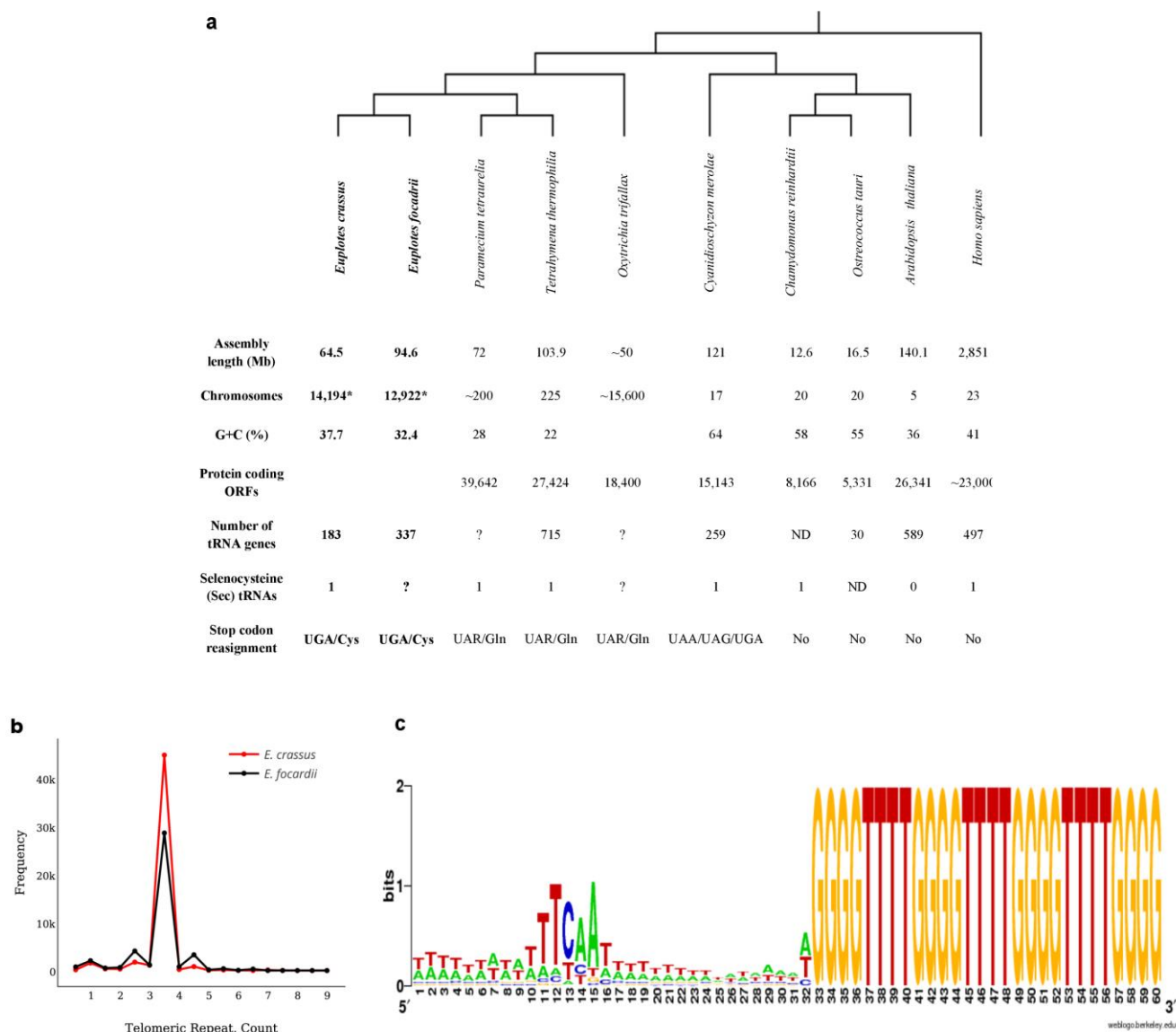


Title	Position-dependent termination and widespread obligatory frameshifting in Euplotes translation
Authors	Lobanov, Alexei V.;Heaphy, Stephen M.;Turanov, Anton A.;Gerashchenko, Maxim V.;Pucciarelli, Sandra;Devaraj, Raghul R.;Xie, Fang;Petyuk, Vladislav A.;Smith, Richard D.;Klobutcher, Lawrence A.;Atkins, John F.;Miceli, Cristina;Hatfield, Dolph L.;Baranov, Pavel V.;Gladyshev, Vadim N.
Publication date	2017
Original Citation	Lobanov, A. V., Heaphy, S. M., Turanov, A. A., Gerashchenko, M. V., Pucciarelli, S., Devaraj, R. R., Xie, F., Petyuk, V. A., Smith, R. D., Klobutcher, L. A., Atkins, J. F., Miceli, C., Hatfield, D. L., Baranov, P. V. and Gladyshev, V. N. (2016) 'Position-dependent termination and widespread obligatory frameshifting in Euplotes translation', Nature Structural and Molecular Biology, 24, pp. 61–68. doi: 10.1038/nsmb.3330
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.nature.com/articles/nsmb.3330 - 10.1038/nsmb.3330
Rights	© 2017, Nature America, Inc., part of Springer Nature. All rights reserved. This is the peer reviewed version of the following article: Lobanov, A. V., Heaphy, S. M., Turanov, A. A., Gerashchenko, M. V., Pucciarelli, S., Devaraj, R. R., Xie, F., Petyuk, V. A., Smith, R. D., Klobutcher, L. A., Atkins, J. F., Miceli, C., Hatfield, D. L., Baranov, P. V. and Gladyshev, V. N. (2016) 'Position-dependent termination and widespread obligatory frameshifting in Euplotes translation', Nature Structural & Molecular Biology, 24, pp. 61–68. doi: 10.1038/nsmb.3330, which has been published in final form at https://doi.org/10.1038/nsmb.3330 .
Download date	2026-08-10 14:42:18
Item downloaded from	https://hdl.handle.net/10468/6518



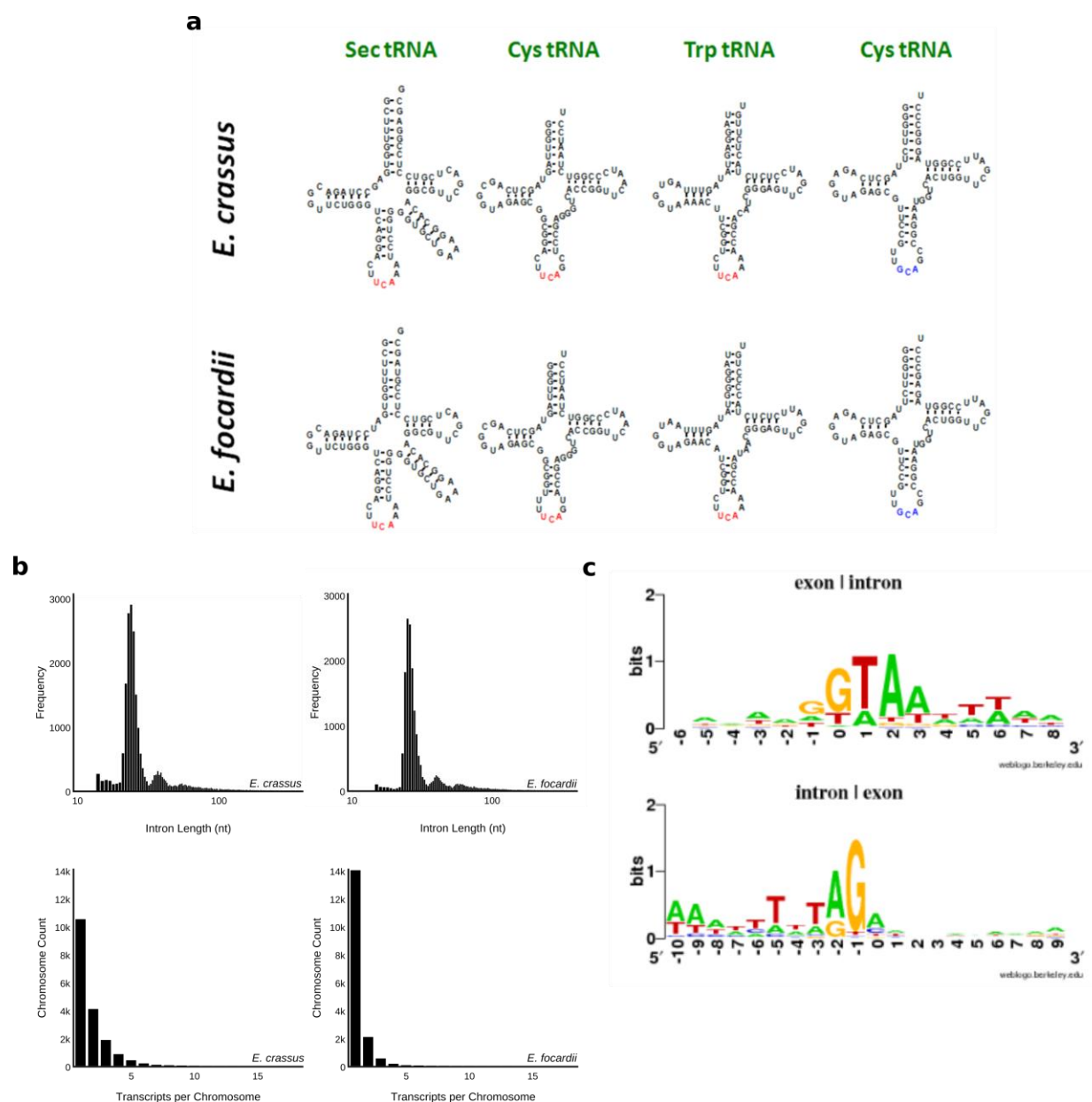
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh



Supplementary Figure 1

Features of *Euplotes* genomes.

(a) Comparison *Euplotes* genomes in comparison with the genomes of other representative eukaryotes. The tree was constructed based on the sequences of 18S rRNA genes, and archaeal 16S rRNA gene (from *Pyrococcus furiosus*) was used as an outgroup. *number of contigs with telomeric repeats at both ends. **(b)** Distribution of telomeric repeat lengths in *E. crassus* (red) and *E. focardii* (black) macronuclear genomes. The X axis indicates the observed telomeric repeat number and the Y axis their frequencies. As expected, *Euplotes* genomes consist of gene-sized chromosomes capped by telomeres. The length of terminal repeats slightly varies; however, most chromosomes in both organisms have a double-stranded telomere length of 3.5 repeats **(c)** Sequence logo of subtelomeric regions at the 3' end of *E. crassus* nanochromosomes. 1000 randomly selected chromosome sequences with telomeric repeat GGGGTTTTGGGGTTTTGGGGTTTTGGGG were chosen for constructing the logo. The logo detects a conserved position-specific sequence motif associated with telomeric repeats. Abundance of high-quality telomeric sequences allowed an unbiased screen for motifs and patterns associated with telomere function. A previously described TCAA motif (Baird S. E. & Klobutcher L. A., *Genes Dev* **3**, 585-597, 1989; Klobutcher, L. A. *et al.*, *Proc Natl Acad Sci USA* **78**, 3015-3019, 1981) was readily detected with Weblogo (Crooks, G. E. *et al.*, *Genome Res* **14**, 1188-1190, 2004) in the subtelomeric region due to its conserved position relative to the telomere repeats. An analysis of sequences in the vicinity of telomeres with a pattern discovery suite MEME (Bailey, T. L. *et al.*, *Nucl Acids Res* **34**, W369-373, 2006) did not reveal additional common motifs.

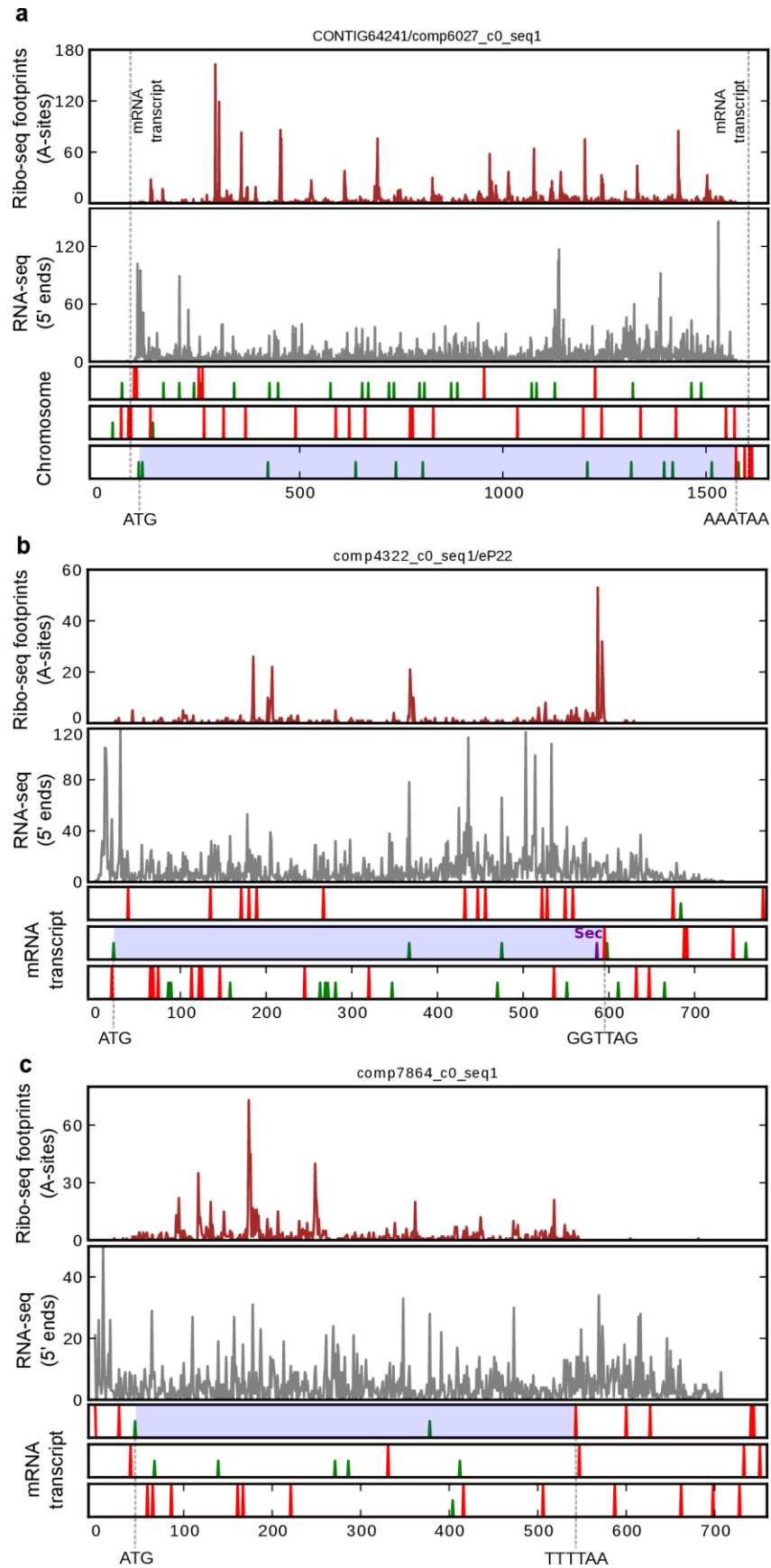


Supplementary Figure 2

Features of the *Euplotes* transcriptome.

(a) *Euplotes* Sec and Cys tRNAs that decode TGA codons. Cys tRNA with the GCA anticodon and mitochondrial Trp tRNA with TCA anticodon are shown for comparison. In total we identified 183 tRNA genes in *E. crassus* and 337 genes in *E. focardii* based on their genomes analysis. **(b)** Frequency of introns of different lengths. The X axis indicates the length of introns in nucleotides, and the Y axis shows how many times they are found in the transcriptomes (log scale). Short introns (~25 nucleotides) is a characteristic feature of *Euplotes* transcriptomes. **(c)** Frequency of chromosomes with different numbers of RNA molecules transcribed from them. The X axis shows a number of transcripts per chromosome, and the Y axis how many such chromosomes are found in the genome. **(d)** *E. crassus* splice sites. Nucleotide conservation around exon-intron junction and intron-exon junctions. *E. crassus*. Transcriptomes were assembled *de novo* using Trinity (Haas, B. J. *et al.*, *Nature Protoc.*, **8**, 1494-1512, 2013); no genomic template was used for the assembly of the transcriptome to ensure independence of the analysis. The assembly procedure produced 33,701 unique transcripts with an average length of 573 nucleotides in *E. crassus*. We obtained the *E. focardii* RNA-seq reads from (Keeling, P. J. *et al.*, *PLoS Biol.*, **12**, e1001889, 2014).; this assembly produced 28,869 unique transcripts with an average length of 667 nucleotides. To identify introns we carried out pairwise alignments between the genome and the transcriptome for each species using FASTA (Pearson, W. *Curr Protoc Bioinf.*, **Chapter 3**, Unit3 9, 2004) In total, we identified 21,798 introns in *E. crassus* and 18,747 in *E. focardii*. The most

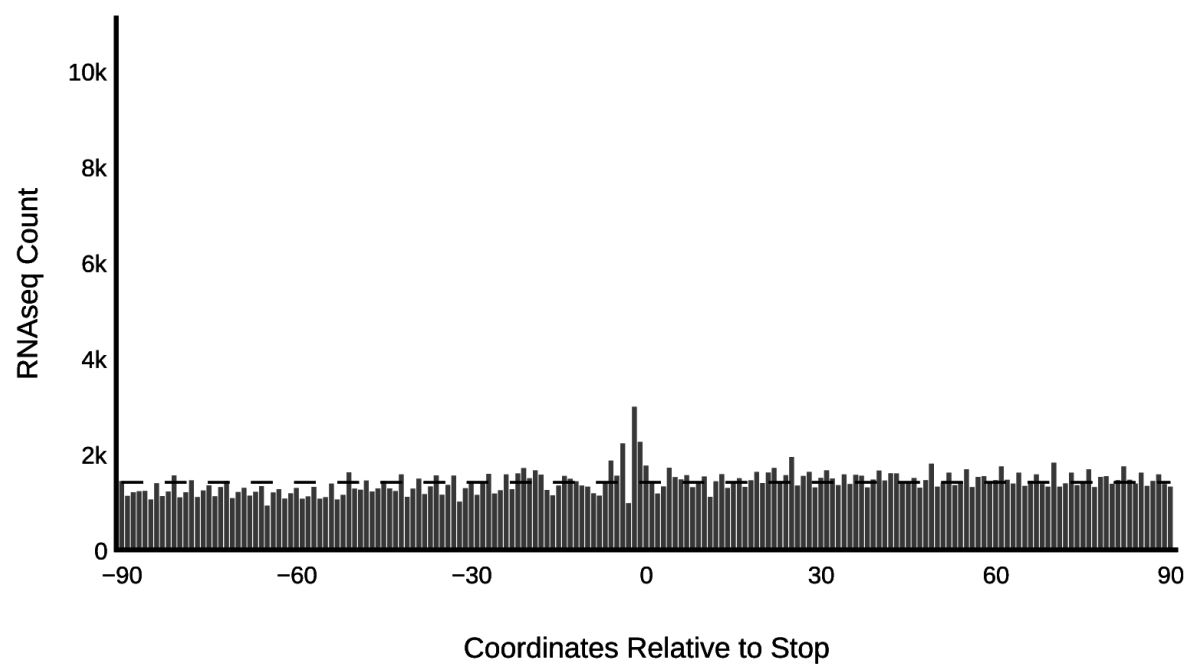
frequent intron length was 25 nucleotides in both *E. crassus* and *E. focardii* with 2,895 and 2,631 occurrences, respectively. Using 10,000 intron sequences from *E. crassus*, we characterized sequence features of the exon-intron donor and intron-exon acceptor sites. We further aligned 32,350 *E. crassus* transcripts or their fragments (96%) to 18,032 genomic contigs, and similarly aligned 21,233 *E. focardii* transcripts (74%) to 16,950 genomic contigs. The majority of chromosomes had a single transcript aligning to them, 10,495 in *E. crassus* and 14,082 in *E. focardii*. Some chromosomes contained two or more predicted transcripts, which could be, at least in part, due to insufficient sequence coverage. Low coverage can result in missassembly of a single transcript as two or more, when reads matching internal positions are missing.



Supplementary Figure 3

Termination at AAATAA and two mRNAs with long 3' UTRs.

In each panel ribosome footprints (top) and mRNA-seq reads (middle) are shown for a transcript whose ORF organization 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 site of termination. Translated segments of ORFs are highlighted in blue. **(a)** An example of mRNA with termination at AAATAA. **(b)** mRNA of selenoprotein P22. The position of UGA Sec codon is shown in dark blue. **(c)** A single detected example of an mRNA with a long 3'UTR not containing SECIS structure.



Supplementary Figure 4

Metagene analysis of RNA-seq density surrounding frameshifting sites.

First nucleotide of a stop codon is shown as a zero coordinate. Only minor alteration of density associated with sequencing biases at specific nucleotides of frameshift sites can be seen.

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

Species	Assembler	Assembly size, kbp	Number of contigs	Number of nanochromosomes*
<i>Euplotes focardii</i>	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
<i>Euplotes crassus</i>	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
MDYSDTPQEESTHSYDYDLFVIGGGSSGLACAKVAQEAGAKVAVADSVKPTPKGTKW **KVGGTCVNVCIPK** KLMHYSALLGNSYHDQVE
SGWEHEKPSHDWG **KMITNVNNHIRG** INFGYKADMRKRGIKFHEKFASFVDPHTVQLVDKKGKTEMITSNYFVIATGGRPLYPDIPGAKE
HAITSDDIFWMKDNPGKTLVVGASYVALEcAGFLHHFGNEVSVCVRSIFLRGFDQDMAQKIA **KDMELSGINFIRDSVPTKIEKDEET** GK
LTcFLTVGGEEETTVEVDTVLFAIGRYAVTADLNLGNAGLIAEKNGKFITDKYQ **KTNVNDNIYAIGDVLHGKLELTPTAIQAGRI** LADRLE
AGGTTTMDFYDVPTTIFTPTLEYGcVGyseEDAREEYGDFI **KVYHTYFQPLEWNFAKS** IYKERNcYVKIIVNTADNDRVIGFHILCPNAG
EITQGIAIAIKVGVTKPQLDNCVGIHPTIAEEMTNLHIDKADNPDPKSDC **US**

>eP22
MESSDDKVGCVQSILVVLEGLNDDSSIITGLEILIKLIKNIKSPHEEKFRNIKKTNKAISTKLSSLGIEDLILALGYKDDNDEFYVF
DIDKYSPLYKLKRAIQEFHDEKRKKYMTPEELEKFEILQEQRKFYEDNKKKAKARKDLENGMKFDREKNQEEIKSS **KANHLNFGANV**
VKE QPPAPASR **UG**

>eSelw2
MDSTTKGHIVVNYCGG **UG** YLPKA **RYVQEAVENRFPGDFSFDLKA** DVGKTGR **LEVTVFVGDDTEGKLVHSDKKGQG** GFVKDSNVDSVLD **SI**
AALLE

>eGPx1
MGAALCFKKRKE **KLETTVESLFEISAEDIDGQEHLLADLAKDKK** C IMVVNVASK **UG** LTKTHYTQMVKIHNKYKDKGFEIFAFCNQFLS
QEPGSNEDIKKFAREKYGAEFQLFS **KIDVNGPNTHEVFRFCRRHSPLYDDETDITQNI PNWNAKFLIDEEGNVVNYSPKSNPDVCVPM**
IEEMLGL

>eGPx2
MGQVFFKSKKEKLATTVKSLFEISAKDIDGQTHLLADLAEGRKCTMVNVASK **UG** LTKTHYQMVKIHNKY **YRDHGFEIFAFCNQFMSQ**
EPGTHEQIKK FAQE **KYGAEFPLFSKVDVNGPDTHEVFKFCR** **RHSPLYDAEKDVVNIPWNFAKFL** IDE **KGQVVEYYTPKQNPDL** CVPKI
EEM LGL

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
MDNIPDYLVRLNGTSFLDRREEILSIEYNSIFTFAEFKMEACRTLVRKYDPKCRFLDKEGIETFEDDLNMLKSHDVLYLASRGEDFDY
SSVLNDYDRKDVLEGGGFGKVYHAVNRETGEDVAIKFMDISHYLTADQIEEYREADALQKLNHSHIISLHKAQVQRKEVILIMEYAG
GGELKDRVEEMKMDDEIYARFIFQQICSAMSZYCHNRGLIHRDLKLENVLFKEKGGMIKIVDFGIAGVCKPQEKEKTDSGTLSYMPPEV
LSGEKLEAGPGIDIWALGVMLYTMIGKLPFYGDTEDEI INCI IKKKPSFKDKKTISKELKDLLIKVLNKDSDKRLSMFDLQNHKWMEM
QDEEILKSIEESKLEQEQUEEEKK **NE** EDELIAFDKLNKDDKKSSKAGSDYNSLSAHSSNPSGGRKKKKMRGSSPRSGMNGTGKKKKKT
IKKKAT

>comp8353_c0_seq1 AAATAA
KKAHAVFNGGVLTALCFSKKRRTTLYTAGDGTFLVWPVGAKPNPNQSVEPADFFSSPDL SNIPQIDNETDDSVKYYKELLQEQLDSEIP
RKKEFKAEISKELEDIRSHIVDLVEDNRKAQEIEQLERHEFVIDVKKQSMEQDGWQQRQLITQAKRKQLEYECLKERVKSATWSTME
THSTACISLSDSLLLLYNGIRLRTPLEKKT LNRLVLFHFRMELRQQITGMETRANKILDQSLFSNHDETYIMNRIRGVQHYEVDEEAPII
TQAKSATKRKNKIAQQEASKSTADGALKKGKRKPKVDLNQFRLGANKPKLLDEDDFDDKLKDRAQNRDESNI AELRWKIVTKKKELEDL
KKDIHILGSWDLLEYPSDLYTDGRKKMQIEMLDVV FALKQEYNKEFERFQRFKDDQIFAIQERSNRITEILEDLKREEELFHPRTHPL
ETPASILEVKPEEVTQKYLTAERAVEDEKHRIEQLERLKALEGDVNGQRGIKNMLGGTYELKKNKGIMEETLEREEWMSKPIEDMTED
EKLKFKFQQREKELQEEKDKRKAQWDELKKNRIEIEEICFKFEDELKKIHKRLYYDMRVYEQELYIIRLTLMLHENKEIKQEAALI
AEKKDKLEQELVESKNTINNFQRMIEDFDSEFKASSAIAEQEKGLRDLFPNAPFRQILAFVRNGGKAANRARGFRGQTEENTLEQEALA

LLCKLDPYIIVDENAVKAKFEQENVKEEYSYDRDKIVNLTPGEFDTLVQERENRNKIDKERKGMEQEIANLSGHKEFCEINANDLEEAY
 EDIKASHTIEIESRMEKLYNF EAVVYMLQGQVEVAQAPVATDYKDAILVNTGVIEDENKKVVOEGNTNVK KLEEITKFKRKLNHETWK
 NDKLKLEIKDLLERAIDVQLYKVTKDTQEIIKGNHRTKDEDEKKRLEDQINNLOENAGARIEVINKKKKKLRKEINEKRKENNELETRA
 RDLQTNVDQRNLIIDLR SKGPSGGGDDRLQDPIKR FKEVATVRKYKEIVDQQKEEIEFLEDELERFRSRTFSPFANMHARQDYAD

>comp7194_c0_seq1 AAATAG
 MDTDLIDQIQKNMDQDPQLKDLFEPSSSEENSQNDHGFTQGSKKFYSDSAMAPPKVSRSKER ELAFLK RAQEIIGLEPYNEYHGKKKTM
 IKKQTPGKDLIFSNIRHKESTSQQRRDHASRDEQQLSNKSLALKGTGTNKEKIQKHKLHQSQRTVGKKTFECKEAKPSNGVAQKRVEVIE
 ISSKTSSSGNYSTPIESCPIENCPSVESCPPVDHKTHEVVSLSNSSDDKNIDDQPLNPKQKKRDKKKEQVDAKNARDFDSNPPKKS
 MVSSTPTVNSEVMDKYLQQTTPDQIEIVEFDSRPKWPTENCNSNFQDQLKLMA SQRKRNKADTLGSNMFMQEMKSTLSNIEASMDSPQGP
 QKEYIHLKESIYPFWQSTFHHLEWNDSDANKIKSREELKERMLDSLQYFSGHKLYRYADPADVKRGLLQNYPFVEDSDSKTEVKLLEG
 KFHMMLKIIVDGITKKIRVSIKKDFT

>comp2566_c0_seq1 AAATAA
 MWGRKKKAEPKKSETKKRAPAKKATGTRKTKPPAKAKFSKLESKEEVKEEIKHDSETIFKVYATDQEMGRPIIGSEGIQN LASDLKLDI
 ASSAELIVFMWHCECEEYQGQISKSEFQKGC DK LGVKDFS HFK KSVPKKLSATLAMQDTPKEFRPFYKFAFTFHR TDGKNVPVETCQVL
 FGLIFSDKYPILKTFFKFLAEKEVTHLTLDQWDSTYDLIRENPENLDNYDEYAAWPTLMDDFYQWYGENK

>comp6054_c0_seq1 AAATAA
 MEDEKNESI QNFKAMAECDDDGIAFYLDSSNNWDLAQAYDQYQNT HQFNQSNPTSTPAPTSFPGGADVDMSPGADAEESA FDDIPDIPN
 IGIPQMDQSPVPQESNAPGSGGLNITSQFSNFASSIQSNLQNL TGGMFSGVMGGGMMGMTDMNTQSNFSNRNLTA AQEFLFQFRKKNGM
 HVILPKFVNNTFEEIGQESKRLRRPVFFYLHNDKGDSCNIVDQSVIGEEMTRMLLNKYICVGVNVNTEEGRKLLTALEIPKAPFIGITY
 IDENGT LQNI GSRSGDEINVMALSEMDEAASGVFNAIFDGD TDTLTFHIEDSNLQLLETEEFKAEISAQMGNRTFDGYNEEP PRRRGPE
 IDPTTGFPVGMTPQQIQDKILKDQQRQEYAEIDEKNKVLIEER KK KQEENNLKR KEELEKKAKIEKLEEEKKEMAEIVRSNLPEEPSE
 GTPDTITIQFRFPDGNHKQVRRFYKTDKVQLLYDYITSFGNENGFEAAH THFSIIQNFPPKFFEDMNKTL EEGLSNCTLMIKEHSHVE

>comp7882_c0_seq1 AAATAA
 MNPKGSKREKRVKSGSNKSAKEAFLKKIDAAMKVYDYVDETKDVKGKSERLNAINELQNL LDQDKSVSQLIIPNLDS CMQMI EKNI FRC
 LPNIKKS NLAFSETGIDQEEETDPAWPHVQGVYEFFLQLIMNDSIEVKLLKGYVTPEFVSRFLELFDSEE AVERDY LKNILHKLYAKLV
 PRRKMIRKAINETFYQLIHEGHKFN GASELLDILASII SGFAVPLREEHVIFNNIIIRLHKVQTCSEFFEQLLRCSMLFLTKDKSLAI
 SLLKGLLKYWPFANCVKETLFLTELQEVLEIVDDDKIGDLVPLFRRI VKCIGGTHLQVADRAMCFFENDYFLTKLRIYKDVTFPMLVP
 VIVELSENHWHKILQESLVALKVIL KEIDSAAFDEAQQISK KDHRRFIVKPNVEKRTELDAKWERLNTTLKSTSAGFTPPDVPFKTSE
 LISNYNGLYRKIYDKEKFIN D

>comp7341_c0_seq1 AACTAA
 MNNMIDCNHSPSLSDSVGADGDKCECFSGERLEGSNNLPEEGITSISPDSL VNKASTLLRALTLFTSFSEFDTQNSPPP KTSKLFNDFS
 LKLNMMKTCLSQA KTS DSEKLPALKEEIESI KASIGEELAL TPLGQALLWN LIEGDN KDSSMKI FDEL DHRCDIANLHEVIAQKDA
 EIQTLSKQIRKLAKFRRTSSLVSEETEDGDSASQSDSGSMTQLSRSSLLNKLNLNTKSRLTCLN KVRDMLVKELKEPLQKINTLSF
 NPGLLALEDAKEFLKNCFPLEVASFHFNKDSLRLNDLEKFLDVLRLTNEYVTDEIVLSNFVIDQDSL VKILSNFKNKEVV SFNSCKMSL
 SNPPEFGDSL D GATLKHLYLNF CGDKSHGDWASNPAHFENLINGLSHSPDLKASLKDIWMEGSGLK KDKARDILDTFGFHSTKIWI LYG

>comp5116_c0_seq1 AATTAA
 MLPKPTNNMEKIKQQEYKYKIRKVFECIAK ESNHNN DITN KNEIAYILR YFSQFPSEAQVTDYVIQKIEDDEPNDFIKLTKFEPYLL
 DVIENNEFEPSPPEHLLAAFRVLDTDKGRIPIDVLNLLTTEGIPFRKEEMDSFQEFALDKSQKFVY YEDYVAKLVEENDKHVEEFLKE
 YPTFKPPINQ

>comp7670_c0_seq1 AAATAG, AATTAA
 MDEETIEKYCQALDLSVSPDNLRLKQAEAFIEGME TPGFIAAMLHISSNPDLNRDRKIDITQAAAIQFKNIVETHWKYKDDEYAKEMR
 EDGYKVIIIPDET KTYVKENILTAYINVHSEKVAKQDFDIVRCITKHDFDKWPD LANKVKDYIESDDLYGSEM FVGLYTLKSICKRYE
 YEFDAKREPLNEIADILFPRLEAITTCVEGDNSDQGSRLKNLIGHCFYISNQISLCKRYLDPSMLDFIVKFNTSALEAEIDNSLTQPT
 SIEEIDHRAESFQWKLKMTAMNFLFRIFQKFSNPQYVNETMKPIAEHCINNYAEGIINLANTLIAKAK SIYIDRQVLSYCFKVSTSI
 NQTSYREMIKPLIPEILTSHCV PAMLLTEKDTEDFEADPVEFIRKARDPNPNIYTARN SVLEMIRNVTQH KSNQDKGALPDFLESFFGF
 LLENLSECIKQDAPDFRIKDALLLCLGQIAPTLLMYDQFHDQLNQVLTGAVFQDLTSENELVKYRALWVYGQCSRVPMEDDHRLEVGKL
 LFQLMNDENTAVKITASTSLYKNLRNNSMKEAFKSELASILEAYLGLMDTIDNEELIAGLEEVVSLYEDCIGPYAIELCSK IVENFN K
 ITGKEQEEEE YGTMGMATSGLVVTIRSIINSCKGDPETLLKLEPVIFPVVRSLSADGCEYLDEAMDCITAILNFTQSATERMWALFPH
 LLIK IIVGGPEDEEGGYAFDYFTSMEDYFR SLIKYGHGDM LTKKIGNDPVMILLIKGIIKILQLVKEGDVNTNAYICIVIVETL LLEFP
 KLDQLLPTFIKILCTELSNKEITKEFRLHALTLVAHCFIYNCTLT LGALTDLVLPVCQNFFSYLKKFSEVEHLRGLIYGITALLRMD

EMPDVIKGSIQKIIIESLIDLMRKYTRERILELRSKFEDKRNRWDEGTDEYNNLDAPFQKLSEWMEEYKDDAYSEDDDDDDNFEEDDYLW
SRSDSCYYKSCLEDKEAPLFFKETLEDFRENKEEVYRGIIEIIPEDSQKLLEMIMERCCEYMQSLQS

>comp5528_c0_seq1 ATATAG

MSSQEILANSITNTVDKEKSAQEEQDDEVIIDDQNPLLEDDLQIDEPEQKVNTDEPDQRNQEDEASENEQNLSDFINNTESYQSSST
TQNLKNLLIQSTIGLALKLKPGLGMLITNSDGGCCDDIRKSLDLNKQLLGEDVADLISVKQITWDESLQVSGTRYDYICITGSHFSQE
FVQILEKIVPTVLSVVDQERVFLGPTSEEDISQFEDNVGDTIFESKKEEIDVPTKNSNSLGQFGDPDNHYSENKDLIDEGIFDDQ
VLQREGHLGLPSLEDDNEDDYFEPIG

>comp8412_c0_seq1 ATATAA

MSKNTKSKQEIDSTSKKLSRKERNLEYIQYAKERKEYQKWEKEEADNLGFEGEESAPPVQNTSTAAGEKKGKKEKKPTDKVTTSP
EQKKAYSETVAKQDALIAELQVKLDQREKRMKDLKKTQEQLSKLDQTMKLTKERDDAKASLSKVEDKCNRLGDYQELIESVNREN
DNEKLIINDLTNDKANLKDIVFDMFEKQKEGNSATEHPVVIPDNLQEEFNRSQTKSSQNNPQLKTLLFANEQIKRLQKEIKEARNH
LPKPELGEIDEATEEKED

>comp6034_c0_seq2 ATATAA

MSKNTKSKKQVTSNAKKGGNKKGKAEPVQPPKEKKELAEWDLEDMPNFGFEPKKIAPTASKGPAVSGDKKKKGKKEKKTVEDTLISIE
EAKRANPEEIRARQETLITELKSQLEQKDQAIADLEKDQKEQFKQLTEQAQQLTEERDETRAALAVAEGQCNQKLDQDFKQTVDRVNRN
FLENEKLISELTSEKSNLKDIVFDMFEKQKEGDSAPEGEEEEITDEITDEIHGEFDRNTRRQAPQDNSQVKTLLDFANEQIKRLQTELKE
VRNQIPESALQELDQIDETLKETED

>comp2483_c0_seq1 ATATAA, AAATAA

MEKFGDLRASRHRVKHSMKSKNHRQYEQEEVKHARSRPDRFPDPPQIDEEKSKYSAGIDEAIIQLVEITEKGVCKINPIAMNIVKGIKTKV
GIISVVGPPYRTGKSFLNRLGQQDGFEGPTVQSCTRGIIWGWKPVKVSSEDMHVLMDTEGLGSCNRTMNIDIKIFTLSVLLSSMFVY
NCLNAIDENALEVLSLVNLAKEYISNQKKNDSMDVYNQANYSFYFMWVVRDFSLQMMPSSELEKAGHDPATYWDKLENQEAKEYLEK
SLEAIDLGTINEENKRTVTKKNEIRKAIKNFFHQREATCLFRPINEEELRIVNKIPYEDLRKPFQVEHLINKIYYNVKPKSINGQT
LTGKMFAQMLEEYTSMMNNGMPEINTAWDRVMDTEIKRVLQESTTKINYQLQEVVIDKMPMLKQLISIERNVRKSALKLLYDPNKN
APKDKLSRLQDKFIENLDEIFEGIFNENEIISKRAKDLPRMYQKIKAMINKGEFETIHDFSIIYGKMAISYFDNTNEPENYKIION
FQINTVFEDLDEIMQTQVQRHESQNEQYETKLETKDHQIEHLNEQLKKEKTKNKDREQLRSKNMNIRSNLEEEIQMIKNQISNKDQ
FESLGTMIKEGWNNSEQVLKEIKAKEIENIKATIALESQKKIKELETTHQEQLHYKKEMKKALEKTLAGLKMSYEDEIRLLKKNVKDQ
DKKITLLKKMCVRKDTQIQMLEEKAQSNEKQDKLQKEHRDILFELARAFKEGTPGKDSTTNESATPY

>comp6951_c0_seq1 GAGTAA

MYSTKFRVMTMAPLLLANPALALCEEPSTADRIRGNYNKIRFFAAPEKIFETFSNIREEDGQVYMSYQDFFHSLTPYNFVASKDDD
DDDDDEENKDKKEKEPEGYFDKFTPEIMTIVDANQDKKIDFNEYIFFITLLQLPEGEVMRIIEKVNPEERKINKAQFAKYLTCLRKCTA
LGLKQMSKSFMPDGRKISTDEDHISKTILLHLFNDKEYITIEDFCELKSKLKHALLHYEFYQFDVDEDETISAESFAKSLLSCLNYTQA
SKYSRRIHSLKLEGRVSFKEYVAFHNLIKADIIMKISTYRFLSLGMFRDLCDDFAKLDPYCNQNKVISISDTQIATFFKVLEDEDENGA
LEYDEVVDILEGKKNIGLGKEDKFKREMMEKIDRYIKKFQKYVGWT

>comp3853_c0_seq1 GTATAA

MSEENKEEVKGTTHTDQYHHGFGNHFSEAEIAGALPKHRNNPQCKFGLYAEQISGTPFTYPRAKMQRSWLYRIMPTVAHPPYKALK
DYNLWIANFARDDDEEVFTTPQQMRWTPIDLPSSEITFVQGIQTVTGAGDPSMKAGINMGVYTCTNSMKNEAFFSSDGDIMIVPQLG
KLSIMTEFGHIEAESWEVVIPRGIKFAVEVNEDCRGYCYCELYDGLHQPDLGPIGTNGSANPRDFAIPKAKYFDETNEFRVIQKYLK
FFEYTI PHNIFDIVAWHGNYPYKYDCHHFTMGSI SYDHPDPSVFTVLTCQTPDHGQAALDFAIFPRWLSMEDTFRPPYFHRNTMNE
FMGNVAGQYDAKEEGFSPGAVSLHSCMSAHGPEAEVVEKASTCELKPQKVGEGLAFMFETCYTMKVTKSFMHDLGATDSYSVNSSKA
VDESYHDCWKGMKRLFDPNPDAGYKKYLSEHKN

>comp5973_c0_seq1 TTATAG

MSHLKNFQFSSVQITEIDTYIEHLYSENMDLKLKGCISILYLCFSAENMEEMIEHESLLPAVSRIILRDYKKSLLDLSLYLLNVFYAYSH
FTEFHPLLIENQIGDTCVKIIEYEIKRYKARVNEYTKTAQLVKQTQQTSPADTDLKELQNNFRKEEKRLSVTIKKQEKVLFVTFHILLN
LAEDLKIERKMKKRRIVPLLVSMLENNPDLLYIVLSFLKKLSVFGSNKDDMLELDIMKKLNRFIPCNALLTQTALRLLFNLSFDNEI
RERVNAIGMIPKLVELLKVAQYRSILLRILYHLSSDDKIKATFAYTSCIPLVYQLVIHFPDAIIGKELIALAINLTNTKNAALISQDD
QLEALIERAFKYNDVLLFRVVRNIAQFGPVTNIDIYEKMDKIIELTKQCGDNTDLQIELIGTLVYINIEKWDTVLSQGDFLDFIHNNL
VSDYSEDDLVLLETIMLIGTMCRSEKCAEAIAGSYIIGMLHELLGAKQEDDEMVOQIILYTYHRLLYRVYRVTREIMLEQTQIVNVILELLND
KNPNIRKLVNSTLDLVQLHDEIWKQEI KTKKFEMHNEVYILGLMEEYEAQAEALDEEALYDYAQAQDPEALAALENGEFGEDDQWLDQND
LAQRIWNGEMDPDQMMDPNQMMDPNQ

>comp4582_c0_seq1 TTATAA

LTVDSFILLADKKNCITLFTFQDLISKIARKKHIFALNKNNEFAPNPMIGFIQNLCDKIYTTITTDKEGKTAKEFLQDFEYCHNEEAET
 DIPKPKIMKKMPPGIKKRLLADYNKVEAAKKEVSKRAKNKVVISSSKTLLKEAFDLQDYHSFEYLFQYVEDKGIGYAEELLHCLRNESN
 RKIFVLILDYVLTTLPEEEFE **LIDVTNTTTQE**ISLRELFPDIDLKEFCLALYDSKNVPGEIPLKSKYLYTKLEVYTKKYSTLDEKNRTK
 FLTTSLLVTGNTNPNEGYECLITKILDIISLYNIEIPIIYEGEVQEALRSNFKKLIFIRNYELVLKLQEIVKKNLRGLYEK**L****SQEHLS**
YISRLVQNSDAELGLEKDLLDSNEDVGVINSRQAVKDTLVFCLEQITLFD SYNQINFNDAPEKIIH MVKGFIHLGGFVNINVGFDSLDN
 KEVNEG DVSEIKRLVTVTQEYKEAKHKFSQRLLLTEFFSTFQKYNDLSYELIDVDSPLRWIIDCPGQESPNFFEYCI EQGNIDLAMKLI
 ESCDISEVISMFP LQERTISNLLNSPHIFEFLKKMSKEEEKIKKLIDRTNILEIPIMKLENSLSIDFDDEEDGNSGKPKYTQDQLTLYY
 FCYLKDSLVPKVGKEIPYFNLLFFNQGF KYSLDELVKVLPLDKIKELNSLGYQIGKII SQKPVNTKDLIEI

>comp7073_c0_seq1 AAATAA, AAATAA, TCCTAA

KVSKESESLSQNK NKPIKRRKITEDDKVEHLLSNSNSNSQQQVNQVKPREEIKQPPQDPHKDQNM DADMARLESLDIPKVGRQPDKHKD
 HPMETDHDQKPQADANQQARDPEKPVRVPEDMRI PPSQPHVNP HLTEAPLRDAPSSQPLRAPQASPIHEIETAKKGKHVAPEVIRPDND
 VDMSKNMFENKSDRPMQQERAQVVTTPQFTEQVPQRKDKAVVHKSISEIKKENDAPNHRDRKGRANDLSAK**L**KLKFIDKYSTSQGRKE
 LGNMIRRISGPQVQGIVRLMRQFHVGNKEGKEFKFSLNTLTPAQCARVGMLIEGISDPGSASSTGRAQTGKPGSHGERSSSAVGSQDAS
 GAGRVSEREREIERRKRAEEEEARYKERR**L**KEHEMKLQERKKDELRRKEQEQRKEENRKFKEQQELLRRQEHERQQESDPHGPSYESKS
 PVPPTTSEQQEAARVKAHQEQLEQKRLEEEK RKQAEAERERIEQERRRAEEDKRRKAEELRKSEEQERQRELAKRLLEEERRKKKEAEE
 QRRREEERKRLELIKQKEEERRRQENSSPSK**KSS****LKACEEERRKRE**QEQLRKREEERRRQOELEQKKKEEQRIREEQERRKREMEELR
 IREEEQRRRQEEEDRKRQELQRK**LKEDEERRLKAEQERKQREEQRRIRREEQERQRR**EEEQRLREEQERKRKQEEERKRKEEEERKR
 KEEEEEERKRKEQEELRLREEQERKRREEEERRRIEEERRMEEEERKRKEEEERKR

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 -XCDA
```

At the second step the peak files were processed with [DtaRefinery](#) to perform post-acquisition recalibration 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 Downloading 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 available 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 leave peptides that resulted only from proper trypsin cleavage events. That is we won't allow peptide resulting from irregular cleavages.

```
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 loosen 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] "SAQEEQDDEVIIDDQNPLEDDLQIDEPEQK" "WTPIDLPSSEEITFVQGIQTVTGAGDPSMK"
## [3] "ESNHNNDITNKNEIAYILR" "KKKQEENNLKR"
```

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

```
meta_trypscX <- 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.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.07e-20	1.10	index=8896	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	7.29e-19	1.10	index=8897	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	2.17e-15	0.94	index=8895	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.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.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_22_13Nov09_Falcon_09-09-15	4.36e-10	3.76	index=9720	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_23_13Nov09_Falcon_09-09-15	2.47e-09	1.64	index=9440	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.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" "WTPIDLPSSEITFVQGIQTVTGAGDPSMK"
## [5] "IIQNFQINTVFEDLDEIMQTQVQR" "KSSKACEEERRKR"
## [7] "LINDLTNDK" "LISELTSEK"
## [9] "IVENFNK" "LSQEHLISYISR"
## [11] "LINDLTNDKANLK"
```

```
meta_tryphprp <- 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_tryphprp, 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.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	2.18e-21	0.78	index=8638	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	3.05e-21	2.11	index=8646	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	4.19e-16	0.82	index=8639	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.19e-21	0.70	index=8806	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.20e-21	1.57	index=8812	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	5.49e-20	1.64	index=8802	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	4.33e-15	1.53	index=8810	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.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.LSQEHLISYISR.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 difference 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      0.0222222 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      0.0222222 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"          "SVNRENLDNEKLINDLTNDKANLKDIVFDMFE"
## [3] "NLDNEKLINDLTNDKANLKDIVFDMFE" "NKIRFFAAPEKIFE"
## [5] "MQDEEILKSIEESKLEQEEEEKNE" "VYLGLMEEYE"
```

```
meta_gluc_hrp <- 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_hrp, 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.SVNRENLDNEKLINDLTNDKANLKDIVFDMFE.K
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.33e-07	0.57	index=9974	4	E.SVNRENLDNEKLINDLTNDKANLKDIVFDMFE.K
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.74e-16	0.44	index=10771	3	E.NLDNEKLINDLTNDKANLKDIVFDMFE.K
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.03e-07	1.11	index=10770	4	E.NLDNEKLINDLTNDKANLKDIVFDMFE.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.MQDEEILKSIEESKLEQEEEEKNE.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.

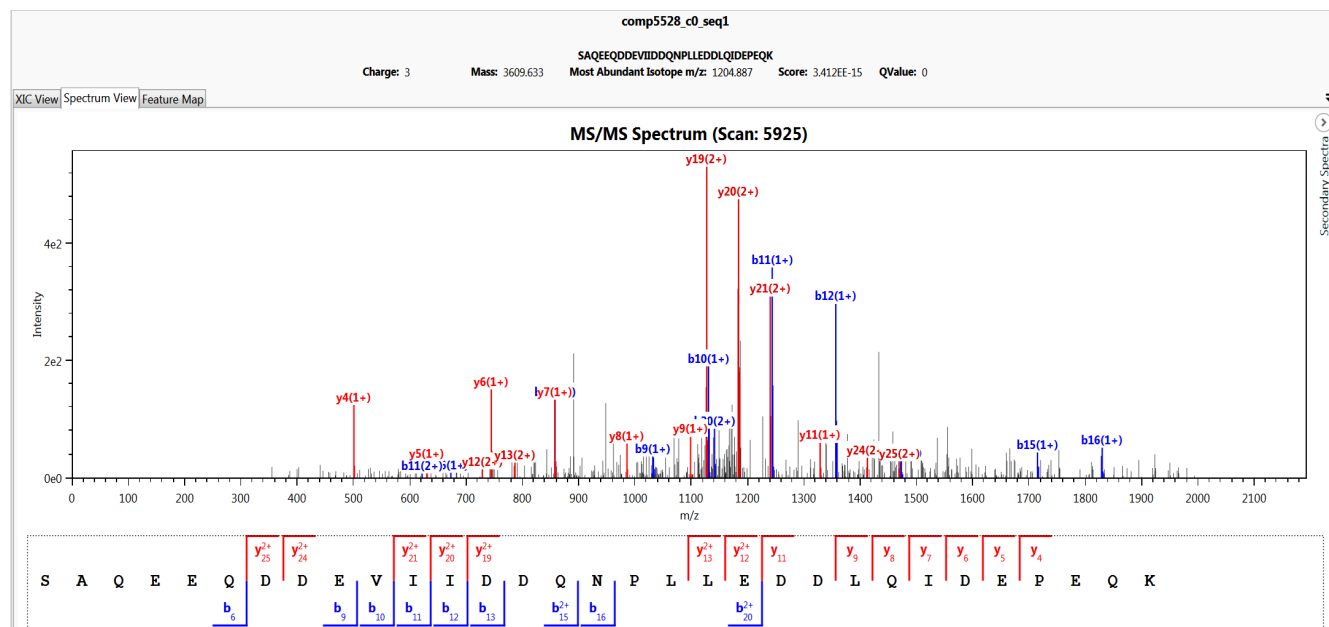
spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide	experiment
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPILLEDLLQIDEPEQK.V	trypsin/SCX
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPILLEDLLQIDEPEQK.V	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.53e-21	0.08	index=8908	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.07e-20	1.10	index=8896	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	7.29e-19	1.10	index=8897	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	2.17e-15	0.94	index=8895	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.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.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_22_13Nov09_Falcon_09-09-15	4.36e-10	3.76	index=9720	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_23_13Nov09_Falcon_09-09-15	2.47e-09	1.64	index=9440	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.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.ELAFKRAQEIGLEPYNEYHGKKK.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.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	2.18e-21	0.78	index=8638	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	3.05e-21	2.11	index=8646	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	4.19e-16	0.82	index=8639	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.19e-21	0.70	index=8806	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.20e-21	1.57	index=8812	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	5.49e-20	1.64	index=8802	2	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	4.33e-15	1.53	index=8810	3	R.WTPIDLPSSEITFVQGIQTVTGAGDPSMK.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.KSSKACEEERRR.K.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.LSQEHLISYISR.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.NFNKITGKEQEEEEE.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.MQDEEILKSIEESKLEQEEEEKKNE.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

###

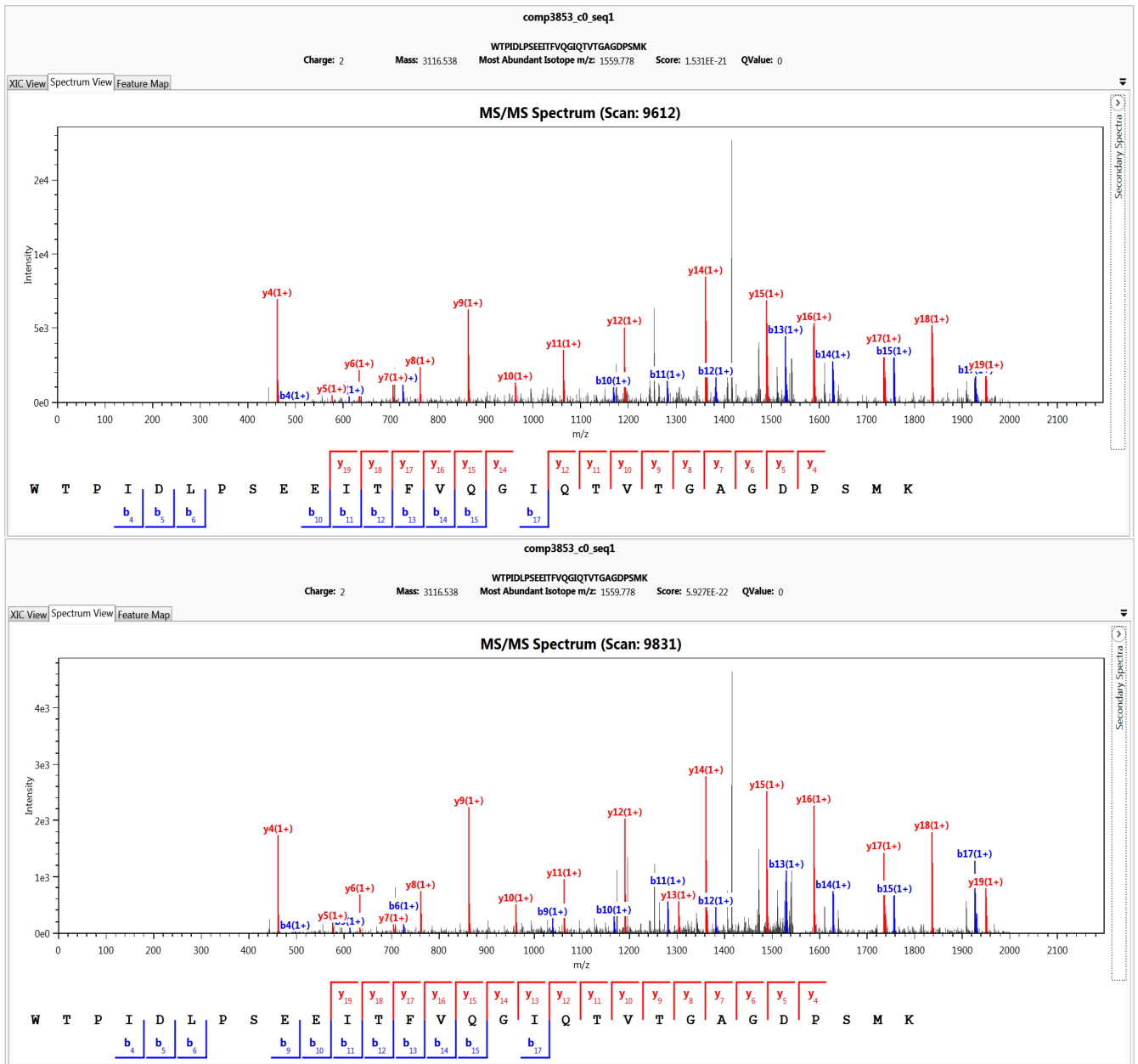
3 Manual Validation

Manual validation was performed by [LCMSSpectator](#). The spectra that have passed the consensus opinion of 5 independent 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

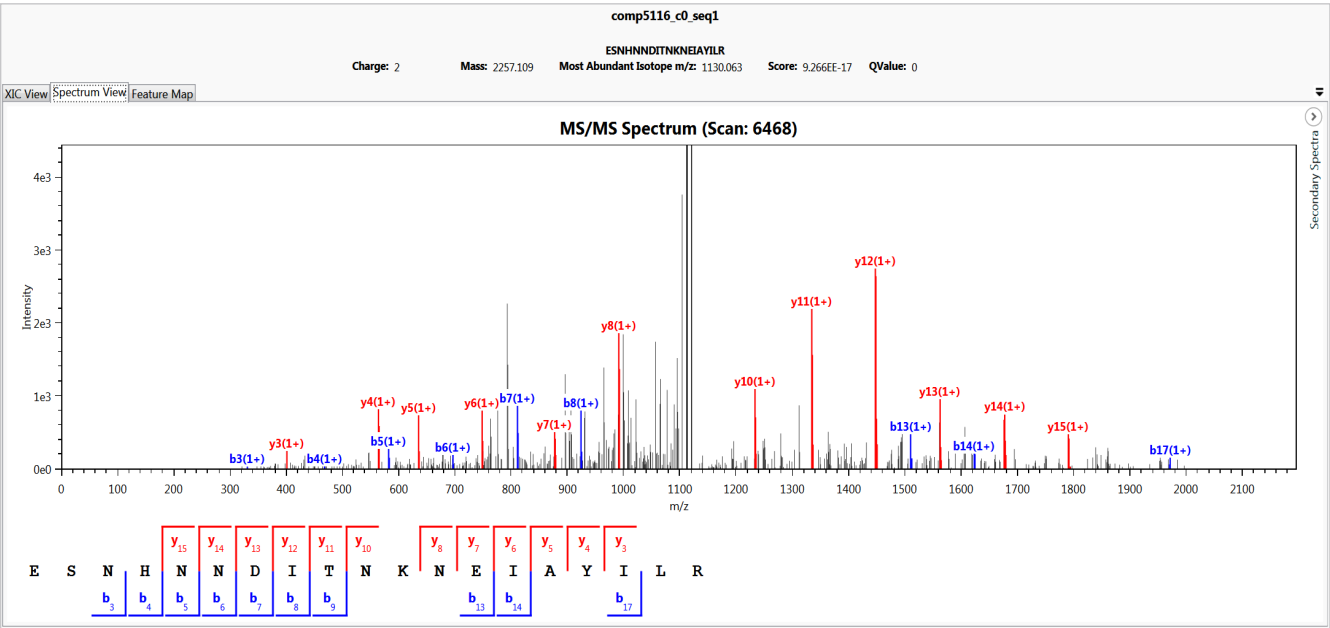


WTPIDLPSSEITFVQGIQTVTGAGDPSMK

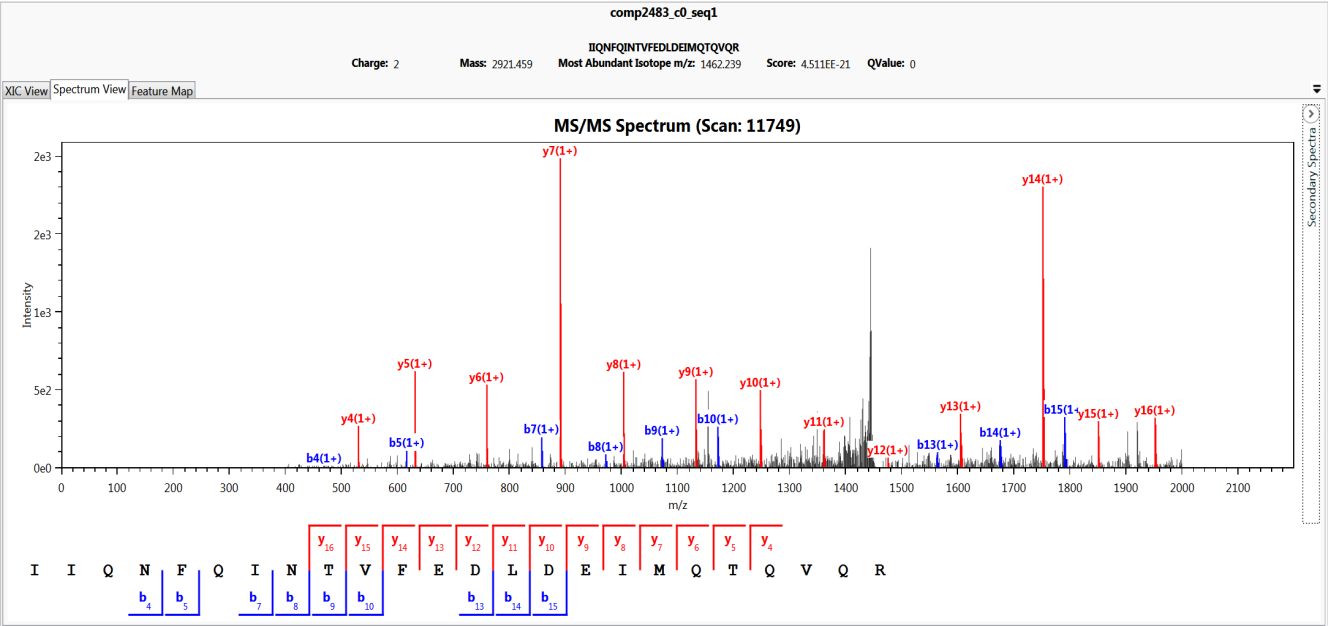


####

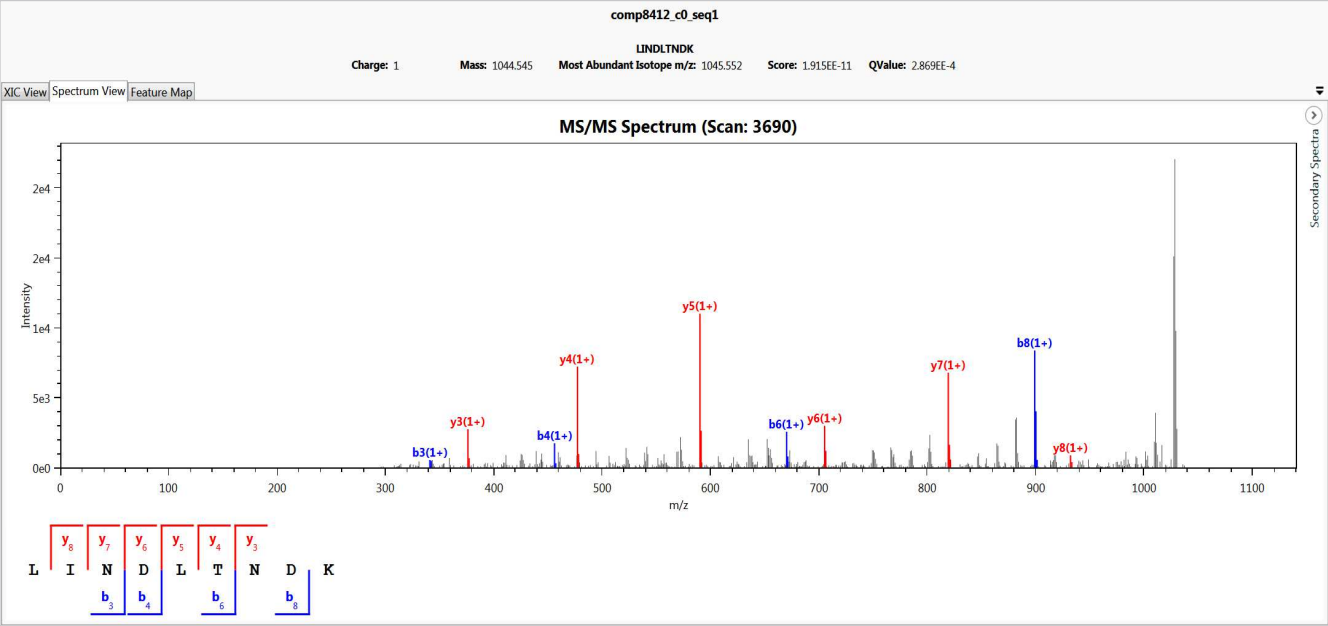
ESNHNNNDITNKNEIAYILR



IIQNFQINTVFEDLDEIMQTQVQR

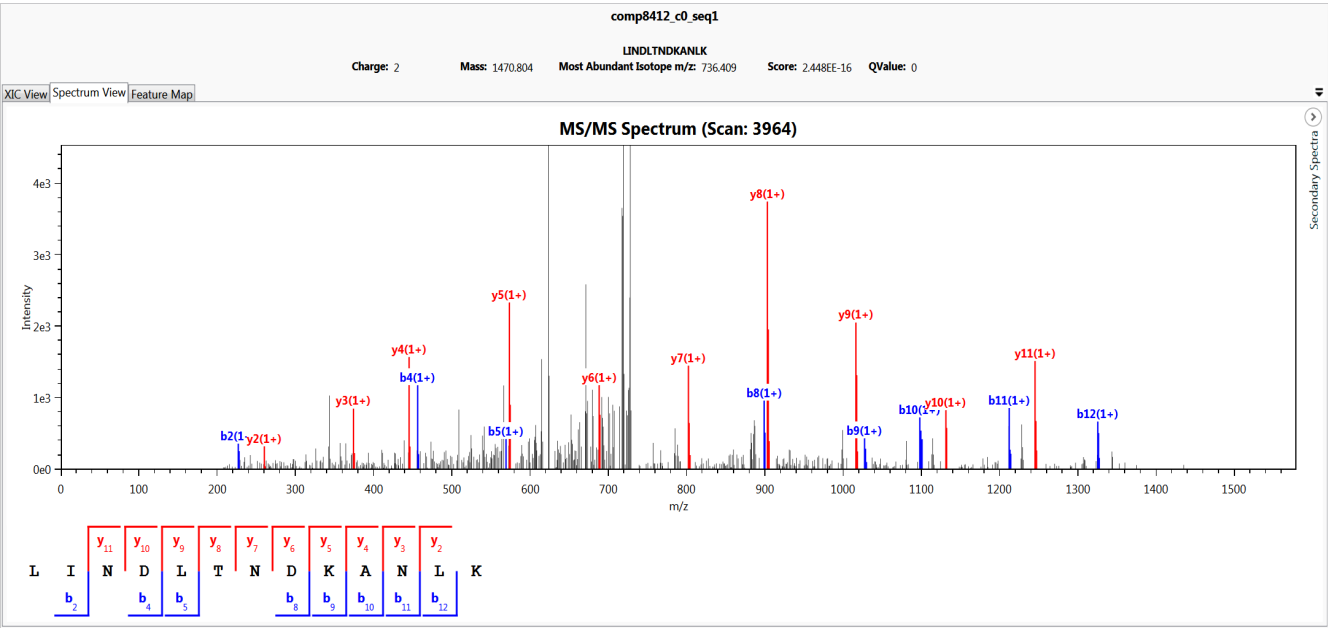


LINDLTNDK

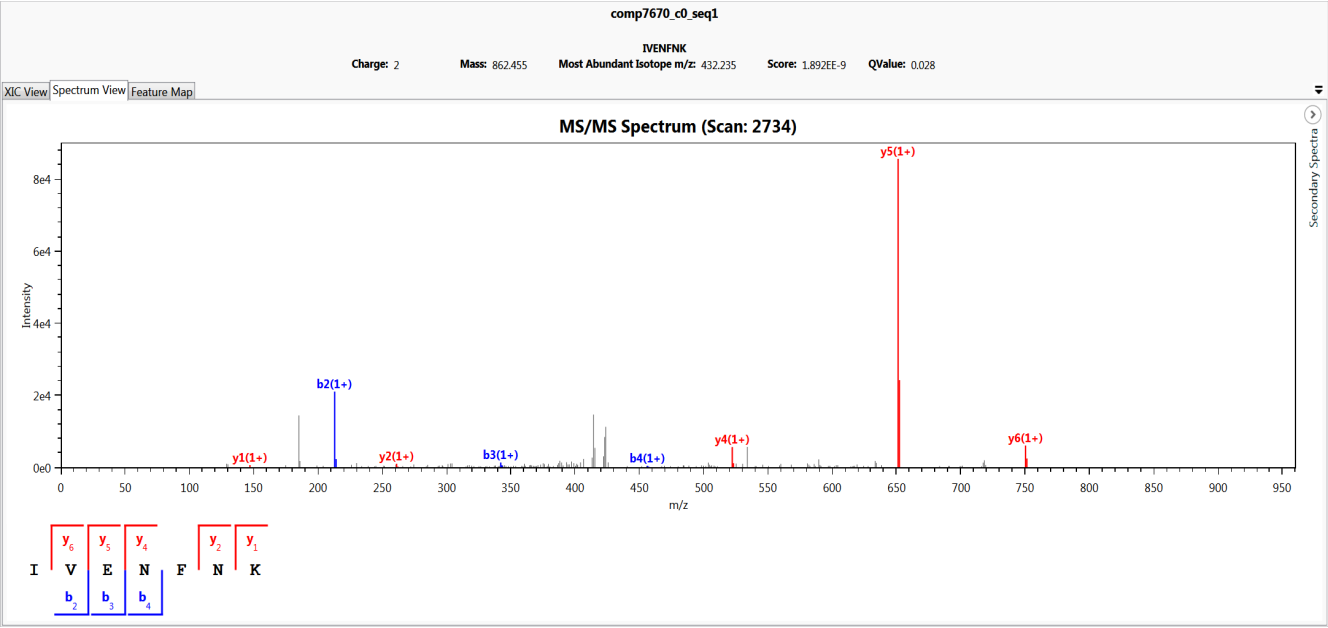


####

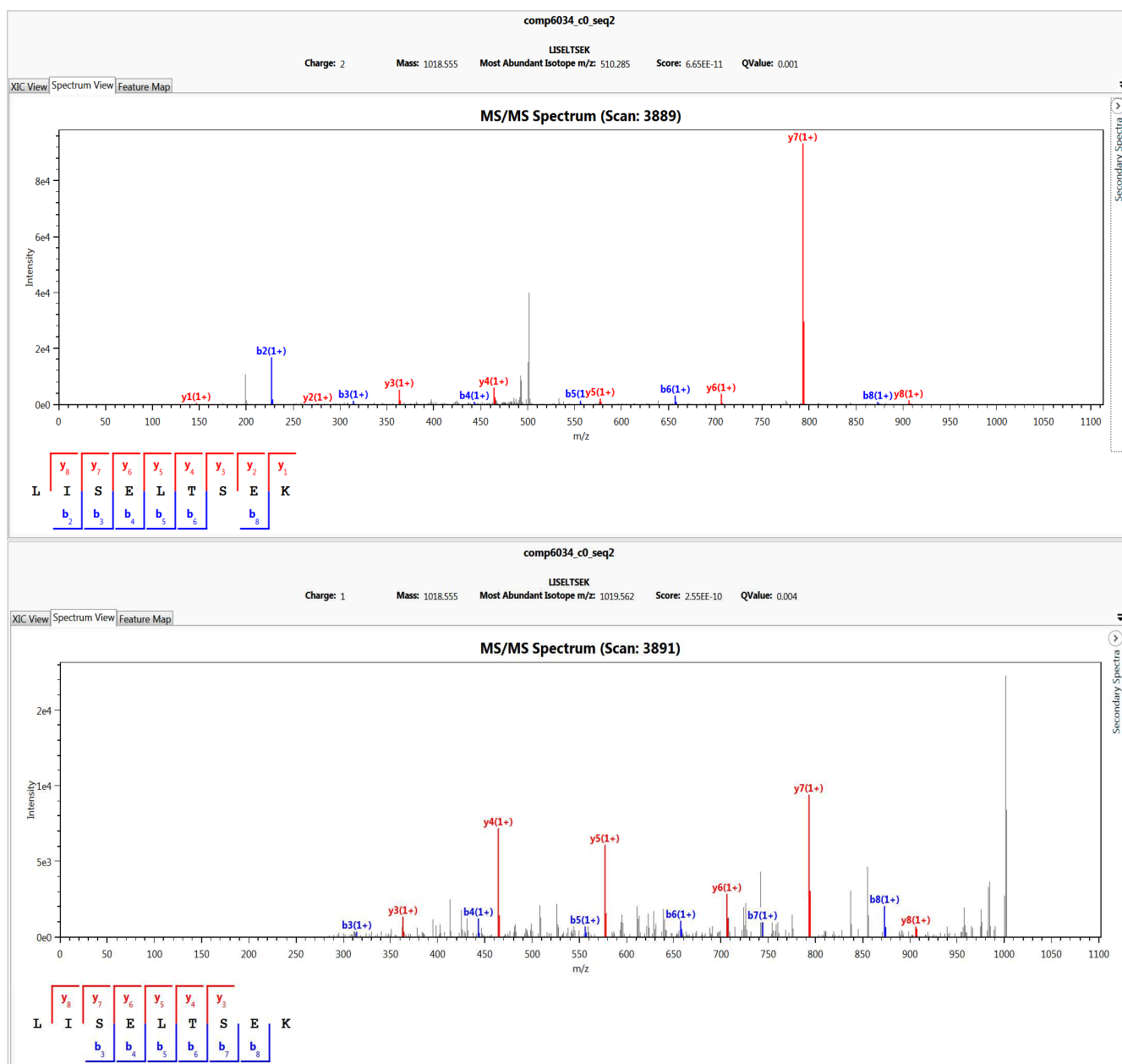
LINDLTNDKANLK



IVENFNK

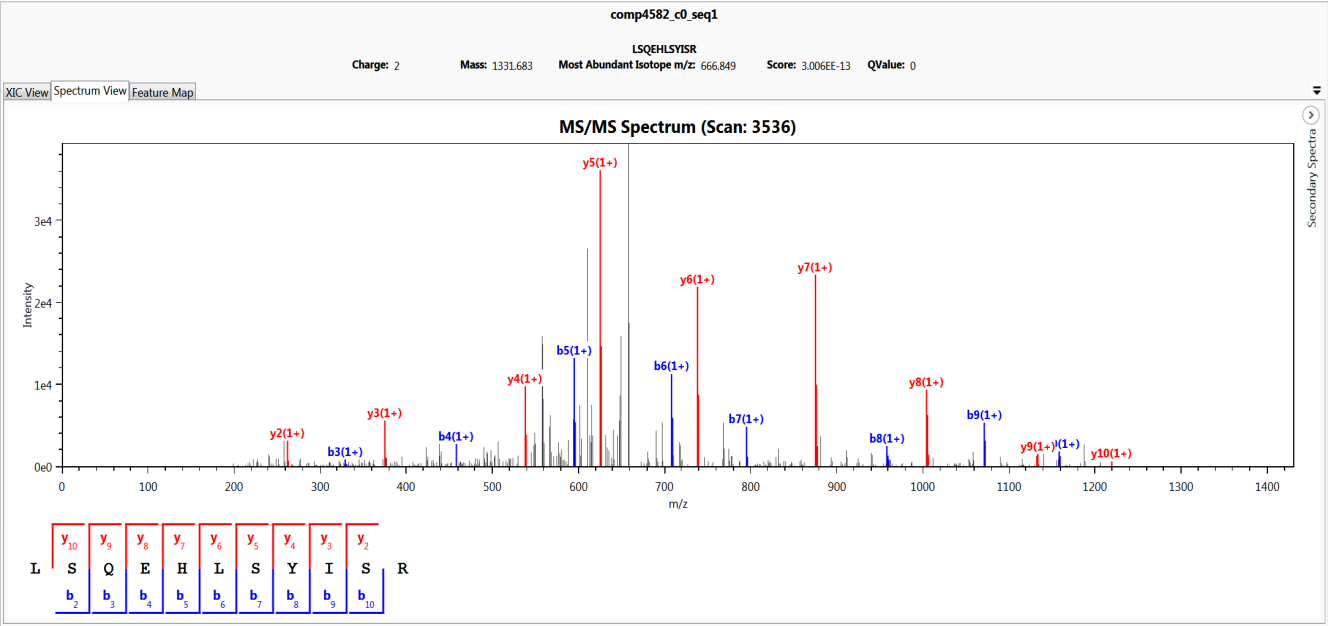


LISELTSEK

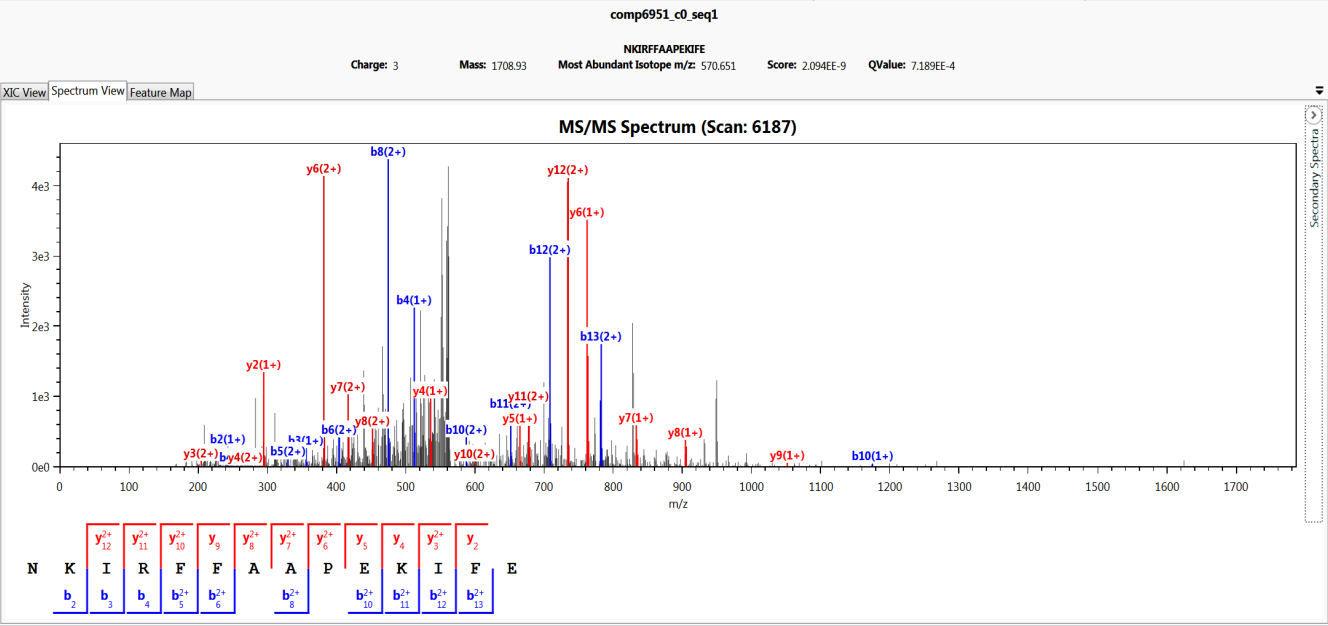


###

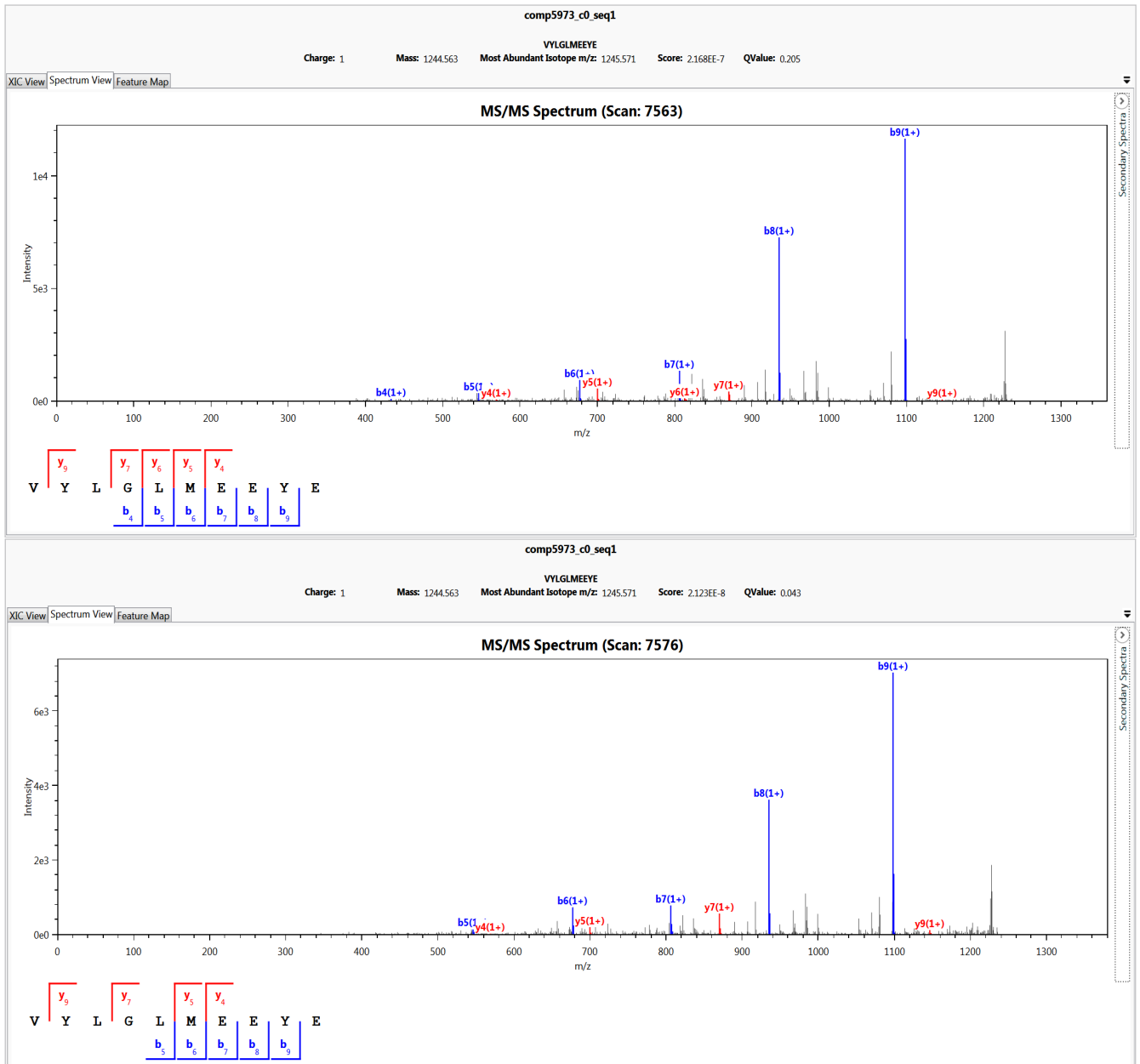
LSQEHLISYISR



NKIRFFAAPEKIFE



VYLGLMEEYE



####

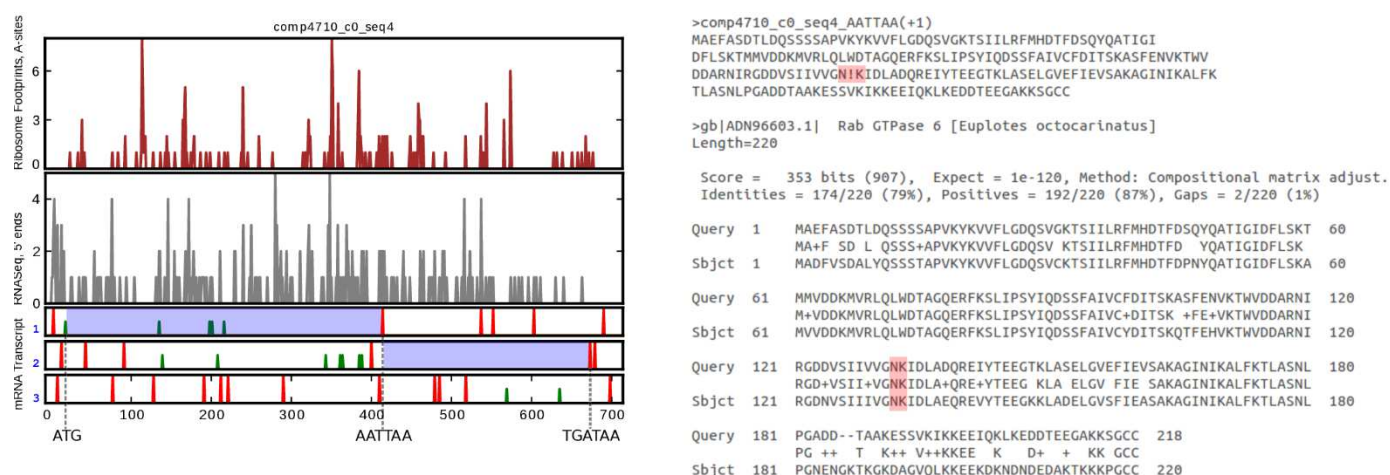
4 Session Information

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

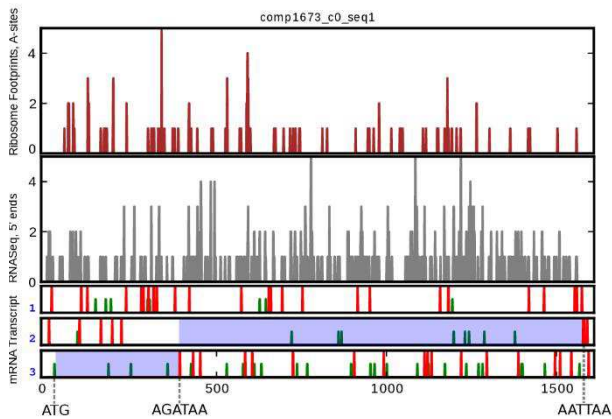
- R version 3.2.4 (2016-03-10), x86_64-apple-darwin13.4.0
- Locale: en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.16.1, BiocStyle 1.8.0, Biostrings 2.38.4, dplyr 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.



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)
MEGNNQGPYNVGEPTKILHISCRKLADLDIITVSDPVCHVVIADSDHPDDW
MLYKGTEQIENNLNPDFVTFYFEMDYFFKIQIKVEFVDVDTLRERIGNFETTLGEIMG
SVNTTLTEGR!ILRTEKVATSNLDIYFSLRINDLVSNKGWFCGSDDPFIFIERARENDQE
EFLRVIQTEPIRNDLNPTWRYLKYEAEICNGDLQCPLKFKVYSWRNSGHHKFFGEFETT
MLIRNGDTQYNLFKDGAAQKQKISCSFEFFIEERASFFDFLHSGWKMMLMVCVDFASNG
EVTVPSSLHYLNPTGEFNDYQNAIRQVGNILELYDYNRQYPCYGGGIPRYSGSNQVSHC
FHLNGLEDPEVDGVNGILESQFSLNCGLYGPTNFGECKRKTVDYIKERMDERMYHILLI
ILTGDGIHDMPTIRDIIVEGSHYPLSIIIGLGESSFDKMIELDGDVVLKNTREATR
DIVQFVKFNDFRHLSKQALAEVLEEVPEQVVSYSQNNIKLDEVN

>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 EKILHISCRKLADLDIITVSDPVCHVVIADSDHPDDWMLYKGTEQIENNLNPDFVTFYFE 75
++++L ISCR L +LD+++ SDP+C VVI D +W L GKTE I NNLNPDF +
Sbjct 19 QRVSLSISCRNLKNDVLSKSDPMCEVVIKDR-KTTNWTLLGKTETINNNLNPDFSSIIY 77

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

Query 122 GRIL-RTEKVATSNLDIYFSLRINDLVSNKGWFCGS-DDPFIERARE-NDQEEFLRVI 178
G+I+ R + V T+ND + LR++ V + CG+ D+P+ I RAR+ N+ ++F+RV
Sbjct 138 GKIIVRLDNTNTNDEV--RLRVSARVQSNAGCCGTQDNPPYIISRARDVNNHDFVRVY 195

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

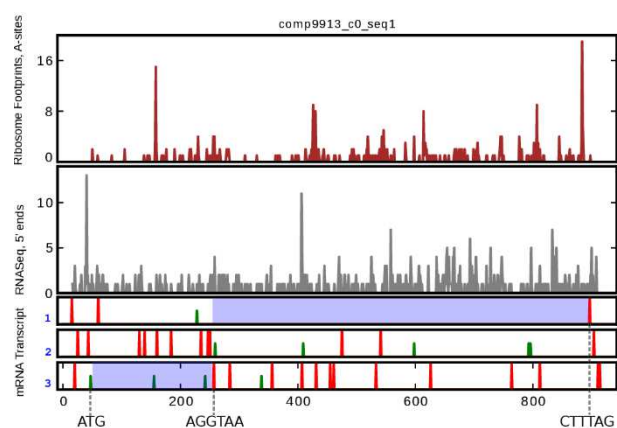
Query 239 GDTQYNLFKDGAAQKQKISCSFEFFIEERASFFDFLHSGWKMMLMVCVDFASNGEVTVP 298
G + + K + + F I E +F ++L SGW +N+ +D+TASNGE T P+
Sbjct 256 GQKSVEITDKKRKIKGSLNIDNFVIREMPNFMEYLRSGWAINMSFAIDYASNGEKTDPN 315

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

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

Query 418 DIHDMPTIRDIIVEGSHYPLSIIIGLGESSFDKMIELDGDVVLKNTREATRDIQV 477
+IHDMP T D+IVE S +P+SIIIG+G F+KM LD D+ L+N++G+ RDIVQF
Sbjct 436 EIHDMPATIDLIVLSRFPVSIIIGVGNEGFEKMNFLDSNQALRNSKGQVAARDIVQF 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)
MSRILERSGVSNNPLNCSTAKQFQSFASKADRFQTMTRGLSSGY
LNLPSSTRSTRSTTFQYGDKIMTNKR!NDSPSPGTYKIPSGFEPGKKGVYSFRCSWKSYS
KVVQKEGFQNTKASDPNVPGPGTYKSVPTFGNKGKFTLKGKLDLPSSVKVPGPAYKD
LTTLPRTGRNFYSKFSVCNGAIGPPTNTRFKDQTQTSAEIPGPQYTPKTLTGTGHYF
LSRFKSSGTTVLGGARRASTHRVSDNGPGGGSYILPSDFGYPTSFHRRRKKSRRAATSQL

>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 MSRILERSGVSNNPLNCSTAKQFQSFASKADRFQTMTRGLSSGYLNLPSSTRSTRSTTFG 60
MS+++ ++ +P+N ST+KQ FSFKA+RFQQ + +LPS +STR+ G
Sbjct 1 MSQVVSTPQQINQHPVNNSTSKQLFSFKAERFGQPKSMNNHRIAYDLPSMKSTRAAGLG 60

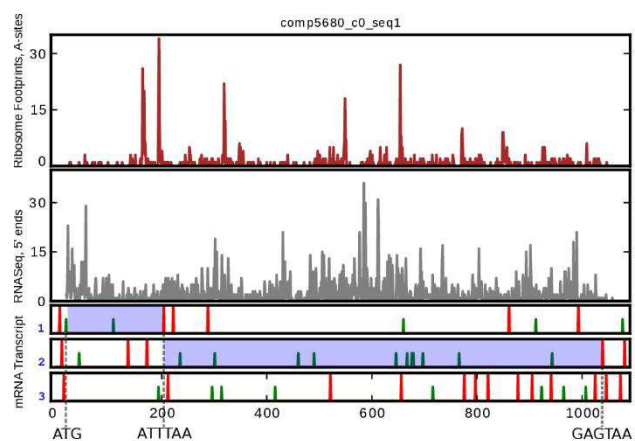
Query 61 YGDKIMTNKRNDSPSPGTYKIPSGFEPGKKGVYSFRCSWKSYSKVVQKEGFQNTKASDP 120
YG + N RNDSPSP Y + S F K +SF S ++YSKVY KE +D
Sbjct 61 YGSRNAFNIRNDSPSPNTYNLKSEFTKSPSNKAFSFGISREAYSKVYIKEN----PLADA 116

Query 121 NVPGPGTYKSVPTFGNKGKFTLKGKLDLPSSVK--VPGPAYKDLTTLPRTRGRNFYSK 178
+VPGPG Y+ P G + K+TL+ K ++ + PGP Y L G F SK
Sbjct 117 SVPGPGQYQIPPIVGKEALKYTLRPKTQNPYTQYKQPGPGQYDTKPALNDKGLYFNSK 176

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

Query 227 RFKSS-----GTTVLGGARRASTHRVSDNGPGGGSYILPSDFGY 265
F+S5 T +LG + + PGP+Y LPDFGY
Sbjct 237 NFQSSMCRSHYHFDRTNILGSTMKG-----PGPNYRLPSDFGY 277
```

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



```
>comp5680_c0_seq1_ATT TAA(+1)
MSLPYYANACVDKPESEYDYTNFELSWGPMDDYEVVQKIGRGKYSEVFDGNT
LNNQKCVI[K]ILKPIKLEKMQREIKILQTLYGKNIKLYDMAQDDVSEVTALVFERVNH
DHKVLYKQFKDFDIRYYIEVLLGLDYCHSLGIMHRDIKPHNIMIDHEQRLRIDWGLA
EYYIPGKEYNVRVASRYKGPPELLVDDRLYNYSLDMWSLGCTMANMMFQRDPHFKGVDND
DQLIKIVEICGIDGLMDYLKTYKLEINRYHQKHLKNWEKVSWEFEITKKNHLVTKEALD
FLEKCLQYDREKRIMPQEAIEHEYFAPVIEYKKKIHGGGEVKEE

>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 MSLPYYANACVDKPESEYDYTNFELSWGPMDDYEVVQKIGRGKYSEVFDGNTLNNQKC 60
      M+LP+YYAN C + P Y DY N+E+ +G + YE+++KIGRGKYSEV++G+NTLNN++
Sbjct 1 MNLPKYYANVCEEMPPEYSDYENYEVKFGSQENYEI[K]KIGRGKYSEVYEGINTLNNERI 60

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

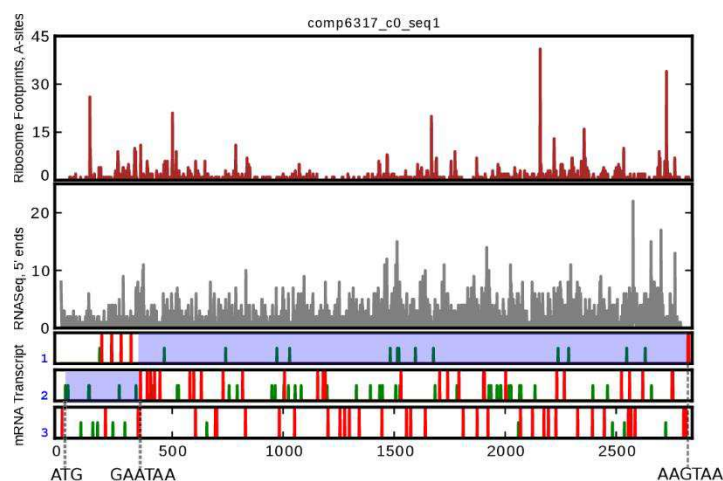
Query 119 LYKQFKDFDIRYYIEVLLGLDYCHSLGIMHRDIKPHNIMIDHEQRLRIDWGLAEYYI 178
      LYK F DFDIRYY++E+L LD+CHS GI HRD+KPHNIMIDH R+LR+IDWGLAE+Y
Sbjct 121 LYKSFTDFDIRYYMFEILKALDFCHSGITHRDVQKPHNIMIDHASRKLRLIDWGLAEFYH 180

Query 179 PGKEYNVRVASRYKGPPELLVDDRLYNYSLDMWSLGCTMANMMFQRDPHFKGVDNDQLI 238
      PG+EYNVRVASRY+KGPPELL+D + Y+YSLD+WSLGC + M+FQ++P F+G DN DQL+
Sbjct 181 PGQEQYNVRVASRYFKGPPELLIDLQTYDYSLDIWSLGCMSGMIFQKEPFFQGDNDYQLV 240

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

Query 299 CLQYDREKRIMPQEAIEHEYFAPVIEYKKK 328
      L+YD +RI+P+++A++H YF PV E+ K
Sbjct 301 MLKYDHAERIVPKDAMHPYFKPVKEFHAK 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)
MSSYMKSSSELEELKIELNSLKVDEQKEAVKQVIAMMTIGKDVSLFPHVTKCI
LSPSIELKKLVLYIINYAKSKPDLTLMVSAFTKDAHEKSNPLRALAVRTMGCIRIEI
KIATYLCESLKDCLVDDDPYVKKTAISVAKIYHTMPEHTKELGFIKLQGLLQDGNVAV
VANAVAALFEISRVAGKNYLKANKETIGKLLNALNETNEWGQIYILESIINYKPKQKEA
EEIIRIMPRLQHANPAVVLGATKNVHFLKFNLSKNKTTILKKLSAPLITLLSSEPEI
QYIALCNILLIQIIPNVFEKNVMMFFCRFSDPIYVKLAKLDVMVGADNTNVDIIITEL
HEYCNIDQDFVRRSVKAIQGVVVKVDRVAKKGVEALREHVNQEQGSDSALQEAVIVASK
ILRKYPPKFEGLVKDQVQDRIDEPEKSAFIWILGEYSKKIEDAGEKLQVYIDSTDE
NINVLQILTSVAVKMFIDSDNYEDMVNVLKLAESSANPDLRDRGYIYRMLSTDPQ
TKDVLAKRPEVEEDTLKMDDETDFIDIFISDTKHSALRPEKTASAPEDSDEVEE
EEKPKSKKKDKKKQKKVKEETKEEIELEPEEVTEDEKPKDLDLDFGLIGDDPVSN
DEPAVDPLAGLIDEGNGGSGTQAASPHDDNGLFGGFGASGSEASLFIKSEHAELVSSST
PGSQNKAAAGLQIKARFYREGTSIKLDMIFYNSTAGIISDFDIMINKNPFGLKPGPISVIP
ISAGQTFITTTVECSIDQSNADLKNPPQCPYVQTAIKNSLDVYVQVPCLLHLTQGTVP
AVTQTCCQMANSIANKHSFTVSSARFAGSASDLKTRMQNSFPYPIYDELNSQIFATSTV
NNIPILLRCTPEGSDIITACTPVAPLYQLEIEAIEKVISK

>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--KSSSELEELKIELNSLKVDEQKEAVKQVIAMMTIGKDVSLFPHVTKCILSPS 58
M++Y + KSSSEL EL+ ELNSLK +E++EA KQVIAMMTIGKDVSL FPH+ KC+ +
Sbjct 1 MTNYFQNMKSSSELAELQHELNSLKPEEKREAAKQVIAMMTIGKDVSLFPHVKKMETTQ 60

Query 59 IELKKLVLYIINYAKSKPDLTLMVSAFTKDAHEKSNPLRALAVRTMGCIRIEKIATY 118
+ELKKLVLYIINYAK KPDLT+MAV++F KD+ + +P++RALAVRTMGCIR+E+I Y
Sbjct 61 MELKKLVLYIINYAKVKPDLTMAVNSFQKDSRDIQSPMMRALAVRTMGCIRVERITEY 120

Query 119 LCESLKDCLVDDDPYVKKTAISVAKIYHTMPEHTKELGFIKLQGLLQDGNVAVVANAV 178
+CESLK+ L D DPYVKKTA+ VAK++ T P K+ IK+LQG+L DGNA+VVANA
Sbjct 121 MCESLKERLNDQDPYVKKTAALGVAKLFQTSPLVKDHSLLIKLQGMLYDGNVAVVANAA 180

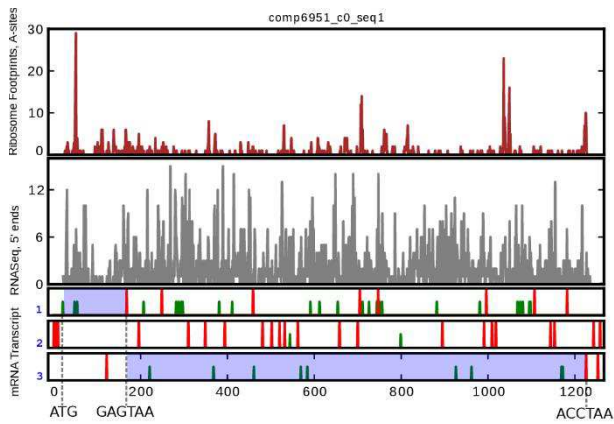
Query 179 AALFEISRVAGKNYLK-ANKETIGKLLNALNETNEWGQIYILESIINYKPKQKEAEIEI 237
A+L EISR +GKNYL+ N + + KLL ALN+ NEWG+IYILE I +Y + KE+E I+
Sbjct 181 ASLLEISRASGKNYLRKNDQGLNKLIALNDANEWGKIYILEGSSYDTSDSKESINIV 240

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

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

Query 358 NNIDQDFVRRSVKAIQGVVVKVDRVAKKGVEALREHVNQEQGSDSALQEAVIVASKILRK 417
+ID + V++SV+AIQ++KVD+ A K VE + E V QG + +QEAVIVA I RK
Sbjct 361 GDIDPELVKKSRAIQIILKVDKAASKAVEIIEIVT--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
MYSTKFRRVMTMAPLLANPALALCEEPSTADIRGNENKIRFFAAPELKIFE
TFSNIREEDGQVYMSYQDFHSLTPYNFVASKDDDDDDDEENKDKKEKEPGYFDKFTP
EIMTIVDANQDKKIDFNEYIFFITLLQLPEGEVMRIIEKVNPEERKINKAQFAKYLTKLR
KCTALGLKQMSKSFMPDGRKISTDEDHISKTILLHLFNDKEYITIEDFCELKSKLKHALL
HYEFYQFDVDEDETISAESFAKSLLSCLNYTQASKYSRRIHSLKLEGRVSFKEYVAFHNL
IEKADIIMKISTYRFLSLGMFRDLCDDFAKLDPYCNQNKVISDTQIATFFKVLDEDEDEN
GALEYDEVVDILEGKNIGLKEDKFKREMMEKIDRYIKKFQKYVGWT

>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--LALCEEPSTADIRGNENKIRFFAAPEKIFETFSNIREED 62
M + PL L+ N L+ CEE DRIRGNENKIRFF+ EKIFETF++ + E
Sbjct 42 AMIITPLAFQMLILNNQHNLSCCEAPRQDRIRGNENKIRFFSPEKIFETFASSKNEK 101

Query 63 GQVYMSYQDFHSLTPYNFVASKDDDDDDDEENKDKKEKEPGYFDKFTPEIMTIVDAN 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 GDGVISFPEFFFFITILQMPLGLIHKEFTKHVKEGGKMNQDEFSKTLTLRKKTLLGKTKQ 204

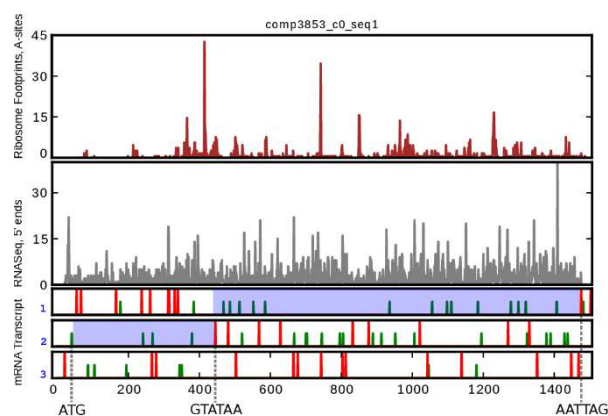
Query 183 MSKSFMPDGRKISTDEDHISKT---ILLHLFNDKEYITIEDFCELKSKLKHALLHYEFYQ 239
++K +PD R IS ED S+T I LF +K + EDF + +LK AL HYEF+Q
Sbjct 205 INKGMVPDARLISATEDDFSQTNNEICNQLFKNKTLSYEDFINFRDELKIALRHYEFHQ 264

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

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

Query 359 DEVVDILEGKNIGLKEDKFKREMMEKIDRYIKKFQKYVG 399
DEV+ +L+ ++ +G GKE++ K + ++ F++ +G
Sbjct 385 DEVIGVLDQRQLLGQGENELKEAIESSFKVKVQWFRETIG 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)
MSEENKEEVKGTHTDQYHGHGSEAEIAGALPKHRNNPQQC
KFLGYAEQISGTPFTYPRAKMQRSLYRIMPTVAHPYKALKDYNNLWIANFARDDDEEV
FTTPQMRWTPIDLPSEETFTVQGIQTVITGAGDPSMKAGINMGVYTCNTSMKNEAFFSS
DGDIMIVPQLGKLSIMTEFGHIEAESWEVVIPRGIKFAVEVNEDCRGYYCELYDGLHI
PDLGPIGTNGSANPRDFAIPKAKYFDETNEFRVIQKYLKFFETIIPHNIFDIVAHGNY
VPYKYDCHHNTGSIYDHPDPSVFTVLTCTPDHQAALDFAIFPPRMLSMEDTFRPP
YFHRNTMNEFMGNVAGQYDAKEEGFSPGAVSLHSCSAHGPEAEVVEKASTCELPQKVG
EGCLAFMFETCYTMKVTKSFMDLEGATDSYSVNSKAVDESVDCKWGMKRLFDPNPD
AGYKKYLSEHK

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 GFGNHFESEAEIAGALPKHRNNPQQCKFLGYAEQISGTPFTYPRAKMQRSLYRIMPTVAH 82
GFGNHFE+E++EGALPK +N+PQ+ GLYAEQ+SGTPFTY R + QRSW YRI PIV H
Sbjct 9 GFGNHFESEAEIAGALPKDQNSPQKVPHGLYAEQLSGTPFTYARHRNQRSWYRIQPTVLH 68

Query 83 PPYKALKDYNNL--WIANFARDDDEEVFTTPQMRWTPIDLPSE-EITFVQGIQTVITGAG 139
P+K + D W+ F +DD + +P Q RW P +P++ ++TF++GIQT+ GAG
Sbjct 69 RPFKPVADQERFKNLITFDKDDSR-LHISPAQYRWRPQTIPTDKQVTFIEGIQTMGGAG 127

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

Query 200 RGIKFAVEVNE-DCRGYYCELYDGLHIQIPDLGPIGTNGSANPRDFAIPKAKYFDETNEFR 258
RGIKF VE+ E RG+ CE+Y H IPDLGPI+NG ANPRDF P A Y D+ E+
Sbjct 188 RGIKFQVELLEGQARGFICEIYKSHFVIPDLGPIGNGMANPRDFQTPIAWYEDKHEEW 247

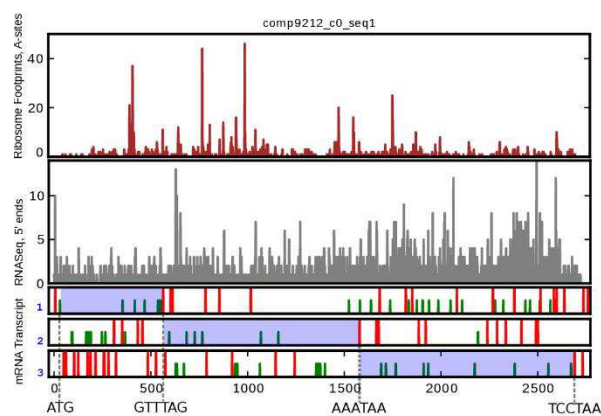
Query 259 VIQKYLKFFETIIPHNIFDIVAHGNYVPYKYDCHHNTGSIYDHPDPSVFTVLTCTQ 318
VI K+ G FF+Y + H+ FDIVAHGNY PYKYD +NT+GSIS+DHPDPS+FTVLT Q
Sbjct 248 VINKFQGSFFQYQVTHSPFDIVAHGNYAPYKYDLRKYNITIGSISFDHPDPSIFTVLTQ 307

Query 319 TPDHQAALDFAIFPPRMLSMEDTFRPPYFHRNTMNEFMGNVAGQYDAKEEGFSPGAVSL 378
T D GQA DF IFPPRML E TFRPPY+HRNTM+EFMGN+ G YDAK E F GA SL
Sbjct 308 TDDPQQAQVDFVIFPPRMLVQEHTFRPPYHHRNTMSEFMGNIQGTYDAK-ESFCAGASSL 366

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

Query 439 VNSKAVDESVDCKWGMKRLF 460
VN V YH+CW+ + R F
Sbjct 415 VNEEN-VYTDYHECWQLPREF 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)
 MSGKEEAKSGIKRRNTGKGHDDKHSAPSGQGNICVVCVRFRPLNQNELNH
 GGSNVCA DFHPNKKSVTIVTEGDVTNKNNEFTFDRVFDINSSQIEVYNQAAPKPIESVME
 GFNGTVFAYGQTGSGKFTTMOGPDIEDIESQGI VPRMVRTVFNRIENSSNIEFTVKVSM
 MEIYMEKVRDLDPTKANMKIKVDKHSAYVHDLTERYIGSOLDVYDIMRIGNNNRKVAS
 TSMNDQSSSRSHSIFVMTVHQNNLDQDTSKTGILYLDLAGSEKVGKTGASGHTLDEAKGI
 NKSLSTLGKVINALTGDKSKHIPYRESKLTILSESLGGNARTALITCSPSVYNDMETI
 STLRFGTTARNIKNPKVNKELTVAEMKLLSKSEKIEVRNRYRVKVLGGCITELGGEVP
 EDEYKDLQLSSKPAPAKEVEKEPAPSKEPEQDDDDDEQDEEKKAIIDELQAKLRQLQDK
 LNIETDKYCNLT EEFVSGNKLAAEQEKCDKLQVQKLEEFTEKINDLVQTKQKLLLL
 EEELRNVSNEKLSKKLDRIEENGCIYLPEDGDSAHDKKRTAQEAQKELIRVSQKL
 KRISLGGSLDEVKQILDIEVRHTDEESKEDPDDEMFKDSSLDMHADLSKFKDKSLSTIF
 LDEEDTNKLIKSIQIELREQKLDERQKHEFLAQIDSLSELENAAKESNPNQAITKKLA
 DEQIMKAENRWAEKVQLRDLERNVSKVQLIEISLDEANEKYRRENTINQCDVPLRKK
 INKLENQAEQITIMYHVQVSEKSLKVDLQVYEKKLRKDEKIASLEKALNKLKEGNVAL
 KKLRLGLKSLMKANDENRILEGNNTGIPSSGRVVKPLRGGRDKPKPARINSSIQS

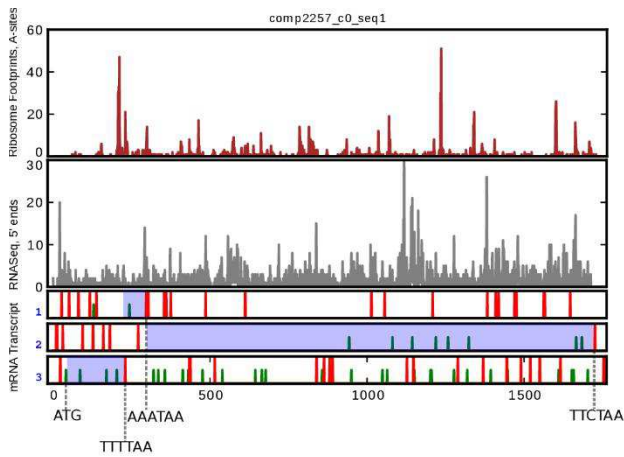
kinesin motor catalytic domain protein [Tetrahymena thermophila 58210]
 Sequence ID: gb|EAR95613.3|Length: 916Number of Matches: 1
 Related Information
 Range 1: 8 to 898GenPeptGraphics Next Match Previous Match
 Alignment statistics for match #1 Score Expect Method Identities Positives Gaps
 541 bits(1393) 3e-174 Compositional matrix adjust. 358/939(38%) 538/939(57%) 144/939(15%)

Query	30	GQGNICVVCVRFRPLNQNELNHGGSNVCA DFHPNKKSVTIVTEGDG---VTNKNNEFTFDRV	86
Sbjct	8	G GNI VVCR RP N++EL GS C +F + +T DG NK F FDRV	66
Query	87	FDINSSQIEVYNQAAPKPIESVMEGFNGTVFAYGQTGSGKFTTMOGPDIEDIESQGI VPR	146
Sbjct	67	FNMTTQEIQIEVAAPKPVVQSVLEGFNGTVFAYGQTSSGKFTTMOGASIDDEKLKGVIPR	126
Query	147	MVRTVFNRIENSSNIEFTVKVSMMEIYMEKVRDLDPTKANMKIKVDKHSAYVHDLTE	206
Sbjct	127	MVKTVFQHISDAPDHIEFRKISIVEIYMEKIRDLDNTQNLVREDKQRGIVIQVTE	186
Query	207	RYIGSOLDVYDIMRIGNNNRKVASTSMNDQSSSRSHSIFVMTVHQNNLDQDTSKTGILYLV	266
Sbjct	187	QVVSNEQDVFLLRIGNQNRAVTATNMNEGSSRSHMLFMMSVSQNNLNDLSAKTGKLLV	246
Query	267	DLAGSEKVGKTGASGHTLDEAKGINKSLSTLGKVINALTGDKSKHIPYRESKLTILSESL	326
Sbjct	247	DLAGSEKVTGA G LDEAK IN+SLS+LG VINALTGDKS HIPYR SKLTR+L ES	306
Query	327	LGGNARTALITCSPSVYNDMETISTLRFGTTARNIKNPKVNKELTVAEMKLLSKSEK	386
Sbjct	307	IGGNSKTLITVTCSPSPFNDELTLSTLRFGIRAKAINKAKVNREVTVAELQVLLAKIEK	366
Query	387	IIIVRNRYRVKVLGGCITELGGEVP-EDEYKDLQLSSKPAPAKEEVKE-PAPSKEPEQDD	444
Sbjct	367	QLEEKTRRVAQLEDYIQQLGSQLPSETNLNQEDQTQIIDSQEEIPQNPQISNIEEIVR	426

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



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)
 MNLEDDTYTFDQLGQGNKAYEKIFDFANEPVVFSDTAIKINMWGI
 QERILLMTTKNIFNFKRKSMMRKIAIEKIGAIISTKINSKELVLHVPSEYDVRQIND
 REVFIKLLKREFVKKRPNHGLRYYVVGSLKEVTTTEKDAKYKISRLPSELRLDQEVY
 GPEYQEESEKTPTESTVAATSFDTKLKAVERSKQDASLEDDEEPDADFDRNSVCVFSH
 KEEEEKVTLSDFEVKGIVGRGTGKVLVRLKTTDTLYAIKSLRKDVLVEAQIENVKLE
 KEILLACNHPFLAGMEYVVFQNDTRLVYFIEFLKGGELYKHLKRRFEDEAKFYATQIA
 MAIGHLHKQNLHRLKLENIMINGOGYIKVIDFLAKIIDDNLAMSFSGTPEYLAPEM
 VKQEGHGRGVDWMSLGLIYEMTIGVTPFFSKSRLTLINNKKEDVIFPGKYKIEYSD
 FVDVVLKLLNKDSKRLGSDGIDEVLEHPYFQSLDLALVNKEIKPPYIPEFTGKDDLG
 SYIKLKSQKDVNDTIPSSKMRKIKKHETDFKNF

Protein kinase domain containing protein [Oxytricha trifallax]
 Sequence ID: gb|EJ372937.1|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 DFANEPVVFSDTAIKINMWGIQERILLMTTKNIFNFKRKSMMRKIAIEKIGAIIS-TK 85
 + ++E +++S+ K N + QERI+++T + IFN K++ ++R IAIEK+ + S ++
 Sbjct 23 EISDETLIYSNRVKKFNRFDWQERIMVITDRIRFVNKKQIKRAIAIEKLLGVTKSLSE 82

Query 86 NSKELVLHVPSEYDVRQINDREVFIKLLKREF--VKRPNHGLRYYVVGK-SLKEVTTT 142
 ++ E VLVH EYDVR + + R+ I++++ F V +P L +Y VK +L E TT+
 Sbjct 83 SNGEFVLHVKEYDVRMRSDTRDQIIEIRKLFISVCGKP---LPFYGVKPKALSEFTTS 139

Query 143 EKDAKYKISRLPSELRLDQEVYGPYQEESEKTPTESTVA--ATSFDTKLKAVERSKQ 200
 +KD K +SR+P R+ D+ ++ T + + F +++ S
 Sbjct 140 KKDLKGLSRIPLELARIKDEG-----DNSTDMQFNLGDDETFSDNQMEERTSSYH 190

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

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

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

Query 379 DFGLAKIIDDNLAMSFSGTPEYLAPEMVKQEGHGRGVDWMSLGLIYEMTIGVTPFFSK 438
 DFGLAK++ +A SFCGTPEYL+PEM+ GHD +DWM+LGIL+YEM IG+ PF+++
 Sbjct 371 DFGLAKML-GEQVANSFCGTPEYLSPEMITGTGHDHTIDWMTLGLIYVEMIIGIPPFYNQ 429

Query 439 SRLTLINNKKEDVIFPG-KKYKIEYSDDFVDVVLKLLNKDSKRLGSDGIDEVLEHPY 497
 ++ + + I+ + +P +K+ E S++ D++ KLL KD KRLG ++E++ HP+
 Sbjct 430 NKHQMYHLIENGPIRNPTMEKHGFEISEESKDLISKLEKDKNKLGRVNGVEEIIISHPH 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

