

Supplementary data for:

**Two single-subunit oligosaccharyltransferases of *Trypanosoma brucei* associate with each other and display different and predictable peptide acceptor specificities.**

Anders Jinnelov, Liaqat Ali, Michele Tinti and Michael A.J. Ferguson\*  
Wellcome Centre for Anti-Infectives Research, School of Life Sciences, University of  
Dundee, Dundee DD1 5EH, U.K.

**Figure legends:**

**Fig. S1. *In situ* tagging of *TbSTT3A*.** Panel A: Predicted digestion pattern for Southern blot analysis of *in situ* tagged *TbSTT3A*. The HA<sub>3</sub> tag is highlighted in green, an intergenic region in purple and the *HYG* resistance gene in grey. When probing against *HYG* following PstI digestion, two bands should be observable at 4.6 and >10 kb and these fragments were visible following Southern blotting. Panel B: VSG221 visualised following PNGaseF or EndoH treatment and Coomassie blue staining of the SDS-PAGE gel. Untreated (lane 1), EndoH (lane 2) and PNGaseF (lane 3) treated samples from a control wild-type cell line showing the typical digestion pattern for native VSG221. Lanes 4-6 show untreated (lane 4), EndoH (lane 5) and PNGaseF (lane 6) treated samples from the *in situ* tagged cell line. Since the digestion pattern resembles that from wild-type cells, we conclude that the product of the single C-terminal HA<sub>3</sub> tagged *TbSTT3A* allele was able to transfer EndoH resistant Man<sub>5</sub>GlcNAc<sub>2</sub> to N263.

**Fig. S2. Glycosylation reporter construct and system validation.** The pLEW82 plasmid (Wirtz et al., 1999) drives expression of both the BipN fusion protein and the phleomycin resistance gene (*BLE*) via a T7 promoter. The plasmid is tetracycline-inducible and has one tetracycline operator and two T7 terminators. *TbBipN* (blue) is fused to an NAT *N*-glycosylation sequon flanked by five variable residues (purple) preceded and followed by Ala<sub>3</sub> linker regions (green), AvrII and MfeI restriction sites (pink) and followed by an HA<sub>3</sub> epitope tag (red). When *TbBipN* is expressed, the protein can carry a biantennary paucimannose or complex *N*-glycans originating from Man<sub>5</sub>GlcNAc<sub>2</sub> transferred by *TbSTT3A* or a triantennary oligomannose *N*-glycans originating from Man<sub>9</sub>GlcNAc<sub>2</sub> transferred by *TbSTT3B* or a mixture of both. Experimental examples showing the EndoH- and PNGaseF-sensitivity of the glycosylated TILKSNYTAEPVR site and the EndoH-

resistance and PNGaseF-sensitivity of the glycosylated TEGLLNATDEIAL site are shown at the bottom. These agree with literature descriptions of these sites (Mehlert et al., 2010, 2012).

**Fig. S3.** Optimisation of peptide length for classifying *TbSTT3A* and *TbSTT3B* catalysed *N*-glycosylation. The receiver operator curve (ROC) area under the curve (AUC) values of the three different machine learning algorithms were used as a proxy to study the relative importance of peptide sequence length surrounding the glycosylated asparagine. The ROC AUC scores (y axes) are plotted against the numbers of amino acids to the right (red), left (purple) and both (blue) used in the training of the (A) Extra Tree classifier, (B) Random Forest classifier and (C) of the Support Vector Machine classifier.

**Fig. S4.** Selection of key features for machine learning training. The amino acid sequence features used by the Extra Tree classifier (n=19), the Random Forest classifier (n=8) and the Support Vector Machine classifier (n=73) are analysed with a Venn diagram. The figure shows the names of the features that are considered important at least by two classifiers. These are: The figure shows the names of the features that are considered important at least by two classifiers. These are: aa\_count:V: the count of valine amino acids in the peptide sequence, acc\_charge\_left@X: accumulating charge from the left side of the peptide with a window of X amino acids, bonus\_all: the sums of all the bonus scores for the peptide, bonus\_max: the highest bonus score for the peptide, bonus\_presence\_D: the presence or absence of Aspartic Acid (D) in the peptide sequence, PH\_Charge\_X: the charge of the peptide at pH X, PI: isoelectric point value of the peptide, scale\_Atch\_X\_window\_Y: Atchely amino acid scale value with a segment of X amino acids and a window of Y amino acids, scale\_polarizability\_window\_17: amino acid polarizability scale value with a window of 17 amino acids.

**Fig S5.** Predictor Scores. The figure shows box plots of the scores (5-fold cross validation) for the ROC AUC values (y axis) of the individual RF, ETC and SVM machine learning algorithms and the final combined (Voting Classifier) machine learning algorithm. Each score is computed 100 times with a different random split of the original dataset.

**Fig. S6.** Amino acid sequence alignment of *TbSTT3A*, B and C. Putative transmembrane domains are indicated in grey and positively charged amino acids found in *TbSTT3A* and *TbSTT3B*, but not found in *TbSTT3C*, are highlighted in red boxes.

## Table legends:

**Supplementary Table S1. Glycoproteomics Data.** Glycopeptides identified from the glycoproteomics experiments described in this paper and the reprocessed data of (Izquierdo et al., 2009b).

**Supplementary Table S2. Peptide count data used to discriminate TbSTT3A from TbSTT3B substrates.** Peptides containing Asn-N-GlcNAc and/or [<sup>18</sup>O]Asp within NXS/T sequons identified  $\geq 3$  times were assigned to being predominantly *TbSTT3A* or *TbSTT3B* substrates when the proportion of [<sup>18</sup>O]Asp containing forms were  $\geq 0.8$  and  $\leq 0.4$ , respectively.

**Supplementary Table S3. Machine learning predictions.** The table reports the Voting Classifier prediction results of the NXS/T ( $X \neq P$ ) sequons in the *T. brucei* proteome for protein with an N-terminal signal peptide. The protein identification numbers (Protein Id), the protein descriptions (Protein Description), and positions (Protein Site) of the sequons are reported along with the peptide sequence window (Predicted Peptide) used for the prediction. Finally, the table reports the score of the Voting Classifier (Predictor Score) with the prediction of the OST enzyme (Predicted Enzyme) that is more likely to modify the NXS/T sequons.

Fig S1

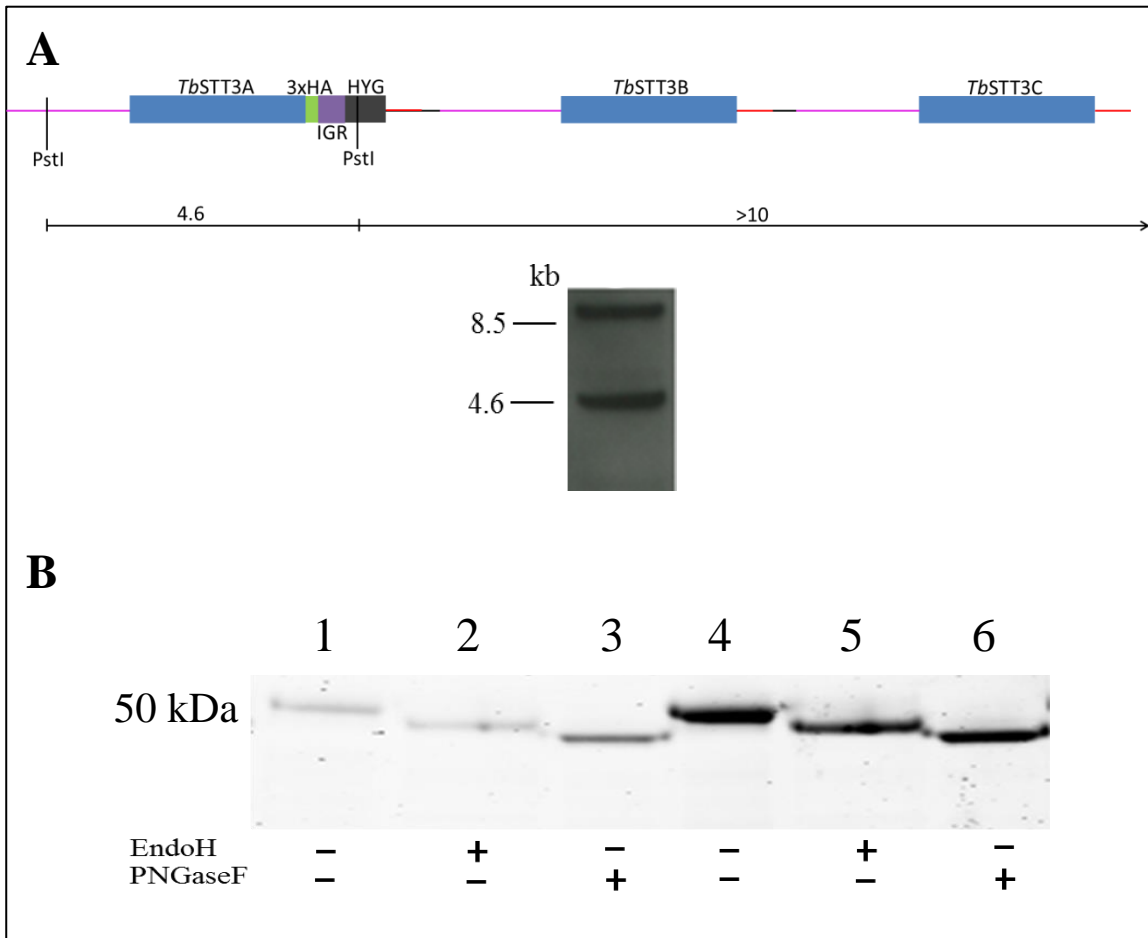


Fig S2

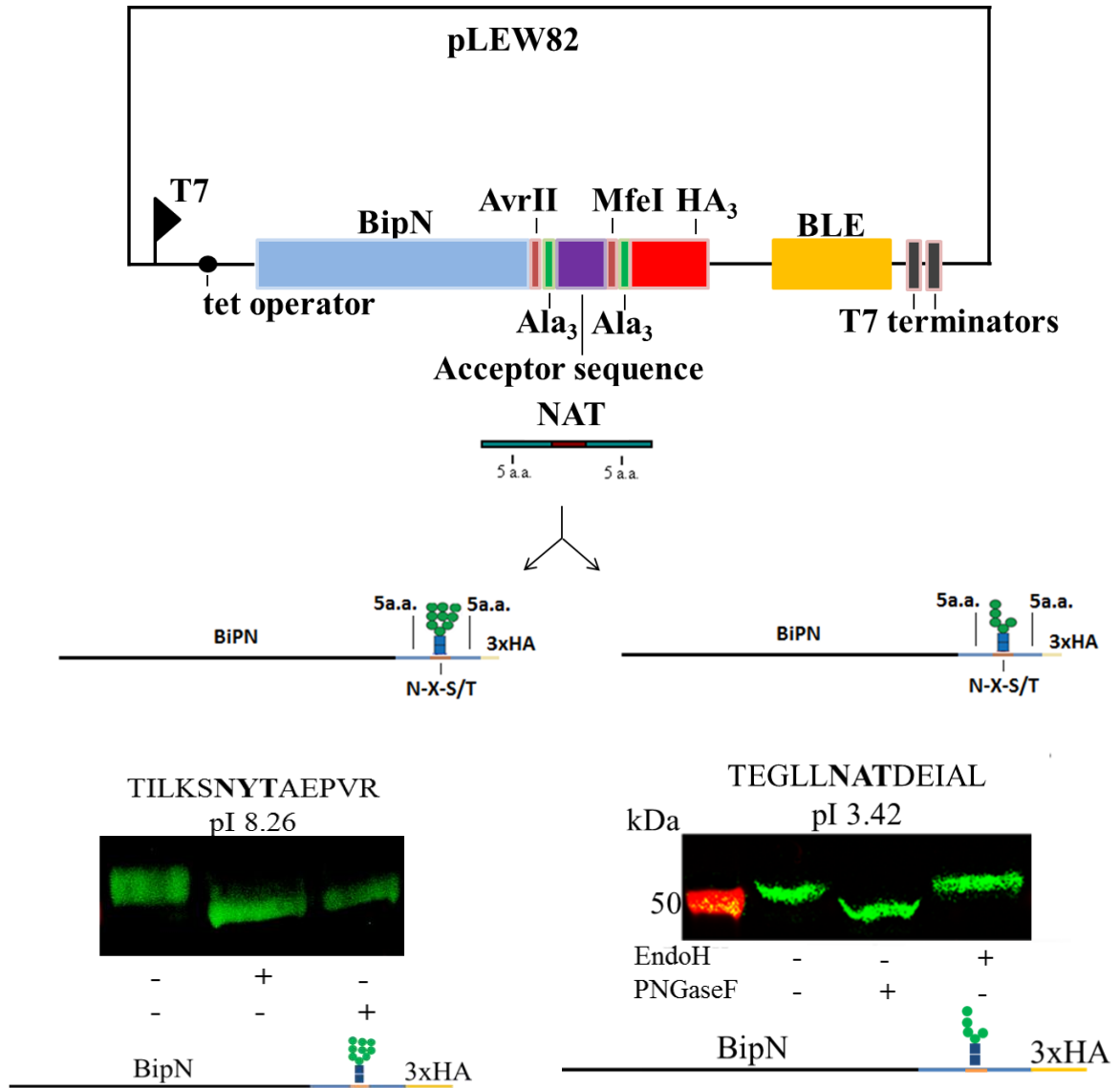
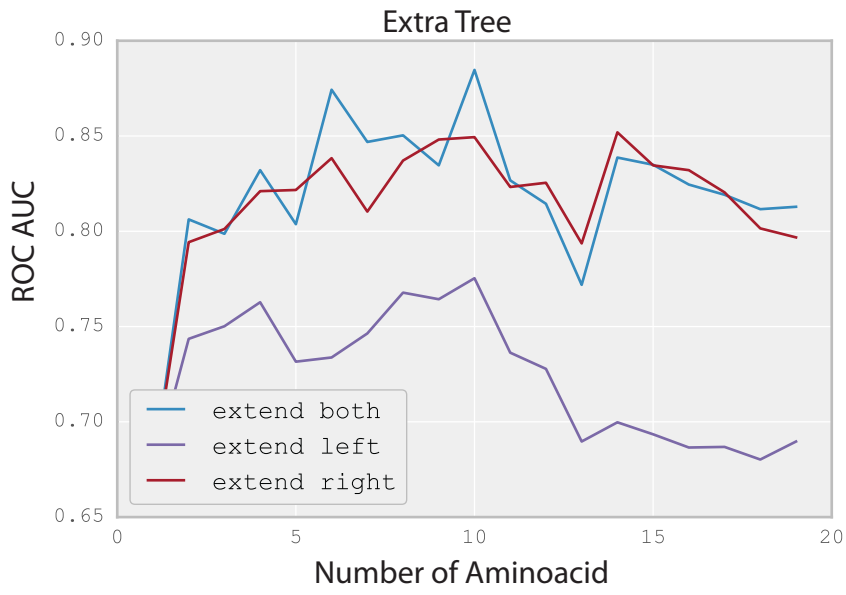
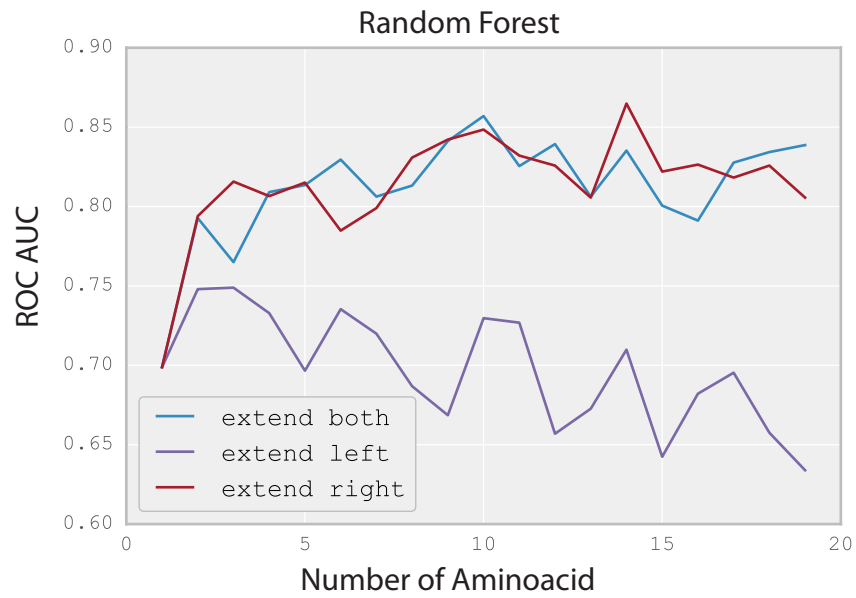


Fig S3

A



B



C

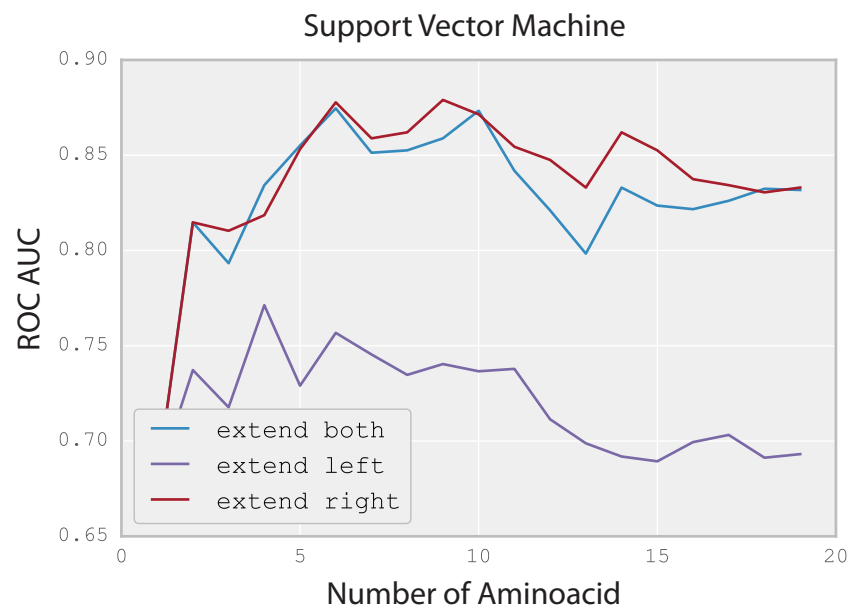


Fig S4

Important features after optimization

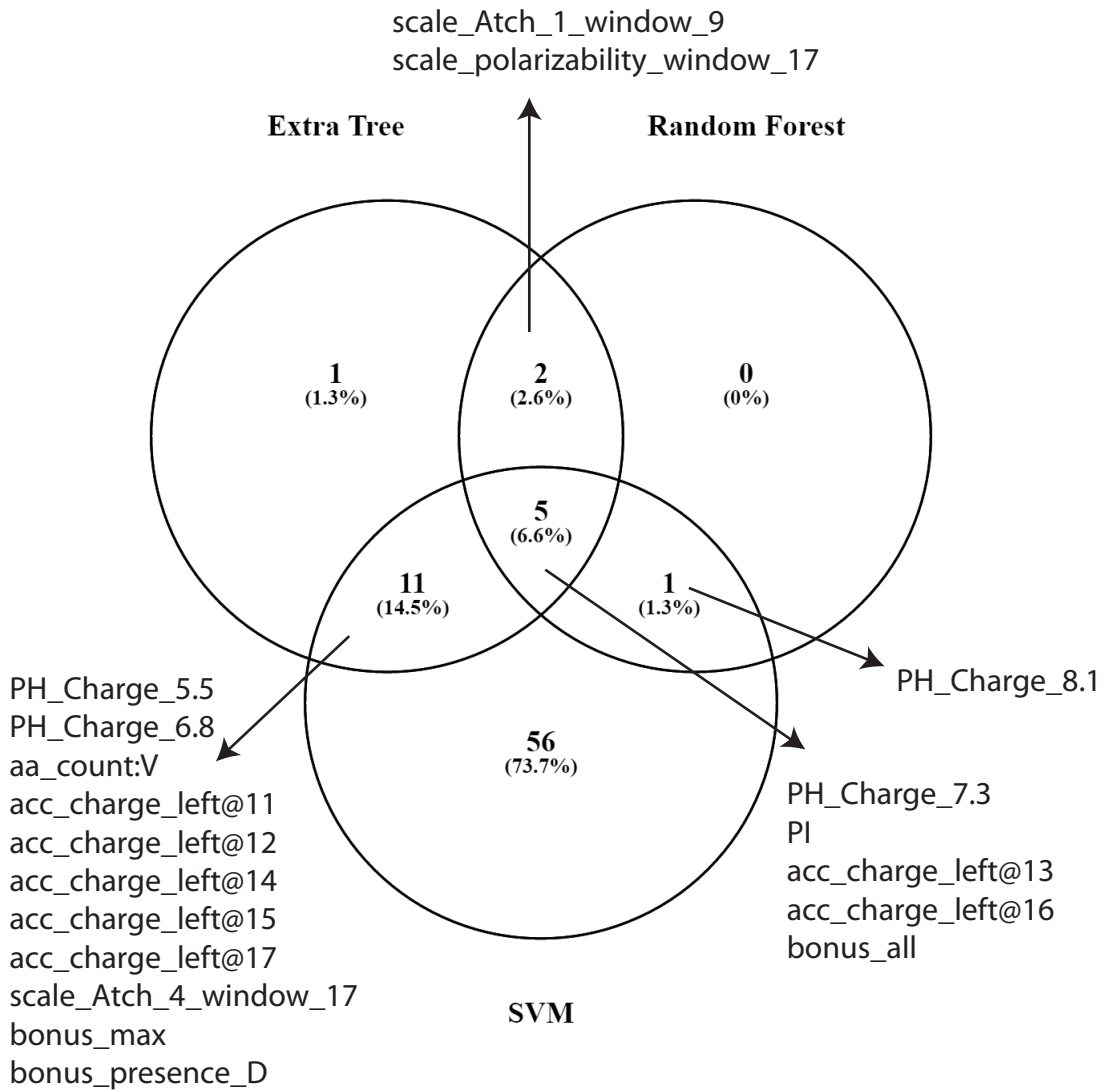


Fig S5

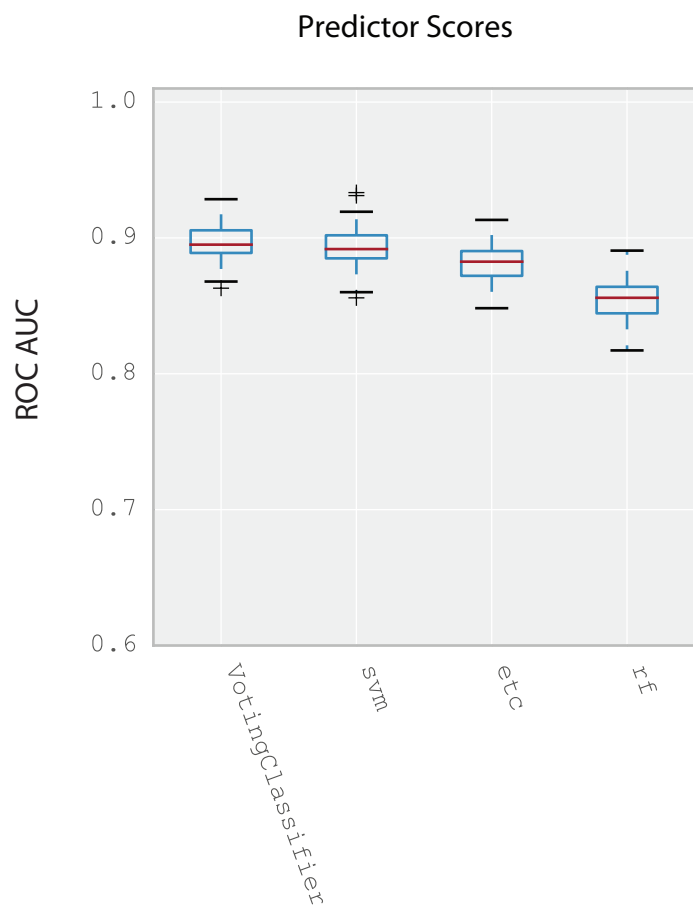




Fig S6.

**Transmembrane Helix Region**

Consensus	MTKGGKVAVTKGSAQSDGAGEGGMKAKSSSTTFVATGGGSLPAWALKAVSTVVS AV I L I YSVHRA YD I R L T S V R L Y G E L I	
STT3C	1	80
STT3B	1	80
STT3A	1	80
Consensus	HEFDPFNFYRATQYLSDNQWRAFFQWYDYMSWYPLGRPVGTTIFPGMQLTGVAIHRVLEMLGRGMSINN ICVYIPAWFGS	
STT3C	81	160
STT3B	81	160
STT3A	81	160
Consensus	IATVLAAL IAYESSNSLSVMAFTA YFFS I VPAHLMRSMAGEFDNECVAMAAMLLTFYMWVRS LRSSSSWP I G A L A G V A Y G	
STT3C	161	240
STT3B	161	240
STT3A	161	240
Consensus	YMVSTWGGYIFV LNMVAFHASVCVLLDWARDGTYYSVLLRAYSLFFVIGTALAICVPPVEWTFPRSLEQLTALFV FVFMWA	
STT3C	241	320
STT3B	241	320
STT3A	241	320
Consensus	LHYSEYLRERARAP IHSSKALQIRARIFMGTLSLLL IVA IYLFSTGYFRPFSSRVRALFVKHTRTGNPLVDSVAEHRPTT	
STT3C	321	400
STT3B	321	400
STT3A	321	400
Consensus	AGAF LRYLHVCYNGWII GFFFMVSCFFHCTP GMSFLLLYSILAYYFSLKMSRLLLLSAPVASILTG YVVGSI V D L A	
STT3C	401	477
STT3B	401	477
STT3A	401	477
Consensus	ADCFAAS GTEHADSKHEHQKARGKGO KEQITVECGCHNPFYKLWCNSFSSRLVVGKFFV VVLS ICGPTFLGS F	
STT3C	478	552
STT3B	478	552
STT3A	478	554
<b>Region homologous to <i>Pfuriosus</i></b>		
Consensus	R CE FA S SSPRII G RVLADDYVYSYLWLRNNTPEDARILSWWDYGYQITGIGNRTTLADGNTWNHEHI	
STT3C	553	631
STT3B	553	631
STT3A	555	634
Consensus	ATIGKMLTSPVKESHALIRHLADYVL IWAGEDRSDDLKSPHVARIGNSVYRDMCSEDDPLCTQFGFYSGDFNKPTPMMQR	
STT3C	632	711
STT3B	632	711
STT3A	635	714
Consensus	SLLYNLHRFGTDGGKTQLDKNMFQLAYVSKYGLVK IYKVMNVS EESKAWADPKNRKCDAPGSWICAGQYPPAKEIQDML	
STT3C	712	791
STT3B	712	791
STT3A	715	794
Consensus	AKRIDYEQL EDFNRRNRSDAYYRAYMRQMG	
STT3C	792	821
STT3B	792	821
STT3A	795	801