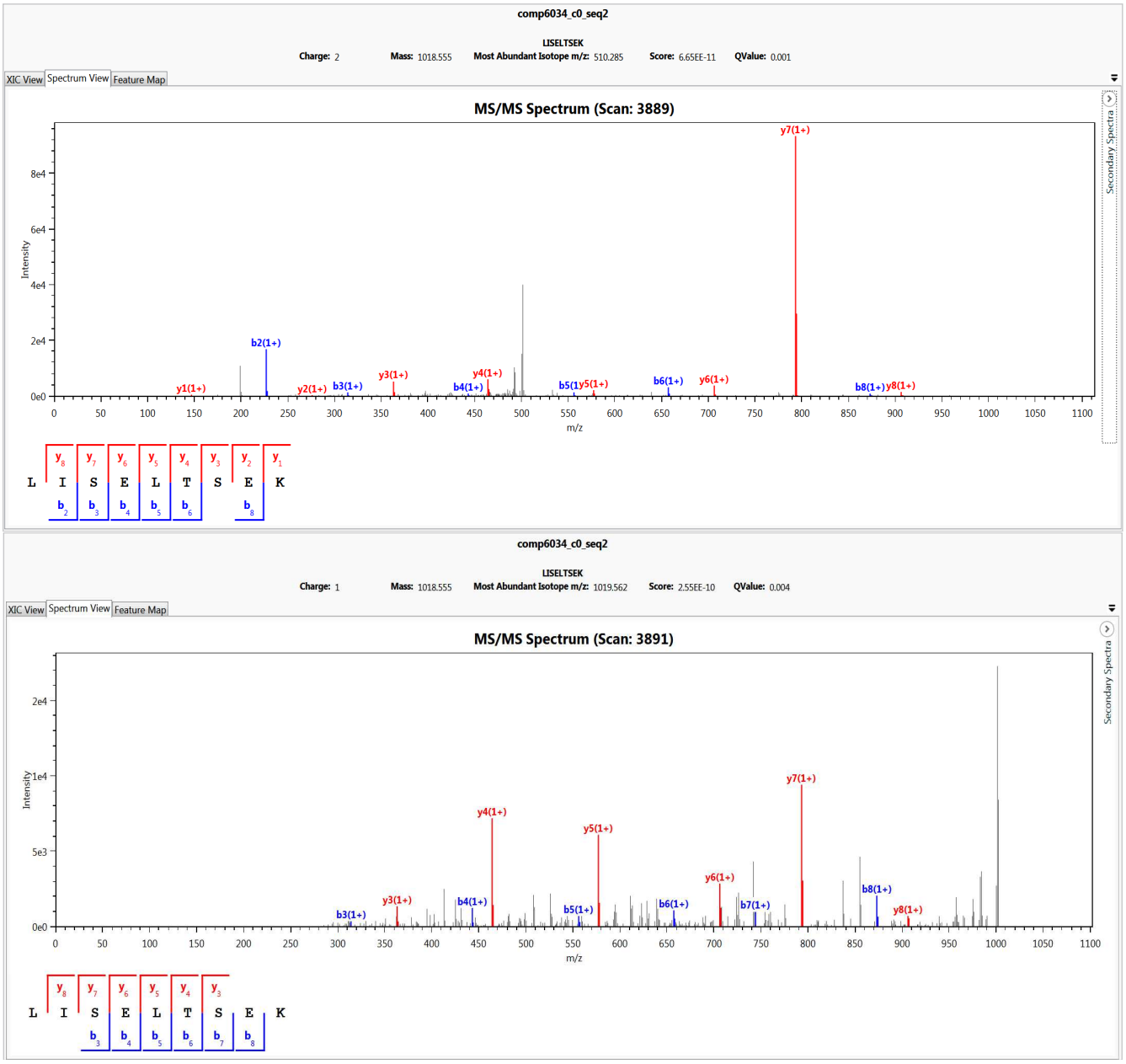
LINDLTNDKANLK



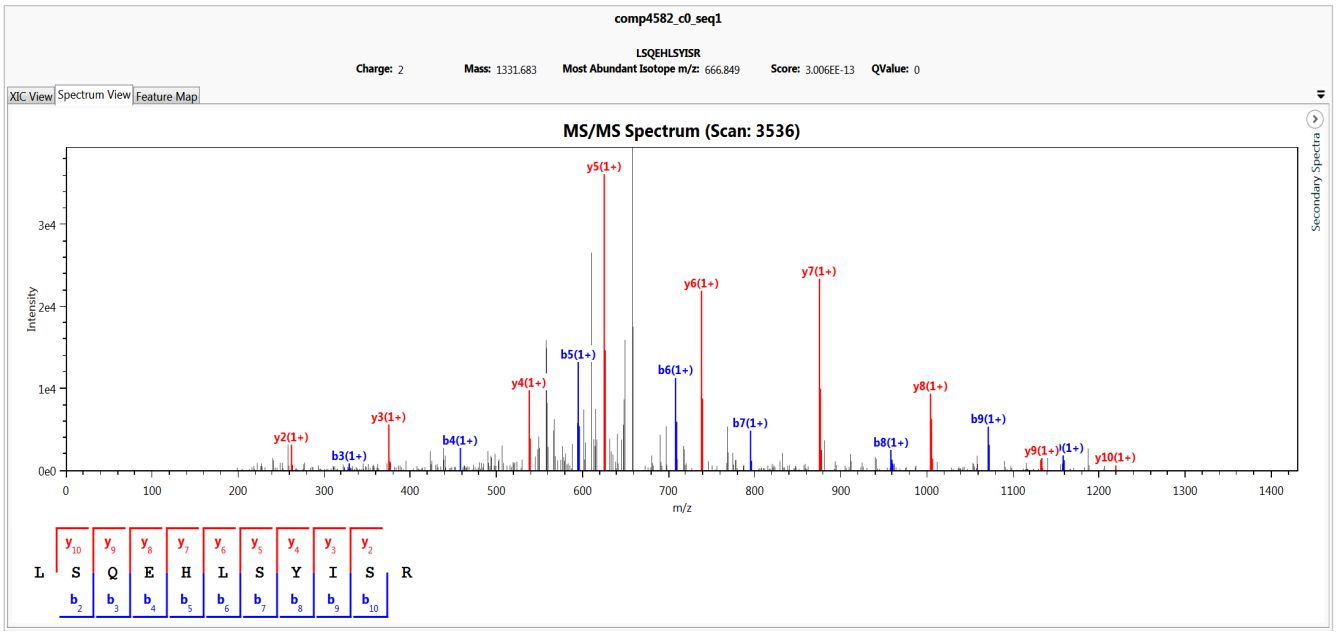
IVENFNK



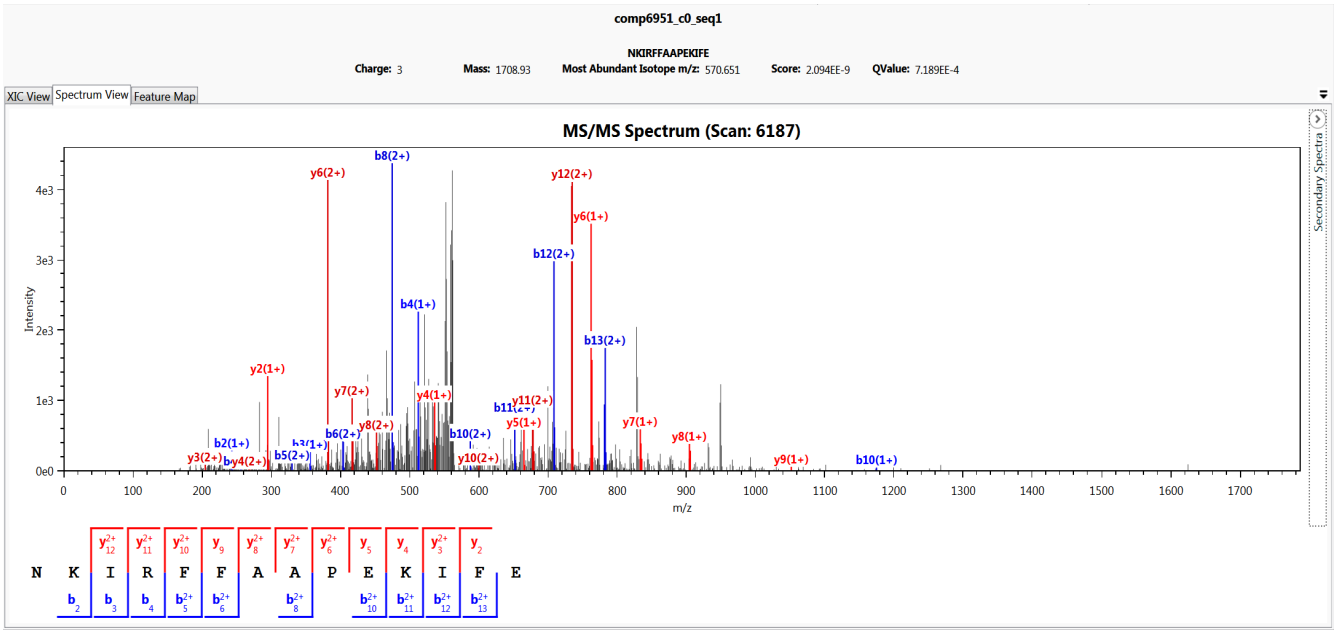
LISELTSEK


comp6034_c0_seq2
LISELTSEK
Charge: 2
Mass: 1018.555
Most Abundant Isotope m/z: 510.285
Score: 6.65EE-11
QValue: 0.001
XIC View
Spectrum View
Feature Map
MS/MS Spectrum (Scan: 3889)
Intensity
m/z
Secondary Spectra
y7(1+)
b2(1+)
y1(1+)
y2(1+)
b3(1+)
y3(1+)
y4(1+)
b4(1+)
b5(1+)
y5(1+)
b6(1+)
y6(1+)
b8(1+)
y8(1+)
L I S E L T S E K
comp6034_c0_seq2
LISELTSEK
Charge: 1
Mass: 1018.555
Most Abundant Isotope m/z: 1019.562
Score: 2.55EE-10
QValue: 0.004
XIC View
Spectrum View
Feature Map
MS/MS Spectrum (Scan: 3891)
Intensity
m/z
Secondary Spectra
y7(1+)
y4(1+)
y5(1+)
y6(1+)
b8(1+)
y3(1+)
b4(1+)
b3(1+)
b5(1+)
b6(1+)
b7(1+)
y8(1+)
L I S E L T S E K

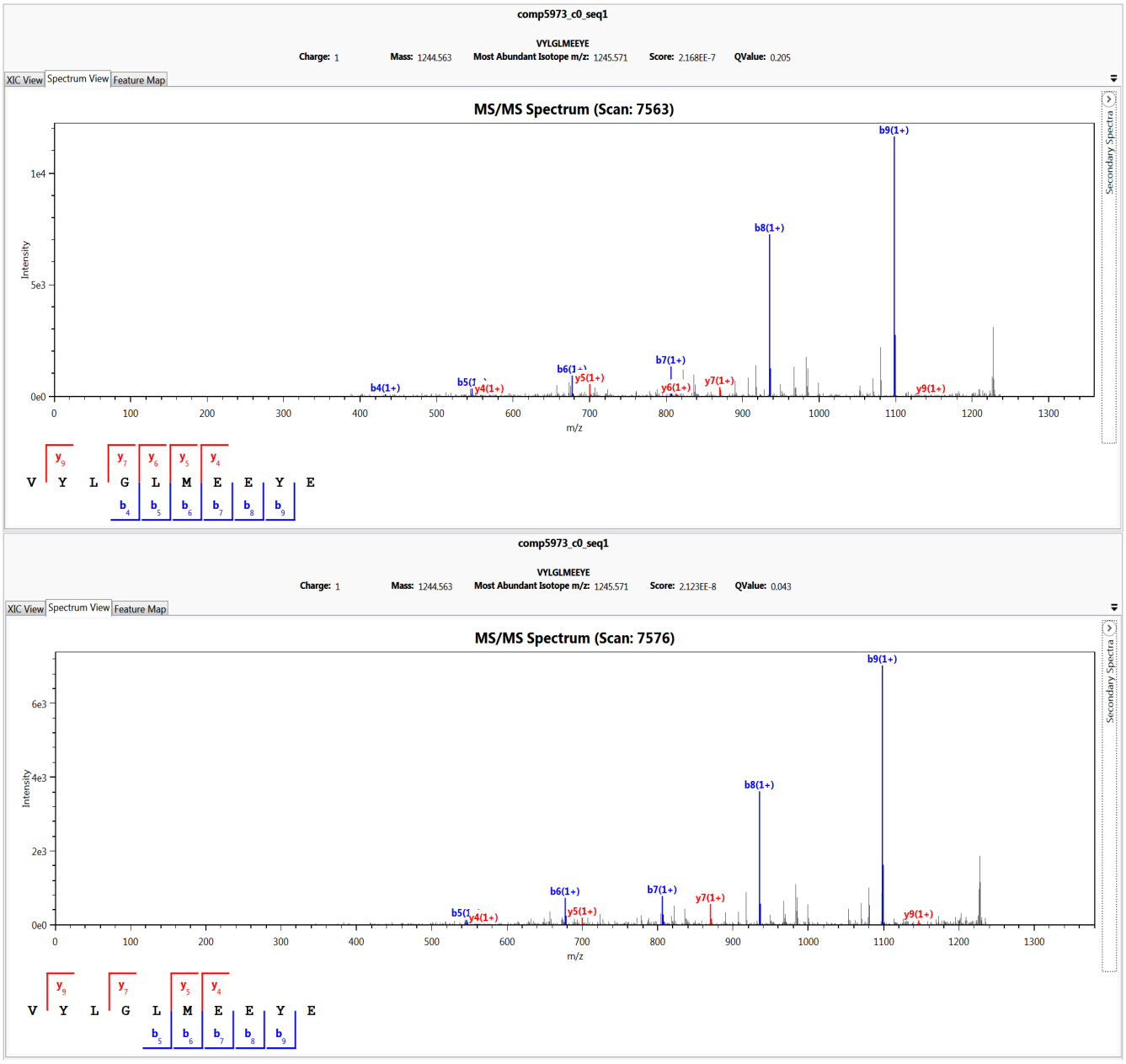
LSQEHLSYISR



NKIRFFAAPEKIFE



VYLGLMEEYE

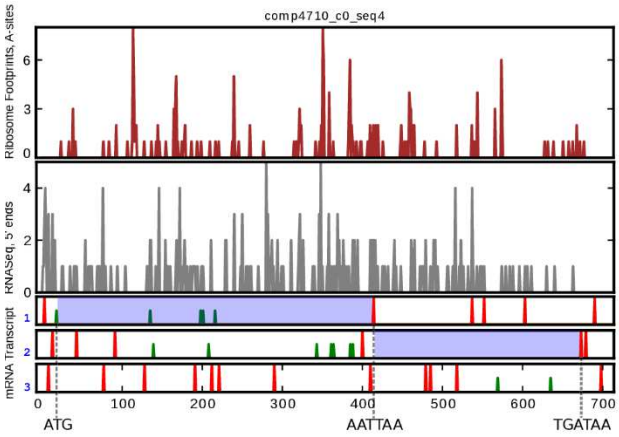

comp5973_c0_seq1
VYLGLMEEYE
Charge: 1
Mass: 1244.563
Most Abundant Isotope m/z: 1245.571
Score: 2.168EE-7
QValue: 0.205
XIC View
Spectrum View
Feature Map
MS/MS Spectrum (Scan: 7563)
Intensity
m/z
Secondary Spectra
b9(1+)
b8(1+)
b7(1+)
b6(1+)
y5(1+)
y6(1+)
y7(1+)
b4(1+)
y4(1+)
y9(1+)
V Y L G L M E E Y E
comp5973_c0_seq1
VYLGLMEEYE
Charge: 1
Mass: 1244.563
Most Abundant Isotope m/z: 1245.571
Score: 2.123EE-8
QValue: 0.043
XIC View
Spectrum View
Feature Map
MS/MS Spectrum (Scan: 7576)
Intensity
m/z
Secondary Spectra
b9(1+)
b8(1+)
b6(1+)
b7(1+)
y7(1+)
y5(1+)
y9(1+)
V Y L G L M E E Y E

4 Session Information

All software and respective versions used in this document, as returned by sessionInfo() are detailed below.

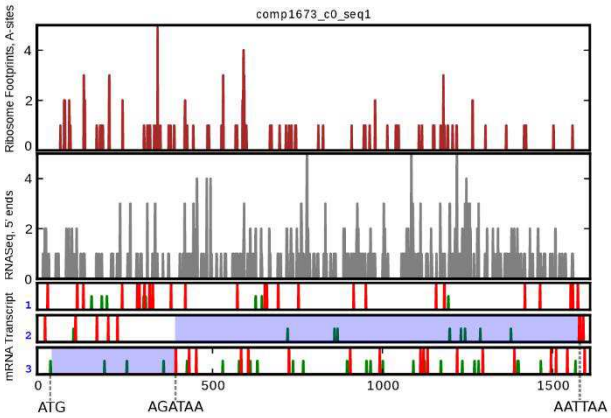
- R version 3.2.4 (2016-03-10), `x86_64-apple-darwin13.4.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.16.1, BiocStyle 1.8.0, Biostrings 2.38.4, dplyr 0.5.0, IRanges 2.4.8, knitr 1.12.3, MSnID 1.7.3, Rcpp 0.12.7, rpx 1.6.0, S4Vectors 0.8.11, xtable 1.8-2, XVector 0.10.0
- Loaded via a namespace (and not attached): affy 1.48.0, affyio 1.40.0, assertthat 0.1, Biobase 2.30.0, BiocInstaller 1.20.3, BiocParallel 1.4.3, bitops 1.0-6, chron 2.3-47, codetools 0.2-14, colorspace 1.2-6, data.table 1.9.6, DBI 0.5-1, digest 0.6.10, doParallel 1.0.10, evaluate 0.8.3, foreach 1.4.3, formatR 1.3, futile.logger 1.4.3, futile.options 1.0.0, ggplot2 2.1.0.9000, grid 3.2.4, gtable 0.2.0, highr 0.5.1, htmltools 0.3.5, impute 1.44.0, iterators 1.0.8, lambda.r 1.1.9, lattice 0.20-33, lazyeval 0.2.0, limma 3.26.9, magrittr 1.5, MALDIquant 1.14, MSnbase 1.18.1, munsell 0.4.3, mzID 1.8.0, mzR 2.4.1, pcaMethods 1.60.0, plyr 1.8.4, preprocessCore 1.32.0, ProtGenerics 1.2.1, R.cache 0.12.0, R.methodsS3 1.7.1, R.oo 1.20.0, R.utils 2.3.0, R6 2.1.2, RCurl 1.95-4.8, reshape2 1.4.1, rmarkdown 0.9.5, scales 0.4.0, stringi 1.1.1, stringr 1.1.0, tibble 1.2, tools 3.2.4, vsn 3.38.0, XML 3.98-1.4, yaml 2.1.13, zlibbioc 1.16.0

SUPPLEMENTARY NOTE 3. Representative profiles of ribosome density mapped to *E. crasus* transcripts and supporting BLAST hits alignments.



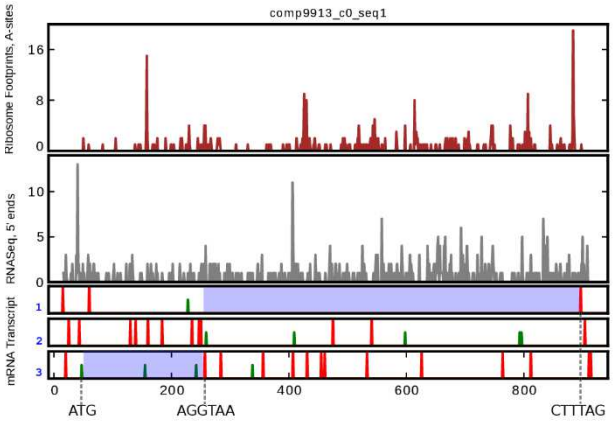

```
>comp4710_c0_seq4_AATTAA(+1)
MAEFASDTLDQSSSSAPVKYKVVFLGDQSVGKTSIILRFMHDTFDSQYQATIGI
DFLSKTMMVDDKMVRLQLWDTAGQERFKSLIPSYIQDSSFAIVCFDITSKASFENVKTWV
DDARNIRGDDVSIIVVGNIKIDLADQREIYTEEGTKLASELGVEFIEVSAKAGINIKALFK
TLASNLPGADDTAAKESSVKIKKEEIQKLKEDDTEEGAKKSGCC

>gb|ADN96603.1|  Rab GTPase 6 [Euplotes octocarinatus]
Length=220

 Score =   353 bits (907),  Expect = 1e-120, Method: Compositional matrix adjust.
 Identities = 174/220 (79%), Positives = 192/220 (87%), Gaps = 2/220 (1%)

Query  1    MAEFASDTLDQSSSSAPVKYKVVFLGDQSVGKTSIILRFMHDTFDSQYQATIGIDFLSKT  60
            MA+F SD L QSSS+APVKYKVVFLGDQSV KTSIILRFMHDTFD  YQATIGIDFLSK
Sbjct  1    MADFVSDALYQSSSTAPVKYKVVFLGDQSVCKTSIILRFMHDTFDPNYQATIGIDFLSKA  60

Query  61   MMVDDKMVRLQLWDTAGQERFKSLIPSYIQDSSFAIVCFDITSKASFENVKTWVDDARNI  120
            M+VDDKMVRLQLWDTAGQERFKSLIPSYIQDSSFAIVC+DITSK +FE+VKTWVDDARNI
Sbjct  61   MVVDDKMVRLQLWDTAGQERFKSLIPSYIQDSSFAIVCYDITSKQTFEHVKTWVDDARNI  120

Query  121  RGDDVSIIVVGNKIDLADQREIYTEEGTKLASELGVEFIEVSAKAGINIKALFKTLASNL  180
            RGD+VSII+VGNKIDLA+QRE+YTEEG KLA ELGV FIE SAKAGINIKALFKTLASNL
Sbjct  121  RGDNVSIIIVGNKIDLAEQREVYTEEGKKLADELGVSFIEASAKAGINIKALFKTLASNL  180

Query  181  PGADD--TAAKESSVKIKKEEIQKLKEDDTEEGAKKSGCC  218
            PG ++  T  K++ V++KKEE  K   D+  +  KK GCC
Sbjct  181  PGNENGKTKGKDAGVQLKKEEKDKNDNDEDAKTKKKPGCC  220
```

Supplementary Note Figure 1. Supporting information for +1 frameshifting at AAT_TAA. Left panel: density of ribosome footprints (top) and mRNA-seq reads (middle) for a transcript whose ORF is shown at the bottom (red lines correspond to stop codons, and green lines to ATG codons). Identity of stop codons and adjacent 5' codons is indicated for the frameshift site and for the site of termination. Translated segments of ORFs are highlighted in blue. Right panel shows protein sequence produced with inferred frameshifting (top) and its alignment to the closest BLAST hit (bottom).



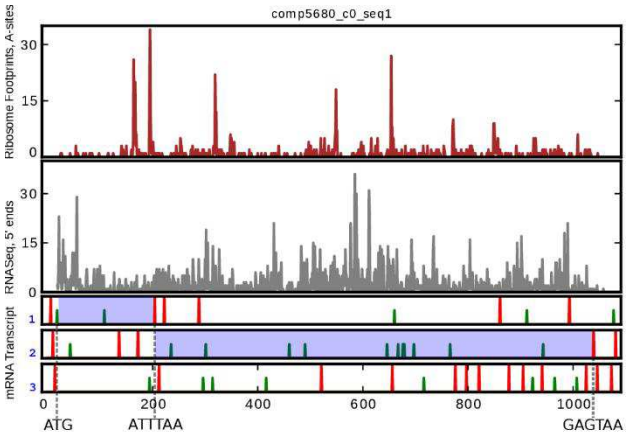

```
>comp1673_c0_seq1_AGATAA(+2
MEGGNQGPYNVGEPTEKITLHISCRKLADLDIITVSDPVCHVYIADSDHPDDW
MLYGKTEQIENNLNPDFVTYFEMDYYFEKIQKIKVEVFDVDVTRLERIGNFETTLGEIMG
SVNTTLTEGR!ILRTEKVATSNDLIYFSLRINDLVSNKGWFCGSDDPFIFIERARENDQE
EFLRVIQTEPIRNDLNPTWRYLKYEAKEICNGDLQCPLKFKVYSWRNSGHHKFFGEFETT
MLRIRNGDTQYNLFKDGAQQKSICSFEEFFIEERASFFDFLHSGWKMNLMVCVDFTASNG
EVTVPSSLHYLNPTGEFNDYQNAIRQVGNILELYDYNRQYPCYGFGGIPRYSGSNQVSHC
FHLNGLEDPEVDGVNGILESYQFSLLNCGLYGPTNFGECMRKTVDYIKERMDERMYHILL
ILTDGDIHDMPITRDIIVEGSHYPLSIIIIGLGESSFDKMIELDGDDVVLKNTRGEATRR
DIVQFVKFNDFRHLSKQALAEEVLEEVPEQVVSYLSQNNIKLDEVN

>emb|CDW78601.1|  copine family protein [Stylonychia lemnae]
Length=554

 Score =   420 bits (1080),  Expect = 3e-137, Method: Compositional matrix adjust.
 Identities = 211/516 (41%), Positives = 330/516 (64%), Gaps = 21/516 (4%)

Query  16   EKITLHISCRKLADLDIITVSDPVCHVYIADSDHPDDWMLYGKTEQIENNLNPDFVTYFE  75
            ++++L ISCR L +LD+++ SDP+C VYI D     +W L GKTE I NNLNPDF +
Sbjct  19   QRVSLSISCRNLKNLDVLSKSDPMCEVYIKDR-KTTNWTLLGKTETINNNLNPDFSSIIY  77

Query  76   MDYYFEKIQKIKVEVFDVDV---TRLERIGNFETTLGEIMGSVN----------TTLTE  121
             DY+FE+ Q IK +++D+D    T  + IG+ ETTLG I+GS+            +T +
Sbjct  78   CDYFFEREQNIKFDLYDIDNQQHTSRDFIGSNETTLGGIIGSMQQTYVADLKDNKSTRSR  137

Query  122  GRIL-RTEKVATSNDLIYFSLRINDLVSNKGWFCGS-DDPFIFIERARE-NDQEEFLRVI  178
            G+I+ R + V T+ND +   LR++  V +    CG+ D+P+  I RAR+ N+ ++F+RV
Sbjct  138  GKIVVRLDNVNTTNDEV--RLRVSARVQSNAGCCGTQDNPYYIISRARDVNNHKDFVRVY  195

Query  179  QTEPIRNDLNPTWRYLKYEAKEICNGDLQCPLKFKVYSWRNSGHHKFFGEFETTMLRIRN  238
            ++  + N   P W   K +  +ICNG    P+KF++YS   SG  + +GE  T++ ++++
Sbjct  196  KSSAMLNSTQPMWNVQKIKLSQICNGINNLPIKFELYSQNISGTDQAYGEGITSIEQLQS  255

Query  239  GDTQYNLFKDGAQQKSICSFEEFFIEERASFFDFLHSGWKMNLMVCVDFTASNGEVTVPS  298
            G     +     + K   + + F I E  +F ++L SGW +N+    +D+TASNGE T P+
Sbjct  256  GQKSVEITDKKRKIKGSLNIDNFVIREMPNFMEYLRSGWAINMSFAIDYTASNGEKTDPN  315

Query  299  SLHYLNPTGE-FNDYQNAIRQVGNILELYDYNRQYPCYGFGGIPRYSGSNQVSHCFHLNG  357
            SLH  +P+G   N Y+ A+  VG ++E Y  N+ +  +GFGGIPR++GSNQ+SHCF+LNG
Sbjct  316  SLHKQDPSGRNLNQYEQALLSVGKVMEPYALNQMFATFGFGGIPRFTGSNQISHCFNLNG  375

Query  358  LEDPEVDGVNGILESYQFSLLNCGLYGPTNFGECMRKTVDYIKERMDERMYHILLILTDG  417
               P++ G+  +  +Y+ ++   GL GPT+F   ++  + Y+++ +  +MYH L+I+TDG
Sbjct  376  SVSPQIQGLQNVYMAYKRTIHQIGLAGPTHFSSVLQSLLVYVQQCLQFQMYHCLMIITDG  435

Query  418  DIHDMPITRDIIVEGSHYPLSIIIIGLGESSFDKMIELDGDDVVLKNTRGEATRRDIVQF  477
            +IHDMP T D+IVE S +P+SIIIIG+G   F+KM  LD D+  L+N++G+   RDIVQF
Sbjct  436  EIHDMPATIDLIVELSRFPVSIIIIGVGNEGFEKMNFLDSDNQALRNSKGQVAARDIVQF  495
```

Supplementary Note Figure 2. Supporting information for +2 frameshifting at AGA_TAA. See Supplementary Fig. S8 for the legend.



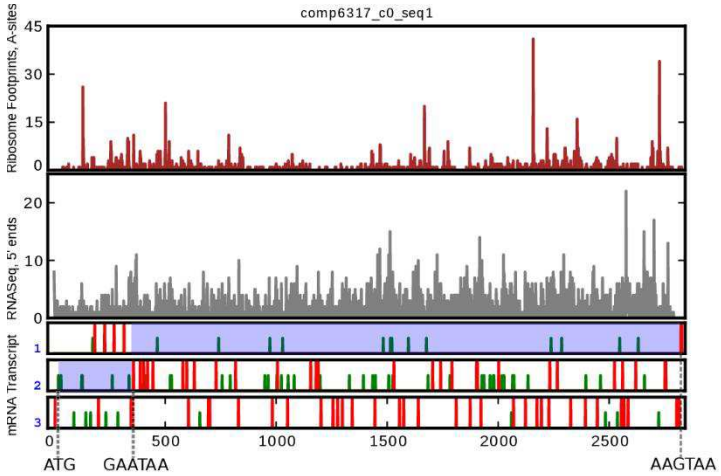

```
>comp9913_c0_seq1_AGGTAA(+1)
MSRILERSGSVSNNPLNCSTAKQQFSFSKADRFGQTMTRGLSSGY
LNLPSTRSTRSTTFGYGDKIMTNKRINDSPSPGTYKIPSGFEPGKKGKVYSFRCSWKSYS
KVYQKEGFQNTKASDPNVPGPGTYKSVPTFGNKGRKFTLKGKLKDLPSSVKVPGPGAYKD
LTTLPRTGRNFYSKFKSVCNGAIGPPTNTRFKDQTQTSAEIPGPGQYTPKTGLTGTGHYF
LSRFKSSGTTVLGGARRASTHRVSNDGPGPGSYILPSDFGYPTSFHRRRKKSRRATSQL

>gb|EJY73919.1|  hypothetical protein OXYTRI_04827 [Oxytricha trifallax]
Length=322

 Score =   170 bits (430),  Expect = 1e-46, Method: Compositional matrix adjust.
 Identities = 111/289 (38%), Positives = 150/289 (52%), Gaps = 36/289 (12%)

Query  1    MSRILERSGSVSNNPLNCSTAKQQFSFSKADRFGQTMTRGLSSGYLNLPSTRSTRSTTFG  60
            MS+++     ++ +P+N ST+KQ FSFSKA+RFGQ  +        +LPS +STR+   G
Sbjct  1    MSQVVSTPQQINQHPVNNSTSKQLFSFSKAERFGQPKSMMNHRIAYDLPSMKSTRAAGLG  60

Query  61   YGDKIMTNKRNDSPSPGTYKIPSGFEPGKKGKVYSFRCSWKSYSKVYQKEGFQNTKASDP  120
            YG +   N RNDSPSP  Y + S F      K +SF  S ++YSKVY KE       +D
Sbjct  61   YGSRNAFNIRNDSPSPTNYNLKSEFTKSPSNKAFSFGISREAYSKVYIKEN----PLADA  116

Query  121  NVPGPGTYKSVPTFGNKGRKFTLKGKLKDLPSSVK--VPGPGAYKDLTTLPRTGRNFYSK  178
            +VPGPG Y+  P  G +  K+TL+ K ++  +      PGPG Y     L   G  F SK
Sbjct  117  SVPGPGQYQIPPIVGKEALKYTLRPKTQNPYTQTYKGQPGPGQYDTKPALNDKGLYFNSK  176

Query  179  FKSVCNGAIGPPTNTRFKD------------QTQTSAEIPGPGQYTPKTGLTGTGHYFLS  226
            FK+    +I PP++ RFKD            Q      +PGPG Y P   +  TG YF+S
Sbjct  177  FKNSSATSIDPPSSVRFKDINGKLVFDVNNIQILGKRSVPGPGTYQPNIEMNKTGSYFVS  236

Query  227  RFKSS----------GTTVLGGARRASTHRVSNDGPGPGSYILPSDFGY  265
             F+SS           T +LG   + +        PGPG+Y LPSDFGY
Sbjct  237  NFQSSMCRSHYHFDRQTNILGSTMKGT--------PGPGNYRLPSDFGY  277
```

Supplementary Note Figure 3. Supporting information for +1 frameshifting at AGG_TAA. See Supplementary Fig. S8 for the legend.



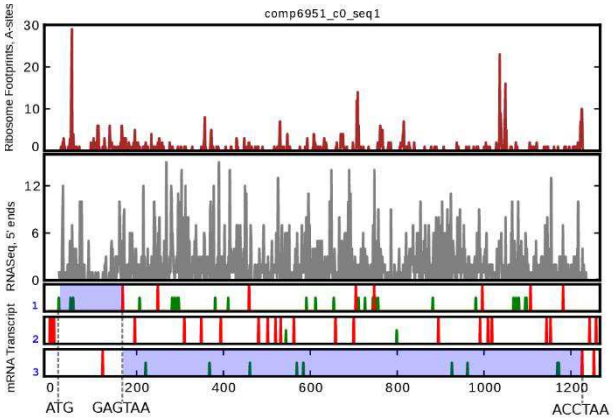

```
>comp5680_c0_seq1_ATTTAA(+1)
MSLPRYYANACVDKPESYYDYTNFELSWGPMDYYEVVQKIGRGKYSEVFDGVNT
LNNQKCVIIKILKPIKLEKMQREIKILQTLYGGKNIIKLYDMAQDDVSEVTALVFERVNHT
DHKVLYKQFKDFDIRYYIYEVLLGLDYCHSLGIMHRDIKPHNIMIDHEQRQLRIIDWGLA
EYYIPGKEYNVRVASRYYKGPELLVDDRLYNYSLDMWSLGCTMANMMFQRDPMFKGVDND
DQLIKIVEICGIDGLMDYLKTYKLEINRYHQKHLKNWEKVSWEEFITKKNKHLVTKEALD
FLEKCLQYDREKRIMPQEAIEHEYFAPVIEYKKKIHGGGEEVKTEE

>gb|EJY69411.1|  Casein kinase II subunit alpha [Oxytricha trifallax]
Length=333

 Score =   392 bits (1008),  Expect = 3e-132, Method: Compositional matrix adjust.
 Identities = 181/330 (55%), Positives = 249/330 (75%), Gaps = 2/330 (1%)

Query  1    MSLPRYYANACVDKPESYYDYTNFELSWGPMDYYEVVQKIGRGKYSEVFDGVNTLNNQKC  60
            M+LP+YYAN C + P  Y DY N+E+ +G  + YE+++KIGRGKYSEV++G+NTLNN++
Sbjct  1    MNLPKYYANVCEEMPPEYSDYENYEVKFGSQENYEIIKKIGRGKYSEVYEGINTLNNERI  60

Query  61   VIKILKPIKLEKMQREIKILQTLYGGKNIIKLYDMAQDDVSEVTALVFERVNH--TDHKV  118
            VIKILKP+K  K++REIKILQTL  G NII L D+ +D +++  AL+ E V+    D +
Sbjct  61   VIKILKPVKKTKIRREIKILQTLKNGINIINLIDVVRDPMTKTPALIMEYVDTGDVDFRT  120

Query  119  LYKQFKDFDIRYYIYEVLLGLDYCHSLGIMHRDIKPHNIMIDHEQRQLRIIDWGLAEYYI  178
            LYK F DFDIRYY++E+L  LD+CHS GI HRD+KPHNIMIDH  R+LR+IDWGLAE+Y
Sbjct  121  LYKSFTDFDIRYYMFEILKALDFCHSKGITHRDVKPHNIMIDHASRKLRLIDWGLAEFYH  180

Query  179  PGKEYNVRVASRYYKGPELLVDDRLYNYSLDMWSLGCTMANMMFQRDPMFKGVDNDDQLI  238
            PG+EYNVRVASRY+KGPELL+D + Y+YSLD+WSLGC  + M+FQ++P F+G DN DQL+
Sbjct  181  PGQEYNVRVASRYFKGPELLIDLQTYDYSLDIWSLGCMFSGMIFQKEPFFQGKDNYDQLV  240

Query  239  KIVEICGIDGLMDYLKTYKLEINRYHQKHLKNWEKVSWEEFITKKNKHLVTKEALDFLEK  298
            KI ++ G + L  Y++ Y + ++ ++   L    K  W +FI   N+HLV++EALD L K
Sbjct  241  KIAKVLGTEELYAYIEKYNVTLDSHYDDILGQHTKKPWHKFINSANEHLVSEEALDLLSK  300

Query  299  CLQYDREKRIMPQEAIEHEYFAPVIEYKKK  328
             L+YD  +RI+P++A++H YF PV E+  K
Sbjct  301  MLKYDHAERIVPKDAMDHPYFKPVKEFHAK  330
```

Supplementary Note Figure 4. Supporting information for +1 frameshifting at ATT_TAA. See Supplementary Fig. S8 for the legend.



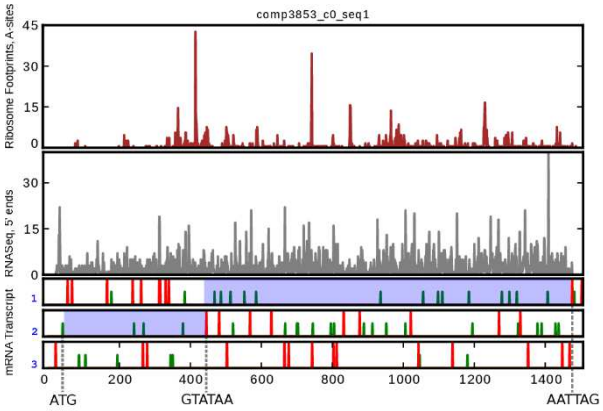

```
>comp6317_c0_seq1_GAATAA(+2/-1)
MSSYMKKSSELEELKIELNSLKVDEQKEAVKQVIAMMTIGKDVSGLFPHVTKCI
LSPSIELKKLVYLYIINYAKSKPDLTLMAVSAFTKDAHEKSNPLIRALAVRTMGCIRIE!
KIATYLCESLKDCLVDDDPYVKKTAAISVAKIYHTMPEHTKELGFIKLLQGLLQDGNAIV
VANAVAALFEISRVAGKNYLKANKETIGKLLNALNETNEWGQIYILESIINYKPKEQKEA
EEIIERIMPRLQHANPAVVLGATKNVLHFLKFVNLKSNKTTILKKLSAPLITLLSSEPEI
QYIALCNILLILQQIPNVFEKNVKMFFCRFSDPIYVKLAKLDVMVGVADNTNVDIIITEL
HEYCNNIDQDFVRRSVKAIGQVVVKVDRVAKKGVEALREHVNQEQGSDSALQEAVIVASK
ILRKYPKKFEGLVKDIVKQQDRIDEPESKSAFIWILGEYSKKIEDAGEKLQVYIDSFTDE
NINVCLQILTSAVKMFIKDSDNYEDMVMNVLKLASESSANPDLRDRGYIYWRMLSTDPSQ
TKDTVLAKRPEVEEDLTKLMDDETRDIFIDIFISDTKHSALRPEKTASAPEPDSDEEVEE
EEKPKKSKKKDKKSKKQKKVKEEETKEEIELPEEVVTEDKPKDDLDDIFGLGIGDDPVSN
DEPAVDPLAGILDEGNGGGESTQAASPWDDNGLFGGFGASGSEASLFIKSEHAEVLSSST
PGSQNKAAGLQIKARFYREGTSIKLDMIFYNSTAGIISDFDIMINKNPFGLKPGPISVIP
ISAGQTFTTTVECSIDQSNADLKNPPQCPYYVQTAIKNSLDVYYFQVPCLLHTLLQGTPV
AVTQTQCQQMANSIANKHSFTVSSARFAGSASDLKTRMQSNSFYPIYDELNSQIFATSTV
NNIPILLRCTPEGSDIQIIACTPVAPLYQLIEEAIKEVISK

>emb|CDW87346.1|  ap-2 complex subunit [Stylonychia lemnae]
Length=1023

 Score =   706 bits (1821),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 428/1024 (42%), Positives = 610/1024 (60%), Gaps = 93/1024 (9%)

Query  1     MSSYMK--KSSELEELKIELNSLKVDEQKEAVKQVIAMMTIGKDVSGLFPHVTKCILSPS  58
             M++Y +  KSSEL EL+ ELNSLK +E++EA KQVIAMMTIGKDVS LFPH+ KC+ +
Sbjct  1     MTNYFQNMKSSELAELQHELNSLKPEEKREAAKQVIAMMTIGKDVSSLFPHMVKCMETTQ  60

Query  59    IELKKLVYLYIINYAKSKPDLTLMAVSAFTKDAHEKSNPLIRALAVRTMGCIRIEKIATY  118
             +ELKKLVYLYIINYAK KPDLT+MAV++F KD+ +  +P++RALAVRTMGCIR+E+I  Y
Sbjct  61    MELKKLVYLYIINYAKVKPDLTIMAVNSFQKDSRDIQSPMMRALAVRTMGCIRVERITEY  120

Query  119   LCESLKDCLVDDDPYVKKTAAISVAKIYHTMPEHTKELGFIKLLQGLLQDGNAIVVANAV  178
             +CESLK+ L D DPYVKKTAA+ VAK++ T P   K+   IK+LQG+L DGNA+VVANA
Sbjct  121   MCESLKERLNDQDPYVKKTAALGVAKLFQTSPRLVKDHSLIKILQGMLYDGNAVVVANAA  180

Query  179   AALFEISRVAGKNYLK-ANKETIGKLLNALNETNEWGQIYILESIINYKPKEQKEAEEII  237
             A+L EISR +GKNYL+  N + + KLL ALN+ NEWG+IYILE I +Y   + KE+E I+
Sbjct  181   ASLLEISRASGKNYLRLKNDQGLNKLLIALNDANEWGKIYILEGISSYDTSDSKESENIV  240

Query  238   ERIMPRLQHANPAVVLGATKNVLHFLKFVNLKSNKTTILKKLSAPLITLLSSEPEIQYIA  297
             ER++P L H NPAV+L A K VL F+  V+ +     I+KKL  PLITLLS+E EIQY+A
Sbjct  241   ERVLPMLTHNNPAVILSAVKTVLKFMNNVSTQDLLKGIIKKLGPPLITLLSTEAEIQYVA  300

Query  298   LCNILLILQQIPNVFEKNVKMFFCRFSDPIYVKLAKLDVMVGVADNTNVDIIITELHEYC  357
             L NI  ILQ+  ++FE+NV++FFC+++DP+YVKL K+D++V VAD+ NV+ I+ EL EY
Sbjct  301   LRNINFILQKYSHLFEQNVRVFFCKYNDPVYVKLEKIDILVKVADDKNVETILAELKEYS  360

Query  358   NNIDQDFVRRSVKAIGQVVVKVDRVAKKGVEALREHVNQEQGSDSALQEAVIVASKILRK  417
              +ID + V++SV+AIGQ+++KVD+ A K VE + E V   QG +  +QEAVIVA  I RK
Sbjct  361   GDIDPELVKKSVRAIGQIILKVDKAASKAVEIIHEIVT--QGGEIGVQEAVIVAKDIFRK  418
```

Supplementary Note Figure 5. Supporting information for +2 frameshifting at GAA_TAA. See Supplementary Fig. S8 for the legend.



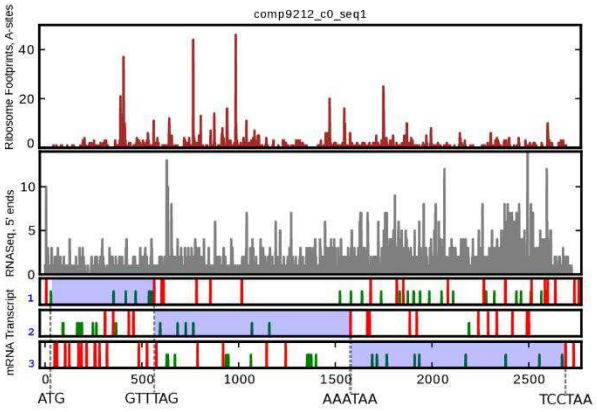

```
>comp6951_c0_seq1_GAGTAA(+2
MYSTKFRRVMTMAPLLLANPALALCEEPSTADRIRGNYENKIRFFAAPE!KIFE
TFSNIREEDGQVYMSYQDFFHSLTPYNFVASKDDDDDDDDEENKDKEKEKEPGYFDKFTP
EIMTIVDANQDKKIDFNEYIFFITLLQLPEGEVMRIIEKVNPEERKINKAQFAKYLTKLR
KCTALGLKQMSKSFMPDGRKISTDEDHISKTILLHLFNDKEYITIEDFCELKSKLKHALL
HYEFYQFDVDEDETISAESFAKSLLSCLNYTQASKYSRRIHSLKLEGRVSFKEYVAFHNL
IEKADIIKMKISTYRFLSLGMFRDLCDDFAKLDPYCNQNKVSISDTQIATFFKVLDEDEN
GALEYDEVVDILEGKKNIGLGKEDKFKREMMEKIDRYIKKFQKYVGWT

>emb|CDW75918.1|  calciumbinding atopy-related autoantigen [Stylonychia lemnae]
Length=426

 Score =   311 bits (798),  Expect = 2e-98, Method: Compositional matrix adjust.
 Identities = 169/401 (42%), Positives = 246/401 (61%), Gaps = 27/401 (7%)

Query  9    VMTMAPL----LLANPA--LALCEEPSTADRIRGNYENKIRFFAAPEKIFETFSNIREED  62
            M + PL    L+ N    L+ CEE    DRIRGNYENKIRFF+ PEKIFETF++ + E
Sbjct  42   AMIITPLAFQMLILNNQHNLSQCEEAPRQDRIRGNYENKIRFFSPPEKIFETFASSKNEK  101

Query  63   GQVYMSYQDFFHSLTPYNFVASKDDDDDDDDEENKDKEKEKEPGYFDKFTPEIMTIVDAN  122
            G + MSY DFF +LTPYN    KD                 +P YFDK+ P+I+ + D+N
Sbjct  102  GDLVMSYSDFFRALTPYNHSEIKDS----------------KP-YFDKYKPDILKVADSN  144

Query  123  QDKKIDFNEYIFFITLLQLPEGEVMRIIEKVNPEERKINKAQFAKYLTKLRKCTALGLKQ  182
             D  I F E+ FFIT+LQ+P G + +   K   E  K+N+ +F+K LT LRK T LG KQ
Sbjct  145  GDGVISFPEFFFFITILQMPLGLIHKEFTKHVKEGGKMNQDEFSKTLTTLRKKTLLGTKQ  204

Query  183  MSKSFMPDGRKISTDEDHISKT---ILLHLFNDKEYITIEDFCELKSKLKHALLHYEFYQ  239
            ++K  +PD R IS  ED  S+T   I   LF +K   + EDF   + +LK AL HYEF+Q
Sbjct  205  INKGMVPDARLISATEDDFSQTNNEICNQLFKNKTLFSYEDFINFRDELKIALRHYEFHQ  264

Query  240  FDVDE-DETISAESFAKSLLSCLNYTQASKYSRRIHSLKLEGRVSFKEYVAFHNLIEKAD  298
            ++V+E +++IS E F KSL+ CL Y +A  Y +R+H LKL+G VSFKE++AF   I+  D
Sbjct  265  YEVNEENDSISMEDFTKSLMVCLPYKEAHMYIKRVHELKLDGEVSFKEFLAFQRFIDDVD  324

Query  299  IIKMKISTYRFLSLGMFRDLCDDFAKLDPYCNQNKVSISDTQIATFFKVLDEDENGALEY  358
             IK K+  YR+++L   + LC +F + D +C +  V I+  Q+    ++LD D NG L++
Sbjct  325  HIKEKVLVYRYITLDQLKSLCKEFCEEDEFCKKENVQINPKQVEALVRLLDLDGNGQLDH  384

Query  359  DEVVDILEGKKNIGLGKEDKFKREMMEKIDRYIKKFQKYVG  399
            DEV+ +L+ ++ +G GKE++ K  +     + ++ F++ +G
Sbjct  385  DEVIGVLDQRQLLGQGKENELKEAIESSFKKVVQWFRETLG  425
```

Supplementary Note Figure 6. Supporting information for +2 frameshifting at GAG_TAA. See Supplementary Fig. S8 for the legend.



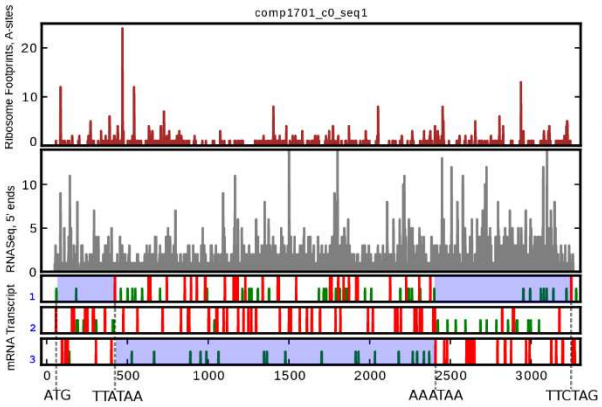

```
>comp3853_c0_seq1_GTATAA(+2)
MSEENKEEVKGTTHTDEDQYHHGFGNHFESEAIEGALPKHRNNPQQC
KFGLYAEQISGTPFTYPRAKMQRSWLYRIMPTVAHPPYKALKDYNNLWIANFARDDDEEV
FTTPQQMRWTPIDLPSEEITFVQGIQTVITGAGDPSMKAGINMGVYTCNTSMKNEAFFSS
DGDIMIVPQLGKLSIMTEFGHIEAESWEVVVIPRGIKFAVEVNEDCRGYYCELYDGHLQI
PDLGPIGTNGSANPRDFAIPKAKYFDETNEFRVIQKYLGKFFEYTIPHNIFDIVAWHGNY
YPYKYDCHHFNTMGSISYDHPDPSVFTVLTCQTPDHGQAALDFAIFPPRWLSMEDTFRPP
YFHRNTMNEFMGNVAGQYDAKEEGFSPGAVSLHSCMSAHGPEAEVVEKASTCELKPQKVG
EGCLAFMFETCYTMKVTKSFMHDLEGATDSYSVNSSKAVDESYHDCWKGMKRLFDPNDPD
AGYKKYLSEHKN

Homogentisate 1,2-dioxygenase [Oxytricha trifallax]
Sequence ID: gb|EJY66813.1|Length: 439Number of Matches: 1
Related Information
Range 1: 9 to 435GenPeptGraphics Next Match Previous Match
Alignment statistics for match #1 Score Expect  Method  Identities      Positives       Gaps
514 bits(1323)  5e-176  Compositional matrix adjust.    249/442(56%)    314/442(71%)    19/442(4%)

Query  23   GFGNHFESEAIEGALPKHRNNPQQCKFGLYAEQISGTPFTYPRAKMQRSWLYRIMPTVAH  82
            GFGNHFE+E++EGALPK +N+PQ+    GLYAEQ+SGTPFTY R + QRSW YRI PTV H
Sbjct  9    GFGNHFETESVEGALPKDQNSPQKVPHGLYAEQLSGTPFTYARHRNQRSWTYRIQPTVLH  68

Query  83   PPYKALKDYNNL--WIANFARDDDEEVFTTPQQMRWTPIDLPSE-EITFVQGIQTVTGAG  139
             P+K + D      W+  F +DD   +  +P Q RW P  +P++ ++TF++GIQT+ GAG
Sbjct  69   RPFKPVADQERFKNWLTIFDKDDSR-LHISPAQYRWRPQTIPTDKQVTFIEGIQTMGGAG  127

Query  140  DPSMKAGINMGVYTCNTSMKNEAFFSSDGDIMIVPQLGKLSIMTEFGHIEAESWEVVVIP  199
             P++K G+ + +Y CN SM+ EAF+SSDGD++IVPQ+G++ + TEFG I  E  E+VV+
Sbjct  128  GPALKNGLGIHLYACNKSMEKEAFYSSDGDLLIVPQVGEILVKTEFGRIYVEPQEIVVVQ  187

Query  200  RGIKFAVEVNE-DCRGYYCELYDGHLQIPDLGPIGTNGSANPRDFAIPKAKYFDETNEFR  258
            RGIKF VE+ E   RG+ CE+Y  H  IPDLGPIG+NG ANPRDF  P A Y D+  E+
Sbjct  188  RGIKFQVELLEGQARGFICEIYKSHFVIPDLGPIGSNGMANPRDFQTPIAWYEDKHEEWT  247

Query  259  VIQKYLGKFFEYTIPHNIFDIVAWHGNYYPYKYDCHHFNTMGSISYDHPDPSVFTVLTCQ  318
            VI K+ G FF+Y + H+ FDIVAWHGNY PYKYD   +NT+GSIS+DHPDPS+FTVLT Q
Sbjct  248  VINKFQGSFFQYQVTHSPFDIVAWHGNYAPYKYDLRKYNTIGSISFDHPDPSIFTVLTAQ  307

Query  319  TPDHGQAALDFAIFPPRWLSMEDTFRPPYFHRNTMNEFMGNVAGQYDAKEEGFSPGAVSL  378
            T D GQA  DF IFPPRWL  E TFRPPY+HRNTM+EFMGN+ G YDAK E F  GA SL
Sbjct  308  TDDPGQAVCDFVIFPPRWLVQEHTFRPPYYHRNTMSEFMGNIQGTYDAK-ESFCAGASSL  366

Query  379  HSCMSAHGPEAEVVEKASTCELKPQKVGEGCLAFMFETCYTMKVTKSFMHDLEGATDSYS  438
            HS MS HGP+AEV +KAS  +LKPQ VG   ++FMFET Y +K+T             Y+
Sbjct  367  HSTMSGHGPDAEVFDKASNADLKPQLVGVDSMSFMFETAYMLKLT------------DYA  414

Query  439  VNSSKAVDESYHDCWKGMKRLF  460
            VN    V   YH+CW+ + R F
Sbjct  415  VNEEN-VYTDYHECWQKLPREF  435
```

Supplementary Note Figure 7. Supporting information for +2 frameshifting at GTA_TAA. See Supplementary Fig. S8 for the legend.



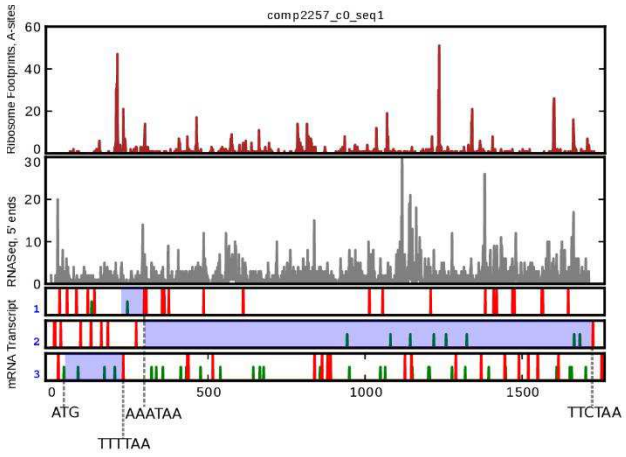

```
>comp9212_c0_seq1_GTTTAG(+1)_AAATAA(+1)
MSGKEEAKSGIKRKNTGKGGHDDKHSAPSGQGNICVVCRFRPLNQNELNH
GGSNVCADFHPNKKSVTIVTEGDGVTNKNEFTFDRVFDINSSQIEVYNQAAKPIIESVME
GFNGTVFAYGQTGSGKTFTMQGPDIEDIESQGIVPRMVRTVFNRIENSSENIEFTVKVSM
MEIYMEKVIRDLLDPTKANMKIKVDKHKSAYVHDLTERYIGSDLDVYDIMRIGNNNRKVAS
TSMNDQSSRSHSIFVMTVHQNNLDDQTSKTGILYLVDLAGSEKVGKTGASGHTLDEAKGI
NKSLSTLGKVINALTDGKSKHIPYRESKLTRILSESLGGNARTALIITCSPSVYNDMETI
STLRFGTTARNIKNKPKVNKELTVAEMKKLLSKSEKIIEVRNYRVKVLEGCITELGGEVP
EDEYKDLQLSSKPAPAKEEVKEPAPSKEPEQDDDDDEQDEEEKKAIDDELQAKLKRLQDK
LNIETDKYCNLTEEFEVSQNKLAASQEKCDKLQVQGQKLEEEIEEK!NDLVQTQKQKLLLL
EEELRNVVESNEKLSKKLDRIEEMGGIYLPEMEDGDSAHDKKRTAQEAMQKELIRVSQKL
KRISELGGLSDEVKQILDIEVRHTDEESKEDPDTDEMFKDSLLDMHADLSKFKDKSLSIF
LDEEDTNKLIKSQIELREQQKLDERQKHEFLAQIDSLSEELENAAKESNPNIQAITKKLA
DEQIMKAENRWAEEKVQILRDLENRVSKVVQLEISLDEANEKYRRLENTINQGDVPLRKK
INKLENQAEQITIMYHQVVSEKSVLKVDLQVYEKKLKRKDEKIASLEKALNKLKEGNVAL
KKILRGLKSLKMKANDENRILEGNNVTGIPSSGRVVKPLRGGRKDKPARIMSSIQS

kinesin motor catalytic domain protein [Tetrahymena thermophila SB210]
Sequence ID: gb|EAR95613.3|Length: 916Number of Matches: 1
Related Information
Range 1: 8 to 898GenPeptGraphics Next Match Previous Match
Alignment statistics for match #1 Score Expect  Method  Identities      Positives       Gaps
541 bits(1393)  3e-174  Compositional matrix adjust.    358/939(38%)    538/939(57%)    144/939(15%)

Query  30   GQGNICVVCRFRPLNQNELNHGGSNVCADFHPNKKSVTIVTEGDG---VTNKNEFTFDRV  86
            G GNI VVCR RP N++EL   GS  C +F   +     +T  DG     NK  F FDRV
Sbjct  8    GSGNIQVVCRVRPFNKSEL-EMGSVPCVEFLDQQTIRVKLTNTDGKEKADNKQLFNFDRV  66

Query  87   FDINSSQIEVYNQAAKPIIESVMEGFNGTVFAYGQTGSGKTFTMQGPDIEDIESQGIVPR  146
            F++ ++Q ++Y  AAKP+++SV+EGFNGTVFAYGQT SGKTFTMQG  I+D + +G++PR
Sbjct  67   FNMETTQEQIYEVAAKPVVQSVLEGFNGTVFAYGQTSSGKTFTMQGASIDDEKLKGVIPR  126

Query  147  MVRTVFNRIENSSENIEFTVKVSMMEIYMEKVRDLLDPTKANMKIKVDKHKSAYVHDLTE  206
            MV+TVF  I ++ ++IEF +K+S++EIYMEK+RDLLD TK N+ ++ DK +  Y+ D+TE
Sbjct  127  MVKTVFQHISDAPDHIEFRIKISIVEIYMEKIRDLLDNTKQNLVVREDKQRGIYIQDVTE  186

Query  207  RYIGSDLDVYDIMRIGNNNRKVASTSMNDQSSRSHSIFVMTVHQNNLDDQTSKTGILYLV  266
            +Y+ ++ DV+D++RIGN NR V +T+MN+ SSRSH +F+M+V QNNL+D ++KTG L LV
Sbjct  187  QYVSNEQDVFDLLRIGNQNRAVTATNMNEGSSRSHMLFMMSVSQNNLNDLSAKTGKLILV  246

Query  267  DLAGSEKVGKTGASGHTLDEAKGINKSLSTLGKVINALTDGKSKHIPYRESKLTRILSES  326
            DLAGSEKV KTGA G  LDEAK IN+SLS+LG VINALTDGKS HIPYR SKLTR+L ES
Sbjct  247  DLAGSEKVAKTGAEGRVLDEAKTINQSLSSLGNVINALTDGKSSHIPYRNSKLTRVLQES  306

Query  327  LGGNARTALIITCSPSVYNDMETISTLRFGTTARNIKNKPKVNKELTVAEMKKLLSKSEK  386
            +GGN++T LI+TCSPS +ND+ET+STLRFG  A+ IKNK KVN+E+TVAE++ LL+K EK
Sbjct  307  IGGNSKTTLIVTCSPSPFNDLETLSTLRFGIRAKAIKNKAKVNREVTVAELQVLLAKIEK  366

Query  387  IIEVRNYRVKVLEGCITELGGEVP-EDEYKDLQLSSKPAPAKEEVKE-PAPSKEPEQDDD  444
             +E +  RV  LE  I +LG ++P E      +  ++   ++EE+ + PA     E+
Sbjct  367  QLEEKTRRVAQLEDYIQQLGSQLPSETNLNQQEDQTQIIDSQEEIPQNPAQISNIEEIVR  426
```

Supplementary Note Figure 8. Supporting information for +1 frameshifting at GTT_TAA. See Supplementary Fig. S8 for the legend.




```
>comp1701_c0_seq1_TTATAA(+2)_AAATAA
MAKYEYFIKRTSKSSQIYYKIIPASSQCAQKGLNIEIGSG
LMHEYDFEDSESSYFSDGKLIYGVLGILHIYPTSYLVVISELNDSKVLRVDGKEIYEIKK
VEILSLSGKKKPKNKSYEDGLISKLLSSGFYFSYELDLTRRLQRGKGDDRYWWNRNMYEDF
KTYGISSSWAIKLIQGYVGYSKIIVNSANLELTLISRRRTGMAGTRYSRRGIDDEGNTAN
FVETEQILKFEDYILSFVQIRGSVPAFWEQTGISADLNLTRTLELDRQAFDKHFNDMIGE
YGKLVWANLLNNKRSYEMKLIRRFEELVKMYAGSNNRYLYFNFHQECSKNNFKVMEQKLK
------------------------------------------------------------
CLKGHDKGRVEALKSHAEDALDDLDPEAGYKFVSQSPMVGLLIMVFAKENLIGRITDVCS
TKVKSGFGGNVGNKGGVMVRFELDNTSICFCCAHLESGQKKEKERRQNLKEVVESGFSDK
PSRYNFMNHDVKIFFGDLNFRINLPYDDVIGIIDDMTEYNKDDKMSILLNNDQLNQCRMD
YSWLYHYKEMPIKFLPTYKYDKIKSDTYDTSKKRRVPAWTDRILWYHNEEESKDKGSQYI
SPILYERRETNFSDHRPVVAYFEIHAHRHCSEKKAKIKNTIASKKATSFIVKKPSYDPAE
TQVKKEFGDDFDLFSSPPKKEQHIHLDLNVNRAHPKTKEEVKRPTPNKDSPLDDLLQFDN
NLIDIREKKQDDGNLIDFSSKTRQMPHPGHRFYPHAHSMHGRNPYVQIQGPIQGVHNGFG
MQNHFPMYTPKPMGFFNSNTGTKGPRMRASVPRKPSADLIGKAAPNNVPQKKAMPRTEEF
FKF

endonuclease exonuclease phosphatase family protein [Stylonychia lemnae]
Sequence ID: emb|CDW87399.1|Length: 878Number of Matches: 1
Related Information
Range 1: 45 to 842GenPeptGraphics Next Match Previous Match
Alignment statistics for match #1 Score Expect  Method  Identities      Positives       Gaps
526 bits(1356)  4e-167  Compositional matrix adjust.    317/823(39%)    459/823(55%)    51/823(6%)

Query  61   IYGVLGILHIYPTSYLVVISELNDSKVLRVDGKEIYEIKKVEILSLSGKKKPKN-KSYED  119
            I G+LG+L+I   +++VVI+E     + ++DG  I+ IK V+++           +++ D
Sbjct  45   ISGLLGVLNIQGQNFIVVITE--KQMIGKIDGANIFTIKNVDLIPFGDDHSVNQIRNFID  102

Query  120  GLSKLLSSGFYFSYELDLTRRLQRG-----KG-------------DDRYWWNRNMYEDFK  161
            G+ KLL+ GFYFSY  D+T   QR      KG             D RY+WN N+ +DF
Sbjct  103  GVKKLLTLGFYFSYNTDITSNRQRSANLRNKGMMGGDSQSIWEQCDKRYFWNFNICQDFL  162

Query  162  TYGISSSWAIKLIQGYVGYSKIIVNSANLELTLISRRRTGMAGTRYSRRGIDDEGNTANF  221
               +   W + +IQGYV YS  I +   LE+ LISRRR  M+GTRY+ RG+DD+GN AN+
Sbjct  163  YQKVDPRWIVPIIQGYVEYSSQIFDGKELEILLISRRRFAMSGTRYNARGLDDDGNVANY  222

Query  222  VETEQILKFEDYILSFVQIRGSVPAFWEQTGISADLNLTRTLELDRQAFDKHFNDMIGEY  281
            VETEQI+ + ++I SFVQIRGSVP FW+Q GI A  ++ R   L  QAFDKH NDM+ +Y
Sbjct  223  VETEQIMSYRNFIYSFVQIRGSVPLFWQQKGIQAHTSIKRDETLTSQAFDKHLNDMVNDY  282
------------------------------------------------------------------------
Query  637  DVCSTKVKSGFGGNVGNKGGVMVRFELDNTSICFCCAHLESGQKKEKERRQNLKEVVESG  696
             + ++KVK+GFGG  GNKG V +RF  DN+S  F   HL SGQK+  ER ++++E+ +
Sbjct  632  SIRTSKVKTGFGGQSGNKGSVAIRFNFDNSSFAFINCHLTSGQKQVSERLEDIREIYKKS  691

Query  697  FSDKPSRYNFM--NHDVKIFFGDLNFRINLPYDDVIGIIDDMTEYNKDDKMSILLNNDQL  754
            F       +FM  +HD K  FGDLNFRI+LPY+ V    D++ + N       L + DQL
Sbjct  692  FDCSQKFQDFMIQSHDYKFIFGDLNFRISLPYETV---KDEIRKRN----YPYLQSKDQL  744

Query  755  NQCRMDYSWLYHYKEMPIKFLPTYKYDKKSDTYDTSKKRRVPAWTDRILWYHNEEESKDK  814
               +   S L  + E  + F PTYKYD  S+TYDTS+K R PAW DRIL+    E+S+
Sbjct  745  LSVKSQNSILNRFSEGILDFDPTYKYDDHSNTYDTSQKLRTPAWCDRILY----EKSQKC  800
```

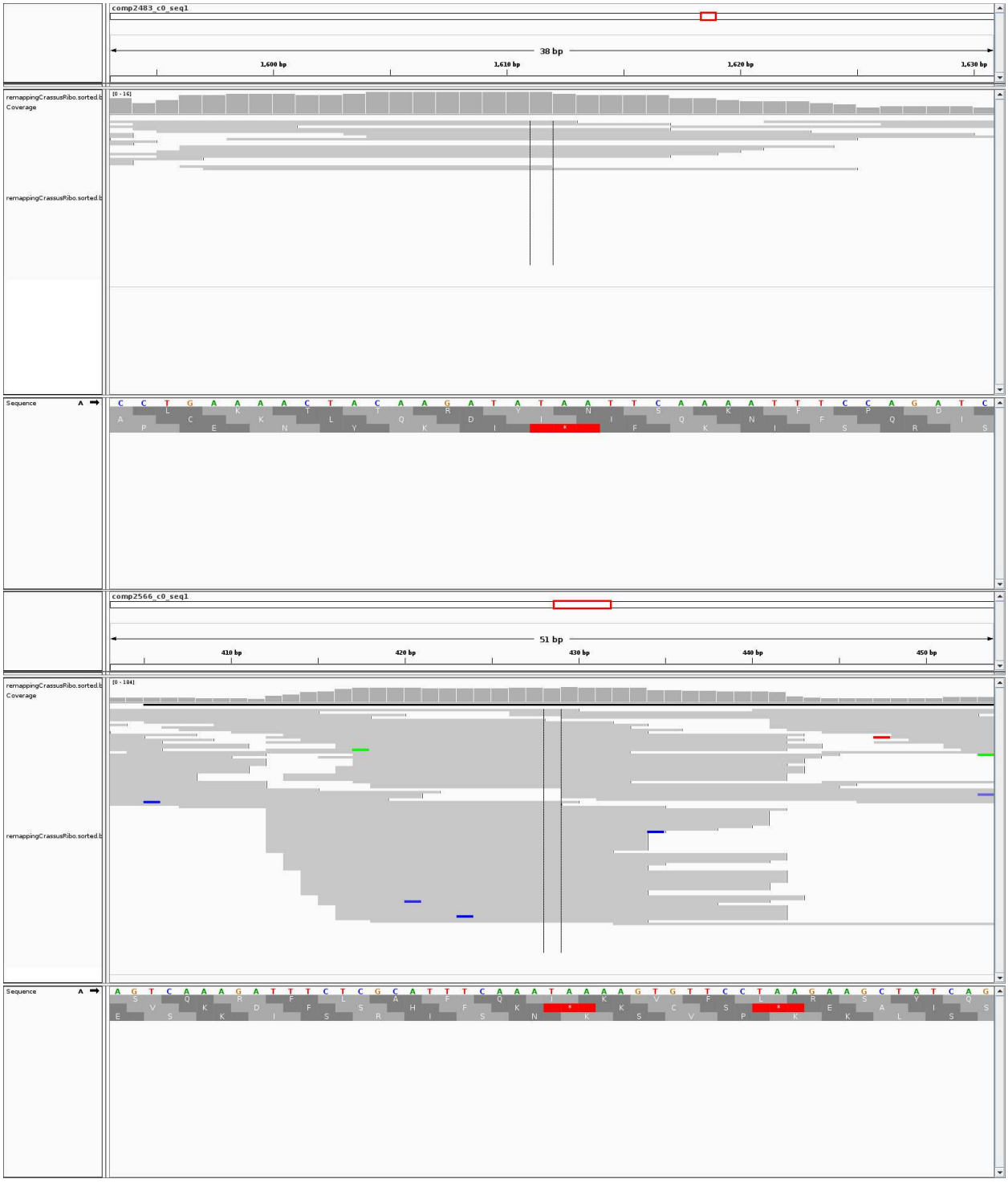
Supplementary Note Figure 9. Supporting information for +2 frameshifting at TTA_TAA. See Supplementary Fig. S8 for the legend.




```
>comp2257_c0_seq1_TTTTAA(+1)_AAATAA(+1)
MNLLEDDTYTFDQLGMQGNKAYEKFIDFANEPVVFSDTAIKINMWGI
KQERILLMTTKNIFNFIKRKSMKRKIAIEKIGAIIISTKINSKELVLHVPSEYDYRYQIND
REVFIKLLKREFVKKRPNGHLRYYVVKGSLKEVTTTEKDAKYKISRLPSSELRLLDQEVY
GPEYQQEESKTPTESTVAATSFDTKLKAVESRKQDASLEDTDEEPDADFDTRNSVCVFSH
KEEEEKVTLSDFEVKGVIGRGTFGKVFLVRLKTTDTLYAIKSLRKDVLVEAGQIENVKLE
KEILLACNHPFLAGMEYVFQNDTRLYFVIEFLKGGELYKHFLKKRRFEEDEAKFYATQIA
MAIGHLHKQNILHRDLKLENIMINGDGYIKVIDFGLAKIIDDDNLAMSFCGTPEYLAPEM
VKQEGHDRGVDWWSLGILIYEMTIGVTPFFSKSRLTLINNIKKEDVIFPGKKYKIEYSDD
FVDVVLKLLNKDSKKRLGSDGDIDEVLEHPYFQSLDLEALVNKEIKPPYIPEFTGKDDLG
SYIKLKSGSKDVNDTMIPSSKMRKIKKHETDFKNF

Protein kinase domain containing protein [Oxytricha trifallax]
Sequence ID: gb|EJY72937.1|Length: 553Number of Matches: 1
Related Information
Range 1: 23 to 553GenPeptGraphics Next Match Previous Match
Alignment statistics for match #1 Score Expect  Method  Identities      Positives       Gaps
408 bits(1049)  6e-132  Compositional matrix adjust.    221/545(41%)    349/545(64%)    24/545(4%)

Query  27   DFANEPVVFSDTAIKINMWGIKQERILLMTTKNIFNFKRKSMKRKIAIEKIGAIIIS-TK  85
            + ++E +++S+   K N +   QERI+++T + IFN K++ ++R IAIEK+  +  S ++
Sbjct  23   EISDETLIYSNRVKKFNRFDWTQERIMVITDRRIFNVKKQKIQRAIAIEKLLGVTKSLSE  82

Query  86   NSKELVLHVPSEYDYRYQINDREVFIKLLKREF--VKKRPNGHLRYYVVKG-SLKEVTTT  142
            ++ E VLHV  EYDYR + + R+  I+++++ F  V  +P   L +Y VK  +L E TT+
Sbjct  83   SNGEFVLHVKDEYDYRMRSDTRDQIIEIIRKLFSIVCGKP---LPFYGVKPKALSEFTTS  139

Query  143  EKDAKYKISRLPSSELRLLDQEVYGPEYQQEESKTPTESTVA--ATSFDTKLKAVESRKQ  200
            +KD K  +SR+P    R+ D+          ++ T  +  +      F  +++   S
Sbjct  140  KKDLKKGLSRIPLELARIKDEG---------DNSTDMQFNLGDDTEDFSNQMEERTSSYH  190

Query  201  DASLED--TDEEPDADFDTRNSVCVFSHKEEEEKVTLSDFEVKGVIGRGTFGKVFLVRLK  258
              S+ D   D+E       R        K  +++ TL DF VK VIG+G+FGKVF+V
Sbjct  191  QNSIPDFSVDKECSQILQGREKSSTLYSKRRDQETTLDDFTVKKVIGQGSFGKVFMVVHN  250

Query  259  TTDTLYAIKSLRKDVLVEAGQIENVKLEKEILLACNHPFLAGMEYVFQNDTRLYFVIEFL  318
             T  +YA+KS+RKDV++++ Q+EN++LEK I+L   HPF+  MEYVFQ D R+YF+++F+
Sbjct  251  ATGNIYAMKSIRKDVVIDSEQLENLRLEKHIMLCVEHPFIISMEYVFQRDYRIYFLMDFI  310

Query  319  KGGELYKHFLKKRRFEEDEAKFYATQIAMAIGHLHKQNILHRDLKLENIMINGDGYIKVI  378
            KGGELY+    K+R +E +AKFYA+Q+A+A+A+G+LHK  I++RDLK ENI+IN DGYIK+
Sbjct  311  KGGELYRQLALKKRLDEYQAKFYASQVALALGYLHKSRIIYRDLKPENILINKDGYIKLA  370

Query  379  DFGLAKIIDDDNLAMSFCGTPEYLAPEMVKQEGHDRGVDWWSLGILIYEMTIGVTPFFSK  438
            DFGLAK++  + +A SFCGTPEYL+PEM+   GHD  +DWW+LGIL+YEM IG+ PF+++
Sbjct  371  DFGLAKML-GEQVANSFCGTPEYLSPEMITGTGHDHTIDWWTLGILVYEMIIGIPPFYNQ  429

Query  439  SRLTLINNIKKEDVIFPG-KKYKIEYSDDFVDVVLKLLNKDSKKRLGSDGDIDEVLEHPY  497
            ++  + + I+   + +P  +K+  E S++  D++ KLL KD  KRLG    ++E++ HP+
Sbjct  430  NKHQMYHLIENGPIRWPTMEKHGFEISEESKDLISKLLEKDKNKRLGRVNGVEEIISHPW  489
```

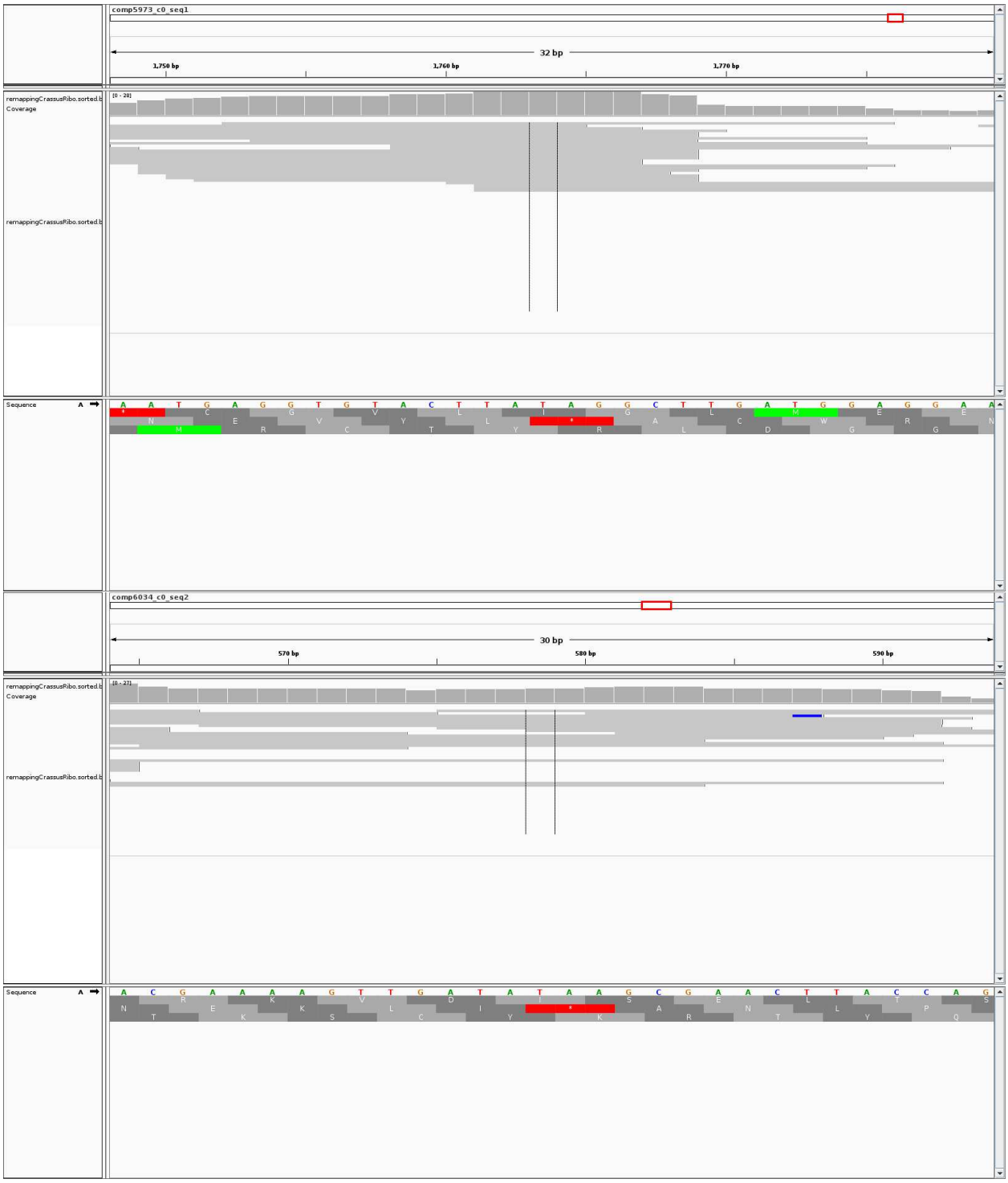
Supplementary Note Figure 10. Supporting information for +1 frameshifting at TTT_TAA. See Supplementary Fig. S8 for the legend.

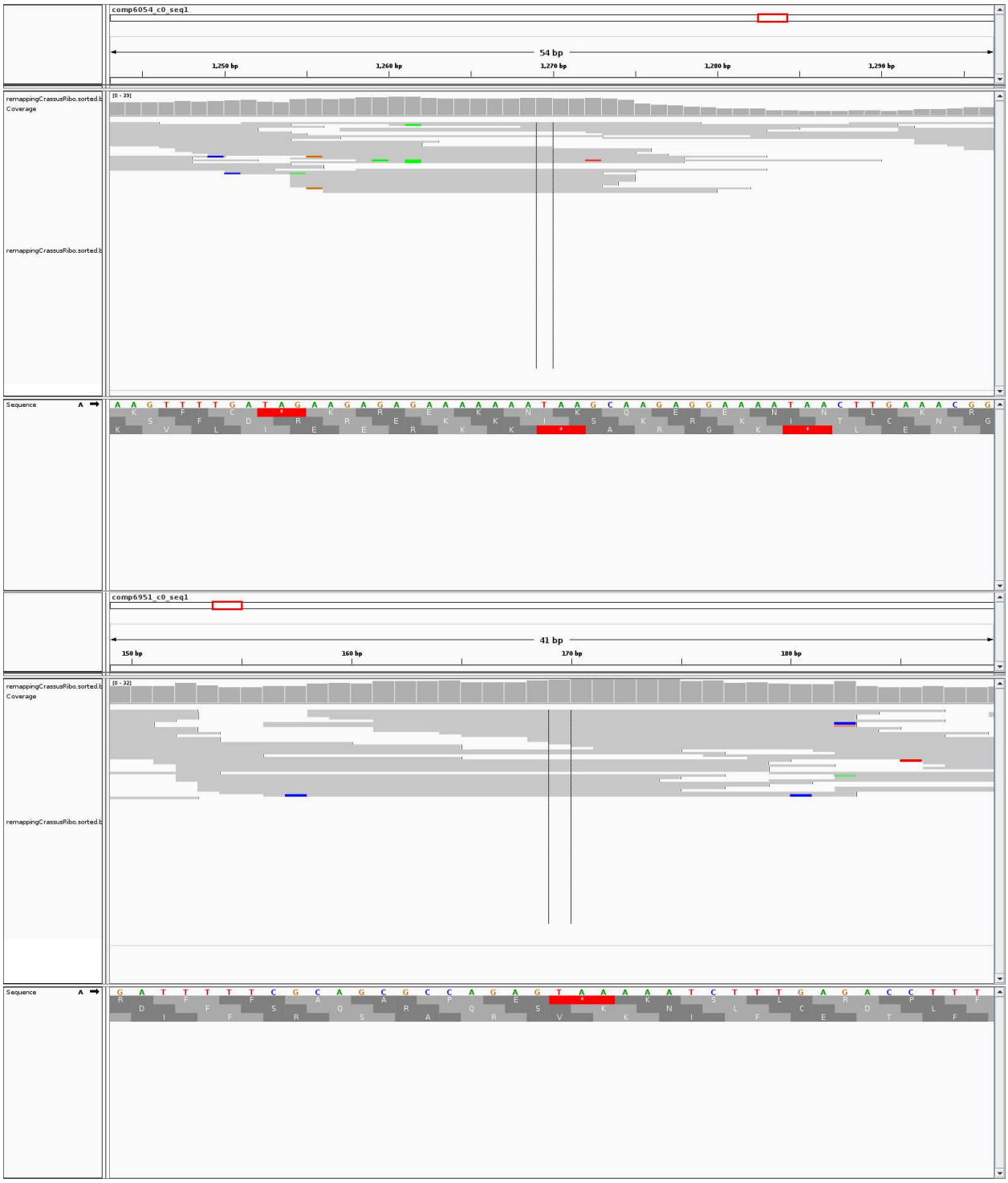
SUPPLEMENTARY NOTE 4. IGV screenshots of ribo-seq reads alignments in the vicinity of selected frameshifting sites




comp3853_c0_seq1
48 bp
430 bp
440 bp
450 bp
460 bp
470 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 34]
remappingCrassusRibo.sorted.b
Sequence
A
comp4582_c0_seq1
34 bp
1,040 bp
1,050 bp
1,060 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 24]
remappingCrassusRibo.sorted.b
Sequence
A


comp5116_c0_seq1
41 bp
130 bp
140 bp
150 bp
160 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 29]
Sequence
comp5528_c0_seq1
30 bp
140 bp
150 bp
160 bp
[0 - 12]

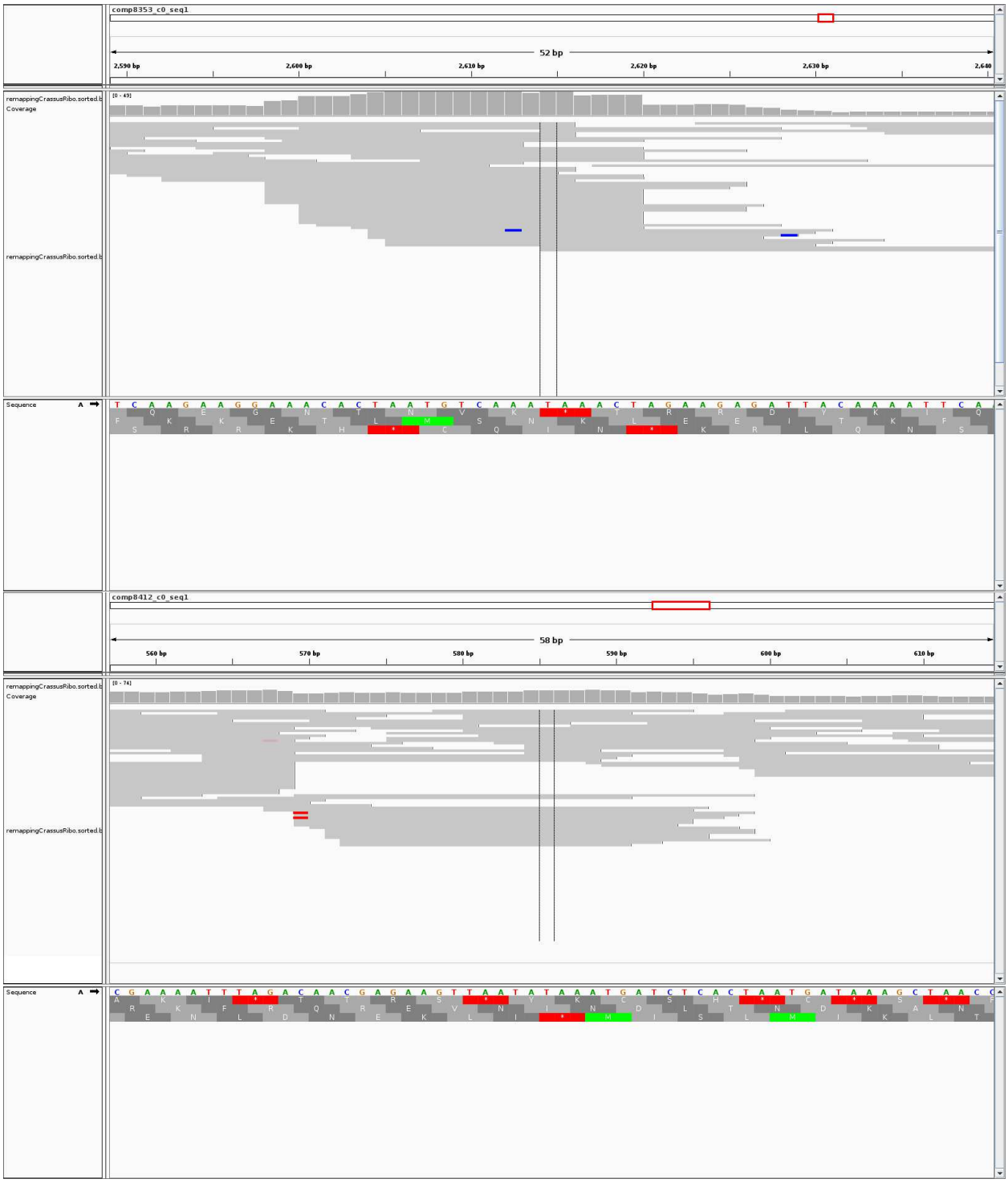

comp5973_c0_seq1
32 bp
3,750 bp
3,760 bp
3,770 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 20]
remappingCrassusRibo.sorted.b
Sequence
A
comp6034_c0_seq2
30 bp
570 bp
580 bp
590 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 27]
remappingCrassusRibo.sorted.b
Sequence
A


comp6054_c0_seq1
54 bp
1,250 bp
1,260 bp
1,270 bp
1,280 bp
1,290 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 35]
remappingCrassusRibo.sorted.b
Sequence
A
comp6951_c0_seq1
41 bp
150 bp
160 bp
170 bp
180 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 32]
remappingCrassusRibo.sorted.b
Sequence
A


comp7073_c0_seq1
37 bp
1,690 bp
1,700 bp
1,710 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 96]
Sequence
comp7194_c0_seq1
52 bp
210 bp
220 bp
230 bp
240 bp
250 bp
[0 - 102]


comp7341_c0_seq1
45 bp
440 bp
450 bp
460 bp
470 bp
480 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 916]
remappingCrassusRibo.sorted.b
Sequence
A
comp7670_c0_seq1
51 bp
1,880 bp
1,890 bp
1,900 bp
1,910 bp
1,920 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 112]
remappingCrassusRibo.sorted.b
Sequence
A


comp7880_c0_seq1
46 bp
1,160 bp
1,170 bp
1,180 bp
1,190 bp
remappingCrassusRibo.sorted.b
Coverage
remappingCrassusRibo.sorted.b
Sequence
A
comp7882_c0_seq1
47 bp
1,190 bp
1,200 bp
1,210 bp
1,220 bp
1,230 bp
remappingCrassusRibo.sorted.b
Coverage
remappingCrassusRibo.sorted.b
Sequence
A


comp8353_c0_seq1
52 bp
2,590 bp
2,600 bp
2,610 bp
2,620 bp
2,630 bp
2,640
remappingCrassusRibo.sorted.b
Coverage
[0 - 45]
remappingCrassusRibo.sorted.b
Sequence
A
comp8412_c0_seq1
58 bp
560 bp
570 bp
580 bp
590 bp
600 bp
610 bp
remappingCrassusRibo.sorted.b
Coverage
[0 - 74]
remappingCrassusRibo.sorted.b
Sequence
A