Identifying *wtf* Genes in the Wild JB1174 Strain of Fission Yeast

Exam number = Y3603988

Introduction

Selfish driver alleles distort Mendelian segregation by killing gametes which do not share these same alleles (Hu et al. 2017; Shropshire and Rokas 2017). The fission yeast *Schizosaccharomyces pombe* is used as a model organism to study the impact of selfish drivers on genome evolution (Hu et al. 2017). The *wtf* multi-gene family encode selfish drivers which use the poison-antidote model of meiotic drive (Hu et al. 2017; Nuckolls et al. 2017). There are 25 *wtf* members in the *S. pombe* reference genome but an unknown number in wild fission yeast strains (Nuckolls et al. 2017).

This project aimed to characterise the *wtf* gene compliment in the JB1174 wild *S. pombe* strain genome to test the following hypotheses: JB22 and *S. pombe* standard reference genome assemblies are identical; JB1174 has more *wtf* genes than JB22; JB1174 *wtf* genes have similar loci to those in JB22; not all JB1174 *wtf* genes have full ORFs with many being truncated/mutated due to their predicted rapid divergence (McLaughlin and Malik, 2017).

Methods

Strain Genome Sequence Information (Table-1)

Table-1. Information on the genome assemblies analysed in this report.

Genome Assembly Name	Bähler laboratory strain name	Strain ID(s)	Sequencing Technology	Information

S. pombe standard reference	JB22	Leupolds 972 (h);
		CBS10395;
		NCYC1430
			Illumina	Short reads. *wtf4* gene sequence used as query sequence in BLAST taken from this genome. (Pombase.org, 2017). Same strain as JB22.

JB22 new reference sequence	JB22	Leupolds 972 (h);
		CBS10395;
		NCYC1430
			Oxford Nanopore	Long reads. Same strain as used for *S. pombe* standard reference genome re-sequenced.

JB1174	JB1174	CECT12918;
		NOTT126;
		IFI2139
			Oxford Nanopore	Long reads. Test strain. Used repolished genome assembly.

BLAST

Linux BLAST+ was used to search for *wtf* genes in JB22, JB1174 and *S. pombe* reference genome assemblies. The *wtf4* nucleotide sequence (including introns and 3' UTR) was used as the BLASTn query sequence to identify potential *wtf* genes. BLASTn hits were filtered to an E-value below 10^{-6} and an identity above 75%. The *wt4* transcript sequence (exons only) was used as the tBLASTn query sequence to identify *wtf* genes with non-

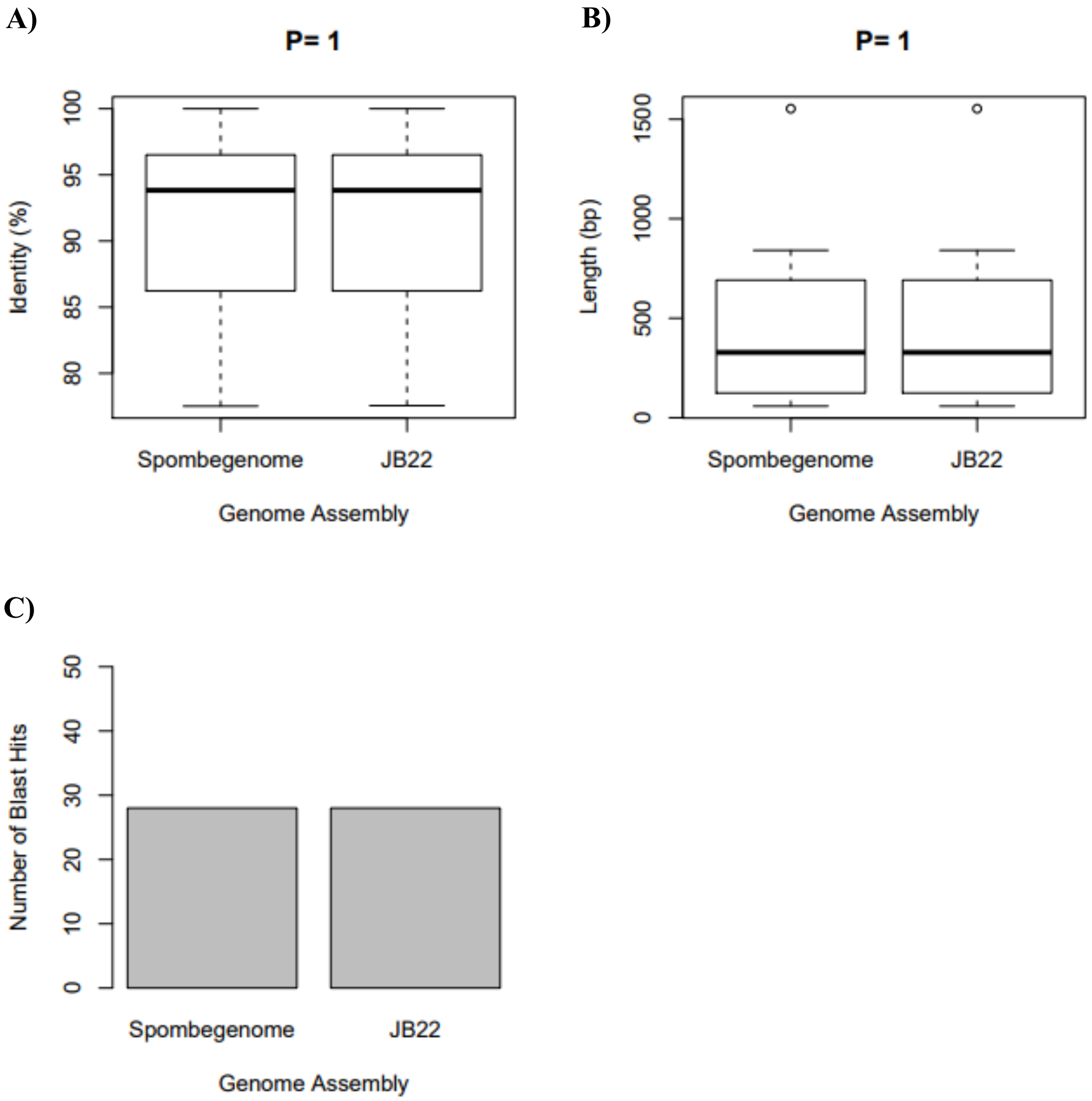


Figure 1. Comparing the BLASTn hits in the standard S. pombe and JB22 reference genomes.

wtf4 gene sequence (including introns and 3'UTR) was queried via BLASTn against the *S. pombe* standard reference and JB22 reference genome assemblies.

A) Average and distribution of **(A)** percentage identity and **(B)** length of BLASTn hits in the *S. pombe* standard and JB22 reference genomes relative to the *wtf4* gene sequence. Thick middle line is the average, box shows interquartile range, whiskers show range. P value displayed above the graph, calculated using Wilcoxon t-test.

C) Number of BLASTn hits in the *S. pombe* standard and JB22 reference genomes.

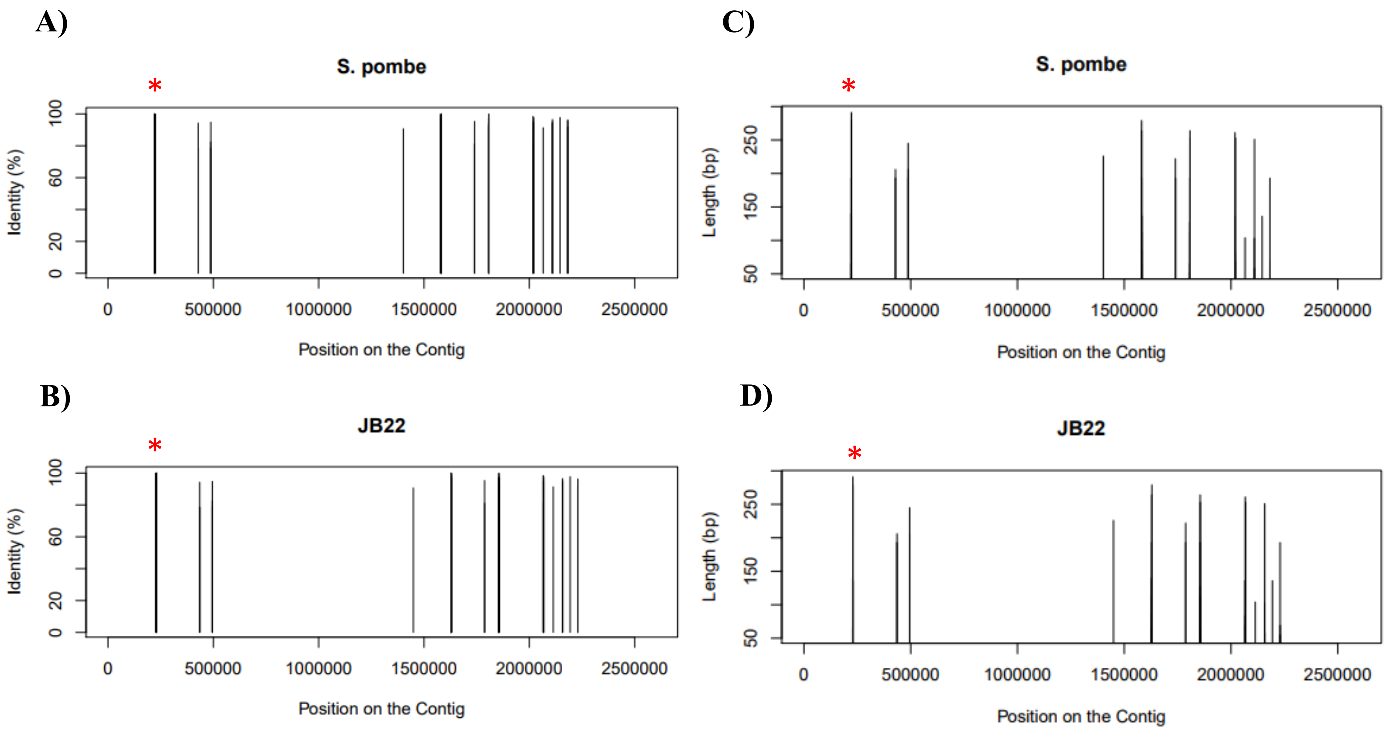


Figure 2. Position and properties of BLASTn hits in the standard S. pombe and JB22 reference genomes.

wtf4 gene sequence (including introns and 3'UTR) was queried via BLASTn against the *S. pombe* standard reference and JB22 reference genome assemblies. The contig lengths in the standard *S. pombe* and JB22 genome assemblies were determined in Linux. Data subsets containing BLAST hits limited to one contig sequence ID were created in R. Negative values of the JB22 hit contig positions were plotted, as the minus strand of the JB22 genome was sequenced. This allows direct comparison with the standard *S. pombe* reference. * = *wtf4* gene used as query sequence. Note: the length of the wtf4 gene (*) in the BLASTn hits is 1553bp, which has been cut off the graph to enable more effective comparison between all other BLAST hits.

A) Percentage identity of BLASTn hits in *S. pombe* standard reference chromosome III plotted at the position on the contig of each hit.

B) Percentage identity of BLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

C) Length (bp) of BLASTn hits in *S. pombe* standard reference chromosome III plotted at the position on the contig of each hit.

D) Length (bp) of BLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

The standard and JB22 reference genome *wtf4*-tBLASTn hits had an identical distribution of identities (Figure-3a; P=1) and lengths (Figure-3b; P=1), at the same contig positions (Figure-4). There are 33 *wtf4*-tBLASTn hits in both reference genome assemblies (Figure-3c).

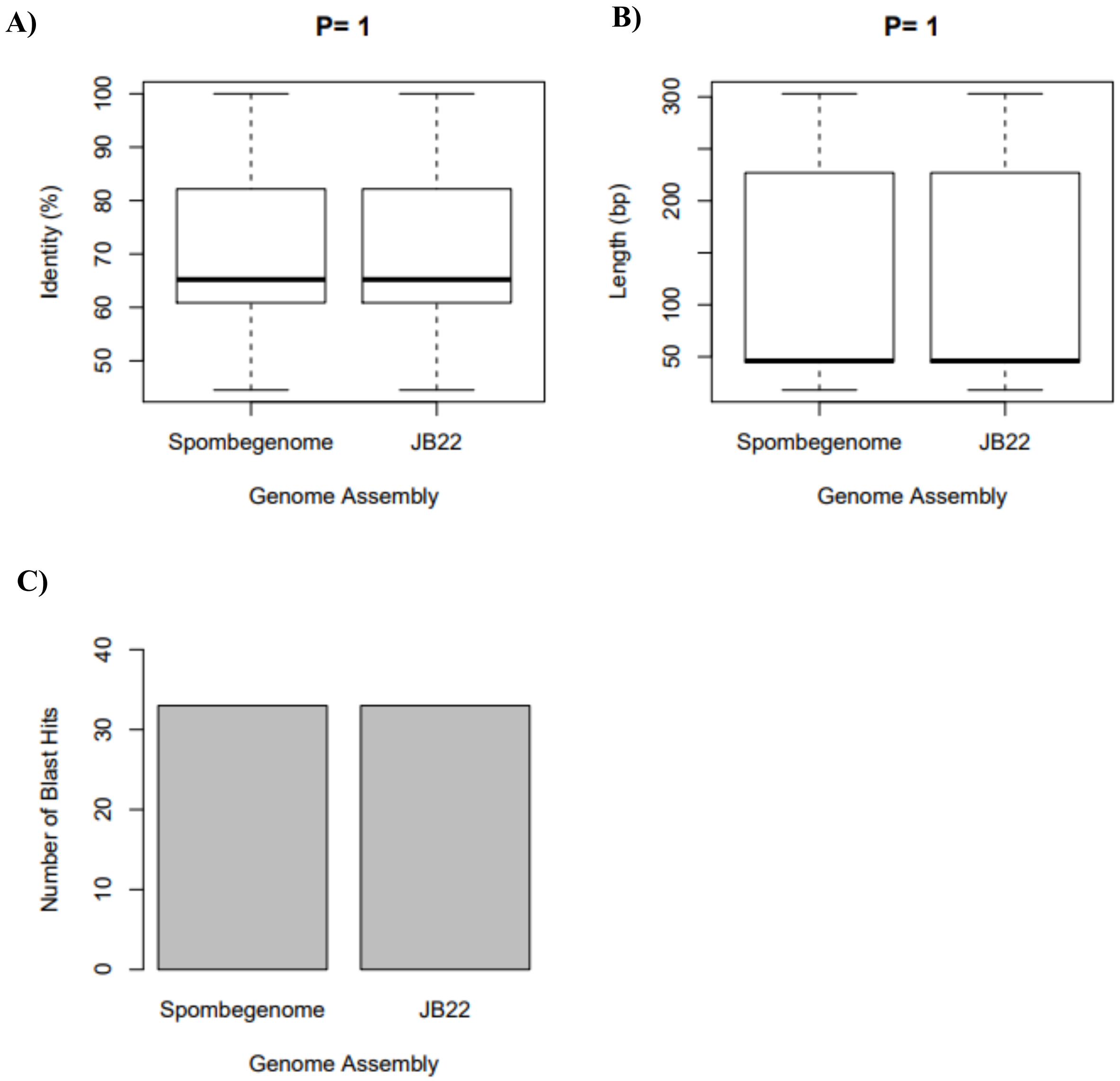


Figure 3. Comparing the tBLASTn hits in the standard S. pombe and JB22 reference genomes.

wtf4 transcript sequence (exons only) was queried via tBLASTn against the *S. pombe* standard reference and JB22 reference genome assemblies.

Average and distribution of **(A)** percentage identity and **(B)** length of tBLASTn hits in the *S. pombe* standard and JB22 reference genomes relative to the *wtf4* gene sequence. Thick middle line is the average, box shows interquartile range, whiskers show range. P value displayed above the graph, calculated using Wilcoxon t-test.

C) Number of tBLASTn hits in the *S. pombe* standard and JB22 reference genomes.

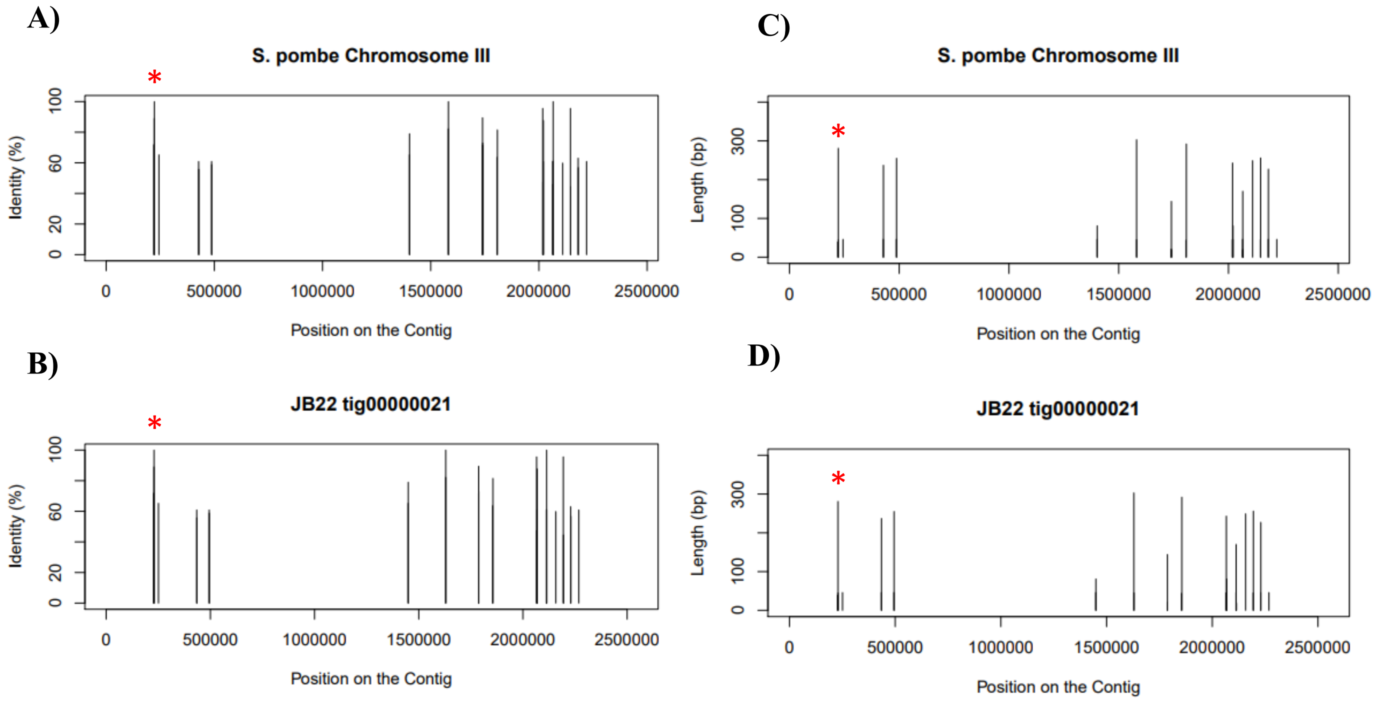


Figure 4. Position and properties of tBLASTn hits in the standard S. pombe and JB22 reference genomes.

wtf4 transcript sequence (exons only) was queried via tBLASTn against the *S. pombe* standard reference and JB22 reference genome assemblies. The contig lengths in the standard *S. pombe* and JB22 genome assemblies were determined in Linux. Data subsets containing BLAST hits limited to one contig sequence ID were created in R. Negative values of the JB22 hit contig positions were plotted, as the minus strand of the JB22 genome was sequenced. * = *wtf4* gene used as query sequence.

A) Percentage identity of tBLASTn hits in *S. pombe* standard reference chromosome III plotted at the position on the contig of each hit.

B) Percentage identity of tBLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

C) Length (bp) of tBLASTn hits in *S. pombe* standard reference chromosome III plotted at the position on the contig of each hit.

D) Length (bp) of tBLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

JB1174 Has More BLAST Hits Than JB22 Reference Genome

JB22 genome assembly was used as the reference for JB1174 BLAST hit analysis. There is no significant difference between the distribution of *wtf4*-BLASTn hit identities and lengths in JB1174 and JB22 (P>0.05) (Figure-5a,b). After removal of sequences shorter than 100bp, JB1174 has 18 more BLASTn hits than JB22 (Figure-5c).

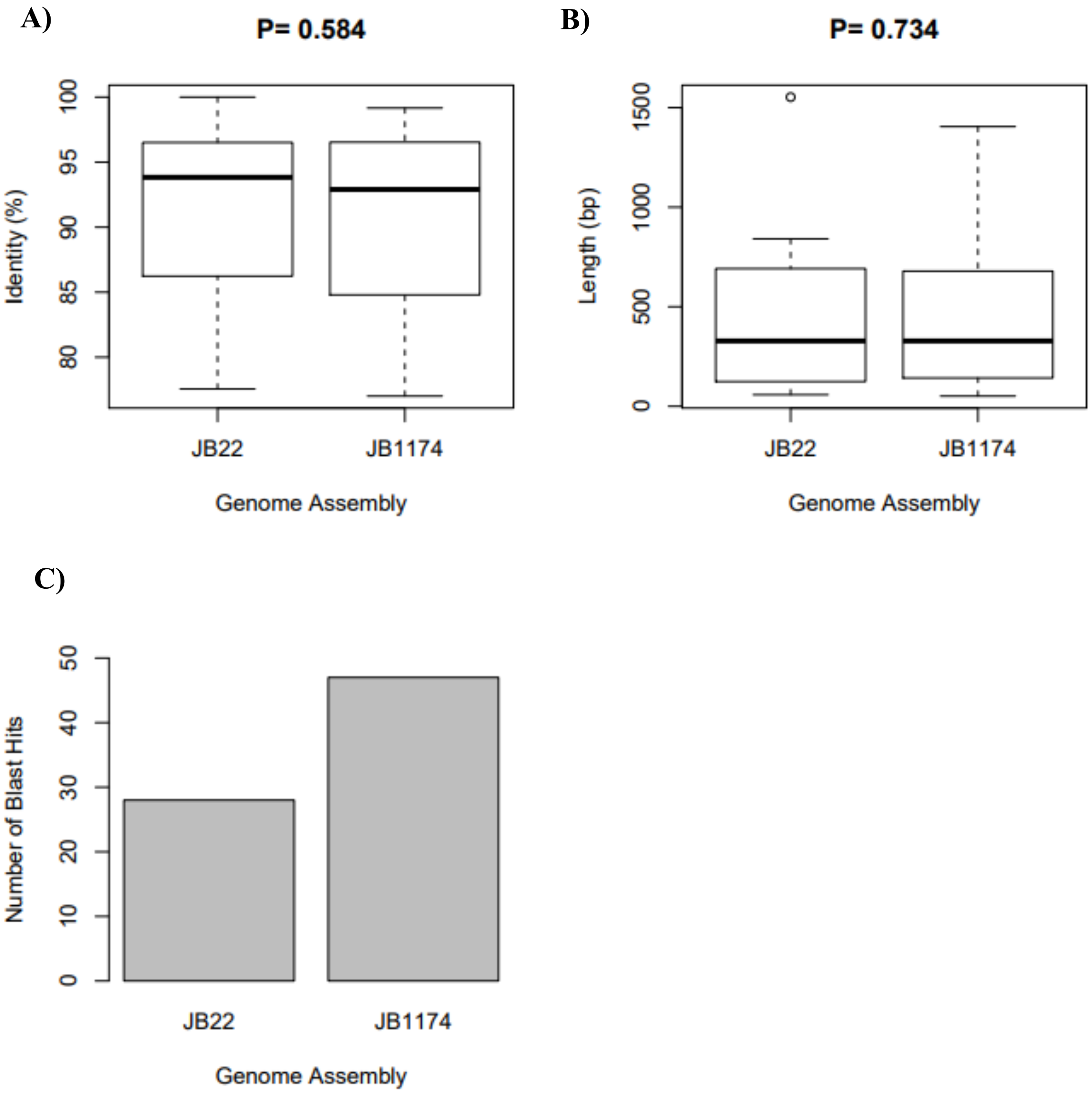


Figure 5. Comparing the BLASTn hits in the JB22 reference and JB1174 genomes.

wtf4 gene sequence (including introns and 3'UTR) was queried via BLASTn against the *S. pombe* standard reference and JB22 reference genome assemblies.

Average and distribution of **(A)** percentage identity and **(B)** length of BLASTn hits in the JB22 reference and JB1174 genomes relative to the *wtf4* gene sequence. Thick middle line is the average, box shows interquartile range, whiskers show range. P value displayed above the graph, calculated using Wilcoxon t-test.

C) Number of BLASTn hits in the JB22 reference and JB1174 genomes.

The identity and length distribution of JB1174 *wtf4*-tBLASTn hits are not significantly different to those of the JB22 tBLASTn hits (P>0.05) (Figure-6a,b). There are 31 more tBLASTn hits in JB1174 than in JB22 (Figure-6c).

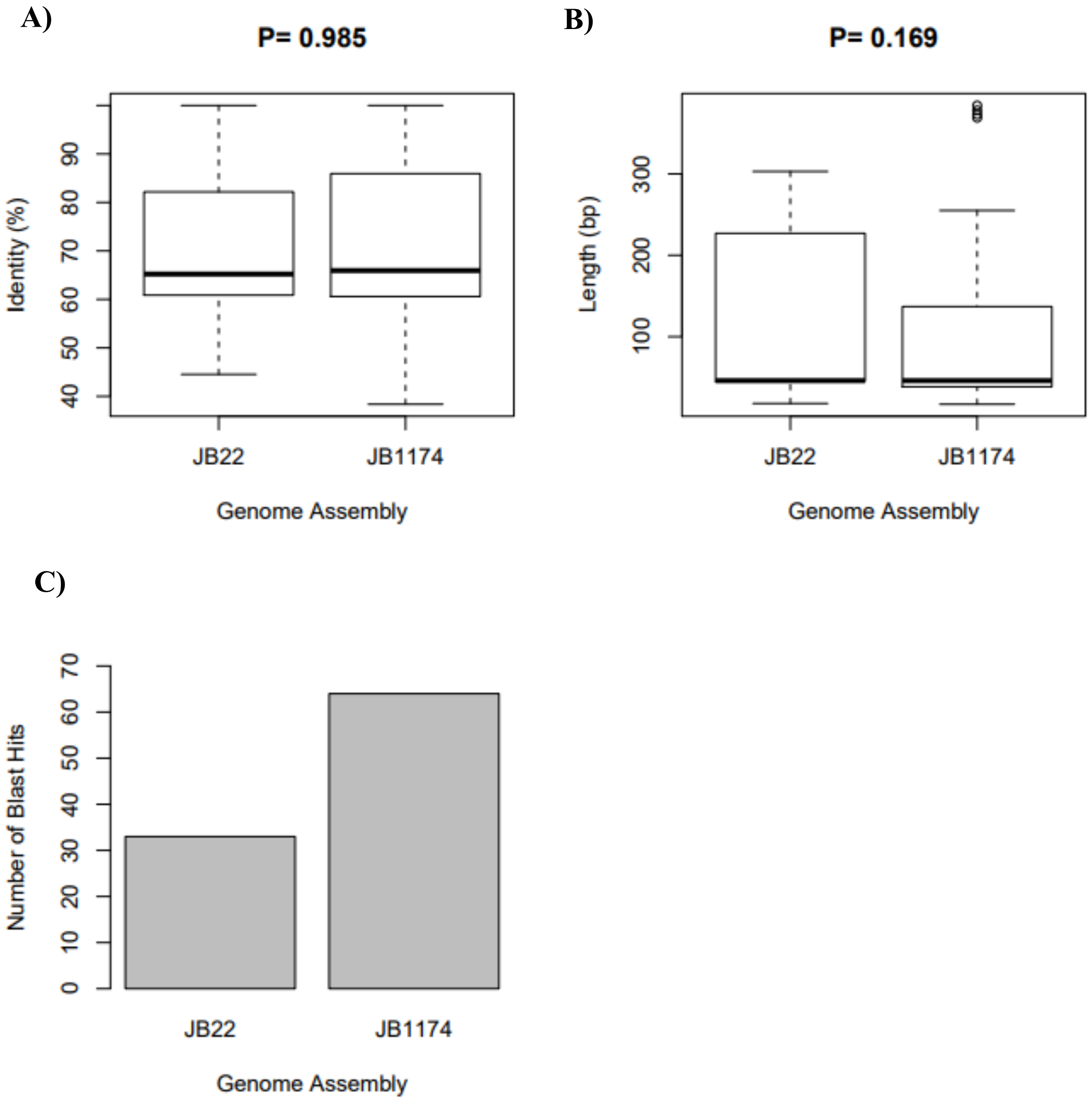


Figure 6. Comparing the tBLASTn hits in the standard JB22 reference and JB1174 genomes.

wtf4 transcript sequence (exons only) was queried via tBLASTn against the JB22 reference and JB1174 genome assemblies.

Average and distribution of **(A)** percentage identity and **(B)** length of tBLASTn hits in the JB22 reference and JB1174 genomes relative to the *wtf4* gene sequence. Thick middle line is the average, box shows interquartile range, whiskers show range. P value displayed above the graph, calculated using Wilcoxon t-test.

C) Number of tBLASTn hits in the JB22 reference and JB1174 genomes.

BLASTn and tBLASTn hits occur in four JB1174 contigs as opposed to one contig like in JB22 (Figure-7,8). Most JB1174 BLAST hits occur on contig 8 (Figure-7b,8b).

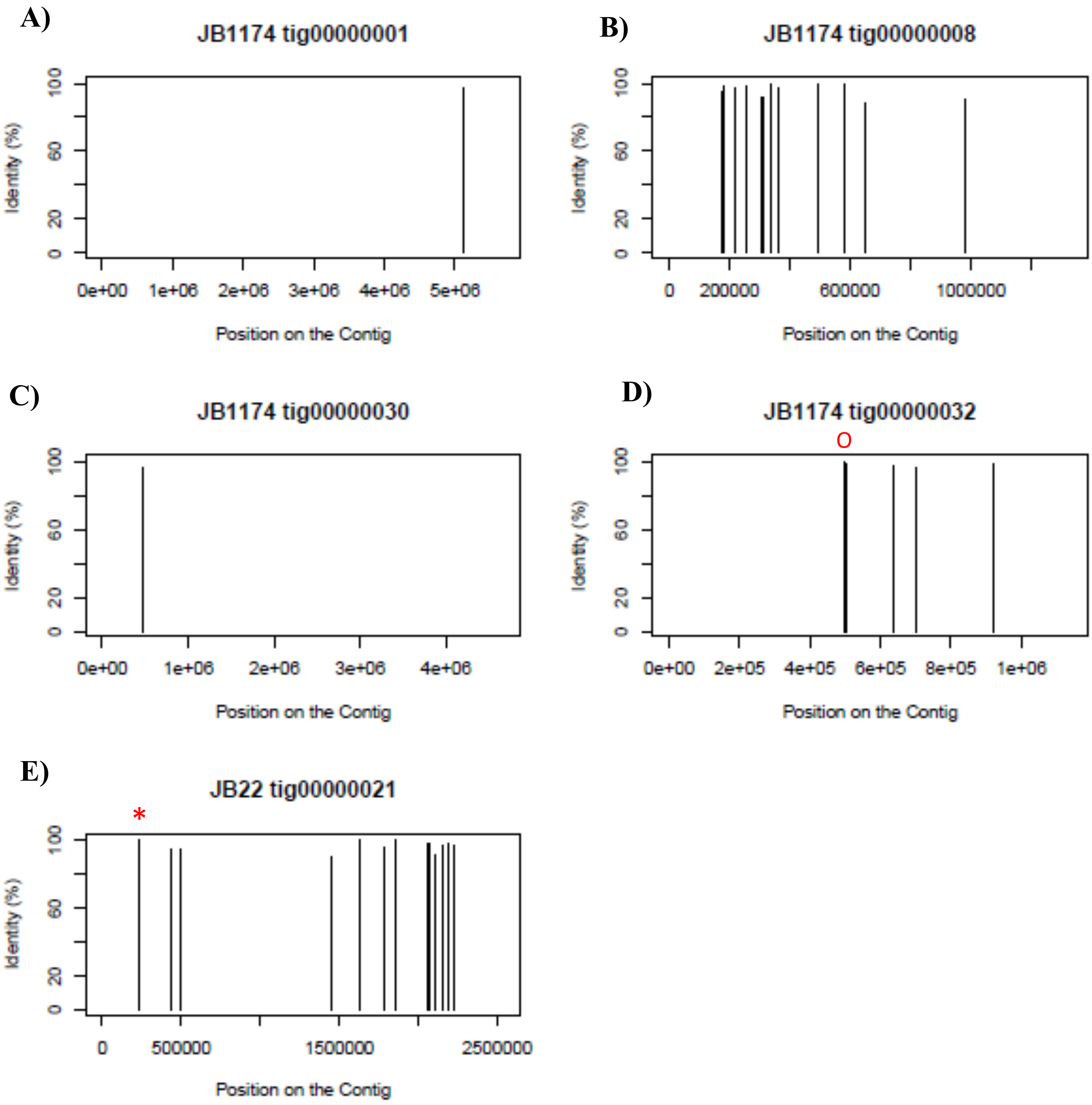


Figure 7. Percentage Identity of BLASTn hits in JB22 reference and JB1174 genomes.

wtf4 gene sequence (including introns and 3'UTR) was queried via BLASTn against the JB22 reference and JB1174 genome assemblies. The contig lengths in the JB22 and JB1174 genome assemblies were determined in Linux. Data subsets containing BLAST hits limited to one contig sequence ID were created in R. Negative values of the JB22 hit contig positions were plotted, as the minus strand of the JB22 genome was sequenced. * = *wtf4* gene used as query sequence. O = *wtf4* gene in JB1174.

Percentage identity of BLASTn hits in JB1174 **(A)** contig 1, **(B)** contig 8, **(C)** contig 30 and **(D)** plotted at the position on the contig of each hit.

E) Percentage identity of BLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

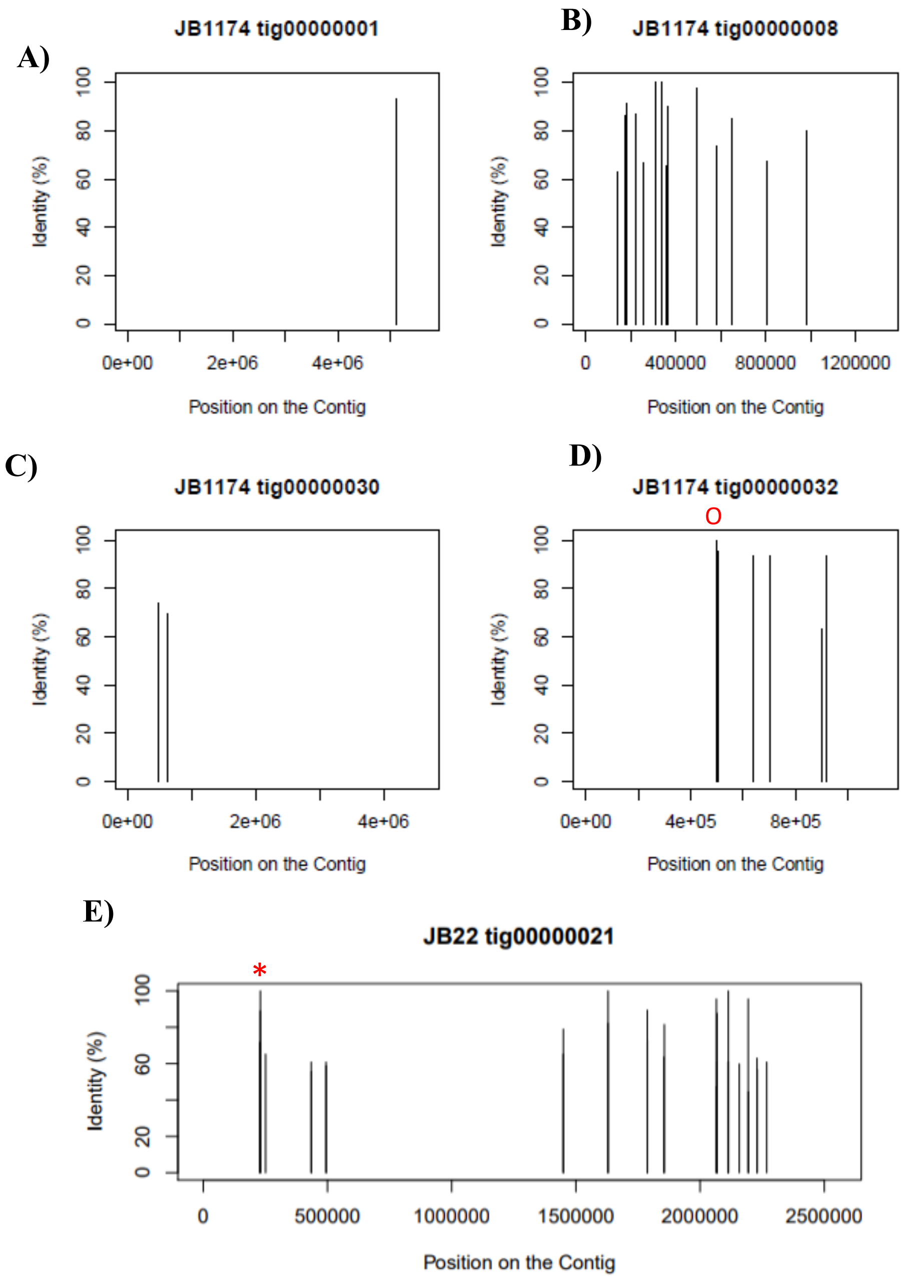


Figure 8. Percentage Identity of tBLASTn hits in JB22 reference and JB1174 genomes.

wtf4 transcript sequence (exons only) was queried via tBLASTn against the JB22 reference and JB1174 genome assemblies. The contig lengths in the JB22 and JB1174 genome assemblies were determined in Linux. Data subsets containing BLAST hits limited to one contig sequence ID were created in R. Negative values of the JB22 hit contig positions were plotted, as the minus strand of the JB22 genome was sequenced. * = *wtf4* gene used as query sequence. O = *wtf4* gene in JB1174.

Percentage identity of tBLASTn hits in JB1174 **(A)** contig 1, **(B)** contig 8, **(C)** contig 30 and **(D)** plotted at the position on the contig of each hit.

E) Percentage identity of tBLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

The JB1174 BLASTn hit encoding *wtf4* is near the middle of contig 32 (Figure-9d). Most extra tBLASTn hits (relative to BLASTn hits) clustering near the beginning of contig 8 are below 100 bps (Figure-10b).

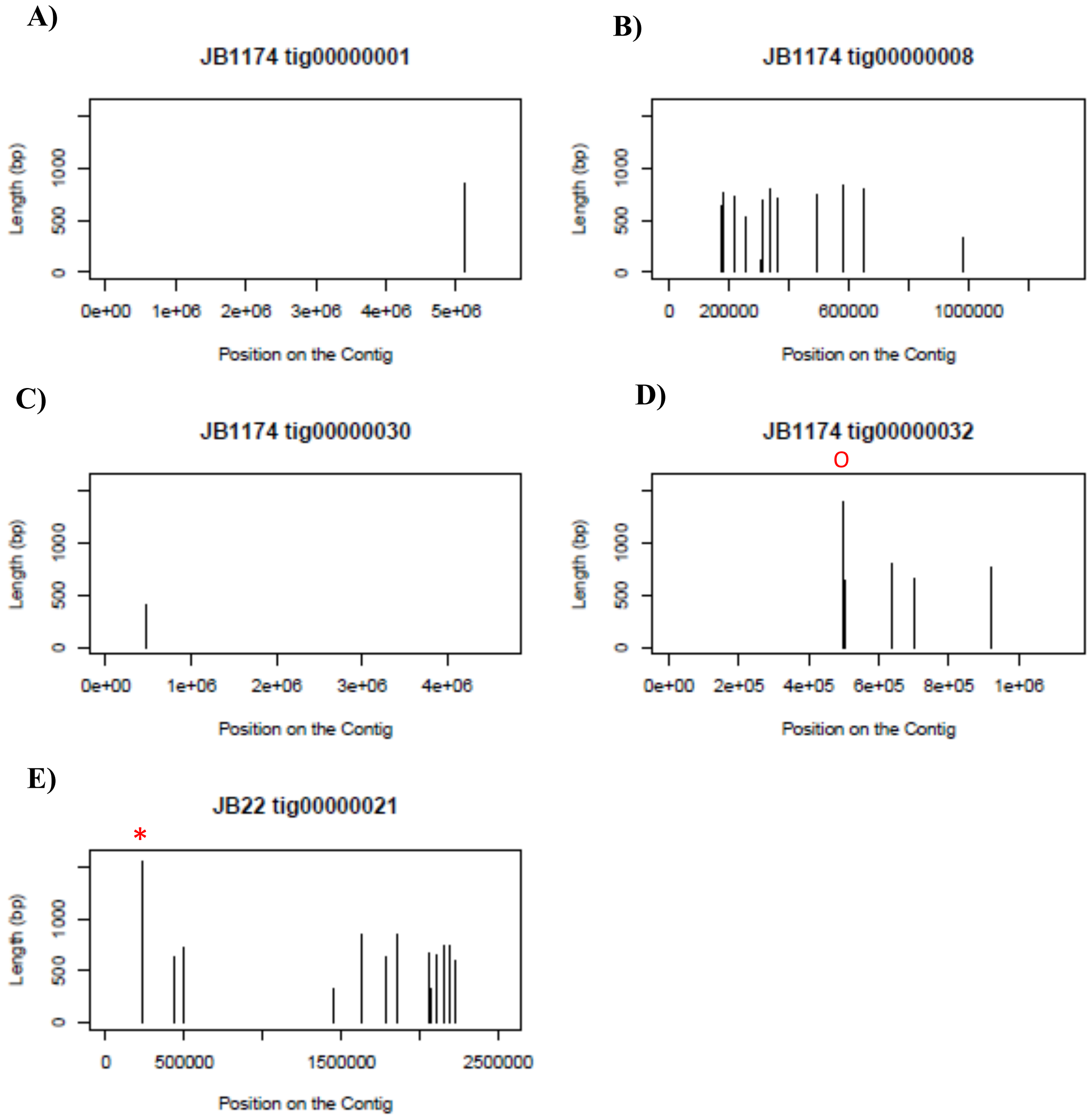


Figure 9. Length of BLASTn hits in JB22 reference and JB1174 genomes.

wtf4 gene sequence (including introns and 3'UTR) was queried via BLASTn against the JB22 reference and JB1174 genome assemblies. The contig lengths in the JB22 and JB1174 genome assemblies were determined in Linux. Data subsets containing BLAST hits limited to one contig sequence ID were created in R. Negative values of the JB22 hit contig positions were plotted, as the minus strand of the JB22 genome was sequenced. * = *wtf4* gene used as query sequence. O = *wtf4* gene in JB1174.

Length of BLASTn hits in JB1174 **(A)** contig 1, **(B)** contig 8, **(C)** contig 30 and **(D)** plotted at the position on the contig of each hit.

E) Length of BLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

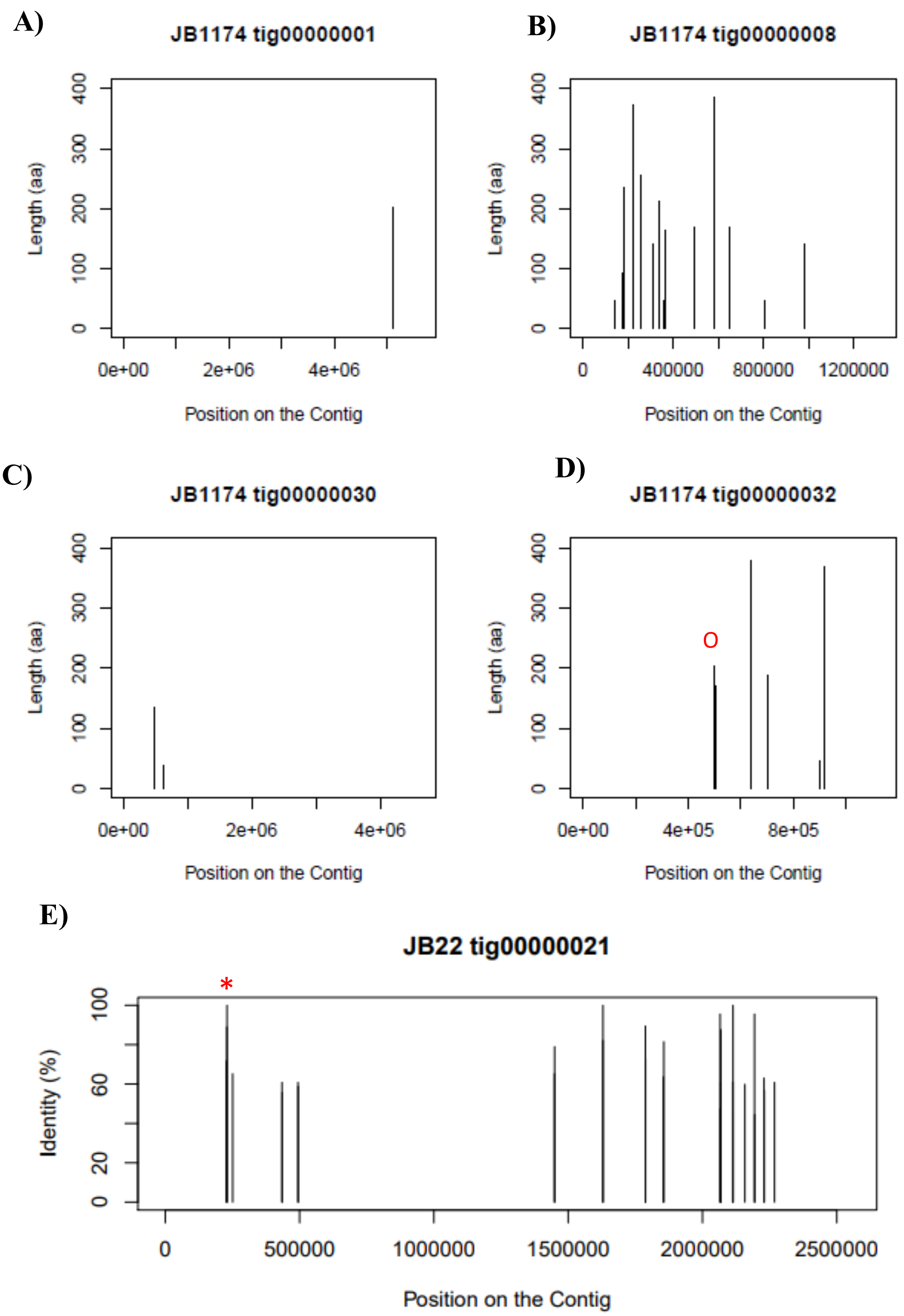


Figure 10. Length of tBLASTn hits in JB22 reference and JB1174 genomes.

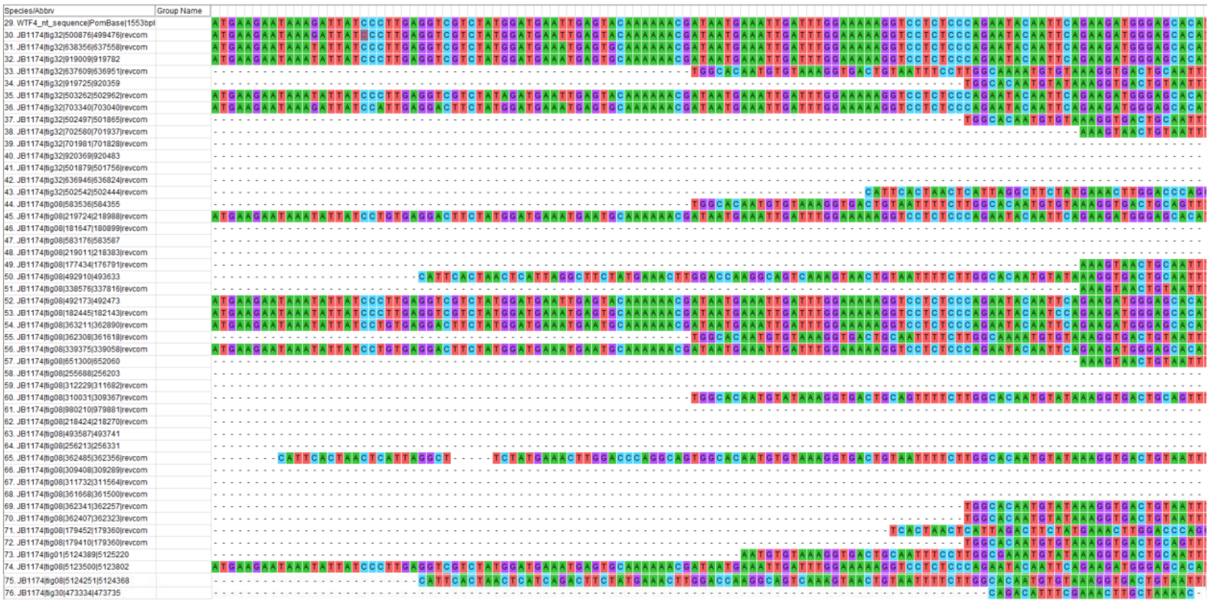
wtf4 transcript sequence (exons only) was queried via tBLASTn against the JB22 reference and JB1174 genome assemblies. The contig lengths in the JB22 and JB1174 genome assemblies were determined in Linux. Data subsets containing BLAST hits limited to one contig sequence ID were created in R. Negative values of the JB22 hit contig positions were plotted, as the minus strand of the JB22 genome was sequenced. * = *wtf4* gene used as query sequence. O = *wtf4* gene in JB1174.

Length of tBLASTn hits in JB1174 **(A)** contig 1, **(B)** contig 8, **(C)** contig 30 and **(D)** plotted at the position on the contig of each hit.

E) Length of tBLASTn hits in JB22 reference contig 21 plotted at the position on the contig of each hit.

Alignment of JB1174 and JB22 BLASTn Hits to *wtf4* Gene Query Sequence

11 of the BLASTn JB1174 hits showed homology to the whole of exon 1 (45bp), of which five are on contig 32 and six are on contig 8 (Figure-11).



Species/Abbrv | Group Name
29. WTF4_nt_sequence|PomBase|1553bpl
30. JB1174|tig32|500876|499476|revcom
31. JB1174|tig32|638356|637558|revcom
32. JB1174|tig32|919009|919782
33. JB1174|tig32|637609|636951|revcom
34. JB1174|tig32|919725|920359
35. JB1174|tig32|503262|502962|revcom
36. JB1174|tig32|703340|703040|revcom
37. JB1174|tig32|502497|501865|revcom
38. JB1174|tig32|702580|701937|revcom
39. JB1174|tig32|701981|701828|revcom
40. JB1174|tig32|920369|920483
41. JB1174|tig32|501879|501756|revcom
42. JB1174|tig32|636946|636824|revcom
43. JB1174|tig32|502542|502444|revcom
44. JB1174|tig08|583536|584355
45. JB1174|tig08|219724|218988|revcom
46. JB1174|tig08|181647|180899|revcom
47. JB1174|tig08|583176|583587
48. JB1174|tig08|219011|218383|revcom
49. JB1174|tig08|177434|176791|revcom
50. JB1174|tig08|492910|493633
51. JB1174|tig08|338576|337816|revcom
52. JB1174|tig08|492173|492473
53. JB1174|tig08|182445|182143|revcom
54. JB1174|tig08|363211|362890|revcom
55. JB1174|tig08|362308|361618|revcom
56. JB1174|tig08|339375|339058|revcom
57. JB1174|tig08|651300|652060
58. JB1174|tig08|255688|256203
59. JB1174|tig08|312229|311682|revcom
60. JB1174|tig08|310031|309367|revcom
61. JB1174|tig08|980210|979881|revcom
62. JB1174|tig08|218424|218270|revcom
63. JB1174|tig08|493587|493741
64. JB1174|tig08|256213|256331
65. JB1174|tig08|362485|362356|revcom
66. JB1174|tig08|309408|309289|revcom
67. JB1174|tig08|311732|311564|revcom
68. JB1174|tig08|361668|361500|revcom
69. JB1174|tig08|362341|362257|revcom
70. JB1174|tig08|362407|362323|revcom
71. JB1174|tig08|179452|179360|revcom
72. JB1174|tig08|179410|179360|revcom
73. JB1174|tig01|5124389|5125220
74. JB1174|tig08|5123500|5123802
75. JB1174|tig08|5124251|5124368
76. JB1174|tig30|473334|473735

Figure 11. Alignment of *wtf4*-BLASTn JB1174 hits with *wtf4* gene sequence – start of exon 1.

MUSCLE was used to align all JB1174 and JB22 BLASTn hits to the *wtf4* gene query sequence from the *S. pombe* standard reference genome. Results views in MEGA7. Alignment of all JB1174 wtf4 BLASTn hits to the start of exon 1 of the *wtf4* reference gene sequence (top sequence) is shown.

All JB1174 *wtf4*-BLASTn hits, except the JB1174 *wtf4* gene, have no *wtf4* homology between the beginning of intron 1 (Figure-12a) to half-way through exon 3 (Figure-12b). The last half of exon 3 is well conserved in these 11 sequences, however SNPs and indels occur (Figure-12b).

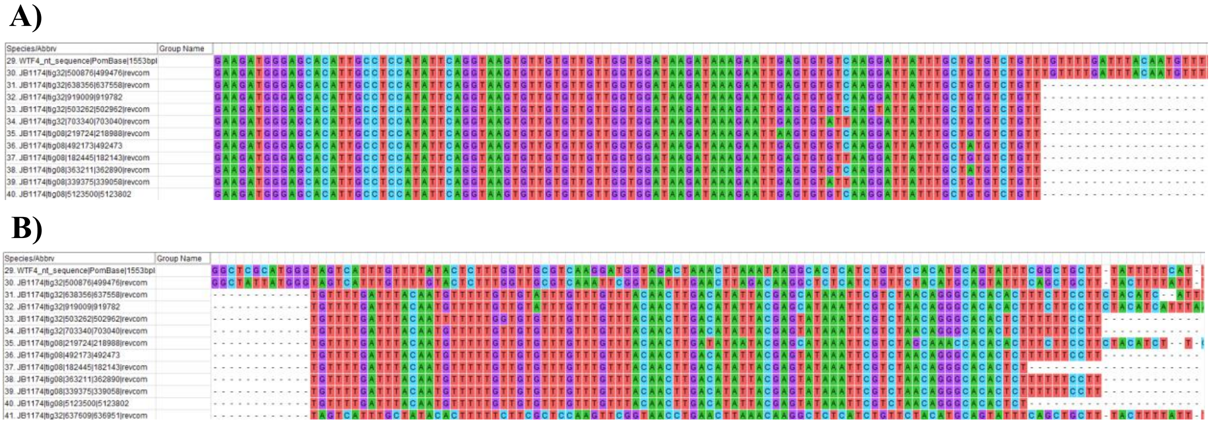

A)
Species/Abbrv
Group Name
29. WTF4_nt_sequence|PomBase|1553bpl
30. JB1174|tig32|500876|499476|revcom
31. JB1174|tig32|638356|637558|revcom
32. JB1174|tig32|919009|919782
33. JB1174|tig32|503262|502962|revcom
34. JB1174|tig32|703340|703040|revcom
35. JB1174|tig08|219724|218988|revcom
36. JB1174|tig08|492173|492473
37. JB1174|tig08|182445|182143|revcom
38. JB1174|tig08|363211|362890|revcom
39. JB1174|tig08|339375|339058|revcom
40. JB1174|tig08|5123500|5123802
B)
Species/Abbrv
Group Name
29. WTF4_nt_sequence|PomBase|1553bpl
30. JB1174|tig32|500876|499476|revcom
31. JB1174|tig32|638356|637558|revcom
32. JB1174|tig32|919009|919782
33. JB1174|tig32|503262|502962|revcom
34. JB1174|tig32|703340|703040|revcom
35. JB1174|tig08|219724|218988|revcom
36. JB1174|tig08|492173|492473
37. JB1174|tig08|182445|182143|revcom
38. JB1174|tig08|363211|362890|revcom
39. JB1174|tig08|339375|339058|revcom
40. JB1174|tig08|5123500|5123802
41. JB1174|tig32|637609|636951|revcom

Figure 12. Alignment of wtf4 BLASTn JB1174 hits with *wtf4* gene sequence – no *wtf4* homology from beginning of intron 1 to half-way through exon 3.

MUSCLE was used to align all JB1174 and JB22 BLASTn hits to the *wtf4* gene query sequence from the *S. pombe* standard reference genome. Results views in MEGA7.
(A) Alignment of the full exon 1-homologous JB1174 *wtf4*-BLASTn hits to the *wtf4* reference gene sequence (top sequence) drops off midway through intron 1 at CTGTT.
(B) Alignment of the full exon 1-homologous JB1174 *wtf4*-BLASTn hits to exon 3 of the *wtf4* reference gene sequence (top sequence) is shown. Homology re-starts half-way through exon 3.

Only six of the 11 exon 1-homologous BLASTn hits show homology to *wtf4* exon 4 (Figure-13a) and exon 5 (Figure-13b).

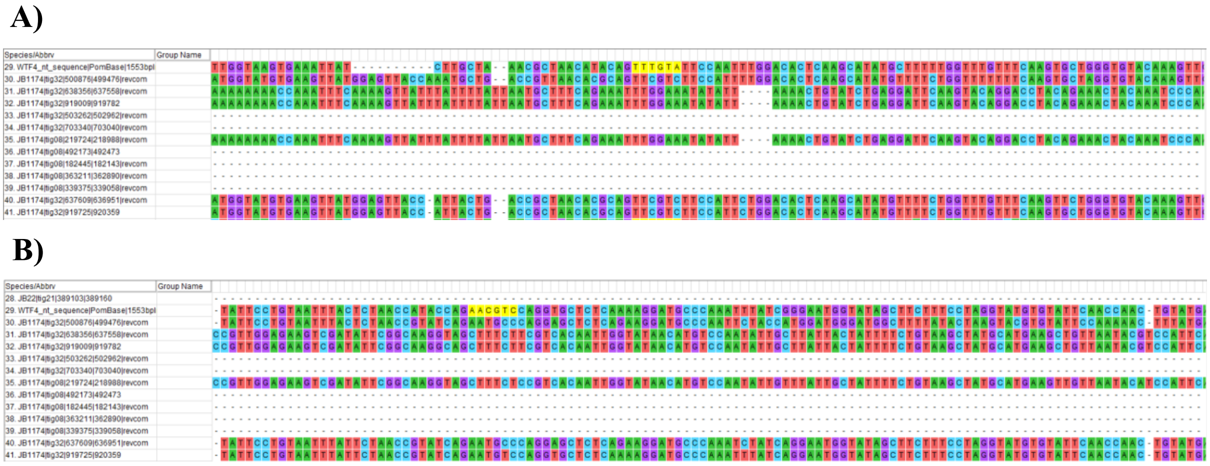
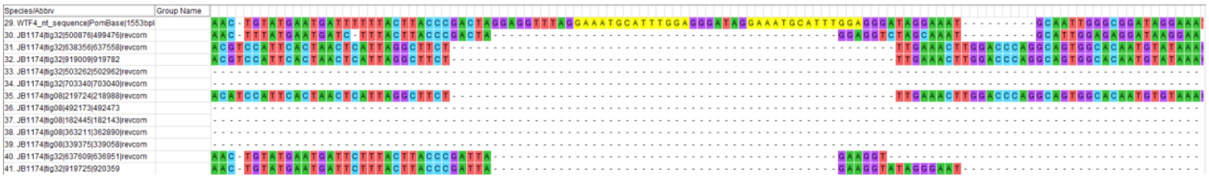

A)
Species/Abbrv
Group Name
29. WTF4_nt_sequence|PomBase|1553bpl
30. JB1174|tig32|500876|499476|revcom
31. JB1174|tig32|638356|637558|revcom
32. JB1174|tig32|919009|919782
33. JB1174|tig32|503262|502962|revcom
34. JB1174|tig32|703340|703040|revcom
35. JB1174|tig08|219724|218988|revcom
36. JB1174|tig08|492173|492473
37. JB1174|tig08|182445|182143|revcom
38. JB1174|tig08|363211|362890|revcom
39. JB1174|tig08|339375|339058|revcom
40. JB1174|tig32|637609|636951|revcom
41. JB1174|tig32|919725|920359
B)
Species/Abbrv
Group Name
28. JB22|tig21|389103|389160
29. WTF4_nt_sequence|PomBase|1553bpl
30. JB1174|tig32|500876|499476|revcom
31. JB1174|tig32|638356|637558|revcom
32. JB1174|tig32|919009|919782
33. JB1174|tig32|503262|502962|revcom
34. JB1174|tig32|703340|703040|revcom
35. JB1174|tig08|219724|218988|revcom
36. JB1174|tig08|492173|492473
37. JB1174|tig08|182445|182143|revcom
38. JB1174|tig08|363211|362890|revcom
39. JB1174|tig08|339375|339058|revcom
40. JB1174|tig32|637609|636951|revcom
41. JB1174|tig32|919725|920359

Figure 13. Alignment of wtf4 BLASTn JB1174 hits with *wtf4* gene sequence – some hits show exon 4 and 5 homology.

MUSCLE was used to align all JB1174 and JB22 BLASTn hits to the *wtf4* gene query sequence from the *S. pombe* standard reference genome. Results views in MEGA7.
A) Alignment of the full exon 1-homologous JB1174 *wtf4*-BLASTn hits to the start of exon 4 of the *wtf4* reference gene sequence (top sequence) is shown. Highlighted yellow are the first 5 nucleotides of *wtf4* exon 4.
B) Alignment of the full exon 1-homologous JB1174 *wtf4*-BLASTn hits to the start of exon 5 of the *wtf4* reference gene sequence (top sequence) is shown. Highlighted yellow are the first 5 nucleotides of *wtf4* exon 4.

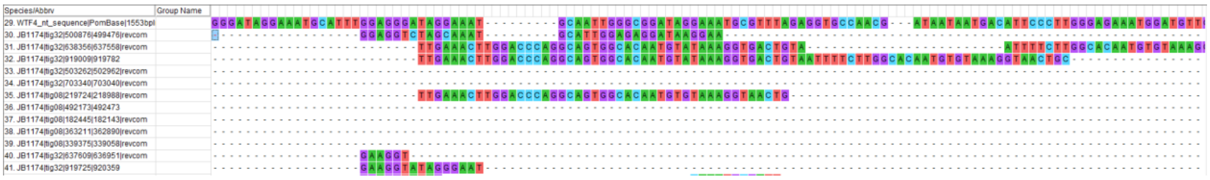
JB1174 BLASTn hits have no homology to the start, and patchy homology throughout, exon 6 (Figure-14a,b).

A)



```
Species/Abbrv                         Group Name
29. WTF4_nt_sequence|PomBase|1553bp|             AAC-TGTATGAATGATTTTTTACTTACCCGACTAGGAGGTTTAGGAAATGCATTTGGAGGGATAGGAAATGCATTTGGAGGGATAGGAAAT---------GCAATTGGGCGGATAGGAAA
30. JB1174|tig32|500876|499476|revcom            AAC-TTTATGAATGATC-TTTACTTACCCGACTA------------------------------------------GGAGGTCTAGCAAAT---------GCATTGGAGAGGATAAGGAA
31. JB1174|tig32|638356|637558|revcom            ACGTCCATTCACTAACTCATTAGGCTTCT------------------------------------------------------TTGAAACTTGGACCCAGGCAGTGGCACAATGTATAAA
32. JB1174|tig32|919009|919782                   ACGTCCATTCACTAACTCATTAGGCTTCT------------------------------------------------------TTGAAACTTGGACCCAGGCAGTGGCACAATGTATAAA
33. JB1174|tig32|503262|502962|revcom            ------------------------------------------------------------------------------------------------------------------------
34. JB1174|tig32|703340|703040|revcom            ------------------------------------------------------------------------------------------------------------------------
35. JB1174|tig08|219724|218988|revcom            ACATCCATTCACTAACTCATTAGGCTTCT------------------------------------------------------TTGAAACTTGGACCCAGGCAGTGGCACAATGTGTAAA
36. JB1174|tig08|492173|492473                   ------------------------------------------------------------------------------------------------------------------------
37. JB1174|tig08|182445|182143|revcom            ------------------------------------------------------------------------------------------------------------------------
38. JB1174|tig08|363211|362890|revcom            ------------------------------------------------------------------------------------------------------------------------
39. JB1174|tig08|339375|339058|revcom            ------------------------------------------------------------------------------------------------------------------------
40. JB1174|tig32|637609|636951|revcom            AAC-TGTATGAATGATTCTTTACTTACCCGATTA------------------------------------------GAAGGT--------------------------------------
41. JB1174|tig32|919725|920359                   AAC-TGTATGAATGATTCTTTACTTACCCGATTA------------------------------------------GAAGGTATAGGGAAT-----------------------------
```

B)



```
Species/Abbrv                         Group Name
29. WTF4_nt_sequence|PomBase|1553bp|             GGGATAGGAAATGCATTTGGAGGGATAGGAAAT---------GCAATTGGGCGGATAGGAAATGCGTTTAGAGGTGCCAACG---ATAATAATGACATTCCCTTGGGAGAAATGGATGTT
30. JB1174|tig32|500876|499476|revcom            ------------------GGAGGTCTAGCAAAT---------GCATTGGAGAGGATAAGGAA----------------------------------------------------------
31. JB1174|tig32|638356|637558|revcom            -------------------------TTGAAACTTGGACCCAGGCAGTGGCACAATGTATAAAGGTGACTGTA------------------------ATTTTCTTGGCACAATGTGTAAAG
32. JB1174|tig32|919009|919782                   -------------------------TTGAAACTTGGACCCAGGCAGTGGCACAATGTATAAAGGTGACTGTAATTTTCTTGGCACAATGTGTAAAGGTAACTGC----------------
33. JB1174|tig32|503262|502962|revcom            ------------------------------------------------------------------------------------------------------------------------
34. JB1174|tig32|703340|703040|revcom            ------------------------------------------------------------------------------------------------------------------------
35. JB1174|tig08|219724|218988|revcom            -------------------------TTGAAACTTGGACCCAGGCAGTGGCACAATGTGTAAAGGTAACTG--------------------------------------------------
36. JB1174|tig08|492173|492473                   ------------------------------------------------------------------------------------------------------------------------
37. JB1174|tig08|182445|182143|revcom            ------------------------------------------------------------------------------------------------------------------------
38. JB1174|tig08|363211|362890|revcom            ------------------------------------------------------------------------------------------------------------------------
39. JB1174|tig08|339375|339058|revcom            ------------------------------------------------------------------------------------------------------------------------
40. JB1174|tig32|637609|636951|revcom            ------------------GAAGGT------------------------------------------------------------------------------------------------
41. JB1174|tig32|919725|920359                   ------------------GAAGGTATAGGGAAT---------------------------------------------------------------------------------------
```

Figure 14. Alignment of wtf4 BLASTn JB1174 hits with *wtf4* gene sequence – patchy exon 6 homology.

MUSCLE was used to align all JB1174 and JB22 BLASTn hits to the *wtf4* gene query sequence from the *S. pombe* standard reference genome. Results views in MEGA7.
A) Alignment of the full exon 1-homologous JB1174 *wtf4*-BLASTn hits to the start of exon 6 of the *wtf4* reference gene sequence (top sequence) is shown. First yellow highlighted section is the start of exon 6.
B) Alignment of the full exon 1-homologous JB1174 *wtf4*-BLASTn hits to the middle of exon 6 of the *wtf4* reference gene sequence (top sequence) is shown.

None of the 11 exon 1-homologous sequences have a *wtf4*-homologous 3'UTR, however many other JB1174 BLASTn hits do (Figure-15).

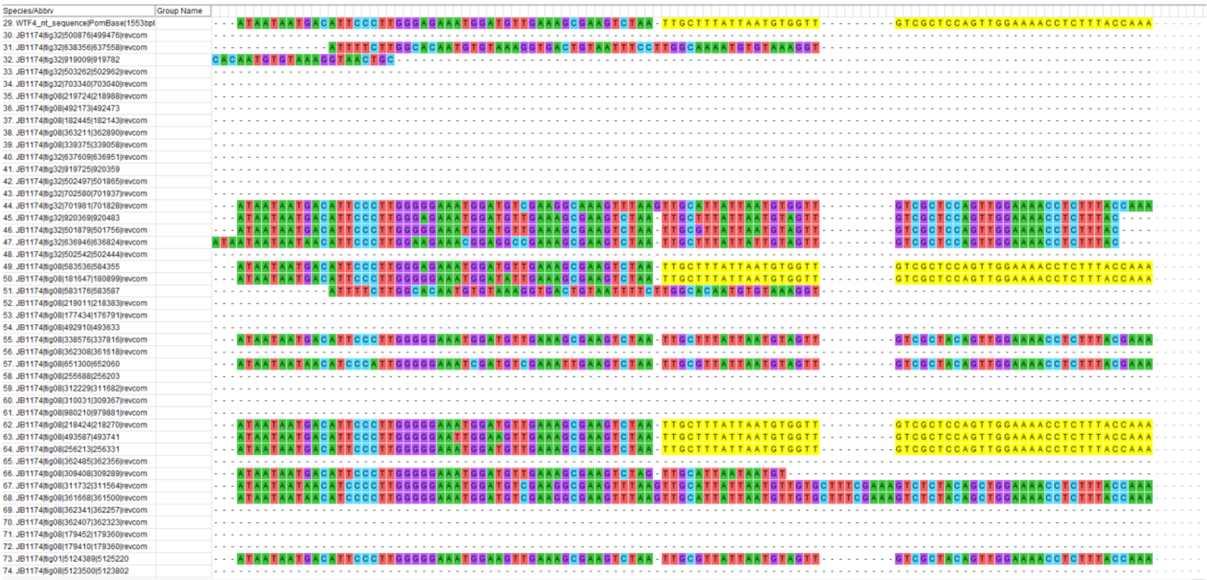


Figure 15. Alignment of wtf4 BLASTn JB1174 hits with *wtf4* gene sequence – 3'UTR homology only in sequences without exon 1-homology.

MUSCLE was used to align all JB1174 and JB22 BLASTn hits to the *wtf4* gene query sequence from the *S. pombe* standard reference genome. Results views in MEGA7. Alignment all JB1174 *wtf4*-BLASTn hits to the wtf4 3'UTR reference gene sequence (top sequence, highlighted yellow) is shown.

JB22 *wtf4*-BLASTn hits have a similar alignment pattern to wtf4 as JB1174, but only 4 hits have full exon 1-homology (Figure-16).

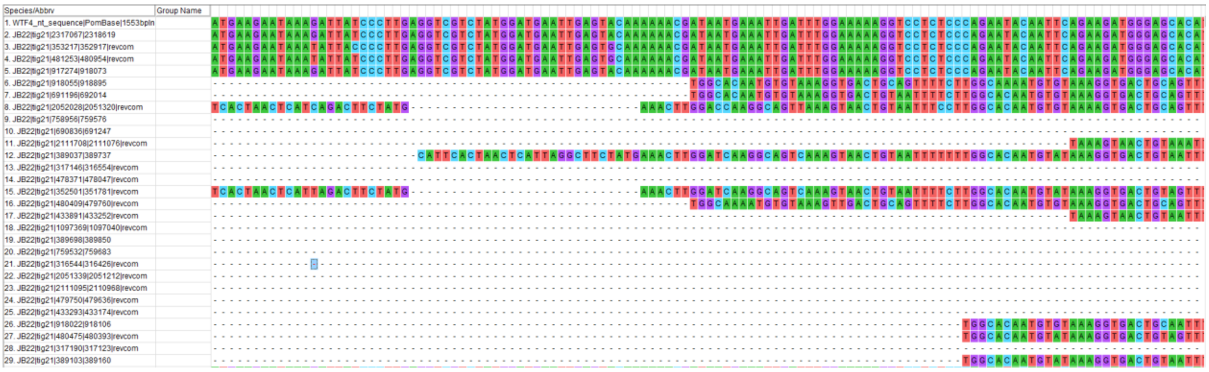


Figure 16. Alignment of *wtf4* BLASTn JB22 hits with *wtf4* reference gene sequence – homology at start of *wtf4* exon 1.

MUSCLE was used to align all JB1174 and JB22 BLASTn hits to the *wtf4* gene query sequence from the *S. pombe* standard reference genome. Results views in MEGA7. Alignment all JB22 *wtf4*-BLASTn hits to the start of *wtf4* exon 1 in the reference gene is shown.

Discussion

Use of JB22 Genome Assembly as a Reference

The 25 *wtf4*-BLASTn hits in JB22 agree with the 25 putative *wtf* genes identified in *S. pombe* (Nuckolls et al. 2017). Contig 21 likely aligns to chromosome III where all *S. pombe wtf* gene loci occur (Nuckolls et al. 2017). JB22 and JB1174 used Oxford Nanopore sequencing which has higher error rates than Illumina sequencing used for standard reference genomes (Laver et al. 2015).

BLASTn vs tBLASTn Hits

The greater number of tBLASTn hits compared to BLASTn hits is likely due to lack of a 75% identity threshold filter. These extra hits are below 100bp, have only 60% *wtf4* identity, and have similar start contig positions to BLASTn hits, so are likely alignments of single *wtf4*-homologous exons in the putative *wtf* genes.

JB22 vs JB1174 BLAST Hits

In contrast to JB22, not all JB1174 BLAST hits locate to the same contig. *wtf4* genes act in a genome-location-independent manner so this should not affect function (Hu et al. 2017). This could be due to *wtf* retro-transposition, which may explain the higher number of putative JB1174 *wtf* genes, or chromosomal rearrangement. Retro-transposition potential can be tested by searching for long terminal repeats flanking *wtf* genes. Given *wtf* gene rapid independent evolution, a more accurate JB1174 *wtf* gene number could be identified using every reference *wtf* gene sequence as a BLASTn query.

wtf4 start codon mutation causes retention of antidote function but loss of poison and therefore meiotic drive function (Nuckolls et al. 2017). JB1174 BLASTn hits without full exon 1-homology may therefore encode *wtf* genes which are no longer drivers. This is consistent with the range of *wtf* gene evolutionary stages in the reference genome. JB1174 *wtf* pseudogenes, formed by degeneration of antidote genes when protection is no longer beneficial, could be identified by BLASTp (Nuckolls et al. 2017). Therefore, there may only be 11 functional *wtf* meiotic driver genes in JB1174. Lack of exon 1 and 2 homology could also be due to rapid poison divergence in different *wtf* genes via C-terminal SNPs, without loss of the self-protecting antidote (Hu et al. 2017; Nuckolls et al. 2017). This considerable *wtf* gene sequence and number variation in JB22 and JB1174 was predicted due to *wtf* rapid divergence.

Word count = 999 words

References

BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279690/

Hu, W., Jiang, Z., Suo, F., Zheng, J., He, W. and Du, L. (2017). A large gene family in fission yeast encodes spore killers that subvert Mendel's law. *eLife*, 6, pp.1-19.

Laver, T., Harrison, J., O'Neill, P., Moore, K., Farbos, A., Paszkiewicz, K. and Studholme, D. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3, pp.1-8.

McLaughlin, R., Malik, S. (2017). Genetic conflicts: the usual suspects and beyond. Journal of Experimental Biology, 220, pp.6-17.

Nuckolls, N., Bravo Núñez, M., Eickbush, M., Young, J., Lange, J., Yu, J., Smith, G., Jaspersen, S., Malik, H. and Zanders, S. (2017). wtf genes are prolific dual poison-antidote meiotic drivers. *eLife*, 6, pp.1-3.

Pombase.org. (2017). *Pombase wtf4*. [online] Available at: https://www.pombase.org/spombe/result/SPCC548.03c [Accessed 1 Nov. 2017].

Shropshire, J. and Rokas, A. (2017). The gene family that cheats Mendel. *eLife*, 6, pp.1-3.

Appendix

Code used for comparative genomics analysis in Linux and R

Prior to this Sequence Analysis module, I had never used Linux or R before and had no previous experience of coding.

```
#Retrieving the genome fasta files
mkdir wtf_data
cd wtf_data
ln -sf /biol/people/dj757/teaching/BIO00058M-data-analysis/wtf/ .
cp wtf/readme .
cp wtf/Spombegenome.fasta .
cp wtf/nanopore/JB1174.repolished.fasta.gz .
cp wtf/*/JB1174.* .
cp wtf/*/JB22.* .
gunzip *.gz
ls
# Downloaded files:
JB1174.primary-contigs.fa
JB1174.scaffolds.fa
JB1174.scaffolds.fa.stats2
JB1174.scaffolds.fa.stats1
JB1174.repolished.fasta
JB22.primary-contigs.fa
JB22.scaffolds.fa
```

```
JB22.scaffolds.fa.stats2
JB22.scaffolds.fa.stats1
JB22.repolished.fasta
Spombegenome.fasta
wtf4-SPCC548.03c-transcript-sequence.fasta
Readme
wtf
```

```
#Creating databases containing standard S. pombe reference, JB22 reference or JB1174 genomes
makeblastdb -in Spombegenome.fasta -parse_seqids -dbtype nucl
makeblastdb -in JB22.repolished.fasta -parse_seqids -dbtype nucl
makeblastdb -in JB1174.repolished.fasta -parse_seqids -dbtype nucl
```

```
#blastn: blast wtf4 nucleotide sequence (including introns and 3’UTR) against each database containing the S. pombe, JB22 or JB1174 genome sequence, producing blast output in tab delimited form.
#Filter results to show only those with a Identity >75% and an E-value >10^6.
blastn -query wtf4_sequence_file.fasta -out blastoutput.fasta.blast -db database.fasta -evalue 1e-6 -perc_identity 75 -outfmt 6
```

```
#tblastn: query wtf4 transcript sequence (excluding introns and 3’UTR) against each database containing the S. pombe, JB22 or JB1174 genome sequence, producing blast output in tab delimited form.
#Filter results to show only those with an E-value >10^6.
tblastn -query wtf4_protein_sequence_file.fasta -out blastoutput.fasta.blast -db database.fasta -evalue 1e-6 -outfmt 6
```

```
#Open R in Linux
#importing blast results tables into R
R-table_name = read.table("blastoutput.fasta.blast")
```

```
#creating headers for R-tables
headers = c(
"seqname",
"chromnum",
"pident",
"qlen",
"mismatch",
"gapopen",
"qstart",
"qend",
"sstart",
"send",
"evalue",
"bitscore")
names(R-table_name) = headers


#view contig numbers
summary (R-table_name)


#Creating boxplots to compare blast results of the S. pombe, JB22 and JB1174 genomes.
wilcox.test(R-table_name1$pident,R-table_name2$pident)
testpid = wilcox.test(R-table_name1$pident, R-table_name2$pident)
ppid = round(testpid$p.value,digits=3)
wilcox.test(R-table_name1$qlen, R-table_name2$qlen)
testlength = wilcox.test(R-table_name1$qlen, R-table_name2$qlen)
plength = round(testlength$p.value,digits=3)
numhits <- c(nrow(R-table_name1), nrow(R-table_name2))
pdf("R-plot_name.pdf")
par(mfrow=c(2,2))
boxplot(
	R-table_name1$pident,
	R-table_name2$pident,
	ylab="Identity (%)",
	xlab="Genome Assembly",
	names=c("Strain 1","Strain 2"),
	main=paste("P=",ppid)
)
boxplot(
	R-table_name1$qlen,
	R-table_name2$qlen,
	ylab="Length (bp)",
	xlab="Genome Assembly",
	names=c("Strain 1","Strain 2"),
	main=paste("P=",plength)
)
barplot(
	numhits,
	ylab="Number of Blast Hits",
	xlab="Genome Assembly",
	ylim = c(0, 50),
	names.arg = c("Strain 1","Strain 2")
)
dev.off()
```

```
#view contig numbers containing blast results
summary(R-table_name1)


#create data subsets each containing blast hits on one contig
subset_name = subset(R-table_name, chromnum == "contig_name")


#exit R, back in Linux
#measuring contig length to determine length of x-axis in R-
plots.
bioawk -c fastx '{ print $name, length($seq)
}'<JB22.repolished.fasta
bioawk -c fastx '{ print $name, length($seq)
}'<JB1174.repolished.fasta
bioawk -c fastx '{ print $name, length($seq)
}'<Spombegenome.fasta


#Create R-plots of blastn and tblastn result properties
against contig position of blast result.
#plot length of blast hits against contig position for one
strain.
pdf("R-plot_name.pdf")
par(mfrow=c(2,2))
plot(
	subset_name1$sstart,
	subset_name1$qlen,
	ty="h",
	xlim=c(1,contig_length),
	ylim=c(1,400),
	ylab="Length (bp)",
	xlab="Position on the Contig",
	main="Strain and contig name"
)
plot(
	subset_name2$sstart,
	subset_name2$qlen,
	ty="h",
	xlim=c(1,contig_length),
	ylim=c(1,400),
	ylab="Length (bp)",
	xlab="Position on the Contig",
	main="Strain and contig name"
)
plot(
	subset_name3$sstart,
	subset_name3$qlen,
	ty="h",
        xlim=c(1,contig_length),
        ylim=c(1,400),
        ylab="Length (bp)",
        xlab="Position on the Contig",
        main="Strain and contig name"
)
plot(
        subset_name4$sstart,
        subset_name4$qlen,
        ty="h",
        xlim=c(1,contig_length),
        ylim=c(1,400),
        ylab="Length (bp)",
        xlab="Position on the Contig",
        main="Strain and contig name"
)
dev.off()


# plot Identity (%) of blast hits against contig position for one strain.
pdf("R-plot_name.pdf")
par(mfrow=c(2,2))
plot(
        subset_name1$sstart,
        subset_name1$pident,
        ty="h",
        xlim=c(1,contig_length),
        ylim=c(1,100),
        ylab="Identity (%)",
        xlab="Position on the Contig",
        main="Strain and Contig name"
)
plot(
        subset_name2$sstart,
        subset_name2$pident,
        ty="h",
        xlim=c(1,contig_length),
        ylim=c(1,100),
        ylab="Identity (%)",
        xlab="Position on the Contig",
        main="Strain and Contig name"
)
plot(
        subset_name3$sstart,
        subset_name3$pident,
        ty="h",
        xlim=c(1,contig_length),
        ylim=c(1,100),
        ylab="Identity (%)",
        xlab="Position on the Contig",
	main="Strain and Contig name"
)
plot(
	subset_name4$sstart,
	subset_name4$pident,
	ty="h",
	xlim=c(1,contig_length),
	ylim=c(1,100),
	ylab="Identity (%)",
	xlab="Position on the Contig",
	main="Strain and Contig name"
)
dev.off()


#plot -log10(E-value) of blast hits against contig position for one strain.
pdf("R-plot_name.pdf")
par(mfrow=c(2,2))
plot(
	subset_name1$sstart,
	-log10(subset_name1$evalue),
	ty="h",
	xlim=c(1,contig_length),
	ylim=c(1,160),
	ylab="E-value (-log10)",
	xlab="Position on the Contig",
	main="Strain and Contig name"
)
plot(
	subset_name2$sstart,
	-log10(subset_name2$evalue),
	ty="h",
	xlim=c(1,contig_length),
	ylim=c(1,160),
	ylab="E-value (-log10)",
	xlab="Position on the Contig",
	main="Strain and Contig name"
)
plot(
	subset_name3$sstart,
	-log10(subset_name3$evalue),
	ty="h",
	xlim=c(1,contig_length),
	ylim=c(1,160),
	ylab="E-value (-log10)",
	xlab="Position on the Contig",
	main="Strain and Contig name"
)
plot(
	subset_name4$sstart,
        -log10(subset_name4$evalue),
        ty="h",
        xlim=c(1,contig_length),
        ylim=c(1,160),
        ylab="E-value (-log10)",
        xlab="Position on the Contig",
        main="Strain and Contig name"
)
dev.off()
```

#Making the blast file to use in perl script – needs to show different columns in results table.

```
blastn -query wtf4_sequence_file.fasta -out
blastoutput.fasta.blast.txt -db database.fasta -evalue 1e-6 -
perc_identity 75 -outfmt "6 qseqid sseqid pident qlen length
mismatch gapopen qstart qend sstart send evalue bitscore
sstrand"
```

#Running the perl script to output blast hits in FASTA format.

```
perl coords-to-fasta4.pl --bla blastoutput.fasta.blast.txt --
fasta Genome used in blast, eg. JB22.repolished.fasta >
output_file.fasta
```

#combining fasta files of all blast results from all genomes

```
cat output_file_1.fasta output_file_2.fasta >
combined_output.fasta
```