bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Article

TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes

Maria Waldl [1,†], Bernhard C. Thiel [1,†], Roman Ochsenreiter [1], Alexander Holzenleiter [2,3], João Victor de Araujo Oliveira [4], Maria Emília M. T. Walter [4], Michael T. Wolfinger [1,5*], Peter F. Stadler [6,7,1,8*]

1 Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; {maria,thiel,romanoch}@tbi.univie.ac.at, michael.wolfinger@univie.ac.at
2 BioInformatics Group, Fakultät CB Hochschule Mittweida, Technikumplatz 17, D-09648 Mittweida, Germany; alexander.holzenleiter@web.de
3 Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
4 Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília; joaovicers@gmail.com, mariaemilia@unb.br
5 Center for Anatomy and Cell Biology, Medical University of Vienna, Währingerstraße 13, 1090 Vienna, Austria
6 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, University Leipzig, Germany
7 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
8 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501
* Correspondence: MTW michael.wolfinger@univie.ac.at; PFS studla@bioinf.uni-leipzig.de
† These authors contributed equally to this work.

Version May 15, 2018 submitted to Genes

1 **Abstract:** The telomerase RNA in yeasts is large, usually $> 1,000$ nt, and contains functional elements
2 that have been extensively studied experimentally in several disparate species. Nevertheless, they
3 are very difficult to detect by homology-based methods and so far have escaped annotation in the
4 majority of the genomes of Saccharomycotina. This is a consequence of sequences that evolve rapidly
5 at nucleotide level, are subject to large variations in size, and are highly plastic with respect to
6 their secondary structures. Here we report on a survey that was aimed at closing this gap in RNA
7 annotation. Despite considerable efforts and the combination of a variety of different methods, it
8 was only partially successful. While 27 new telomerase RNAs were identified, we had to restrict our
9 efforts to the subgroup *Saccharomycetacea* because even this narrow subgroup was diverse enough to
10 require different search models for different phylogenetic subgroups. More distant branches of the
11 Saccharomycotina still remain without annotated telomerase RNA.

12 **Keywords:** non-coding RNA; telomerase RNA; secondary structure; synteny; homology search; yeast

13 1. Introduction

14 The linear chromosomes of eukaryotes require a specialized mechanism for completing
15 duplication. Most commonly this is achieved by a special reverse transcriptase, telomerase, that
16 carries a specific RNA the template with telomeric sequence [1]. Most likely, this constitutes the
17 ancestral state in eukaryotes. Despite its crucial function, telomerase has been lost several times in both
18 animals (in particular insects) and possibly also in some plants [2]. In some cases, the ancestral telomere
19 structure has been replaced by tandem arrays of DNA sequences that look much like heterochromatin

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

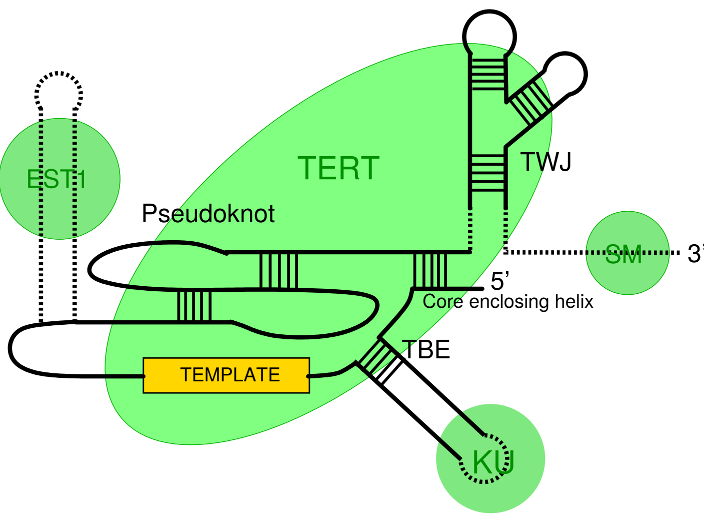

Figure 1. Schematic organization of TER. Contact regions for important binding sites are indicated by green circles (EST1, SM, KU). The green ellipse denotes the contact region with the reverse transcriptase (TERT). Other major features are the template, the pseudoknot region, the template boundary element (TBE) and the three-way junction (TWJ). Adapted from [8].

and can be elongated by gene conversion. Specialized telomere-specific retrotransposons are at work in Drosophila [3].

The telomerase (holo)enzyme consists of two main components, a specialized reverse transcriptase (TERT) and a RNA component (TER) that provides the template sequence. In addition, there are usually multiple clade-specific accessory protein components [4]. Four conserved regions in TER, Fig. 1, are essential for telomerase activity: the template boundary element (TBE), the pseudoknot, and the template sequence itself are part of the the catalytic core. The fourth region, the trans activating domain, is involved in binding of TERT [5]. The three-way junction (TWJ) structure of this region region is widely conserved at least between animal and fungal telomerase RNAs, where it is crucial for proper functioning [6]. The precisely defined template within TER is processively copied by TERT and regenerated, releasing a single-stranded DNA product [7].

Telomerase RNA is highly divergent. The TER in ciliates [9], human [10], and budding yeast [11] have a length of about 150 nt, 438 nt, and ∼1.3 kb, respectively. A TER more than 2kb in length has been reported for *Candida glabrata* [12], which, interestingly, seems to lack a TWJ. TERs in other kingdoms of eukaryotes have been discovered only quite recently in plants [13,14], excavates [15,16] and alveolates [17,18].

Despite their deeply conserved primary function and architectural similarities that seem to extend across eukaryotic kingdoms, TERs have turned out be very difficult to find by homology search even within phylogenetically relatively narrow groups. Within the animal kingdom, even surveys of vertebrates turned out to be non-trivial [19]. Echinoderm TERs were found by deep sequencing of *Strongylocentrotus purpuratus* RNA pulled down with the TERT protein [20] after homology based searches remained unsuccessful. This opened the door to identifying TERs from other sea urchins, brittle stars, and a crinoid [21]. Still, no TER from a protostome is known.

Within Fungi, the situation is similar: So far, TERs have been reported only for Ascomycota, while no candiates are known in Basidiomycota and any of the basal divisions. The TERs of Pezizomycotina and Taphrinomycotina share core features of vertebrate TERs. In particular, they have a fairly well-conserved secondary structure of the pseudoknot and the TWJ, and at least in these regions the sequence is sufficiently conserved for successful homology-based identification of TERs within these clades [22–24]. The TERs known for Saccharomycetes, the relatives of budding yeast, on the

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

49 other hand, are sometimes remarkably large and present little similarity in sequence and secondary
50 structure to vertebrate or ciliate TERs.
51 To-date, yeast TERs have been reported for three phylogenetically narrow subgroups
52 (*Saccharomyces spp.*[11,25], *Kluyveromyces spp.*[6,26,27], and *Candida spp.*[28,29]), as well as some
53 individual species such as *Candida glabrata* [12] and *Hansenula polymorpha* [30]. These sequences
54 are already too diverse for reliable sequence alignments. It is not surprising, therefore, that simple
55 sequence-based homology searches have not been successful in identifying TER in the majority of the
56 saccharomycete genome sequences to-date. Even protein binding sites that are functionally important
57 in budding yeast [31] are not widely conserved. For instance, Ku or Sm binding sites seem to be absent
58 in the TERs of filamentous fungi [4,22].
59 The obvious alternative is to increase the set of known TERs by finding homologs that are
60 sufficiently similar to one of known yeast TERs, to allow the construction of multiple alignments of
61 phylogenetically narrow subgroup. From these alignments, conserved elements can be extracted,
62 which in turn form the basis for searches with tools such as `fragrep` [32] or `infernal` [33]. This
63 strategy has been successful in previous searches for TER genes in both animals [19] and fungi [22],
64 but so-far has not been successfully applied to Saccharomycetes.
65 Until very recently, a phylogenetically local approach to homology search was also hampered by
66 the lack of a trustworthy phylogeny of the Saccharomycotina. Recent updates in the International Code
67 of Nomenclature for algae, fungi and plants [34,35] have substantially restructured the classification of
68 fungi in general and of Saccharomycotina in particular. With large-scale efforts to sequence fungal
69 genomes underway, first phylogenomic studies provide a trustworthy backbone of Saccharomycotina
70 phylogeny [36], which we largely confirmed with an independent analysis.

71 2. Materials and Methods

72 *2.1. Phylogenomics of Ascomycotes*

73 Annotated protein sequences for 72 yeast species were downloaded from RefSeq. Initially,
74 `ProteinOrtho` [37,38] was used to identify an initial set of 21.289 ortholog groups. Only 193 of these
75 contained representatives of all 72 species. We therefore included all 1666 ortholog groups that covered
76 at least 67 species. We used `OMA` (2.2.1) [39,40] to decompose the `ProteinOrtho` groups further into
77 clusters of 1-1 orthologs. This resulted in 6.295 groups of which 841 contained at least 67 species. This
78 conservatively filtered data set was then processed with `Gblocks` [41] to remove uninformative and
79 potentially error-prone parts of the alignment, resulting in a data set comprising 72 species and 248,581
80 characters. Phylogenetic trees were estimated with `RAxML` [42].

81 *2.2. Ascomycote Telomerase RNAs*

82 Telomerase RNA regions have been published for several *Saccharomyces* [11,25], *Kluyveromyces*
83 [6,26,27], and *Candida* [12,28,29] species. Most of these published TER regions are collected in the
84 telomerase database [43], which therefore provided a good starting point for our research. These
85 sequences, however, are too diverse to construct multiple sequence alignments beyond the three
86 genera individually. This effectively prohibits the automated discovery of novel TERs beyond close
87 relatives with the help of either `blast` [44] (using sequence information alone) or `infernal` (relying on
88 a combination of sequence and secondary structure information).
89 Therefore, we explored different strategies to overcome the limitations imposed by the extremely
90 poor sequence conservation of saccaromycete telomerase RNAs. The basic idea is to use common
91 features of the TERs to extract candidates from the genomes that can be analyzed and then inspected
92 further using different techniques.
93 First, we attempted to learn TER-specific sequence patterns using `MEME/GLAM2` [45], and also
94 several machine learning techniques using *k*-mer distributions within sequence windows of the size

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the known TERs. All attempts to learn from a training set covering the *Saccharomycetaceae* or all *Saccharomycotina* species failed.

There are several possible reasons. Machine learning methods crucially depend on a training and test sets, both positive and negative. In our case we have few positive samples, these have poorly defined features, and are very diverse as far as their sequences are concerned. It is unclear in this setting how a negative training set should be properly designed. The obvious choice of picking genomic sequence at random may be confounded by unintended strong signals, such as coding potential or repetitive sequence elements. It would appear that at the very least a more a careful construction of the positive and negative sets, and an appropriate normalization or scaling of the feature data will be required to make progress in this direction. Restricting the training phase to a more narrow phylogenetic range to reduce the inherent diversity of the training data, on the other hand, is infeasible due to the small number of known TER sequences.

The `EDeN` motif finder [46] was applied to 24 known TERs as positive set and 48 shuffled sequences as negative data. Only trivial sequence motifs such a poly-U stretch, presumably corresponding to part of the U-rich pseudoknot region, were found. Unsupervised clustering also remained unsuccessful.

2.3. Synteny-Based Homology Search

As an alternative strategy, we established a semi-automated workflow that aims at first extracting partially conserved RNA sequence-structure elements, which are then used to identify candidate loci. In response to the negative results of a direct pattern-based approach, we systematically used synteny to narrow down the search space in the initial phase. Starting from a whole genome alignment of phylogenetically related species, we used the positions of protein coding genes whose homologs are known to be adjacent in a closely related species to delimit the syntenic regions that are likely to contain a TER gene. These candidate regions were then analyzed in detail by means of pairwise or multiple sequence alignments. Whenever a global alignment of the entire candidate syntenic region did not yield a plausible alignment, we attempted to identify conserved motifs inside the syntenic region (usually the SM binding site and/or the template region, which is sometimes conserved between close relatives). Typically, these motifs were also sufficient to determine the correct reading direction of the TER candidate.

To identify known features in the candidate TER regions, we first constructed `infernal` [33] covariance models restricted to subgroups of Saccharomycetaceae covering only substructures, such as the Ku hairpin, Est1 binding site, and TWJ in the Saccharomyces and Kluyveromyces species. The alignments underlying the `infernal` models were constructed with the help of many software tools, including `locarna` [47], `mafft` [48], `mauve` [49], `MEME` [45] and `fragrep` [32], as well as manual curation. These models were then used for precise localization of conserved TER elements in species that were (a) taxonomically closely related, but not/only partially annotated in literature (*Saccharomyces uvarum*, *Saccharomyces sp. 'boulardii'*, *Saccharomyces sp. M14*, *Saccharomyces eubayanus* or (b) phylogenetically located in the subtree spanned by the Saccharomyces and Kluyveromyces species (see Fig. 2). Both the `ViennaNGS` [50] suite and custom Perl/Python scripts were used for handling and conversion of genomic annotation data.

We then extracted a sequence corresponding to the most closely related TER sequence as initial estimate of the full-length TER gene. We used `mafft` [48] to produce initial sequence-based alignments of candidate regions, which were then realigned with `locarna` [47] to obtain RNA structural alignments. The latter was used with its free-end-gaps option, in particular in those cases where `mafft` was not sensitive enough to reliably estimate the TER boundaries. Conversely, `mafft` was able to identify and correctly align highly conserved subsequences, providing reliable anchors for the more divergent sequence regions. While `locarna` is good at finding locally conserved structures in the whole alignment, we expected only parts of the TER sequences to be structurally conserved. Typically multiple iterations of refinement of the TER boundaries were required to obtain the final TER candidate sequence.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

143 Following this approach, we could localize TER regions for several members of the
144 *Saccharomycetacea* clade. Subsequent alignment of candidate regions with known TERs allowed
145 for exact localization of TERs.

146 *2.4. Search for Candidates Using Telomere Template Sequences*

147 The scope of the synteny-based approach is limited because fungal genomes are subject to frequent
148 genome rearrangements at the time-scales of interest. We therefore attempted to identify candidate
149 regions containing the template sequence for the telomere repeats. (See [51] for a comprehensive
150 review of the characteristics of different telomeric repeats.) In genomes for which these sequences
151 have not been reported, we searched chromosome ends for telomeric repeats. Unfortunately, most
152 genome assemblies are not on chromosome level or do not include the telomere regions, hence we
153 only succeeded to newly identify the template region of *Ashbya aceri* and *Eremothecium cymbalariae* this
154 way. For the latter species, the pertinent information is available in [52], although the telomeric repeat
155 is not explicitly reported. In addition, we used the published telomere sequences from the telomerase
156 database [43].
157 We used the concatenation of two copies of telomeric repeat sequence as query for a `blast` [44]
158 search against the whole genome (in case of longer, complex repeats) or against the syntenic region for
159 shorter repeats. Other template regions were identified with by aligning them to known sequences
160 and/or `blast` searches of known template regions in closely related species. A typical feature of the
161 template region, which helped us to verify our hits, is the fact that it usually contains a few nucleotides
162 repeated at both the beginning and the end of the template region [12].

163 *2.5. `Blast` Pipeline*

164 `blast` [44] is by far the most commonly used tool for homology search. While it has been reported
165 to have limited sensitivity for telomerase RNAs in previous studies [19,20,32], it has contributed
166 significantly to the identification of the TER sequences in other ascomycete clades [22,24]. Here we
167 used a set of known TER regions as `blast` queries that comprises all Saccharomycetales TER regions
168 that we found in literature, as well as all TERs newly identified in the contribution. As targets for
169 `blastn` (with default parameters) we used the full genomes of species that are featured at the `NCBI`
170 `refseq` database within the Saccharomycetales group (Taxonomy ID: 4892). The resulting `blast` hits
171 were then filtered for E-values ($E < 0.1$), a minimum alignment length of 25nt and a minimum identity
172 of 60%. In addition, all hits on known telomeric regions were excluded. From the hits in genomes
173 with known TERs we computed the empirical false positive rate and found that the alignment length
174 proved to be the most informative parameter. It has therefore been used to evaluate the reliability of
175 hits, given their score.
176 The `blast` pipeline also contributed to the identification of the TER boundaries in some of the
177 unannotated genomes. In cases were we initially chose the boundaries of our queries too generously
178 and included neighboring coding regions or regulatory elements, the blast pipeline returned "false
179 positive" hits. Thus, whenever multiple false positive hits in the beginning or the end of the query
180 sequence occurred, we rechecked and, if necessary, improved the boundaries of the TER region.

181 **3. Results**

182 *3.1. Phylogenomics of Saccharomycotina*

183 The phylogenetic trees obtained of our phylogenomic analysis of the Saccharomycetales is
184 essentially congruent with the one reported by Shen *et al.*[36], see the Appendix for more details. For
185 consistency, we adopted the phylogenetic tree published by Shen *et al.*[36] as the basis for presenting
186 our results.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2. Survey of TER Genes in Saccharomycotina

We initially screened 52 ascomycote genomes. Predominantly sequence-based methods (`blast`, but also `meme`, `glam2`, and `infernal`) only contributed TERs from close relatives of baker's yeast. The `blast` pipeline was applied to all 185 NCBI genomes Saccharomycetales, the subclade containing all known Saccharomycotina genomes. With the exception of the TER in *Ogataea parapolymorpha*, a very close relative of the known *Ogataea polymorpha* TER [30] all new sequences we found within the Saccharomycetaceae. We therefore restricted a more detailed analysis to this clade.

We found credible TER sequences in 46 of the 53 Saccharomycetaceae. Most of these TER sequences could be detected only after a short candidate region had been identified based on synteny. To our knowledge, at least 27 of these have not been reported previously.

3.3. Features of TER in Saccharomycetacea

In order to better understand the TER and its evolutionary constraints at least within the Saccharomycetacea we performed a detailed analysis of their structural features. Table 1 summarizes the results of the homology search and the functional features of the candidate TER genes. A graphical overview is given in Fig. 2.

The exact genomic positions marking the 3′ and 5′ ends of the TER RNA are difficult to determine without additional experimental evidence. The 5′ ends are therefore approximate. The 3′ end of the mature TER is produced by splicing in most Ascomycota [24,59,60]. This mechanism, however, was lost at some point during the evolution of the Saccharomycotina. It has been reported in the Candida group and for *Ogataea angusta* (previously *Hansenula polymorpha*), but it is missing in Saccharomyces and Kluyveromyces [24]; hence we expect that the splicing-based 3′-end processing was lost prior to the divergence of Saccharomycetacea. Indeed, no indication of a splice site was found for any of the TER sequences included in Table 1. We therefore used a position 10 nt downstream of the SM binding motif as approximation of the 3′ end in Table 1.

Several of the features listed in Table 1 have been discussed in some detail in the literature. Not all of them were found in all the candidates reported here. This may, in some cases, be explained by sequences that are too divergent to be detected. In other cases, most likely the function is not preserved. Unfortunately, many studies report neither complete sequences nor coordinates, making it effectively impossible to accurately compare their results with the annotation reported here. References are included in Table 1 if sufficient information was included to locate the features unambiguously.

No Ku binding hairpin was recovered in *Kluyveromyces* or the *Eremothecium* species. This is not unexpected since there is experimental evidence that neither the Ku binding hairpin nor its function is present in *K. lactis* [53]. The putative Ku binding hairpin reported for *Candida glabrata* in [12] lacks experimental support and contains long insertions that made it impossible to include it in our covariance model. Furthermore, this region of the TER sequence is very poorly conserved in the closest relatives of *C. glabrata*. While the TER of *C. glabrata* is among the longest known members of this gene family [12], its close relative *C. castellii* features a TER that has been shortened drastically in its 3′ half, with only ~ 200 nt separating the EST1 and SM1 binding sites. Furthermore, the sequence `GCUA`, which is conserved in most known Ku binding sites, is not present within 600nts upstream of the template region. The most likely explanation is that the TER of *Candida castellii* (which like *Candida glabrata* does not belong to the monophylogenetic genus *Candida*, see Appendix) does not bind Ku. Of course, we cannot rule out without further experimental data that the motif has diverged beyond our ability to recognize it.

In a few species we failed to identify the template region. In these cases (*Lachancea*, *Zygosaccharomyces* and *Torulaspora* species and *Nakaseomyces bacillisporus*) the telomeric repeat sequence is not known and seems to be very different from both the fungal consensus sequence `TTAGGG` [22] and the telomeric sequences found in closely related species.

The EST1 binding site could not be identified in *Eremothecium* species, *Lachancea dasiensis* and in the *Candida glabrata* group, even though it has been published for *Candida glabrata*. While an EST1

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Species	Accession	Strand	TER coordinates	Ku binding site	Template region	Est1 binding site	TWJ	SM1
K. aestuarii	AEAS01000245.1	neg	16338-17322 [26]		16940-16966	16794-16862 [26]	16378-16485 [6]	16350-16359
K. wickerhamii	AEAV01000432.1	pos	250-1327 [26]		662-693	765-858 [26]	1183-1290 [6]	1307-1316
K. marxianus	NC_036029.1	pos	506443-507711 [26]		506855-506888	506967-507049 [26]	507518-507671 [6]	507691-507700
K. dobzhanskii	CCBQ010000012.1	pos	461805-463090 [26]		462224-462257	462337-462499 [26]	462905-463051 [6]	463070-463079
K. lactis	NC_006038.1	pos	611456 - 612727 [27]	absent[53]	611890-611919 [54]	612006-612090 [26]	612532-612687 [6]	612708-612716 [54]
E. coryli	AZAH01000001.1	neg	269038-270368		269938-269968			
E. cymbalariae	NC_016454.1	pos	54147 - 54960		54451-54480			
E. gossypii	NC_005782.2	neg	677871-679048[55]		678276-678305 [29]			
Ashbya aceri	CP006020.1	neg	693543 - 694708		693942-693973			
L. kluyveri	CM000690.1	pos	348600 - 349844	348876-348930	348957-348982 [56]	349129-349208		349825-349833
L. lanzarotensis	NW_019212880.1	pos	854162 - 855236	854389-854444		854754-854820		855217-855225
L. waltii	AADM01000270.1	neg	134961 - 136000	135698-135756 [12]	135613-135636	135409-135470		134973-134981
L. thermotolerans	NC_013079.1	pos	702500 - 703549	702730-702791 [12]	702853-702876	703022-703083		703530-703538
L. dasiensis	LT598456.1	pos	682034 - 682916	682124-682181	682261-682283			682900-682905
L. sp. CBS 6924	LT598470.1	neg	441802 - 442700	442582-442638		442229-442292		441811-441820
L. fermentati	LT598488.1	neg	306329 - 307150	307076-307129		306786-306850		306339-306348
L. meyersii	LT598477.1	pos	575851 - 576676	575886-575941		576233-576294		576657-576666
L. mirantina	LT598468.1	pos	690800 - 691797			691218-691282		691777-691786
L. nothofagi	LT598449.1	pos	388401 - 389382	388567-388624		388937-389004		389362-389371
T. delbrueckii	NC_016504.1	pos	709007 - 709780	709057-709086		709267-709336		709761-709770
T. microellipsoides	FYBL01000005.1	neg	426211 - 427050	427000-427028		426726-426817		426221-426229
Z. bailii	HG316456.1	neg	712655 - 713400			712902-712974		712665-712673
Z. rouxii	NC_012990.1	pos	297087 - 297883			297527-297616		297865-297873
Z. parabailii	CP019499.1	pos	455564 - 455975[57]			455656-455728		455957-455965
T. blattae	NC_020193.1	neg	404150 - 405050	405003-405033		404650-404733		404165-404173
N. castellii	NC_016499.1	pos	381827 - 383194[54]	382404-382432 [12]	382506-382519 [54]	382647-382710 [54]	382994-383155[54]	383176-383184[54]
N. dairenensis	NC_016479.1	neg	1519837 - 1521377	1520648-1520678	1520550-1520562	1520303-1520369	1519864-1520027	1519849-1519857
C. castellii	CAPW01000002.1	neg	272769 - 274000		273158-273179	272992-273085		272781-272789
N. bacillisporus	CAPX01000073.1	pos	1230 - 2215					2197-2204
C. glabrata	NC_006032.2	neg	419194 - 421150 [12]	421007-421081 [12]	420914-420932 [12]	420657-420852[12]		419206-419214[12]
C. bracarensis	CAPU01000044.1	pos	2586 - 4361		2836-2854			4342-4350
N. delphensis	CAPT01000167.1	neg	254761 - 256469		256151-256169			254773-254781
C. nivariensis	CAPV01000033.1	pos	87530 - 89215		87780-87798			89196-89204
S. uvarum	NOWY01000011.1	pos	45720 - 46940	45996-46050	46193-46203	46301-46377	46703-46848	46921-46929
S. eubayanus	NC_030979.1	pos	476134 - 477336	476392-476446	476588-476598	476694-476770	477100-477240	477317-477325
S. arboricola	NC_026172.1	pos	287410 - 288645	287705-287739	287888-287898	288019-288096	288417-288558	288626-288634
S. kudriavzevii	AY639012.1	pos	1 - 1215 [25]	284-320 [25]	424-434	585-662	981-1128 [6]	1201-1209
S. mikatae	AABZ01000048.1	neg	18591 - 19809[25]	19497-19532 [25]	19349-19356	19156-19232	18687-18833 [6]	18603-18611
S. paradoxus	CP020294.1	pos	307733 - 308897[25]	308010-308045 [25]	308154-308161	308281-308353	308660-308803 [6]	308878-308886
S. cerevisiae	NC_001134.8	pos	307597 - 308757[11]	307880-307914 [58]	308057-308064 [54]	308185-308256 [26]	308563-308682 [6]	308737-308746[54]
S. pastorianus	AZCJ01000004.1	neg	478773 - 479970[25]	479664-479718 [25]	479512-479520	479340-479417	478866-479012 [6]	478785-478793
S. cer. x S. kud.	AGVY01000004.1	pos	284183 - 285344	284465-284501	284645-284655	284772-284843	285150-285269	285325-285333
S. bayanus	AACG02000058.1	pos	58142 - 59362[25]	58418-58472 [25]	58613-58620	58723-58799	59125-59270 [6]	59343-59351
S. sp. 'boulardii'	CM003558.1	pos	287536 - 288696	287818-287854	287998-288008	288124-288195	288502-288621	288677-288685
S. sp. M14	MVPU01000005.1	neg	473800 - 474997	474691-474745	474537-474547	474368-474444	473894-474038	473812-473820
S. cariocanus	AY639010.1	pos	1 - 1163[25]	278-313 [25]	424-434	549-621	928-1072 [6]	1147-1155

Table 1. Overview of conserved telomer substructures in Saccharomycetacea, as identified by the combined synteny/covariance model pipeline. The 3′ end is defined as 10 nt downstream of the SM binding site. The 5′ end is approximate. Citations refer to publication in which the sequence and/or the coordinates of respective features are reported explicitly. These annotations form the basis of Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

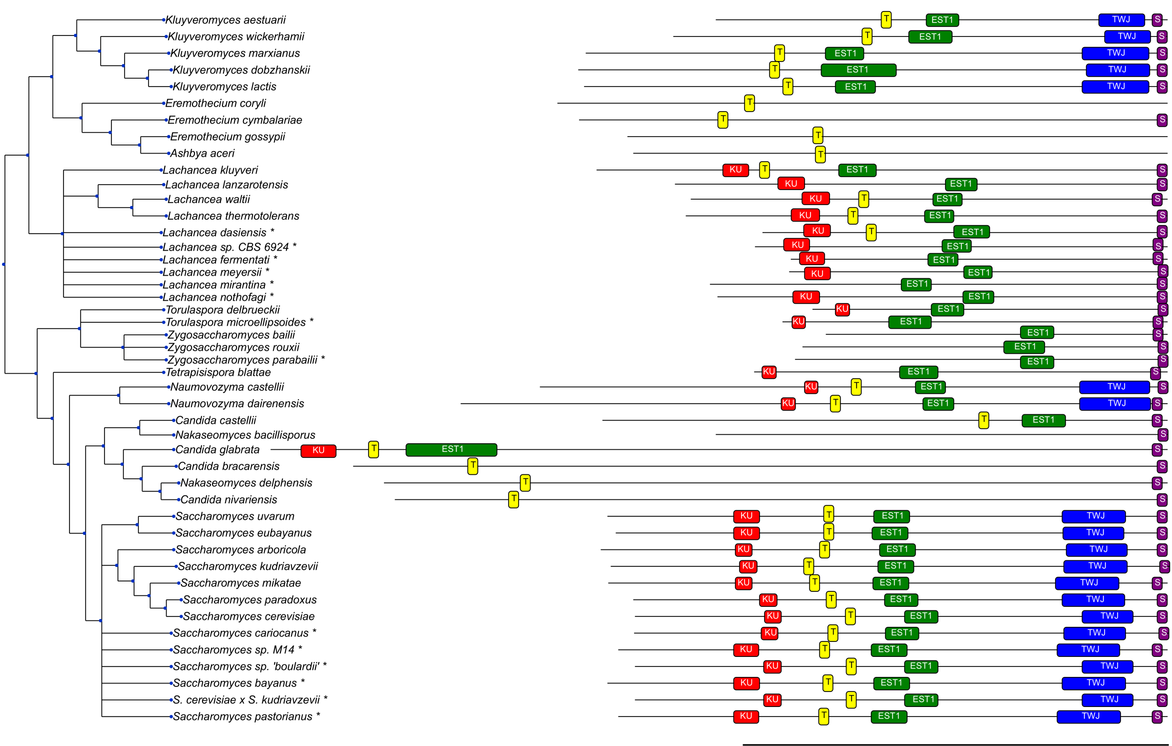

Figure 2. Features identified in TER sequences. **KU)** ku binding hairpin, **T)** template region, **EST1)** Est1 binding site, **TWJ)** three-way junction, **SM1)** SM1 binding site. Elements not shown are either not present in the corresponding species (e.g. the TWJ in *C. glabrata*) or could not be located with reasonable certainty. Species marked by * are not part of the phylogenetic tree and were placed next to their closest related neihbour based on the similarity of their TER sequences.

236 binding site is present even in the more distantly related genus *Candida* [29], this motif is intrisically
237 too variable to be unambiguously recognizable in distant relatives. This pertains to both its sequence
238 and the its base-pairing patterns.

239 Consistent with [12], we found no plausible secondary structure for the TWJ in *C. glabrata*,
240 although the respective region of the sequence contains the highly conserved sequence `AATA`. It is
241 worth noting in this context that the telomerase of the ciliate *Tetrahymena* has a stem-loop structure
242 in place of the threeway junction [61]. TERs of the *C. glabrata* group thus may also have a functional
243 trans-activation domain, albeit with an aberrant structure. Our TWJ covariance model, which was
244 constructed from *Kluyveromyces* and *Saccharomyces* sequences only, also failed to detect a TWJ in
245 *Eremothecium* and *Lachancea*. It remains an open question whether TERs of these species have a TWJ
246 with a diverged structure that is just beyond our ability to detect, or whether trans-activation is
247 achieved by different means.

248 The sequence of the SM binding motif `AATTTTTGG` is perfectly conserved throughout much of the
249 Saccharomycetaceae, with the notable exception of *K.lactis* [54] and additional small variations in other
250 *Kluyoveromyces* species, see Fig. 3. We could not find this motif in species of the genus *Eremothecium*
251 and the highly related species *Ashba acerii*.

252 ## 4. Discussion

253 Although we succeeded in detecting 27 previously unknown TER sequences in
254 Saccharomycetaceae, the main take-home message is of this contribution is that homology
255 search can be a terribly difficult problem. Although yeast TERs are quite long and fulfil a
256 well-conserved function, their sequences are very poorly conserved. In this respect, yeast TER behaves

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

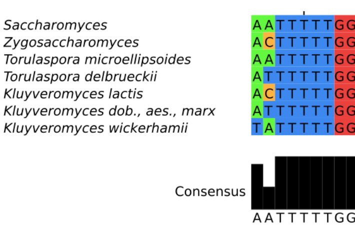

Figure 3. Alignment of the core SM-binding site motif. The common pattern of most Saccharomycetaceae is shown on top, species-specific variants are listed below.

257 much like the majority of long non-coding RNAs, which are also poorly conserved in sequence but
258 often are evolutionary quite well conserved as functional entitied, see [62] for a recent review.
259 The "blast graph" in Figure 4 highlights the practical problem. Sequence comparison methods
260 identify homology only in closely related species. A comparison of Figure 4 and a corresponding
261 graph based on the previously published TER sequences only (see Online Supplemental Material)
262 shows that the larger set of queries identifies many additional connections and thus improves the
263 situation at least within the Saccharomycetacea. Even within the clade, however, we have been unable
264 to confirm the candidate hits in *Kasachstania*. The tree in Figure A1 indicates longer branch lengths
265 leading to *Kasachstania*; it appears that the accelerated evolution of these genomes is already sufficient
266 to hide the TER genes from our homology search methods.
267 While the direct sequence-based search against complete genomes was not very successful, we
268 observed that the synteny-based approach worked remarkably well. This is not entirely unexpected,
269 since the restriction to the interval between a pair of coding genes effectively reduces the size of
270 the target from several million nucleotides to a few thousand. Unfortunately, the applicability of
271 synteny-based methods is limited to relatively narrow phylogenetic scales. On longer time-scales,
272 genome rearrangments are likely to disrupt syntenic conservation. A systematic exploitation of synteny
273 similar to the work described here for Saccharomycetacea would most likely be successful in a survey
274 for TER in the Candida group. In fact synteny has been employed to find some of the known TERs in
275 this clade [29].
276 The study presented here was largely conducted using publicly available tools complemented by
277 some custom scripting. It also highlights the need for customized tools to conduct difficult homology
278 searches. In particular, specific alignment tools and viewers to efficiently evaluate the synteny-based
279 candidates relative to known template sequences and alignments of the better conserved regions
280 would facilitate the manual curation efforts, which we found to be indispensible.
281 Finally, it remains on open question whether direct machine learning methods can be adapted
282 as homology search tools, and if so, whether such a strategy can be more effective than sequence
283 comparison methods. It is likely that such efforts failed so far because of the difficulties inherent
284 in the construction of a suitable negative training set that is not confounded by frequent genomic
285 features such as coding sequence. Furthermore, the small number of positive samples was presumably
286 insufficient to capture the full variability of TER sequences.
287 Complementarily, a phylogenetically dense sample of TERs that are sufficiently similar to support
288 global sequence aligments might help to better understand the rapid divergence of TER sequences.
289 This may be helpful not only to identify informative features for machine learning applications, but
290 may also help to design modified sequence comparison algorithms that better reflect the peculiarities
291 of rapidly evolving long non-codign RNAs. In this contribution we have provided such a set of TERs
292 for the Saccharomycetaceae.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

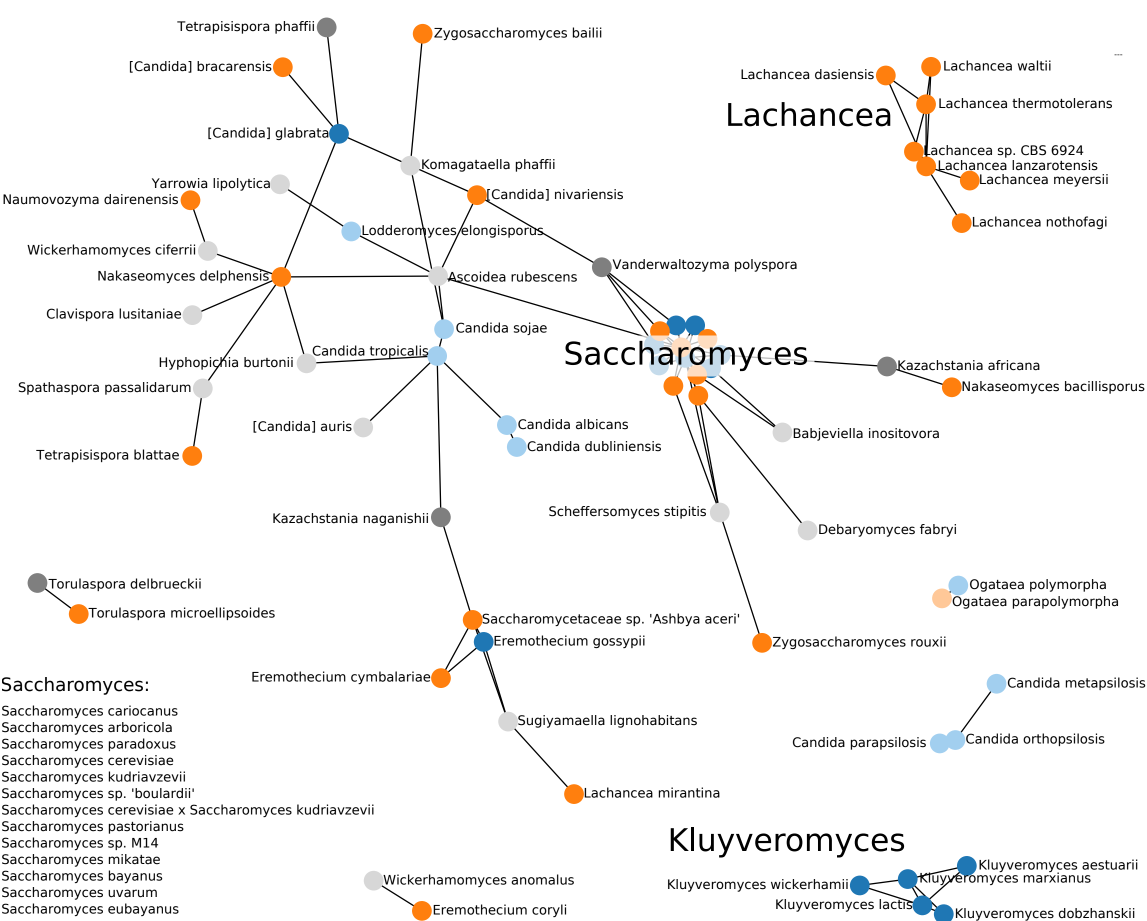

Figure 4. Summary of the `blast`-based survey of TER genes. Blue nodes show TERs described in literature, orange nodes represent TERs that we identified, and grey nodes are additional candidates for which we could not validate characteristic features. TERs outside the Saccharomycetaceae group are presented in light colors. The length of the edges are weighted by the inverse of the length of the `blast` hit. Note that distances in drawing between nodes not connected by an edge are not indicative of their evolutionary distance.

Supplementary Materials: Machine readable Supplemental Information, in particular accession numbers, TER sequences, alignments of conserved features, and covariance models are available at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/18-048/.

Acknowledgments: The research reported here is the outcome of a PhD level seminar taught by PFS at Univ. Vienna in the fall term 2017/18. It was funded in part by the Doktoratskolleg RNA Biology at Univ. Vienna, the Austrian Science Fund (SFB 4305-B09, I 2874-N28, I-1303 B21), Sinergia (CRSII3_154471/1), and the German Federal Ministery for Education and Research (BMBF 031A538A, de.NBI/RBC).

Author Contributions: PFS and MTW conceived the study. MW, BT, RO, and MTW analyzed the TER sequences and structures, AH conducted the phylogenomic analysis, JVdAO and MEMTW contributed machine learning approaches. All authors contributed to writing the paper and approved of the submitted version.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Phylogenomics of the Saccharomycetales

The maximum likelihood tree obtained from 841 orthologous groups of proteins present in at least 67 of the 72 species is shown in Fig. A1. The phylogeny is nearly identical to the tree reported in [36]. In particular, it provides strong support for monophyletic Saccharomycetacea (comprising in particular the genera *Saccharomyces* and *Kluyveromyces*), and the Candida group. Noteworthy, "*Candida glabrata*"

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

309 is nested within the Saccharomycetacea as a close relative of *Saccharomyces* rather than appearing as
310 member of the Candida clade.

311

312 1. Greider, C.W.; Blackburn, E.H. Identification of a specific telomere terminal transferase activity in
313 *Tetrahymena* extracts. *Cell* **1985**, *43*, 405–413.
314 2. Mason, J.M.; Reddy, H.M.; Frydrychova, R.C. Telomere Maintenance in Organisms without Telomerase. In
315 *DNA Replication – Current Advances*; Seligmann, H., Ed.; InTech: Rijeka, HR, 2011; chapter 15.
316 3. Pardue, M.; Rashkova, S.; Casacuberta, E.; DeBaryshe, P.G.; George, J.A.; Traverse, K. Two retrotransposons
317 maintain telomeres in Drosophila. *Chromosome Res.* **2005**, *13*, 443–453.
318 4. Podlevsky, J.D.; Chen, J. Evolutionary perspectives of telomerase RNA structure and function. *RNA Biol*
319 **2016**, *13*, 720–732.
320 5. Chen, J.; Greider, C.W. An emerging consensus for telomerase RNA structure. *Proc. Natl. Acad. Sci. USA*
321 **2004**, *101*, 14683–14684.
322 6. Brown, Y.; Abraham, M.; Pearl, S.; Kabaha, M.M.; Elboher, E.; Tsfati, Y. A critical three-way junction is
323 conserved in budding yeast and vertebrate telomerase RNAs. *Nucleic Acids Res.* **2007**, *35*, 6280–6289.
324 7. Wu, R.A.; Upton, H.E.; Vogan, J.M.; Collins, K. Telomerase Mechanism of Telomere Synthesis. *Annu Rev*
325 *Biochem* **2017**, *86*, 439–460.
326 8. Webb, C.J.; Zakian, V.A. Telomerase RNA is more than a DNA template. *RNA Biol.* **2016**, *13*, 683–689.
327 9. Greider, C.W.; Blackburn, E.H. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for
328 telomere repeat synthesis. *Nature* **1989**, *337*, 331–337.
329 10. Feng, J.; Funk, W.D.; Wang, S.S.; Weinrich, S.L.; Avilion, A.A.; Chiu, C.P.; Adams, R.R.; Chang, E.; Allsopp,
330 R.C.; Yu, J. The RNA component of human telomerase. *Science* **1995**, *269*, 1236–1241.
331 11. Singer, M.S.; Gottschling, D.E. TLC1: template RNA component of *Saccharomyces cerevisiae* telomerase.
332 *Science* **1994**, *266*, 404–409.
333 12. Kachouri-Lafond, R.; Dujon, B.; Gilson, E.; Westhof, E.; Fairhead, C.; Teixeira, M.T. Large telomerase
334 RNA, telomere length heterogeneity and escape from senescence in *Candida glabrata*. *FEBS Lett.* **2009**,
335 *583*, 3605–3610.
336 13. Cifuentes-Rojas, C.; Kannan, K.; Tseng, L.; Shippen, D.E. Two RNA subunits and POT1a are components
337 of *Arabidopsis telomerase*. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 73–78.
338 14. Beilstein, M.A.; Brinegar, A.E.; Shippen, D.E. Evolution of the *Arabidopsis* telomerase RNA. *Front Genet.*
339 **2012**, *3*, 188.
340 15. Gupta, S.K.; Kolet, L.; Doniger, T.; Biswas, V.K.; Unger, R.; Tzfati, Y.; Michaeli, S. The *Trypanosoma brucei*
341 telomerase RNA (TER) homologue binds core proteins of the C/D snoRNA family. *FEBS Lett* **2013**,
342 *587*, 1399–1404.
343 16. Sandhu, R.; Sanford, S.; Basu, S.; Park, M.; Pandya, U.M.; Li, B.; Chakrabarti, K. A trans-spliced telomerase
344 RNA dictates telomere synthesis in *Trypanosoma brucei*. *Cell Res* **2013**, *23*, 537–551.
345 17. Chakrabarti, K.; Pearson, M.; Grate, L.; Sterne-Weiler, T.; Deans, J.; Donohue, J.P.; Ares Jr, M. Structural
346 RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA
347 analysis. *RNA* **2007**, *13*, 1923–1939.
348 18. Religa, A.A.; Ramesar, J.; Janse, C.J.; Scherf, A.; Waters, A.P. *P. berghei* Telomerase Subunit TERT is Essential
349 for Parasite Survival. *PLoS ONE* **2014**, *9*, e108930.
350 19. Xie, M.; Mosig, A.; Qi, X.; Li, Y.; Stadler, P.F.; Chen, J.J.L. Size Variation and Structural Conservation of
351 Vertebrate Telomerase RNA. *J. Biol. Chem.* **2008**, *283*, 2049–2059.
352 20. Li, Y.; Marz, M.; Qi, X.; Podlevsky, J.D.; Hoffmann, S.; Stadler, P.F.; Chen, J.J.L. Identification of Purple Sea
353 Urchin Telomerase RNA using a Next-Generation Sequencing Based Approach. *RNA* **2013**, *19*, 852–860.
354 21. Podlevsky, J.D.; Li, Y.; Chen, J.J. Structure and function of echinoderm telomerase RNA. *RNA* **2016**,
355 *22*, 204–215.
356 22. Qi, X.; Li, Y.; Honda, S.; Hoffmann, S.; Marz, M.; Mosig, A.; Podlevskya, J.D.; Stadler, P.F.; Selker, E.U.;
357 Chen, J.J.L. The common ancestral core of vertebrate and fungal telomerase RNAs. *Nucleic Acids Res.* **2013**,
358 *41*, 450–462.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

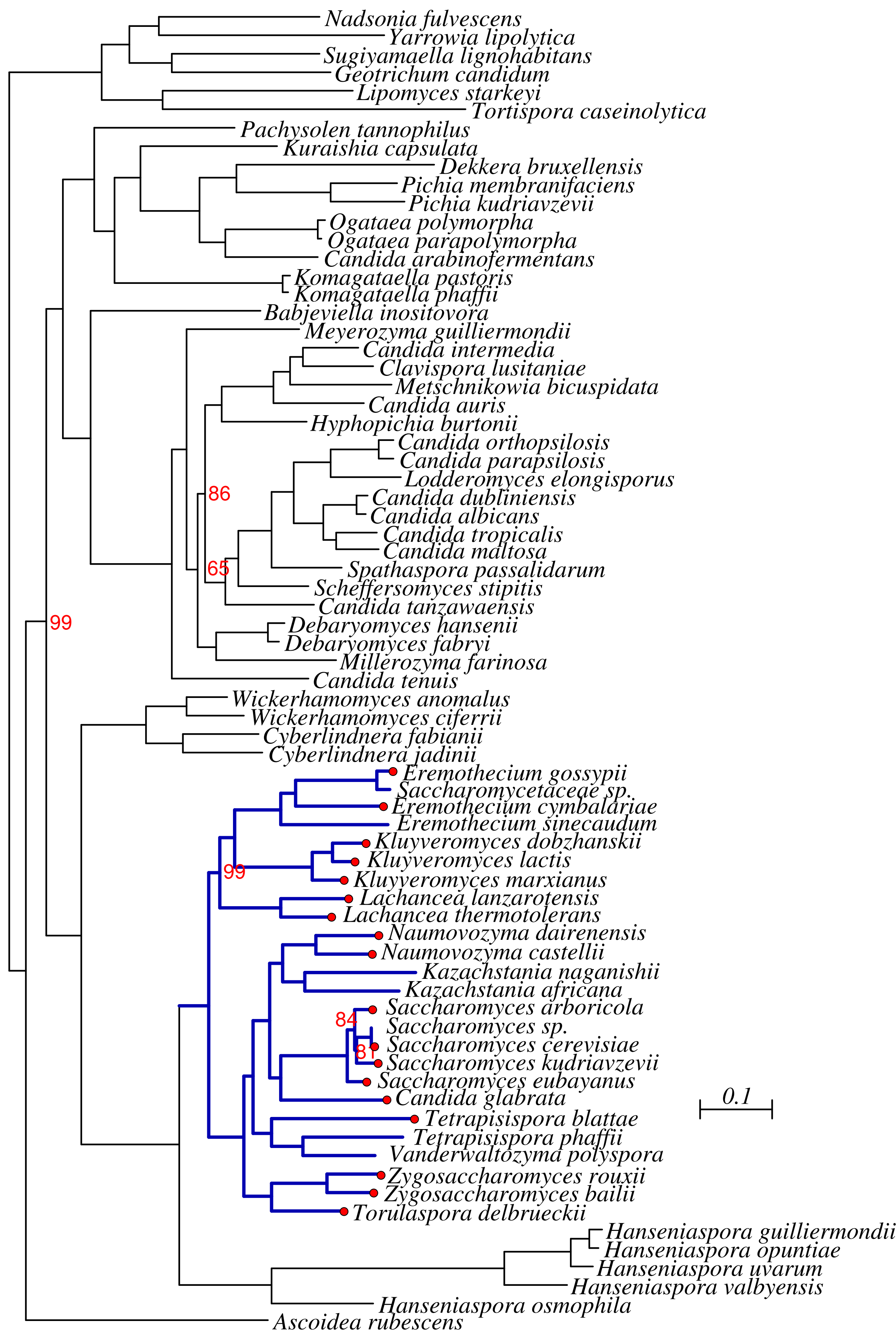

Figure A1. Phylogeny of the Saccharomycetales. Bootstrap support is 100% unless otherwise indicated. The Saccharomycetacea are indicted in dark blue. A red dot at tip of the tree indicates a TER sequences listed in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

359 23. Kuprys, P.V.; Davis, S.M.; Hauer, T.M.; Meltser, M.; Tzfati, Y.; Kirk, K.E. Identification of telomerase RNAs
360 from filamentous fungi reveals conservation with vertebrates and yeasts. *PLoS One* **2013**, *8*, e58661.
361 24. Qi, X.; Rand, D.P.; Podlevsky, J.D.; Li, Y.; Mosig, A.; Stadler, P.F.; Chen, J.J. Prevalent and distinct
362 spliceosomal 3′-end processing mechanisms for fungal telomerase RNA. *Nat Commun.* **2015**, *6*, 6105.
363 25. Dandjinou, A.T.; Lévesque, N.; Larose, S.; Lucier, J.F.; Abou Elela, S.; Wellinger, R.J. A phylogenetically
364 based secondary structure for the yeast telomerase RNA. *Curr Biol* **2004**, *14*, 1148–1158.
365 26. Seto, A.G.; Livengood, A.J.; Tzfati, Y.; Blackburn, E.H.; Cech, T.R. A bulged stem tethers Est1p to telomerase
366 RNA in budding yeast. *Genes Dev* **2002**, *16*, 2800–2812.
367 27. McEachern, M.J.; Blackburn, E.H. Runaway telomere elongation caused by telomerase RNA gene mutations.
368 *Nature* **1995**, *376*, 403–409.
369 28. Hsu, M.; McEachern, M.J.; Dandjinou, A.T.; Tzfati, Y.; Orr, E.; Blackburn, E.H.; Lue, N.F. Telomerase core
370 components protect Candida telomeres from aberrant overhang accumulation. *Proc Natl Acad Sci USA*
371 **2007**, *104*, 11682–11687.
372 29. Gunisova, S.; Elboher, E.; Nosek, J.; Gorkovoy, V.; Brown, Y.; Jean-François, L.; Laterreur, N.; Wellinger, R.J.;
373 Tzfati, Y.; Tomaska, L. Identification and comparative analysis of telomerase RNAs from Candida species
374 reveal conservation of functional elements. *RNA* **2009**, *15*, 546–559.
375 30. Smekalova, E.M.; Malyavko, A.N.; Zvereva, M.I.; Mardanov, A.V.; Ravin, N.V.; Skryabin, K.G.; Westhof, E.;
376 Dontsova, O.A. Specific features of telomerase RNA from *Hansenula polymorpha*. *RNA* **2013**, *19*, 1563–1574.
377 31. Zappulla, D.C.; Goodrich, K.J.; Arthur, J.R.; Gurski, L.A.; Denham, E.M.; Stellwagen, A.E.; Cech, T.R. Ku
378 can contribute to telomere lengthening in yeast at multiple positions in the telomerase RNP. *RNA* **2011**,
379 *17*, 298–311.
380 32. Mosig, A.; Chen, J.L.; Stadler, P.F. Homology Search with Fragmented Nucleic Acid Sequence Patterns.
381 Algorithms in Bioinformatics (WABI 2007); Giancarlo, R.; Hannenhalli, S., Eds.; Springer Verlag: Berlin,
382 Heidelberg, 2007; Vol. 4645, *Lecture Notes in Computer Science*, pp. 335–345.
383 33. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold Faster RNA Homology Searches. *Bioinformatics* **2013**,
384 *29*, 2933–2935.
385 34. Hawksworth, D.L. A new dawn for the naming of fungi: impacts of decisions made in Melbourne in July
386 2011 on the future publication and regulation of fungal names. *IMA Fungus* **2011**, *2*, 155–162.
387 35. McNeill, J.; Barrie, F.R.; Buck, W.R.; Demoulin, V.; Greuter, W.; Hawksworth, D.L.; Herendeen, P.S.; Knapp,
388 S.; Marhold, K.; Prado, J.; Prud'homme van Reine, W.F.; Smith, G.F.; Wiersema, J.H.; Turland, N.J., Eds.
389 *International Code of Nomenclature for algae, fungi and plants*; Vol. 154, *Regnum Vegetabile*, Koeltz Scientific
390 Books: Oberreifenberg, D, 2012.
391 36. Shen, X.X.; Zhou, X.; Kominek, J.; Kurtzman, C.P.; Hittinger, C.T.; Rokas, A. Reconstructing the Backbone
392 of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3 (Bethesda)* **2016**, *6*, 3927–3939.
393 37. Lechner, M.; Findeiß, S.; Steiner, L.; Marz, M.; Stadler, P.F.; Prohaska, S.J. `Proteinortho`: Detection of
394 (Co-)Orthologs in Large-Scale Analysis. *BMC Bioinformatics* **2011**, *12*, 124.
395 38. Lechner, M.; Hernandez-Rosales, M.; Doerr, D.; Wieseke, N.; Thévenin, A.; Stoye, J.; Hartmann, R.K.;
396 Prohaska, S.J.; Stadler, P.F. Orthology Detection Combining Clustering and Synteny for Very Large Datasets.
397 *PLoS ONE* **2014**, *9*, e105015.
398 39. Roth, A.C.J.; Gonnet, G.H.; Dessimoz, C. Algorithm of OMA for large-scale orthology inference. *BMC*
399 *Bioinformatics* **2008**, *9*, 518.
400 40. Altenhoff, A.M.; Gil, M.; Gonnet, G.H.; Dessimoz, C. Inferring Hierarchical Orthologous Groups from
401 Orthologous Gene Pairs. *PLOS One* **2013**, *8*, e53786.
402 41. Talavera, G.; Castresana, J. Improvement of phylogenies after removing divergent and ambiguously
403 aligned blocks from protein sequence alignments. *Syst Biol* **2007**, *56*, 564–577.
404 42. Stamatakis, A. `RAxML` Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.
405 *Bioinformatics* **2014**, *30*, 1312–1313.
406 43. Podlevsky, J.D.; Bley, C.J.; Omana, R.V.; Qi, X.; Chen, J.J.L. The Telomerase Database. *Nucleic Acids Res*
407 **2007**, *36*, D339–D343.
408 44. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J Mol Biol*
409 **1990**, *215*, 403–410.
410 45. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S.
411 MEME SUITE: tools for motif discovery and searching. *Nucleic acids res* **2009**, *37*, W202–W208.

bioRxiv preprint doi: https://doi.org/10.1101/323675; this version posted May 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

412 46. Costa, F.C.; De Grave, K. Fast neighborhood subgraph pairwise distance Kernel. Proceedings of the
413 27th International Conference on International Conference on Machine Learning (ICML'10); Fürnkranz, J.;
414 Joachims, T., Eds.; Omnipress: Madison, WI, 2010; pp. 255–262.
415 47. Will, S.; Reiche, K.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. Inferring noncoding RNA families and classes
416 by means of genome-scale structure-based clustering. *PLoS Comp. Biol.* **2007**, *3*, e65.
417 48. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in
418 performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.
419 49. Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss
420 and rearrangement. *PloS ONE* **2010**, *5*, e11147.
421 50. Wolfinger, M.T.; Fallmann, J.; Eggenhofer, F.; Amman, F. ViennaNGS: A toolbox for building efficient
422 next-generation sequencing analysis pipelines. *F1000Research* **2015**, *4*.
423 51. Teixeira, M.T.; Gilson, E. Telomere maintenance, function and evolution: the yeast paradigm. *Chromosome*
424 *Research* **2005**, *13*, 535–548.
425 52. Wendland, J.; Walther, A. Genome Evolution in the *Eremothecium* Clade of the Saccharomyces Complex
426 Revealed by Comparative Genomics. *G3: Genes, Genomes, Genetics* **2011**, *1*, 539–548.
427 53. Kabaha, M.M.; Zhitomirsky, B.; Schwartz, I.; Tzfati, Y. The 5′ Arm of *Kluyveromyces lactis* Telomerase RNA
428 Is Critical for Telomerase Function. *Mol Cell Biol* **2008**, *28*, 1875–1882.
429 54. Telomerase Database - Secondary structures. http://telomerase.asu.edu/structures.html#secondary, last
430 accessed April 30, 2018.
431 55. Dietrich, F.S. The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient *Saccharomyces cerevisiae*
432 Genome. *Science* **2004**, *304*, 304–307.
433 56. Genome resources for yeast chromosomes database - TLC1 (Telomerase RNA template). http://gryc.inra.
434 fr/index.php?page=locus&seqid=SAKL0D04356r, last accessed Apr 30, 2018.
435 57. Ortiz-Merino, R.A.; Kuanyshev, N.; Braun-Galleani, S.; Byrne, K.P.; Porro, D.; Branduardi, P.; Wolfe, K.H.
436 Evolutionary restoration of fertility in an interspecies hybrid yeast, by whole-genome duplication after a
437 failed mating-type switch. *PLoS Biol* **2017**, *15*, e2002128.
438 58. Peterson, S.E.; Stellwagen, A.E.; Diede, S.J.; Singer, M.S.; Haimberger, Z.W.; Johnson, C.O.; Tzoneva, M.;
439 Gottschling, D.E. The function of a stem-loop in telomerase RNA is linked to the DNA repair protein Ku.
440 *Nature genetics* **2001**, *27*, 64.
441 59. Box, J.A.; Bunch, J.T.; Tang, W.; Baumann, P. Spliceosomal cleavage generates the 3′ end of telomerase
442 RNA. *Nature* **2008**, *456*, 910–914.
443 60. Kannan, R.; Helston, R.M.; Dannebaum, R.O.; Baumann, P. Diverse mechanisms for spliceosome-mediated
444 3′ end processing of telomerase RNA. *Nat Commun.* **2015**, *6*, 6104.
445 61. Singh, M.; Wang, Z.; Koo, B.K.; Patel, A.; Cascio, D.; Collins, K.; Feigon, J. Structural Basis for Telomerase
446 RNA Recognition and RNP Assembly by the Holoenzyme La Family Protein p65. *Mol Cell* **2012**, *47*, 16–26.
447 62. Nitsche, A.; Stadler, P.F. Evolutionary Clues in lncRNAs. *Wiley Interdiscip Rev RNA* **2017**, *8*, 1.

448 © 2018 by the authors. Submitted to *Genes* for possible open access publication under the terms and conditions
449 of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).