Final scientific report of PD 140347 (previously 128617) entitled „Comparative molecular genetic analysis of *Candida* species derived from animal and human source by whole genome sequencing method"

The fungal kingdom includes as many as 6 million species and is remarkable in terms of the breadth and depth of its impact on global health, agriculture, biodiversity, ecology, manufacturing, and biomedical research. More than 600 fungal species are associated with humans and animals, either as commensals and members of the microbiome or as pathogens that cause some of the most lethal infectious diseases. Although underestimated as a cause of infection in humans, fungi are associated with approximately 1.5 million deaths and 1.7 billion superficial infections yearly, resulting in an enormous economic burden. However, compared with mycoses in humans, fungal diseases in animals have received much less attention. The aim of our study was to extend the whole genome sequencing based epidemiologic and evolutionary analyses of *Candida* species to animal-derived strains and to identify differences in genomic features that shape adaptive processes.

During the first year we processed animal samples collected by the Department of Microbiology and Infectious Diseases at the University of Veterinary Medicine. Ninety isolates collected from animals and 6 human isolates were involved in the study. Samples were inoculated on Sabouraud dextrose agar and incubated at 37°C for 3-4 days. All colonies with a yeast-like morphology were isolated and used for DNA extraction and identification. Although species level identification was carried out previously by biochemical or MALDI-TOF mass spectrometer, we performed molecular-based method as well by sequencing the entire internal transcribed spacer region (ITS) of the fungal rDNA. Generated nucleotide sequences were deposited in GenBank.

Fifty-one isolates originated from goose and ducks diagnosed with oesophageal mycosis were selected for analysis. The most prevalent pathogen was *C. albicans*; but *Saccharomyces cerevisiae*, *C. kefyr*, and *Kazachstania bovina* were also frequently isolated, whereas other yeasts such as *C. lambica*, *C. inconspicua*, *C. rugosa*, *C. pelliculosa*, *C. krusei*, *Magnusiomyces capitatus*, and *Trichosporon asahii* were rarely isolated species. Historic data indicate *C. albicans* takes the primary role in the etiology of crop mycosis of poultry, but other yeast-like species have not been mentioned in the literature. The hypothesis of the etiology of non-*albicans* yeasts in oesophageal disease needed formal demonstration, given that histopathologic investigations that could have confirmed macroscopic observations was not included in this study. According to our observations along with other authors, sequencing the ITS region could

be a widely available alternative method for yeast identification for veterinary diagnostic laboratories. The results were published a peer-reviewed scientific journal (Domán et al. 2020).

Multilocus sequence typing (MLST) approach was formerly developed for four *Candida* species (*C. albicans*, *C. glabrata*, *C. tropicalis*, *C. krusei*), therefore we used this method for the preliminary molecular characterisation of all *C. albicans* strains collected in this period (n=30). The genotyping was based on partial amplification and sequencing seven housekeeping genes. Distinct alleles and diploid sequence types (DSTs) were identified and numbered by comparing the sequences with those available in the *C. albicans* MLST database (https://pubmlst.org/organisms/candida-albicans). Novel alleles and allelic combinations (new DTSs) were submitted to the central MLST database where sequences were validated by the curator. Eight known and six new MLST genotypes (DST numbers: 3595-3600) were determined. Our work yielded a significant increase in the number of submitted isolates from Hungary in the database (30 out of 34 isolate).

In the second year, phylogenetic and population structure analyses of the collected *C. albicans* isolates was performed by unweighted pair group method with arithmetic averages (UPGMA) algorithm and goeBURST algorithm. The *C. albicans* isolates grouped into 8 clades. Concerning particularly human isolates the five major clades assigned by UPGMA (Clade 1, 2, 3, 4 and 11) have proved the most consistent over several years of rapid expansion of the MLST global database. In our analysis, Clade 4 was consisted of two human isolates and one bird isolate. Two human isolates clustered to Clade 1, while no isolate was found in Clade 2 and 11 suggesting that *C. albicans* isolates originating from animal source rather belong to minor clades. The most prevalent genotype was DST 172 which clustered into Clade 15. To reveal the evolutionary relationships between isolates, further cluster analysis was carried out by goeBURST algorithm. Isolates in the same eBURST clonal complexes were grouped together in the respective clades determined by UPGMA clustering, which was consistent with former observations. Interestingly, one waterfowl isolate and one human isolate was genetically closely related. Moreover, the Im-12 isolate derived from ostrich and the human isolate 14362 were shared the same DST, confirming that there are no host specificity of *C. albicans* strains in certain genotypes. Our study revealed that *C. albicans* subpopulations from birds and humans presumably develop relatively independently, while still maintaining some common features enabling the transfer of several genotypes between humans and animals. The results were published a peer-reviewed scientific journal (Domán et al. 2021a).

We continued to collect oesophageal samples originated from ducks and geese held in distinct flocks in the south-eastern region of Hungary. Based on gross pathologic findings, 58 isolates were obtained from the oesophageal lesions of birds. The isolates were identified by macroscopic morphology on Sabouraud dextrose agar and sequencing the ITS region. By sequence analysis, the most frequently isolated species was *C. albicans* (n=29), followed by *K. bovina* (n=21), *Trichosporon* species (n=6) and *S. cerevisiae* (n=2). In ducks, the mechanical damage caused by feeding-tube was more frequently noticed than in geese and typical signs of oesophageal mycosis were rarely seen. Herewith, we performed histopathologic examinations of oesophageal samples that confirmed macroscopic observations. The results were published in a Hungarian journal (Domán et al. 2021b).

In the last period, we continued to analyse the polymorphisms within DNA sequences of *C. albicans* isolates collected from ducks and geese enabling genotyping and phylogenetic studies. In order to investigate the genetic diversity of isolates causing infection compared to commensals of normal digestive microbiota, molecular typing of isolates from oesophageal mucosa of healthy birds were carried out as well. According to our results, three genotypes was responsible for the oesophageal mycosis of geese originated from different flocks (DST 840, DST 605, DST 656). Furthermore, DST 605 was isolated from ducks as well. Interestingly, isolates from healthy birds assigned to DST 840 and new MLST genotypes (DST 3670, DST 3671). Cluster analysis showed that DST 605 was putatively evolved from DST 656 and clustered to Clade 4. These genotypes were closely related to DST 3596, a new MLST genotype determined from a human sample in the previous year. DST 840 was clustered to Clade 7, while commensal isolates were clustered to another clade (Clade 14). Although isolates from animals belonged to minor clades in contrast with the majority of human isolates, no host specificity was observed. Besides understanding the genetic composition of strains, we have gained insight into the molecular epidemiology and population evolution of *C. albicans* in birds (Vásárhelyi, 2021).

Eight *C. krusei* isolates were also collected in the third year from the uterus of cows diagnosed with reproductive disorders. Isolates were identified by MALDI-TOF and by sequencing the ITS region of fungal rDNA. Relatively little genetic or genomic investigation has been carried out on *C. krusei* isolates so far, therefore we used MLST method to explore strain-level differences within this species as well. One isolate from our culture collection was also involved in the study, which was originated from the oesophagus of a duck diagnosed with gastrointestinal mycosis. The MLST scheme employed for *C. krusei* genotyping was based on

partial amplification and sequencing of six protein-coding genes (*HIS3*, *LEU2*, *NMT1*, *TRP1*, *ADE2*, and *LYS2D*). Isolates were assigned to DST 19, DST 24 and DST 67. However, some isolates were classified in new MLST genotypes (DST numbers: 201-203). At that time, *C. krusei* MLST database did not receive any submissions due to the retirement of the curator. We thought that sequence data from multiple loci facilitates determinations of population structures and epidemiological correlates of properties, such as geographical and anatomical origins of isolates and their transmission within and between hosts, hence I applied for the role of curator. Considering my experience with *Candida* species and MLST method, the head of public MLST database accepted my application (https://pubmlst.org/organisms/candida-krusei).

We performed phylogenetic and population structure analysis of all *C. krusei* DSTs available in the database (n=203) using the same algorithms as mentioned in case of *C. albicans*. Overall, 75 polymorphic sites were found among all examined loci. The *NMT1* locus produced the highest number of alleles (n=35), while *HIS3* displayed the lowest (n=23). The most commonly encountered strain types were DST 17 (21 isolate) and DST 67 (25 isolate). By UPGMA method, *C. krusei* DSTs could be subgrouped into five clusters with an approach described by Jacobsen et al. (2007). The largest group was subtype 1 containing 52.06% of DSTs, followed by subtype 2 (28.35%), subtype 4 (15.46%), subtype 3 (3.61%) and subtype 5 (0.51%) (Figure 2). DSTs from human blood and animal source were the main components of group subtype 1, although, isolates from animals were submitted only from two countries (France and Hungary). Subtype 2 predominantly consisted of DSTs from the oropharynx (including sputum and bronchoalveolar lavage) of human patients. The anatomical origin of DST 49 and DST 87 was known among the members of the group subtype 3 (oropharynx/other superficial source and blood, respectively), while subtype 4 was prevailingly composed of blood isolates. One DST obtained from an animal formed group subtype 5 alone. Evaluation of genotypic relationship of strains with eBURST algorithm resulted 18 clonal complexes (CCs) and 94 singletons (DSTs that could not be assigned to any group). Our isolates with known DSTs clustered to CC-3 (ST 19 and ST 24) and CC-4 (ST 67), while newly identified DSTs were singletons. The DST 24 was putatively evolved from the DST 19 group founder due to loss of heterozygosity considering nucleotide positions 2467 and 2659 in *LYS2D*. In CC-4, DST 67 or DST 164 was predicted as founder of the group. Interestingly, DST 67 was the most prevalent genotype that was isolated globally (including North America, South America, Asia, and Europe). Besides Hungarian isolates, *C. krusei* MLST data from animals were reported only from France. Of note, DST 17, DST 19, DST 24 and DST 67 were detected from animal samples in both countries. Moreover, DST 17 and DST 67 were the most prevalent genotypes identified even

from humans and animals worldwide, raising the possibility that strains belong to these genotypes are better adapted to colonise or infect different hosts. A manuscript, that contains these results, was submitted to a peer-reviewed scientific journal (Domán et al. 2021c).

We elaborated the genome-based bioinformatical analysis of *Candida* species as next-generation sequencing-related bioinformatics used in our laboratory was applied mainly for characterisation of viruses and bacteria. Through genome complexity (e.g. diploid genome), our knowledge of pathogenesis of fungal infections falls away that for other microbial diseases. The comparison of animal- and human-derived strains might contribute to better understand the route of pathogenesis and promote the expansion of new diagnostic methods. We selected eight *C. albicans* isolates with remarkable difference in genotypic patterns (mostly new MLST genotypes) for whole genome sequencing. The DNA concentration of two isolates was too low for library preparation for Illumina sequencing, thus libraries were prepared only from the genomic DNA of the remaining six isolates. Illumina NextSeq 500 platform was used to generate 150 bp single reads. Quality control of raw sequence data was checked with FastQC software. Sequences were mapped to the genome of *C. albicans* reference strain SC5314 available from Candida Genome Database using the Burrows–Wheeler Alignment tool with the BWA-MEM algorithm. SNPs were called using Genome Analysis Toolkit (GATK). Poor quality SNPs and indels were filtered using the GATK VariantFiltration module. To investigate the phylogenetic relationship applying high-throughput sequencing results, we used alignment and assembly free (AAF) method by Skmer tool that constructs phylogenies directly from unassembled genome sequence data. The same phylogenetic relations were noticed between isolates as with UPGMA method based on seven housekeeping genes (Figure 1).

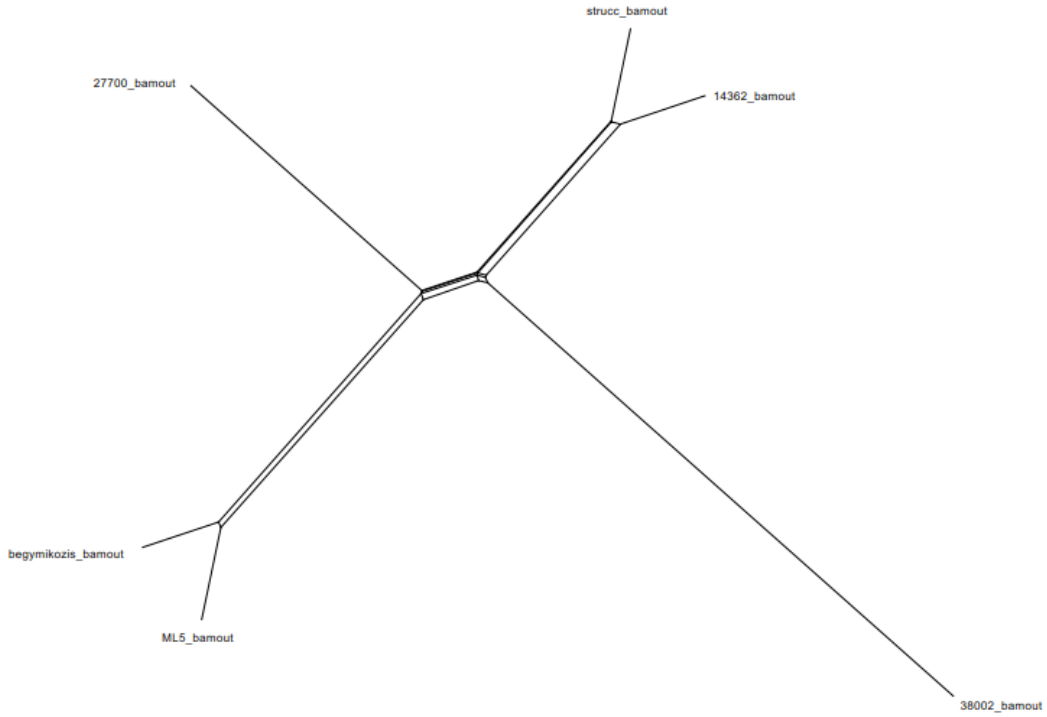

Figure 1. AAF phylogenetic network analysis of six *Candida albicans* isolates

Isolate	Source	MLST genotype	UPGMA clade	SNP/indel	SNP-filtered	Heterozygous SNP
14362	human blood	DST 3598	7	17105	14747	310
27700	human cervix	DST 3600	12	18907	16411	355
38002	human cervix	DST 3597	1	6490	4593	44
ML-5	goose oesophagus	DST 3599	10	19868	16868	448
Im-12	ostrich intestine	DST 3598	7	17378	15093	270
Om-8	duck oesophagus	DST 3595	10	21414	18684	512

Table 1. Genetic characteristics of *Candida albicans* isolates selected for whole genome sequencing

Interestingly, during whole genome analysis the highest number of SNP was found in Om-8, while 38002 displayed the lowest number of SNP. Moreover, the frequency of SNPs was 3-fold lower (and more than 6-fold lower in case of heterozygous SNPs) in 38002 than that was seen

in other isolates (Table 1). To understand this large difference considering SNP frequency in the genome of *C. albicans* isolates, further evaluation of sequencing results needed which is still in progress. The diploid genome of *C. albicans* contains a relatively high density of heterozygous positions even when compared to other *Candida* clade species. However, the frequency of large loss of heterozygosity (LOH events) increases in response to environmental stressors (oxidative stress, high temperature, antifungal drugs) that may be responsible for the variation in SNP density between strains. LOH events will be defined by at least 2 successive losses of heterozygous SNP positions (transition from a heterozygous to homozygous position between two genome comparisons) to assess the rate of this key mechanism by which isolates can evolve and adapt to their environment.

Unfortunately, we did not have the chance to sequence more *Candida* strains due to difficulties in procurement of consumables correlating to pandemic and institutional change. The current project covered no consumable costs; additionally, the transfer of the budget of our research group between institutions was delayed preventing us to perform more complex investigations by analysing significant amount of new complete genome sequences. Despite the unforeseen difficulties, the project provided valuable data about genetic and epidemiologic features of *Candida* species detected from animals:

-We assessed the prevalence of *Candida* species causing disease in waterfowls. *C. albicans* proved to be the main etiologic agent of oesophageal mycosis, however, other yeasts might contribute to the development of infection as well.

-We performed preliminary molecular characterisation of *Candida* strains isolated from different hosts by MLST method. No host specificity was observed, although dominant genotypes were identified indicating that certain genotypes are better adapted to diverse environmental niches and potential inter-species transmission may occur. Our results may help the development of disease prevention strategies.

-New MLST genotypes were detected and data from Hungary were submitted to public databases contributing to better understand the genetic variability of *Candida* strains.

-Phylogenetic relationship between strains could be accurately determined with both MLST and genome sequencing method. The number of SNPs in the genome of animal-derived *C. albicans* isolates is similar to SNP frequency in isolates from human source. Our results may provide good basis for identification of potential new targets for development of antimicrobials.

Overall, two article have been published in peer-reviewed, international scientific journals (Scimago journal rank Q1 and Q2) and one article have been published in a Hungarian journal yet. A manuscript discussing the molecular phylogenetics of *C. krusei* strains including isolates originated from cattle is currently under review. We also intend to submit a manuscript about the comparative genome analysis of *C. albicans* strains isolated from different source to an international scientific journal. A thesis was written in Hungarian including some results of this study by a veterinary student, Balázs Vásárhelyi (Thesis title: Molecular epidemiological study of *Candida albicans* strains in fattened goose and ducks).

References

Domán M, Makrai L, Bali K, Lengyel G, Laukó T, Bányai K. Unexpected Diversity of Yeast Species in Esophageal Mycosis of Waterfowls. Avian Dis. 2020; 64(4):532-535. doi: 10.1637/aviandiseases-D20-00053.

Domán M, Makrai L, Lengyel G, Kovács R, Majoros L, Bányai K. Molecular Diversity and Genetic Relatedness of *Candida albicans* Isolates from Birds in Hungary. Mycopathologia. 2021a; 186(2):237-244. doi: 10.1007/s11046-021-00527-3.

Domán M, Vásárhelyi B, Balka Gy, Jantyik T, Laukó T, Bányai K, Makrai L. Nyelőcsőmikózis magyarországi lúd- és kacsaállományokban. Magyar Állatorvosok Lapja. 2021b; 143(11): 667-675.

Domán M, Makrai L, Bányai K. Molecular phylogenetic analysis of *Candida krusei*. 2021c. (Submitted for publication).