The FASTA and the Furious: *wtf* is going on?

Summary

In 2002, the sequenced genome of the fission yeast *Schizosaccharomyces pombe* was published (Wood et al., 2002). Most fission yeast research since has focused on this strain, but *S. pombe* is a widely distributed organism, with many different strains isolated.

Since 2002, sequencing technologies have advanced and it is now possible to quickly sequence and assemble genomes. Nanopore sequencing (Jain et al., 2016) is a recently implemented technology that allows for rapid generation of long sequence reads. Although high quality genome sequences with Nanopore can be achieved (Loman et al., 2015), it is less accurate than other, more developed techniques such as ILLUMINA (Jain et al., 2016). Recent work has sequenced other strains (Jeffares et al., 2015) to understand more about the evolutionary background of this model organism.

The *wtf* gene family are a family of 25 genes in *S. pombe,* some of which are involved in meiotic drive by killing spores not inheriting the driver genes (Hu et al., 2017; Nuckolls et al., 2017). As few meiotic drivers are known, characterisation and understanding of *wtf* genes in such a tractable model organism informs study of gene transmission that disobeys Mendel's laws.

This study evaluates the comparability of results between the original Sanger sequenced reference genome and a Nanopore assembly (JB22) of the reference strain's genome, and compares JB22 to a different Nanopore sequenced strain (JB873). As the reference genome and JB22 are the same strain, results should be identical. As JB873 was isolated from a different location (Poland; JB22 was isolated in France: Jeffares et al., 2015) and the *wtf* gene family is rapidly evolving (Nuckolls et al., 2017), more divergence is expected.

Methods

BLAST+ (Camacho et al., 2009) was used to run nucleotide BLAST searches against the *S. pombe* standard reference genome (REF), and the new Nanopore assemblies of the reference strain (JB22) and JB873. As a query, *wtf*13 was selected as it had the longest exon sequence of the active genes predicted to be involved in meiotic drive (Nuckolls et al., 2017). The *wtf*13 query was downloaded (including introns) from PomBase.org (Wood et al., 2012).

Overlapping and very closely situated hits (within 12 bases) were merged using BEDtools (Quinlan and Hall, 2010) and sequences were extracted from the genome based on the start

and end positions of hits. A merge distance of 12 was chosen as there were some hits in the blast outputs that were close together on REF, but did not cause merging across *wtf* gene boundaries.

Gene names were assigned to hits on REF using known *wtf* gene locations (retrieved from PomBase.org). Extracted FASTA sequence files were concatenated (REF+JB22 and JB22+JB873) and alignment was performed using MUSCLE (Edgar, 2004). As this is preliminary work, and there is evidence that alignment filtering may worsen phylogenies (Tan et al., 2015), it was decided not to filter alignments. Midpoint-rooted maximum-likelihood trees were built in MEGA7 (Kumar et al., 2016) with 1000 bootstrap replicates. Trees were imported into R (R Core Team, 2017) and visualised using ggtree (Yu et al., 2017).

Linux and R code is included with this submission.

Results

Following processing, 13 of the 25 *wtf* genes were recovered from REF (table 1). *wtf*19 and *wtf*23 are represented by two hits each in table 1. Table 2 shows the hits obtained from JB22 and has the same number of hits, identical e-values and strand patterning as REF. There were 24 hits recovered from JB873 (table 3).

Table 1 - Summary of processed hits from a BLAST search of the *Schizosaccharomyces pombe* standard reference strain using *wtf*13 as a query. Following comparison to known *wtf* gene locations on the reference genome, the corresponding hit for each gene was appended for meaningful comparison to JB22. *wtf*19 and *wtf*23 appear twice.

Hit	S.Start	S.End	E.Value	Strand	Gene
REF:chr3:*wtf*4	221199	222701	0	(-)	*wtf*4
REF:chr3:*wtf*8	428073	428763	0	(+)	*wtf*8
REF:chr3:*wtf*9	487796	488519	0	(+)	*wtf*9
REF:chr3:*wtf*12	1401922	1402251	$3e^{-95}$	(+)	*wtf*12
REF:chr3:*wtf*13	1580123	1581694	0	(-)	*wtf*13
REF:chr3:*wtf*16	1739338	1740019	0	(-)	*wtf*16
REF:chr3:*wtf*18	1807004	1808136	0	(-)	*wtf*18
REF:chr3:*wtf*19	2017717	2018016	$2e^{-142}$	(+)	*wtf*19
REF:chr3:*wtf*19	2018495	2019288	$3e^{-164}$	(+)	*wtf*19
REF:chr3:*wtf*20	2020599	2020923	$9e^{-155}$	(+)	*wtf*20
REF:chr3:*wtf*21	2065079	2065780	$1e^{-99}$	(+)	*wtf*21
REF:chr3:*wtf*22	2109178	2109947	$1e^{-177}$	(-)	*wtf*22
REF:chr3:*wtf*23	2145761	2146061	$4e^{-143}$	(+)	*wtf*23
REF:chr3:*wtf*23	2146521	2147197	$1e^{-123}$	(+)	*wtf*23
REF:chr3:*wtf*24	2181784	2182504	$9e^{-175}$	(+)	*wtf*24

Table 2 - Summary of processed hits from a BLAST search of the *Schizosaccharomyces pombe* newly assembled reference genome (JB22) using *wtf*13 as a query.

Hit	S.Start	S.End	E.Value	Strand
JB22:tig21:H1	316476	317196	$9e^{-175}$	(-)
JB22:tig21:H2	351781	352457	$1e^{-123}$	(-)
JB22:tig21:H3	352917	353217	$4e^{-143}$	(-)
JB22:tig21:H4	389031	389800	$2e^{-177}$	(+)
JB22:tig21:H5	433190	433891	$1e^{-99}$	(-)
JB22:tig21:H6	478047	478371	$1e^{-154}$	(-)
JB22:tig21:H7	479682	480475	$3e^{-164}$	(-)
JB22:tig21:H8	480954	481253	$2e^{-142}$	(-)
JB22:tig21:H9	690836	691968	0	(+)
JB22:tig21:H10	758956	759637	0	(+)
JB22:tig21:H11	917274	918845	0	(+)
JB22:tig21:H12	1097040	1097369	$3e^{-95}$	(-)
JB22:tig21:H13	2051262	2051985	0	(-)
JB22:tig21:H14	2111018	2111708	0	(-)
JB22:tig21:H15	2317067	2318569	0	(+)

Table 3 - Summary of processed hits from a BLAST search of the *Schizosaccharomyces pombe* JB873 strain using *wtf*13 as a query.

Hit	S.Start	S.End	E.Value	Strand
JB873:tig32:H1	218905	220332	0	(-)
JB873:tig32:H2	241425	241744	$2e^{-122}$	(+)
JB873:tig34:H3	11690	12016	$2e^{-131}$	(+)
JB873:tig34:H4	12421	13229	$1e^{-138}$	(+)
JB873:tig34:H5	205652	205952	$1e^{-139}$	(+)
JB873:tig34:H6	206412	207113	$6e^{-107}$	(+)
JB873:tig34:H7	265698	266000	$2e^{-141}$	(+)
JB873:tig34:H8	266452	267180	0	(+)
JB873:tig34:H9	396679	396979	$1e^{-144}$	(+)
JB873:tig34:H10	397393	398139	$1e^{-113}$	(+)
JB873:tig34:H11	399065	400465	0	(+)
JB873:tig34:H12	1501076	1501696	0	(-)
JB873:tig34:H13	1568799	1570336	0	(-)
JB873:tig34:H14	1661632	1662366	$1e^{-148}$	(-)
JB873:tig34:H15	1662850	1663150	$1e^{-144}$	(-)
JB873:tig34:H16	1787195	1787513	$2e^{-137}$	(+)
JB873:tig34:H17	1787915	1788846	$4e^{-168}$	(+)
JB873:tig34:H18	1811616	1812469	$7e^{-176}$	(+)
JB873:tig34:H19	1837481	1837799	$4e^{-144}$	(+)
JB873:tig34:H20	1838228	1839068	0	(+)
JB873:tig34:H21	1882275	1883050	0	(-)
JB873:tig34:H22	1883481	1883781	$2e^{-141}$	(-)
JB873:tig34:H23	1923369	1924083	0	(+)
JB873:tig1639:H24	387053	387454	0	(+)

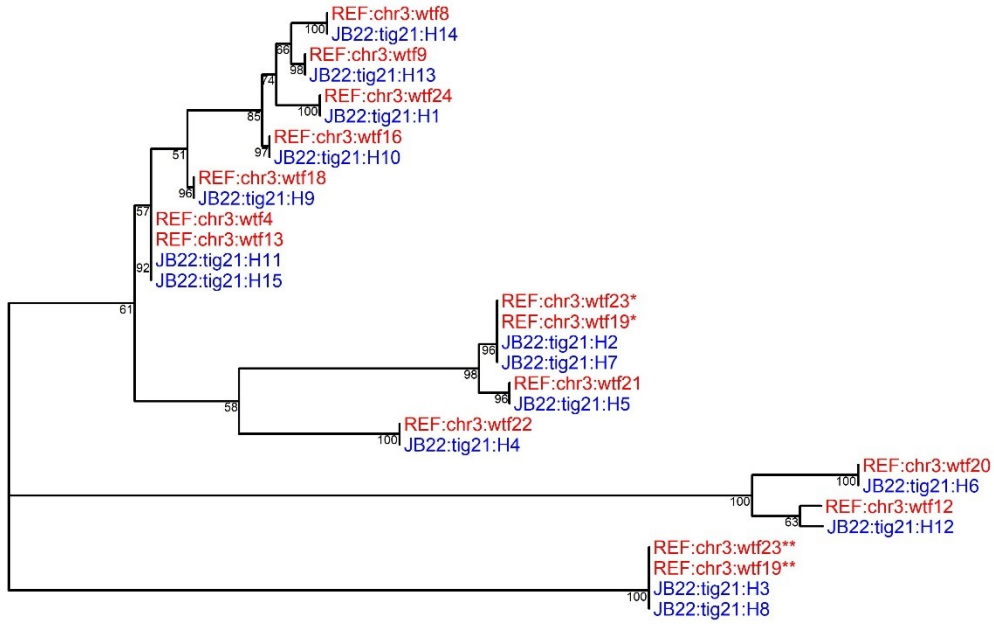

Figure 1 - Maximum likelihood tree produced using MEGA7 (Kumar et al., 2016) with 1000 bootstrap replicates and midpoint rooting. Red tips are blast hits from the standard reference genome for *Schizosaccharomyces pombe* and blue tips are hits from the Nanopore reassembly of the *S. pombe* reference strain (JB22). Bootstrap support values below 50 are omitted.

The gene tree in figure 1 pairs most of the *wtf* genes with a hit from the JB22 genome with high bootstrap support, as would be expected in two assemblies of the same genome. Three regions of the tree are not well resolved to individual pairs, corresponding to *wtf*4 and *wtf*13, and the two regions covering *wtf*19 and *wtf*23. In these regions, pairs from REF are sisters to pairs from JB22 but bootstrap support is low.

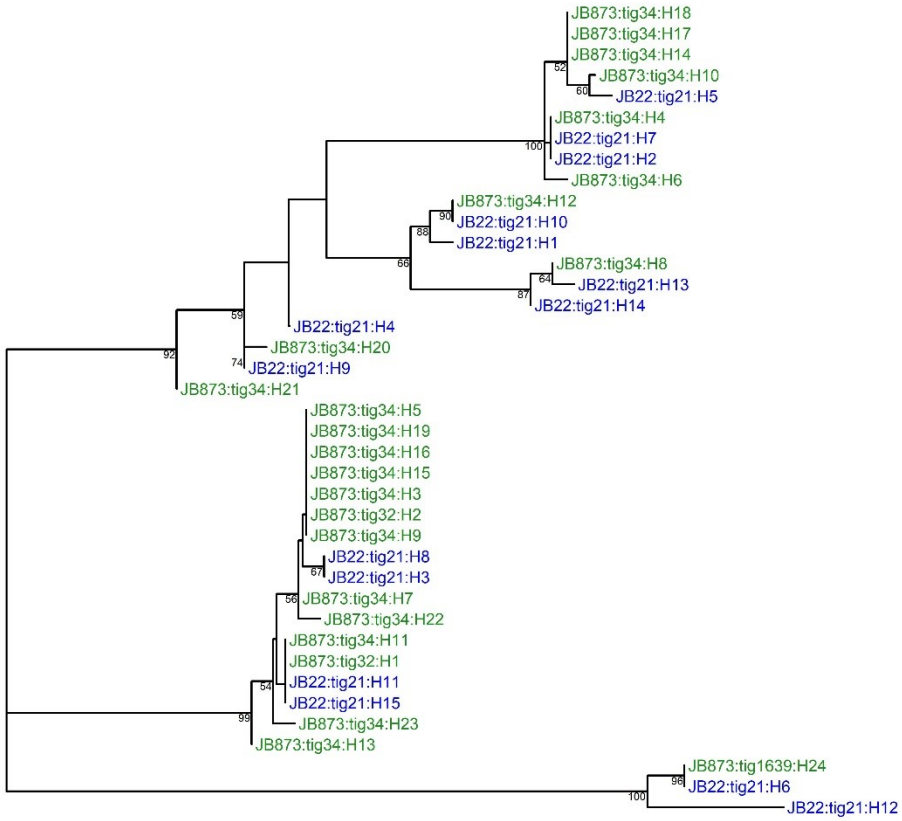

Figure 2 – Maximum likelihood tree produced using MEGA7 (Kumar et al., 2016) with 1000 bootstrap replicates and midpoint rooting. Blue tips are blast hits from the JB22 Nanopore reassembly of the *Schizosaccharomyces pombe* reference genome and green tips are from the JB873 Nanopore assembly. Bootstrap support values below 50 are omitted.

In figure 1, JB22 H11 and H15 are paired to *wtf*4 and *wtf*13 with high bootstrap support which suggests that H11 and H15 may represent these genes. Though still paired, bootstrap support for the relationship between H11 and H15 is low in figure 2. JB22 H3 and H8 are grouped with portions of *wtf*19 and *wtf*23 in figure 1. In figure 2, JB22 H3 and H8 are paired, and are sister to a group of seven hits from JB873, though this pairing is not well supported. There are only 5 clear pairings of JB22 and JB873 hits in figure 2 (table 4).

Table 4 – Well-supported blast hit pairings between JB22 and JB873, with bootstrap support from the maximum-likelihood tree (fig 2). Inferred genes are the genes on REF that were paired to hits on JB22 (fig 1).

JB22 hit	Inferred gene	JB873 hit	Bootstrap support (%)
H6	*wtf*21	H24	96
H9	*wtf*18	H20	74
H13	*wtf*9	H8	64
H10	*wtf*16	H12	90
H5	*wtf*20	H10	60

Conclusions

Only 13 of the 25 different *wtf* genes known to be in the reference genome were recovered in REF and JB22, suggesting that *wtf*13 may not be the most effective query for picking up these genes. The genes picked up by the *wtf*13 query were all clustered together as 6-exon genes by Hu et al. (2017). *wtf*19 and *wtf*23 are represented twice in REF, and though these two genes pair together, the pairs are not closely related in figure 1 as you would not expect two regions of the same gene to appear closely related to each other.

It is interesting that in the JB873 strain (figure 2), the JB22 hits corresponding to *wtf*4, *wtf*13 and one of the *wtf*19-*wtf*23 pairings (inferred from comparison to REF in fig.1) are all clustered together. These are the four genes predicted to be meiotic drive genes predicting to encode both poison and antidote (Hu et al., 2017). Within this cluster there are many more hits from the JB873 strain, which may represent a rapid proliferation of these drive genes, or the regions of *wtf*19 and *wtf*23 to which they are paired.

As meiotic driver *wtf* genes appear to code poison and antidote with an alternative transcription system (Nuckolls et al., 2017) and are able to lose the antidote protection (Hu et al., 2017), further investigation may find that these shorter hits (JB873: H2, H3, H5, H9, H15, H16, H19) are a proliferation of poison coding regions. Findings such as these support the work of Jeffares et al. (2015), suggesting that broadening understanding of the genetic diversity in *S. pombe* can help to better elucidate unusual evolutionary processes such as meiotic drive.

(*954 words*)

References

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+ : architecture and applications. *Bmc Bioinformatics,* 10.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32**,** 1792-1797.

Hu, W., Jiang, Z. D., Suo, F., Zheng, J. X., He, W. Z. & Du, L. L. 2017. A large gene family in fission yeast encodes spore killers that subvert Mendel's law. *Elife,* 6.

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology,* 17**,** 11.

Jeffares, D. C., Rallis, C., Rieux, A., Speed, D., Prevorovsky, M., Mourier, T., Marsellach, F. X., Iqbal, Z., Lau, W., Cheng, T. M. K., Pracana, R., Mulleder, M., Lawson, J. L. D., Chessel, A., Bala, S., Hellenthal, G., O'fallon, B., Keane, T., Simpson, J. T., Bischof, L., Tomiczek, B., Bitton, D. A., Sideri, T., Codlin, S., Hellberg, J., Van Trigt, L., Jeffery, L., Li, J. J., Atkinson, S., Thodberg, M., Febrer, M., Mclay, K., Drou, N., Brown, W., Hayles, J., Salas, R. E. C., Ralser, M., Maniatis, N., Balding, D. J., Balloux, F., Durbin, R. & Bahler, J. 2015. The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nature Genetics,* 47**,** 235-+.

Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution,* 33**,** 1870-1874.

Loman, N. J., Quick, J. & Simpson, J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods,* 12**,** 733-U51.

Nuckolls, N. L., Nunez, M. A. B., Eickbush, M. T., Young, J. M., Lange, J. J., Yu, J. S., Smith, G. R., Jaspersen, S. L., Malik, H. S. & Zanders, S. E. 2017. *wtf* genes are prolific dual poison-antidote meiotic drivers. *Elife,* 6.

Quinlan, A. R. & Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M. & Dessimoz, C. 2015. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology,* 64**,** 778-791.

R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., Mcdonald, S., Mclean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature,* 415**,** 871-880.

Wood, V., Harris, M. A., Mcdowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., Aslett, M., Lock, A., Bahler, J., Kersey, P. J. & Oliver, S. G. 2012. PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Research,* 40**,** D695-D699.

Yu, G. C., Smith, D. K., Zhu, H. C., Guan, Y. & Lam, T. T. Y. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution,* 8**,** 28-36.