Parker Christenson Assignment 5 Topic modeling

Instructions

1. Pre-process the data by removing stop words and applying stemming or lemmatization, then vectorize the raw text with the CountVectorizer function from Sklearn.
    - Optionally create bigrams and/or trigrams from the available words.
2. Build topic models with 8, 9, and 10 different topics.
3. Evaluate the different topics using perplexity and/or topic coherence.
4. Select a single model, and describe the different topics generated. You can leverage the helper functions provided for this purpose.
    - List the most common word in each topic and optionally create word clouds for each topic.
    - Write a sentence or two to describe your perspective on the composition of each topic. What makes this topic stand out from the others?
    - Take one or two sentences and summarize all the topics together.
    - Note that all the items in Step 4 should be in text form in a Markdown cell in your submitted Jupyter Notebook. For reference, there is no example of this in the COVID19.ipynb Download COVID19.ipynbassignment lab Jupyter Notebook.

In [ ]:
```
# library imports
import numpy as np
import pandas as pd
import re
import json
import sys
import os
import ast
import random
pd.set_option('display.max_columns', 40)
import nltk
import gensim
import wordcloud
import faiss
from nltk.corpus import stopwords
from helper_functions import *

# having some trouble with the nltk stopwords, so I'm going to download them again
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\tehwh\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

Out[ ]: True

Data pre-processing

In [ ]:
```
root_location = 'E:/Data_sets' # I offloaded all of the data to a 4TB external driv

file_locations = [f'{root_location}/biorxiv_medrxiv/biorxiv_medrxiv',
                  f'{root_location}/noncomm_use_subset/noncomm_use_subset',
                  f'{root_location}/comm_use_subset/comm_use_subset',
                  f'{root_location}/custom_license/custom_license']

## Set up Stop Words
## Add Any other relevant options manually

stop_words = stopwords.words('english') + ['et', 'al', 'fig', 'etal', 'et al', 'et-

processed_articles = process_articles(file_locations, stop_words)
processed_articles.read_files()

print('Example File Name')
print(processed_articles.root_files[0])
print('Number of files')
print(len(processed_articles.root_files))
print('Example Article Information')
print(processed_articles.title_text[2])
```

Processing Files at the requested location
There are 177 files to process
There were 885 files in the dataset
Processing Files at the requested location
There are 470 files to process
There were 2353 files in the dataset
Processing Files at the requested location
There are 1823 files to process
There were 9118 files in the dataset
Processing Files at the requested location
There are 3391 files to process
There were 16959 files in the dataset
Example File Name
200b85eea010e695ac7281e5b8758c9fbbd6d0ad.json
Number of files
3391
Example Article Information
['Data_sets', 'Generating genomic platforms to study Candida albicans pathogenesis',
['Over the last decade, there has been an exponential growth in the quantity of avai
lable genome sequence data due to the very rapid progress in sequencing technology.
In 2004, the genome sequence of the human fungal pathogen Candida albicans was relea
sed as Assembly 19 (1). With the challenge of working with a heterozygous diploid or
ganism, new computational methods had to be developed and resulted in the release in
2013 of Assembly 22 , an assembly of a completely phased diploid genome sequence for
the standard C. albicans reference strain SC5314 (2) . This opened new perspectives
to understand the genetic basis and functional mechanisms that underlie pathogenesis
and evolution in this organism. Although C. albicans gene sequences have been availa
ble to the community for more than a decade (3, 4) , only 1,670 out of the 6,198 pre
dicted protein-coding genes have been characterized as of January 31st, 2018 accordi
ng to the Candida Genome Database (5) . Nowadays, the growing availability of whole-
genome datasets has encouraged a shift towards the development of functional genomic
s and systems biology, enabling analysis of highthroughput whole-genome assays to be
tter understand biological networks. In this context, major efforts have been made t
o generate large-scale deletion (6) (7) (8) (9) or overexpression (10) (11) (12) mut
ant collections. Genome-wide ORF libraries or ORFeomes represent useful resources fo
r the implementation of approaches used to elucidate gene function (13) . Besides th
e development of collections of overexpression mutants (11) , ORFeomes facilitate ap
proaches aimed at evaluating protein subcellular localization and identifying protei
n-protein interactions (yeast two-hybrid) both at steady state and in response to en
vironmental stimuli (14) . Large-scale cloning projects, with the goal of cloning al
l predicted ORFs into flexible recombinational vectors, have been described for seve
ral model organisms, including powerful association of the C. albicans ORFeome and a
C. albicans-adapted two-hybrid system (32) that will allow systematic protein-protei
n interaction screening in C. albicans. C. albicans strains used in this study are l
isted in Table 1 . C. albicans strains were routinely cultured at 30°C in YPD medium
(1% yeast extract, 2% peptone, 2% dextrose), synthetic dextrose (SD) medium (0.67% y
east nitrogen base, 2% dextrose) or nourseothricincontaining YPD medium (YPD + 200 µ
g/ml Nourseothricin). Solid media were obtained by adding 2% agar. E. coli strains w
ere routinely cultured at 30°C or 37°C in LB or 2YT supplemented with 10 µg/ml genta
micin, 50 µg/ml kanamycin or 50 µg/ml ticarcillin. Solid media were obtained by addi
ng 2% agar. DH5α E. coli cells competent for transformation with DNA were prepared a
ccording to Hanahan et al. (33) . The ORFeome collection is being distributed by the
Centre International de Ressources Microbiennes (CIRM). The 49 expression vectors ar
e available from Addgene, while twohybrid strains and vectors are now accessible fro
m the Belgian Co-ordinated Collections of Micro-organisms (BCCM). Primer design. The
DNA sequence of C. albicans was obtained from the Candida Genome Database Release 21

(www.candidagenome.org), before Assembly 22 was released. Genespecific forward prime rs were designed by adding the sequence 5\'-GGGGACAAGTTTGTACAAAAAAGCAGGCTTG-3\' to t he 5\' end of the first 30 author/funder. All rights reserved. No reuse allowed with out permission. The copyright holder for this preprint (which was not peer-reviewed) is the . https://doi.org/10.1101/261628 doi: bioRxiv preprint nucleotides of each OR F. Gene-specific reverse primers were designed by adding the sequence 5\'-GGGGACCACT TTGTACAAGAAAGCTGGGTC-3\' to the 5\' end of the last 30 nucleotides of each ORF exclu ding the Stop codon. PCR fragments amplified from these primers are compatible for C -terminal tagging with destination vectors containing the Gateway™ cassette A. A tot al of 6,205 primer pairs were obtained from Invitrogen in a 96well format and are li sted in Supplemental Table S1 . Gateway™ cloning of the C. albicans ORFeome. The det ailed method for the cloning of C. albicans ORFs in the pDONR207 vector has been des cribed (34) . Briefly, ORFs, ranging from 90 bp to 5,295 bp, were amplified from gen omic DNA of C. albicans strain SC5314 in 96-well plates using the Thermo Scientific Phusion High-Fidelity DNA Polymerase and 30 cycles of amplification, with elongation time varying from 1 to 3 min according to the ORF size. After ethanol precipitation, Gateway™-compatible amplified ORFs were recombined into pDONR207 (Invitrogen) using the Gateway™ BP Clonase™ II Enzyme Mix (Invitrogen). Reaction mixes containing pDONR 207, the PCR products and the BP Clonase™ were incubated overnight in 96-well plate s, at room temperature. After adding proteinase K (Invitrogen) and incubating 10 min at 37°C, the BP reactions were directly used for bacterial transformation. 45-50 µl of chemically competent E. coli DH5α were added to the BP reactions and incubated fo r 30 min at 4°C. After heat-shock at 42°C for 35 s, 150 µl of SOC medium (20 ml YPD + 2 ml LB + 1 ml Hepes) was added to the transformation reactions and the samples we re covered with breathing films and incubated for 1.5 h at 37°C with shaking. Then 1 00 µl of each sample were plated onto LB agar containing 10 µg/ml Gentamicin and inc ubated overnight at 37°C. The remainder of the transformation reactions were stored at -80°C in 30% glycerol. A single colony from each transformation reaction was inoc ulated in 400 µl 2YT + 10 µg/ml Gentamicin in 96 deep-well microplates. After 36hgro wth at 37°C, the cultures were used for plasmid extraction and for -80°C storage. Va lidation of entry clones by DNA sequencing and bioinformatics analysis. Sanger seque ncing of the 5\'-end of each BP clone was performed with the 207VER-F oligonucleotid e (see Supplemental Table 2 for oligonucleotide sequences except those used to ampli fy ORFs) to confirm that the expected ORF had been cloned. The 3\'-ends were also se quenced with the 207VER-R oligonucleotide to ensure that the oligonucleotides did no t carry deletions. All inserts validated by Sanger sequencing were systematically su bjected to full-length sequence analysis using Illumina technology as described in C hauvel et al. (11) . If nonsense or frameshift mutations were observed in a cloned O RF, another colony was checked, or the ORF reamplified and cloned again. If these at tempts were unsuccessful, cloning of the ORF was abandoned. Clones containing contig uous deletions or insertions of multiples of 3 bp were accepted. Missense mutations and those located within introns were accepted. It should be noted that further anal ysis of each mutation-containing ORF is required to determine whether or not mutatio ns affect the function of the ORF. CandidaOrfDB was created to integrate the data of the C. albicans ORFeome project and is available at http://candidaorfeome.eu. The da tabase is a relational database using the SQL database server Oracle with a web inte rface developed using J2EE. The alignment algorithm uses biojava3-alignment Java lib rary, implemented with a NeedlemanWunsch aligner, configured with a gap open penalty of 5 and a gap extension penalty of 2. The substitution matrix used is named "nuc-4_ 4" and was created by Todd Lowe (https://github.com/sbliven/biojava/blob/master/bioj ava3-alignment/src/main/resources/nuc-4_4.txt). The alignment algorithm was specific ally customized for CandidaOrfDB to work on segments of 500 codons, which provides a n optimized alignment performance for C. albicans average sequence length. author/fu nder. All rights reserved. No reuse allowed without permission. The copyright holder for this preprint (which was not peer-reviewed) is the . https://doi.org/10.1101/261 628 doi: bioRxiv preprint The C. albicans ORFeome clones are available at the Centre

International de Ressources Microbiennes (CIRM -https://www6.inra.fr/cirm_eng/). All destination plasmids constructed in this study were derived from CIp-P TET -GTW (11) and were confirmed by Sanger sequencing. Plasmid sequences have been submitted to GenBank. Accession numbers are given in Table 2 . All oligonucleotides used for PCR and/or sequencing are listed in Supplemental Table 2. SP cloning. First, a spacer sequence (SP) was amplified from the E. coli kanR gene borne on pCR-topo-blunt plasmid (Invitrogen) with primers SP3 and SP4. The PCR product was cloned in the SacII site of CIp-P TET -GTW, yielding CIp-P TET -GTW-SP, also referred to as pCA-Dest1100. This 390 bp-long sequence is flanked by the SP1 and SP2 Illumina pairedend sequencing primers 1 and 2 and can be used to insert molecular barcodes if expression plasmids are to be used in signature-tagged mutagenesis approaches. Epitope tags. We then inserted in frame tags either upstream of attR1 or downstream of attR2 to allow N-terminal or C-terminal protein tagging, respectively. For the cloning of N-terminal tags, we used pUC-attR1-CmR, containing a PciI-BglII fragment from CIp-GTW cloned into the PciI and BamHI sites of pUC18. CIp-GTW was obtained after deleting P TET from CIp-P TET -GTW with Acc65I and HpaI. For cloning of the 3xHA coding sequence, two oligonucleotides (oligo1_PciHABsrG and oligo2_PciHABsrG) containing three HA epitopes were hybridized, gel purified, and cloned into pUC-attR1-CmR cut with PciI and BsrGI. The resulting plasmid was then used as a PCR template with primers GTW02 and GTW03. The resulting PCR product was cut with BspEI and HpaI, and cloned into the same sites of pCA-Dest1100 to yield pCA-Dest1110. For cloning of the GFP tag, the GFP gene was amplified from pFA-GFP-URA3 (35) with primers GTW04 and GTW05. The PCR product was cut with PciI and EcoRV and ligated into the same sites of pUC-attR1-CmR. The resulting plasmid was then used as a PCR template with primers GTW07 and GTW03. The resulting PCR product was cut with ScaI and BspEI and ligated into pCA-Dest1100 cut with HpaI and BspEI, yielding pCA-Dest1120. For cloning of the TAP-tag, the TAP-tag coding region was PCR amplified from pFA-TAP-URA3 (11) with primers GTW13 and GTW14. The resulting PCR product was then cut with PciI and BsrGI, and ligated into the same sites of pUC-attR1-CmR. The resulting plasmid was then digested with EcoRV and BspEI and the smallest restriction fragment was ligated into pCA-Dest1100 cut with HpaI and BspEI to yield pCA-Dest1130. For the cloning of C-terminal tags, we generated pUC-attR2 by inserting a SalI-HindIII fragment from pCA-Dest1100 into the same sites of pUC18. For cloning of the 3xHA epitope, pCaMPY-3xHA (36) was used as a PCR template with primers GTW20 and GTW21, and the resulting PCR product cut with BsrGI and NsiI was ligated in the same sites of pUC-attR2. The resulting plasmid was cut with SalI and NsiI and the fragment cloned in the same sites of pCA-Dest1100, yielding pCA-Dest1101. For cloning the GFP tag, the GFP gene was amplified from pFA-GFP-URA3 (35) with primers GTW11 and GTW12. The resulting PCR product was cut with NsiI and EcoRV and ligated into the same sites of pUC-attR2. The resulting plasmid was then cut with SalI and NsiI and the GFP-bearing fragment ligated into the same sites of pCA-Dest1100, yielding pCA-Dest1102. For cloning of the TAP-tag, the TAP-tag coding region was amplified from CIp10-P PCK1 -GTW-TAPtag (11) with primers GTW15 and GTW16. The PCR product was cut with BsrGI and NsiI and cloned in the same sites of pUC-attR2. The resulting plasmid was digested with SalI and NsiI, and the TAP-containing fragment ligated in the same sites of pCA-Dest1100, yielding pCA-Dest1103. author/funder. All rights reserved. No reuse allowed without permission. The copyright holder for this preprint (which was not peer-reviewed) is the . https://doi.org/10.1101/261628 doi: bioRxiv preprint Promoters. We replaced P TET with either P PCK1 (obtained from CIp10-P PCK1 -GTW-TAPtag, (11)), yielding the pCA-Dest12xx series; P TDH3 (cut from pKS-P TDH3 ; P TDH3 was PCR amplified from SC5314 genomic DNA with oligos SZ11 and SZ12, and the fragment inserted into pBluescript-KS (+) cut with XhoI and EcoRV) yielding the pCA-Dest13xx series; or P ACT1 , cut from CIp10::P ACT1 -gLUC59 (37), yielding the pCA-Dest14xx series. Selection markers. In all plasmids except the P TET series, the auxotrophic marker URA3 was replaced with the NAT1 marker conferring nourseothricin resistance. Briefly, a XbaI-DraIII fragment or SpeI-NaeI fragment was cut from pUC-NAT1 and inserted into the similarly cut vectors. pUC-NAT1 was built by amplifying NAT1 u

nder the control of P TEF1 promoter with oligonucleotides UZ50 and UZ51, and using p FA6-SAT1 (38) as a template, and cloning the AatII-cut fragment into the same site o f pUC18. The 49 expression vectors are available from Addgene (https://www.addgene.o rg). Construction of C. albicans UME6-overexpression strains. Detailed methods for t he transfer of C. albicans ORFs from pDONR207 into the expression plasmids as well a s the integration of the resulting expression plasmids at the RPS1 locus have been d escribed (34) . Briefly, the UME6 ORF was transferred from the entry clone into each one of 15 expression vectors by using the Gateway™ LR Clonase™ II Enzyme Mix (Invitr ogen). After E. coli transformation, the plasmids were verified by EcoRV digest. The URA3-and NAT1-bearing expression plasmids were digested by StuI and transformed into C. albicans strain CEC161 or CEC369, respectively, according to Walther and Wendland (39) . Transformants were selected for prototrophy or nourseothricin resistance, res pectively, and verified by PCR using primer CIpUL with (i) primer CIpUR for the URA3 -bearing plasmids and (ii) primer author/funder. All rights reserved. No reuse allow ed without permission. The copyright holder for this preprint (which was not peer-re viewed) is the . https://doi.org/10.1101/261628 doi: bioRxiv preprint CgSAT1-rev for the SAT1-bearing plasmids, that yield 1 kb and 1.6 kb products, respectively, if int egration of the OE plasmid has occurred at the RPS1 locus. Induction of the Tet-On s ystem. Overexpression from P TET was achieved by the addition of anhydrotetracycline (ATc, 3 µg/ml; Fisher Bioblock Scientific) in YPD at 30°C (40) . Overexpression expe riments were carried out in the dark, as ATc is light sensitive. Microscope analysis for filamentation. Cells were observed with a Leica DM RXA microscope (Leica Microsy stems) with an x40 oil-immersion objective. For the P TET , P TDH3 and P ACT1 strain s, a 30 ml culture in YPD or YPD+ATc3 was inoculated at OD600 = 0.2 with a freshly g rown colony and incubated at 30°C with shaking until OD600 reaches ~1. For the P PCK 1 strain, cells were scraped off the agar of Petri dish cultures and resuspended in 1 ml dH 2 O after centrifugation. The cell suspension was used to inoculate a 40 ml SD culture at OD600 = 0.2. 20 ODs of exponentially growing cells were collected by c entrifugation and resuspended in lysis buffer (9M Urea, 1.5% w/v dithiothreitol, 2- 4% w/v CHAPS, and 1.5 M Tris pH9.5). Homogenization of the cells was achieved using a FastPrep bead beater. After homogenization, the lysed cells were centrifuged at 1 3,000 rpm for 10 min at room temperature, and the supernatant containing the solubil ized proteins was used directly or stored at -80 °C. Proteins were separated on an I nvitrogen 8% NuPage gel, transferred onto nitrocellulose. TAP-and 3HA-tagged protein s were detected using peroxidase-coupled anti-peroxidase and anti-HA-peroxydase anti bodies (Sigma and Roche, respectively) and an ECL kit (GE Healthcare). author/funde r. All rights reserved. No reuse allowed without permission. The copyright holder fo r this preprint (which was not peer-reviewed) is the . https://doi.org/10.1101/26162 8 doi: bioRxiv preprint Mating-compatible strains were generated by deletion of MTLa or MTLα locus using the SAT1 flipper cassette (38) in the two-hybrid strain backgrou nd SC2H3 (32) . To delete the MTLa locus, flanking regions were amplified with prime rs MTLa-5\'F and MTLa-5\'R, and with MTLa-3\'F and MTLa-3\'R. To delete the MTLα loc us, homologous regions were amplified with primers MTLα-5\'F and MTLα-5\'R, and with MTLα-3\'F and MTLα-3\'R. Both fragments for each locus were cloned into pSAT1 (38) , at SacI/NotI and XhoI/KpnI sites respectively. SacI/KpnI deletion constructs were tr ansformed in SC2H3 (32) and transformants were selected on YPD medium containing 200 µg/ml of nourseothricin. Positive SC2H3a and SC2H3α strains were subsequently grown in maltose medium for induction of the recombinase and loss of the SAT1 flipper cass ette. For induction of opaque switching, a WOR1 fragment was amplified with primers WOR1F and WOR1R and cloned into pNIM1 vector (40) at SalI and BglII sites. SC2H3a an d SC2H3α strains were transformed with pNIM1-WOR1 with selection on nourseothricinco ntaining medium. The resulting strains, SC2H3a-pWOR1 and SC2H3α–pWOR1, were grown in presence of doxycycline (50 µg/ml) to induce the opaque state (41) . Opaque mating-c ompatible strains were finally transformed with prey or bait plasmid and selected on SC-ARG or SC-LEU respectively. Bait and prey two-hybrid vectors, pC2HB and pC2HP res pectively (32) were converted into Gateway™ destination vectors using the Gateway™ V

ector Conversion System to generate pC2HB-GC and pC2HP-GC respectively. The genes en coding the bait protein Hst7 (Orf19.469) and the prey Cek1 (Orf19.2886) were transfe rred by LR reactions from the donor collection of vectors into pC2HB-GC and pC2HP-GC destination vectors respectively. author/funder. All rights reserved. No reuse allow ed without permission. The copyright holder for this preprint (which was not peer-re viewed) is the . https://doi.org/10.1101/261628 doi: bioRxiv preprint Opaque cells o f SC2H3a and SC2H3α, expressing VP16-Cek1 (Arg+) and lexA-Hst7 (Leu+) respectively, were crossed as follows. Opaque cells of both types were mixed at 1 x 10 6 cells eac h in Spider medium (42) and incubated at 23°C for 24h. Resulting tetraploid cells we re selected on SC-LEU-ARG medium for 48h incubation at 30°C, and transferred to SC-H IS-MET medium for protein-protein interaction detection. Chromosome loss was induced on pre-sporulation medium at 37°C for 10 days as described previously (43) . Our obj ective was to develop a C. albicans ORFeome encompassing as many of the 6,205 ORFs p redicted in Assembly 21 of the C. albicans genome as possible (44) . To this aim, fo rward and reverse oligonucleotides with, respectively, attB and attP sequences at th eir 5\' ends were synthesized for each of the 6,205 ORFs (Supplemental Table S1), u sed in independent PCR reactions with genomic DNA of C. albicans strain SC5314 (45) and the resulting PCR products were cloned in pDONR207 using Gateway™ recombinationa l cloning (30) . The resulting plasmids were then subjected to Sanger sequencing at the 5\' and 3\' ends of the cloned ORFs and to Illumina sequencing throughout the cl oned ORFs. 46)). We defined 4 groups: (i) 4,209 sequences that are 100% identical w ith a reference author/funder. All rights reserved. No reuse allowed without permiss ion. HapA/B; Fig. 1B) . When nucleotide differences were present in the two alleles of the reference ORF sequence, we did not notice any bias in the haplotype that was cloned: haplotype A ORFs were cloned in 33% of cases (HapA; Fig. 1B) and haplotype B ORFs were cloned in 31% of cases (HapB; Fig. 1B) . On chromosomes R, 1, 2, 3, 4, 5 a nd 7, ORF cloning was uniformly successful, with a cloning success rate ranging from ~82% to ~86%. Chromosome 6 ORFs were slightly underrepresented (76.9%) (Fig. 1C). Although ORFs up to 1.5kb could be cloned with a success rate around 87% and ORFs be tween 1.5 and 4kb could be cloned with a success rate around 77%, we observed a decr eased success rate for the longer ORFs (Fig. 1D) . The analysis of the 6,880,620 nuc leotides of the 5,102 sequence-validated ORFs revealed 2077 SNPs. These nucleotide s ubstitutions could be real SNPs between our strain and the reference strain, or due to poor quality of some reference sequences, or mutations in the primers, or misinco rporation by the DNA polymerase. Overall, the resulting maximum error rate amounts t o 3 x 10 -4 or 1 SNP every 3,312 cloned nucleotides. Notably, the C. albicans SC5314 genome sequence available at CGD (version_A22-s07-m01-r18) contained 272 ORFs with s equence ambiguities for at least one of the alleles. These ambiguities included stre tches of \'Ns\' or IUPAC code nucleotides (Y, R, S, M, K, W). The ORFeome-generated sequencing data resolved sequence ambiguities for 110 of these ORFs (Supplemental Ta bles author/funder. All rights reserved. No reuse allowed without permission. The co pyright holder for this preprint (which was not peer-reviewed) is the . https://doi. org/10.1101/261628 doi: bioRxiv preprint 3 and 4). In this study, we also detected r ecombinant haplotypes for 116 of the cloned ORFs (Supplemental Table 5) that displa y characteristics of both reference haplotypes. Taken together, our study provides a n almost comprehensive, extremely high quality C. albicans ORFeome. CandidaOrfDB was created to integrate the data of the C. albicans ORFeome project and is available at http://candidaorfeome.eu. CandidaOrfDB enables the scientific community to search fo r availability and quality of the clones. All data pertinent to the cloning process are stored in the database. This information includes the name and sequence of the p rimers used to amplify the ORFs from C. albicans SC5314 genomic DNA, the coordinates of the primers in their 96-well storage plates, and any sequencing information avail able on the clones. SNPs and DIPs, their location in the ORF and their influence on the corresponding amino acid sequences are shown (Fig. 2) . Nucleotide sequence of t he cloned ORF and the corresponding amino acid sequence are displayed with highlight s of SNPs or DIPs relative to the closest haplotype sequence (Fig. 2) . The referenc

e sequence data have been extracted from CGD (46) . The database can be queried thro ugh the ORF number (orf19.xxxx or 19.xxxx; (5)), the Assembly 22 name (Ci_XXXXXW or Ci_XXXXXC; (5)) or the gene name. The C. albicans ORFeome described above was develo ped in pDONR207 as this allows subsequent transfer of the cloned ORFs in Gateway™-ad apted destination vectors. While destination vectors for ORF expression in hosts suc h as E. coli or S. cerevisiae are available (47, 48) , only a few destination vector s for ORF expression in C. albicans have been reported (11) . Therefore, we set out to establish a collection of Gateway™-adapted destination vectors for constitutive or conditional expression of untagged or tagged OR Fs in C. albicans. Our primary collection of 15 C. albicans destination vectors deri ved from CIp10S (11) provides a choice of two promoters, either the constitutive pro moter P TDH3 or the inducible promoter P TET , and the option for N-or C-terminal fu sion to various epitope tags (3xHA or TAP) (Table 2 and Fig. 3) . Integration of the destination vectors and their derivatives at the RPS1 locus in the C. albicans genom e is promoted by StuI or I-SceI linearization. These CIp10derived vectors carry eith er the auxotrophic marker URA3 or the NAT1 marker that confers resistance to nourseo thricin and can be used with clinical isolates. However, P TET -bearing plasmids can not carry the NAT1 marker since the transactivator needed for tetracyclinemediated i nduction of P TET is borne on the NAT1-carrying pNIMX plasmid (11) . All vectors har bor a so-called spacer sequence (SP) originating from the E. coli kanamycin resistan ce gene, for barcode insertion and subsequent Illumina-based barcode sequencing. A s econd set of 34 plasmids was also generated with the constitutive promoters P ACT1 o r the inducible promoter P PCK1 , and the option for N-or C-terminal fusion to GFP (Table 2). All 49 destination vectors were designated with a standardized nomenclatu re, pCA-DESTijkl, whereby i stands for the transformation marker (1, URA3; 2, NAT1); j stands for the promoter (1, P TET ; 2, P PCK1 ; 3, P TDH3 ; 4, P ACT1); k stands for N-terminal tagging (0, no tag; 1, 3xHA; 2, GFP; 3, TAP); and l stands for C-term inal tagging (0, no tag; 1, 3xHA; 2, GFP; 3, TAP) (Table 2 and Fig. 3) . All destin ation vectors are available from Addgene (https://www.addgene.org). In order to vali date that the primary set of 15 destination vectors could drive constitutive or cond itional expression of untagged or tagged ORFs, the UME6 ORF was transferred into eac h of them using LR Clonase™. UME6 encodes a transcription factor whose overexpressio n has been shown to force filamentation in C. albicans (49, 50) . The validation cri teria for the destination vectors were (i) success of the LR reaction, (ii) integrat ion in the C. albicans genome at the RPS1 locu s, (iii) filamentation phenotype and/or (iv) detection of the 3xHA-or TAP-tagged Ume 6 protein by western blot. The LR reaction was successfully used to transfer UME6 fr om the pDONR207::UME6 plasmid to all 15 destination vectors. The resulting expressio n plasmids were successfully integrated at the C. albicans RPS1 locus. The 5 strains containing P TET were induced overnight at 30°C by adding 3 µg/ml anhydrotetracyclin e (ATc3) to the culture medium, while the 10 strains containing P TDH3 were grown ov ernight at 30°C in YPD. As described in Chauvel et al. (11) , overexpression of Ume6 led to hyperfilamentation. This phenotype was observed for all strains, regardless t he tags (Fig. 4A and 4B) . 3xHA-or TAP-tagged Ume6 was detected in the relevant str ains and under the appropriate growth conditions (Fig. 5A and 5B) . Taken together, these data present a collection of 49 available destination vectors and validate 15 of those for application with the C. albicans ORFeome. The availability of the ORFeo me collection is a prerequisite to large, systematic two-hybrid (2H) screenings in C. albicans. The Candida 2H (C2H) system was developed for targeted one to one prote in-protein interaction detection (32) . Hence, there was a need to engineer the syst em for high throughput screening based on rapid and convenient recombination cloning systems for the generation of prey/bait arrays of proteins, and to use a mating appr

oach to circumvent the low efficiency of transformation of this fungus. The former w as obtained by adapting the prey and bait recipient vectors for Gateway™-mediated tr ansfer of the ORFeome library, yielding pC2HB-GC and pC2HP-GC, respectively. The lat ter involved the combined processes of white-opaque switching and the preparation of mating-compatible 2H strains. The pleiomorphic fungus C. albicans is characterized b y a parasexual cycle, with the author/funder. All rights reserved. No reuse allowed without permission. The copyright holder for this preprint (which was not peer-revie wed) is the . https://doi.org/10.1101/261628 doi: bioRxiv preprint possible mating o f diploids, and generation of tetraploids, the absence of meiosis and the reversion to the diploid state by chromosome loss (reviewed in (51)). An essential prerequisit e to mating in C. albicans is the phenotypic switch from white to opaque, a stable b ut reversible switch regulated by environmental stimuli such as low temperatures, ca rbon dioxide and nutrients. The epigenetic switch is tightly linked to mating since the a1/α2 complex, encoded at the mating type like (MTL) loci, acts as a r epressor of opaque switching (52) . In that context, the two-hybrid strain SC2H3 ori ginating from the diploid a/α SN152 strain was deleted for the MTLa or MTLα locus to construct SC2H3α and SC2H3a strains. The efficient switch to opaque in the hemizygot e strains was achieved by the doxycyclinecontrolled expression of WOR1, a major regu lator of the white to opaque switch (41, 53) . A proof of principle of the whole pro cedure for protein-protein detection is shown in Figure 6 . The MAP kinase kinase Hst7 was used as bait and linked to the DNA binding domain lexA, as part of the pC2H B-GC vector. Similarly, the Cek1 kinase was expressed as a prey protein, fused to th e activation domain VP16, component of the pC2HP-GC vector. pC2HB-Hst7 and pC2HP-Cek 1 were transformed into opaque SC2H3α and SC2H3a strains, respectively. Mating of t he resulting transformants was achieved through selection for leucine and arginine p rototrophy. Upon interaction of the two proteins, the reconstituted transcriptional module promoted the expression of the HIS4 marker, allowing only the mating-derived strains to grow on a histidine-free medium, whereas the original diploid strains cou ld not (Fig. 6B) . Through a mechanism of chromosome loss, the crossed tetraploid st rains return back to a diploid state. We showed that the plasmids were stably expres sed in the diploid offspring strains as these strains conserved their ability to gro w in absence of arginine and leucine, indicating the presence of the prey and bait v ectors respectively. In addition, all these strains were able to grow in absence of histidine, indicative of the targeted protein-protein interaction. author/funder. Al l rights reserved. No reuse allowed without permission. The copyright holder for thi s preprint (which was not peer-reviewed) is the . https://doi.org/10.1101/261628 do i: bioRxiv preprint DISCUSSION A partial C. albicans Gateway™-adapted ORFeome (ORFeo meV1; 644 ORFs) and the corresponding C. albicans overexpression strains have alread y been successfully used to investigate morphogenesis, biofilm formation and genome dynamics in C. albicans (11, 31, 54) . In this study, we have now generated the C. a lbicans ORFeomeV2 within the (15, 16, 18, 21, 23, 29) . Given these high standards, the success rate of 83% that we achieved with the C. albicans ORFeomeV2 is quite rem arkable. In eukaryotes, ORFeomes are usually cloned from full-length cDNA libraries to account for the presence of introns and therefore splicing variants. However, onl y 6% of C. albicans genes have introns (55) , rendering the generation of the C. alb icans ORFeome more straightforward than for pluricellular eukaryotes. Indeed, ORFs w ere simply amplified from the genomic DNA of the reference strain SC5314 and therefo re, 360 ORFs in the C. albicans ORFeomeV2 harbor introns, a matter that should be ta ken into consideration when expressing ORFs in a prokaryotic host and, possibly, oth er eukaryotic hosts than C. albicans. A second aspect to be taken into consideration when using the ORFeome in hosts other than author/funder. All rights reserved. No re use allowed without permission. The copyright holder for this preprint (which was no t peer-reviewed) is the . https://doi.org/10.1101/261628 doi: bioRxiv preprint C. al bicans is the unusual codon usage in this species (56) . Indeed, in C. albicans, CUG is decoded as a serine instead of a leucine the majority of the time. Hence, improper translation in hosts with a standard genetic code may distort protein structure an

d function. Sanger sequencing of both 5\' and 3\' ends was performed to confirm ORF identity and exclude clones containing primer or recombination errors. Previous studies have reported mutation rates in primers of 3-10% (26, 29) . Similar rates were observed in this work. Unlike other Gateway™ recombinational cloning projects (15, 29) , we did not see major differences in cloning efficiency for ORFs up to 4 kb. A decrease in success rate was observed only for ORFs >4 kb. This size bias could be attributed to an increased difficulty to amplify the ORF by PCR, a reduced efficiency of the Gateway™ BP Clonase™ reactions with long PCR products, and the error rate of the PCR polymerase, which increases with longer products. In addition, we also corrected the sequences of 110 ORFs with sequence ambiguities (N-tracts and IUPAC code nucleotides) according to the Candida Genome Database and identified 116 ORFs that display a recombinant haplotype. Recombinant haplotypes could be explained by template switching during PCR cycles or by incorrectly phased SNPs in reference sequences. This could be addressed by performing long-read sequencing of the C. albicans SC5314 genome. However, many challenges remain to be addressed. For heterozygous genes, only one allele is present in the ORFeome. Because functional differences have been reported between the 2 alleles of a heterozygous gene (57, 58) , it would be relevant to also clone and validate the second allele in these instances. In addition, oligonucleotides have been designed based on annotation of the haploid set of Assembly 19 of the C. albicans genome. As a consequence, genes of the MTLα locus have not been included in the ORFeome. Despite our efforts, the collection is still missing 1,081 clones. The next version of the C. albicans ORFeome could extend gene coverage by adding the missing ORFs, allelic variants or strain-specific variations. ORFeomes are essential to bridge the gap between genome annotation and systems biology and allow large-scale gene and protein characterization. The development of a collection of C. albicans constitutive or conditional overexpression strains is one of the applications of the C. albicans ORFeome that we have successfully explored (11, 31, 54) . In this respect, the C. albicans ORFeome project is currently completing its second and third phases, which consist of transferring the 5,102 cloned ORFs into barcoded destination vectors and generating a collection of C. albicans barcoded overexpression mutants, each mutant carrying one of the 5,102 cloned ORFs under the control of the inducible P TET promoter. CandidaOrfDB already provides information on the available overexpression plasmids and strains. In this study, we have paved the way for a second application of the C. albicans ORFeome through the development of tools for a 2H matrix approach of protein-protein interaction detection via mating in C. albicans. Our data show that mating of diploid C. albicans expressing a bait fused to the LexA DNA binding domain and a prey fused to the VP16 activation domain, respectively, allows protein-protein interactions to be tested. In S. cerevisiae, large-scale 2H screens can be performed in a matrix design whereby haploid strains expressing baits and preys are mated and protein-protein interactions are scored in the resulting diploids. Our toolkit now enables the implementation of such a matrix design for large-scale 2H screens in C. albicans. To this aim, a collection of 1500 prey clones has already been generated. As mentioned above, C. albicans codon usage is unusual, which limits the development of applications of the C. albicans ORFeome in species that use standard decoding such as S. cerevisiae. Our development of tools for large-scale 2H screens in C. albicans circumvents this limitation, opening the path for the characterization of the C. albicans interactome. Defining the C. albicans interactome will undoubtedly impact our understanding of C. albicans pathogenesis in humans. In summary, high-level gene coverage coupled with the versatility of the Gateway™ recombinational cloning, C. albicans-adapted functional genomics tools and full access to clones, make the C. albicans ORFeomeV2 a unique and valuable resource for the scientific community th

```
at should greatly facilitate future functional studies in C. albicans. We are gratef
ul to other members of our groups for their support throughout the development of th
e C. albicans ORFeome. author/funder. All rights reserved. No reuse allowed without
permission. The copyright holder for this preprint (which was not peer-reviewed) is
the N-ter N-ter C-ter C-ter Promoter: Tag position: Promoter: Tag position: author/f
under. All rights reserved. No reuse allowed without permission. The copyright holde
r for this preprint (which was not peer-reviewed) is the . https://doi.org/10.1101/2
61628 doi: bioRxiv preprint ']]
```

In [ ]:
```
processed_articles.process_text()
```

```
Cleaning out Junk
Tokenizing words
Converting to list of words and removing stop words
Creating word stems
```

As provided in the covid 19 lab, we will be creating the `bigrams and trigrams` from the available words.

In [ ]:
```
processed_articles.trigrams([art[3] for art in processed_articles.processed_article
processed_articles.processed_trigrams[2][4]
```

```
Creating Bigrams
Creating Trigrams from Bigrams
```

Out[ ]:
"['last_decad', 'exponenti_growth', 'quantiti', 'avail', 'genom_sequenc', 'data',
'due', 'rapid_progress', 'sequenc', 'technolog', 'genom_sequenc', 'human', 'fungal
_pathogen', 'candida_albican', 'releas', 'assembl', 'challeng', 'work', 'heterozy
g', 'diploid', 'organ', 'new', 'comput', 'method', 'develop', 'result', 'releas',
'assembl', 'assembl', 'complet', 'phase', 'diploid', 'genom_sequenc', 'standard',
'albican', 'refer', 'strain', 'sc', 'open_new', 'perspect', 'understand', 'genet_b
asi', 'function', 'mechan_underli', 'pathogenesi', 'evolut', 'organ', 'although',
'albican', 'gene', 'sequenc', 'avail', 'commun', 'decad', 'predict', 'protein', 'c
ode', 'gene', 'character', 'januari_st', 'accord', 'candida', 'genom', 'databas',
'nowaday', 'grow', 'avail', 'whole_genom', 'dataset', 'encourag', 'shift_toward',
'develop', 'function', 'genom', 'system', 'biolog', 'enabl', 'analysi', 'highthrou
ghput', 'whole_genom', 'assay', 'better_understand', 'biolog', 'network', 'contex
t', 'major', 'effort_made', 'gener', 'larg_scale', 'delet', 'overexpress', 'mutan
t', 'collect', 'genom_wide', 'orf', 'librari', 'orfeom', 'repres', 'use', 'resour
c', 'implement', 'approach', 'use', 'elucid', 'gene', 'function', 'besid', 'develo
p', 'collect', 'overexpress', 'mutant', 'orfeom', 'facilit', 'approach', 'aim', 'e
valu', 'protein', 'subcellular_local', 'identifi', 'protein', 'protein', 'interac
t', 'yeast', 'two_hybrid', 'steadi_state', 'respons', 'environment_stimuli', 'larg
_scale', 'clone', 'project', 'goal', 'clone', 'predict', 'orf', 'flexibl', 'recomb
in', 'vector', 'describ', 'sever', 'model', 'organ', 'includ', 'power', 'associ',
'albican_orfeom', 'albican', 'adapt', 'two_hybrid', 'system', 'allow', 'systemat',
'protein', 'protein', 'interact', 'screen', 'albican', 'albican', 'strain', 'use',
'studi', 'list_tabl', 'albican', 'strain', 'routin', 'cultur', 'ypd', 'medium', 'y
east_extract_pepton', 'dextros', 'synthet', 'dextros', 'sd', 'medium', 'yeast_nitr
ogen', 'base', 'dextros', 'ypd', 'medium', 'ypd', 'ml', 'nourseothricin', 'solid_m
edia', 'obtain', 'ad', 'agar', 'coli_strain', 'routin', 'cultur', 'lb', 'yt', 'sup
plement', 'ml_gentamicin', 'ml_kanamycin', 'ml', 'ticarcillin', 'solid_media', 'ob
tain', 'ad', 'agar', 'dh_coli', 'cell', 'compet', 'transform', 'dna', 'prepar', 'a
ccord', 'hanahan', 'orfeom', 'collect', 'distribut', 'centr', 'intern', 'de_ressou
rc', 'microbienn', 'cirm', 'express_vector', 'avail', 'addgen', 'twohybrid', 'stra
in', 'vector', 'access', 'belgian', 'co_ordin', 'collect', 'micro_organ', 'bccm',
'primer_design', 'dna', 'sequenc', 'albican', 'obtain', 'candida', 'genom', 'datab
as', 'releas', 'www', 'candidagenom', 'org', 'assembl_releas', 'genespecif', 'forw
ard_primer', 'design', 'ad', 'sequenc', 'end', 'first', 'author_funder_right_reser
v', 'reus_allow_without_permiss', 'copyright_holder_preprint_peer', 'review_doi_bi
orxiv_preprint', 'nucleotid', 'orf', 'gene', 'specif', 'revers_primer', 'design',
'ad', 'sequenc', 'end', 'last', 'nucleotid', 'orf', 'exclud', 'stop_codon', 'pcr',
'fragment_amplifi', 'primer', 'compat', 'termin_tag', 'destin_vector', 'contain',
'gateway', 'cassett', 'total', 'primer_pair', 'obtain', 'invitrogen', 'well', 'for
mat', 'list', 'supplement_tabl', 'gateway_clone', 'albican_orfeom', 'detail', 'met
hod', 'clone', 'albican', 'orf', 'pdonr', 'vector', 'describ_briefli', 'orf', 'ran
g', 'bp_bp', 'amplifi', 'genom', 'dna', 'albican', 'strain', 'sc', 'well_plate',
'use', 'thermo_scientif', 'phusion_high_fidel', 'dna_polymeras', 'cycl_amplif', 'e
long', 'time', 'vari', 'min', 'accord', 'orf', 'size', 'ethanol_precipit', 'gatewa
y', 'compat', 'amplifi', 'orf', 'recombin', 'pdonr', 'invitrogen', 'use', 'gateway
_bp', 'clonas_ii', 'enzym', 'mix', 'invitrogen', 'reaction_mix', 'contain', 'pdon
r', 'pcr_product', 'bp_clonas', 'incub_overnight', 'well_plate', 'room_temperatu
r', 'ad', 'proteinas', 'invitrogen', 'incub_min', 'bp', 'reaction', 'directli', 'u
se', 'bacteri', 'transform', 'chemic_compet', 'coli_dh', 'ad', 'bp', 'reaction',
'incub_min', 'heat_shock', 'soc', 'medium', 'ml', 'ypd', 'ml', 'lb', 'ml', 'hepe',
'ad', 'transform', 'reaction', 'sampl', 'cover', 'breath', 'film', 'incub_shake',
'sampl', 'plate', 'onto_lb', 'agar', 'contain', 'ml_gentamicin', 'incub_overnigh
t', 'remaind', 'transform', 'reaction', 'store', 'glycerol', 'singl', 'coloni', 't
ransform', 'reaction', 'inocul', 'yt', 'ml_gentamicin', 'deep', 'well_micropl', 'h
growth', 'cultur', 'use', 'plasmid', 'extract', 'storag', 'valid', 'entri', 'clon
e', 'dna', 'sequenc', 'bioinformat_analysi', 'sanger_sequenc', 'end', 'bp', 'clon

e', 'perform', 'ver', 'oligonucleotid', 'see_supplement', 'tabl', 'oligonucleoti
d', 'sequenc', 'except', 'use', 'amplifi', 'orf', 'confirm', 'expect', 'orf', 'clo
ne', 'end', 'also', 'sequenc', 'ver', 'oligonucleotid', 'ensur', 'carri', 'delet_i
nsert', 'valid', 'sanger_sequenc', 'systemat', 'subject', 'full_length', 'sequen
c', 'analysi', 'use', 'illumina', 'technolog', 'describ', 'chauvel', 'nonsens_fram
eshift', 'mutat', 'observ', 'clone', 'orf', 'anoth', 'coloni', 'check', 'orf', 're
amplifi', 'clone', 'attempt', 'unsuccess', 'clone', 'orf', 'abandon', 'clone', 'co
ntain', 'contigu', 'delet_insert', 'multipl', 'bp', 'accept', 'missens_mutat', 'lo
cat_within', 'intron', 'accept', 'note', 'analysi', 'mutat', 'contain', 'orf', 're
quir', 'determin_whether', 'mutat', 'affect', 'function', 'orf', 'candidaorfdb',
'creat', 'integr', 'data', 'albican_orfeom', 'project', 'avail', 'databas', 'rela
t', 'databas', 'use', 'sql', 'databas', 'server', 'oracl', 'web_interfac', 'develo
p', 'use', 'jee', 'align_algorithm', 'use', 'biojava', 'align', 'java', 'librari',
'implement', 'needlemanwunsch', 'align', 'configur', 'gap', 'open', 'penalti_gap',
'extens', 'penalti', 'substitut_matrix', 'use', 'name', 'nuc', 'creat', 'todd', 'l
ow', 'align_algorithm', 'specif', 'custom', 'candidaorfdb', 'work', 'segment', 'co
don', 'provid', 'optim', 'align', 'perform', 'albican', 'averag', 'sequenc', 'leng
th', 'author_funder_right_reserv', 'reus_allow_without_permiss', 'copyright_holder
_preprint_peer', 'review_doi_biorxiv_preprint', 'albican_orfeom', 'clone', 'avai
l', 'centr', 'intern', 'de_ressourc', 'microbienn', 'cirm', 'destin', 'plasmid_con
struct', 'studi', 'deriv', 'cip_tet', 'gtw', 'confirm', 'sanger_sequenc', 'plasmi
d', 'sequenc', 'submit_genbank_access', 'number', 'given', 'tabl', 'use', 'pcr',
'sequenc', 'list', 'supplement_tabl', 'sp', 'clone', 'first', 'spacer', 'sequenc',
'sp', 'amplifi', 'coli', 'kanr', 'gene', 'born', 'pcr_topo', 'blunt', 'plasmid',
'invitrogen', 'primer', 'sp_sp', 'pcr_product', 'clone', 'sacii', 'site', 'cip_te
t', 'gtw', 'yield', 'cip_tet', 'gtw', 'sp', 'also', 'refer', 'pca_dest', 'bp_lon
g', 'sequenc_flank', 'sp_sp', 'illumina', 'pairedend', 'sequenc', 'primer', 'use',
'insert', 'molecular', 'barcod', 'express_plasmid', 'use', 'signatur', 'tag', 'mut
agenesi', 'approach', 'epitop_tag', 'insert', 'frame', 'tag', 'either', 'upstrea
m', 'attr', 'downstream', 'attr', 'allow', 'termin', 'termin', 'protein', 'tag',
'respect', 'clone', 'termin_tag', 'use', 'puc_attr', 'cmr', 'contain', 'pcii', 'bg
lii', 'fragment', 'cip', 'gtw', 'clone', 'pcii', 'bamhi_site', 'puc', 'cip', 'gt
w', 'obtain', 'delet', 'tet', 'cip_tet', 'gtw', 'acci', 'hpai', 'clone', 'xha', 'c
ode_sequenc', 'two', 'oligo', 'pcihabsrg', 'oligo', 'pcihabsrg', 'contain', 'thre
e', 'ha', 'epitop', 'hybrid', 'gel_purifi', 'clone_puc', 'attr_cmr', 'cut', 'pci
i', 'bsrgi', 'result', 'plasmid', 'use', 'pcr', 'templat', 'primer_gtw_gtw', 'resu
lt', 'pcr_product', 'cut', 'bspei', 'hpai', 'clone', 'site', 'pca_dest', 'yield_pc
a_dest', 'clone', 'gfp_tag', 'gfp', 'gene', 'amplifi', 'pfa', 'gfp', 'ura', 'prime
r_gtw_gtw', 'pcr_product', 'cut', 'pcii', 'ecorv', 'ligat', 'site', 'puc_attr', 'c
mr', 'result', 'plasmid', 'use', 'pcr', 'templat', 'primer_gtw_gtw', 'result', 'pc
r_product', 'cut', 'scai', 'bspei', 'ligat', 'pca_dest', 'cut', 'hpai', 'bspei',
'yield_pca_dest', 'clone', 'tap_tag', 'tap_tag', 'code_region', 'pcr_amplifi', 'pf
a', 'tap', 'ura', 'primer_gtw_gtw', 'result', 'pcr_product', 'cut', 'pcii', 'bsrg
i', 'ligat', 'site', 'puc_attr', 'cmr', 'result', 'plasmid', 'digest_ecorv', 'bspe
i', 'smallest', 'restrict', 'fragment_ligat', 'pca_dest', 'cut', 'hpai', 'bspei',
'yield_pca_dest', 'clone', 'termin_tag', 'gener', 'puc_attr', 'insert', 'sali_hind
iii', 'fragment', 'pca_dest', 'site', 'puc', 'clone', 'xha', 'epitop', 'pcampi',
'xha', 'use', 'pcr', 'templat', 'primer_gtw_gtw', 'result', 'pcr_product', 'cut',
'bsrgi', 'nsii', 'ligat', 'site', 'puc_attr', 'result', 'plasmid', 'cut', 'sali',
'nsii', 'fragment_clone', 'site', 'pca_dest', 'yield_pca_dest', 'clone', 'gfp_ta
g', 'gfp', 'gene', 'amplifi', 'pfa', 'gfp', 'ura', 'primer_gtw_gtw', 'result', 'pc
r_product', 'cut', 'nsii', 'ecorv', 'ligat', 'site', 'puc_attr', 'result', 'plasmi
d', 'cut', 'sali', 'nsii', 'gfp', 'bear', 'fragment_ligat', 'site', 'pca_dest', 'y
ield_pca_dest', 'clone', 'tap_tag', 'tap_tag', 'code_region', 'amplifi', 'cip', 'p
ck', 'gtw', 'taptag', 'primer_gtw_gtw', 'pcr_product', 'cut', 'bsrgi', 'nsii', 'cl
one', 'site', 'puc_attr', 'result', 'plasmid', 'digest_sali', 'nsii', 'tap', 'cont

ain', 'fragment_ligat', 'site', 'pca_dest', 'yield_pca_dest', 'author_funder_right
_reserv', 'reus_allow_without_permiss', 'copyright_holder_preprint_peer', 'review_
doi_biorxiv_preprint', 'promot', 'replac', 'tet', 'either', 'pck', 'obtain', 'ci
p', 'pck', 'gtw', 'taptag', 'yield', 'pca', 'destxx', 'seri', 'tdh', 'cut', 'pk',
'tdh_tdh', 'pcr_amplifi', 'sc', 'genom', 'dna', 'oligo', 'sz_sz', 'fragment', 'ins
ert', 'pbluescript_ks', 'cut', 'xhoi', 'ecorv', 'yield', 'pca', 'destxx', 'seri',
'act', 'cut', 'cip', 'act', 'gluc', 'yield', 'pca', 'destxx', 'seri', 'select', 'm
arker', 'plasmid', 'except', 'tet', 'seri', 'auxotroph_marker', 'ura', 'replac',
'nat', 'marker', 'confer', 'nourseothricin', 'resist', 'briefli', 'xbai', 'draii
i', 'fragment', 'spei', 'naei', 'fragment', 'cut', 'puc', 'nat', 'insert', 'simila
rli', 'cut', 'vector', 'puc', 'nat', 'built', 'amplifi', 'nat', 'control', 'tef',
'promot', 'uz', 'uz', 'use', 'pfa', 'sat', 'templat', 'clone', 'aatii', 'cut', 'fr
agment', 'site', 'puc', 'express_vector', 'avail', 'addgen', 'construct', 'albica
n', 'ume', 'overexpress', 'strain', 'detail', 'method', 'transfer', 'albican', 'or
f', 'pdonr', 'express_plasmid', 'well', 'integr', 'result', 'express_plasmid', 'r
p', 'locu', 'describ_briefli', 'ume', 'orf', 'transfer', 'entri', 'clone', 'one',
'express_vector', 'use', 'gateway', 'lr_clonas', 'ii', 'enzym', 'mix', 'invitroge
n', 'coli_transform', 'plasmid', 'verifi', 'ecorv_digest', 'ura', 'nat', 'bear',
'express_plasmid', 'digest', 'stui', 'transform', 'albican', 'strain', 'cec', 'ce
c', 'respect', 'accord', 'walther', 'wendland', 'transform', 'select', 'prototroph
i', 'nourseothricin', 'resist', 'respect', 'verifi', 'pcr', 'use', 'primer', 'cipu
l', 'primer', 'cipur', 'ura', 'bear', 'plasmid', 'ii', 'primer', 'author_funder_ri
ght_reserv', 'reus_allow_without_permiss', 'copyright_holder_preprint_peer', 'revi
ew_doi_biorxiv_preprint', 'cgsat', 'rev', 'sat', 'bear', 'plasmid', 'yield', 'kb_k
b', 'product', 'respect', 'integr', 'oe', 'plasmid', 'occur', 'rp', 'locu', 'induc
t', 'tet', 'system', 'overexpress', 'tet', 'achiev', 'addit', 'atc', 'ml', 'fishe
r', 'bioblock', 'scientif', 'ypd', 'overexpress', 'experi_carri', 'dark', 'atc',
'light', 'sensit', 'microscop', 'analysi', 'filament', 'cell', 'observ', 'leica_d
m', 'rxa', 'microscop_leica', 'microsystem', 'oil_immers_object', 'tet', 'tdh', 'a
ct', 'strain', 'ml', 'cultur', 'ypd', 'ypd', 'atc', 'inocul', 'od', 'freshli', 'gr
own', 'coloni', 'incub_shake', 'od_reach', 'pck', 'strain', 'cell', 'scrape', 'aga
r', 'petri_dish', 'cultur', 'resuspend_ml', 'dh', 'centrifug', 'cell', 'suspens',
'use', 'inocul', 'ml', 'sd', 'cultur', 'od_od', 'exponenti_grow', 'cell', 'collec
t', 'centrifug_resuspend', 'lysi_buffer', 'urea', 'dithiothreitol', 'chap', 'tri_p
h', 'homogen', 'cell', 'achiev', 'use', 'fastprep', 'bead_beater', 'homogen_lyse',
'cell', 'centrifug_rpm', 'min_room_temperatur', 'supernat', 'contain', 'solubil',
'protein', 'use', 'directli', 'store', 'protein', 'separ', 'invitrogen', 'nupag_ge
l', 'transfer_onto', 'nitrocellulos', 'tap', 'ha_tag', 'protein', 'detect', 'use',
'peroxidas', 'coupl', 'anti', 'peroxidas', 'anti_ha', 'peroxydas', 'antibodi', 'si
gma', 'roch', 'respect', 'ecl', 'kit_ge_healthcar', 'author_funder_right_reserv',
'reus_allow_without_permiss', 'copyright_holder_preprint_peer', 'review_doi_biorxi
v_preprint', 'mate', 'compat', 'strain', 'gener', 'delet', 'mtla', 'mtl_locu', 'us
e', 'sat', 'flipper', 'cassett', 'two_hybrid', 'strain', 'background', 'sch', 'del
et', 'mtla', 'locu', 'flank_region', 'amplifi_primer', 'mtla', 'mtla', 'mtla', 'mt
la', 'delet', 'mtl_locu', 'homolog', 'region', 'amplifi_primer', 'mtl', 'mtl', 'mt
l', 'mtl', 'fragment', 'locu', 'clone', 'psat', 'saci', 'noti_xhoi', 'kpni_site',
'respect', 'saci_kpni', 'delet', 'construct', 'transform', 'sch', 'transform', 'se
lect', 'ypd', 'medium_contain', 'ml', 'nourseothricin', 'posit', 'scha', 'sch', 's
train', 'subsequ', 'grown', 'maltos', 'medium', 'induct', 'recombinas', 'loss', 's
at', 'flipper', 'cassett', 'induct', 'opaqu', 'switch', 'wor', 'fragment_amplifi',
'primer', 'worf', 'worr', 'clone', 'pnim', 'vector', 'sali', 'bglii_site', 'scha',
'sch', 'strain', 'transform', 'pnim', 'wor', 'select', 'medium', 'result', 'strai
n', 'scha', 'pwor', 'sch', 'pwor', 'grown', 'presenc', 'doxycyclin', 'ml', 'indu
c', 'opaqu', 'state', 'opaqu', 'mate', 'compat', 'strain', 'final', 'transform',
'prey_bait', 'plasmid', 'select', 'sc', 'arg', 'sc', 'leu', 'respect', 'bait_pre
y', 'two_hybrid', 'vector', 'pchb', 'pchp', 'respect', 'convert', 'gateway_desti

n', 'vector', 'use', 'gateway', 'vector', 'convers', 'system', 'gener', 'pchb_gc',
'pchp_gc', 'respect', 'gene_encod', 'bait', 'protein', 'hst', 'orf', 'prey', 'ce
k', 'orf', 'transfer', 'lr_reaction', 'donor', 'collect', 'vector', 'pchb_gc', 'pc
hp_gc', 'destin_vector', 'respect', 'author_funder_right_reserv', 'reus_allow_with
out_permiss', 'copyright_holder_preprint_peer', 'review_doi_biorxiv_preprint', 'op
aqu', 'cell', 'scha', 'sch', 'express', 'vp', 'cek', 'arg', 'lexa', 'hst', 'leu',
'respect', 'cross', 'follow', 'opaqu', 'cell', 'type', 'mix', 'cell', 'spider', 'm
edium', 'incub', 'result', 'tetraploid', 'cell', 'select', 'sc', 'leu_arg', 'mediu
m', 'incub', 'transfer', 'sc', 'met', 'medium', 'protein', 'protein', 'interact',
'detect', 'chromosom', 'loss', 'induc', 'pre', 'sporul', 'medium', 'day', 'describ
_previous', 'object', 'develop', 'albican_orfeom', 'encompass', 'mani', 'orf', 'pr
edict', 'assembl', 'albican', 'genom', 'possibl', 'aim', 'forward_revers', 'respec
t', 'attb', 'attp', 'sequenc', 'end', 'synthes', 'orf', 'supplement_tabl', 'use',
'independ', 'pcr_reaction', 'genom', 'dna', 'albican', 'strain', 'sc', 'result',
'pcr_product', 'clone', 'pdonr', 'use', 'gateway', 'recombin', 'clone', 'result',
'plasmid', 'subject', 'sanger_sequenc', 'end', 'clone', 'orf', 'illumina_sequenc',
'throughout', 'clone', 'orf', 'defin', 'group', 'sequenc_ident', 'refer', 'author_
funder_right_reserv', 'reus_allow_without_permiss', 'hapa', 'nucleotid', 'differ',
'present', 'two', 'allel', 'refer', 'orf', 'sequenc', 'notic', 'bia', 'haplotyp',
'clone', 'haplotyp', 'orf', 'clone', 'case', 'hapa', 'haplotyp', 'orf', 'clone',
'case', 'hapb', 'chromosom', 'orf', 'clone', 'uniformli', 'success', 'clone', 'suc
cess', 'rate', 'rang', 'chromosom', 'orf', 'slightli', 'although', 'orf', 'kb', 'c
ould', 'clone', 'success', 'rate', 'around', 'orf', 'kb', 'could', 'clone', 'succe
ss', 'rate', 'around', 'observ', 'decreas', 'success', 'rate', 'longer', 'orf', 'a
nalysi', 'nucleotid_sequenc', 'valid', 'orf', 'reveal', 'snp', 'nucleotid_substitu
t', 'could', 'real', 'snp', 'strain', 'refer', 'strain', 'due', 'poor_qualiti', 'r
efer', 'sequenc', 'mutat', 'primer', 'dna_polymeras', 'overal', 'result', 'maximu
m', 'error_rate', 'amount', 'snp', 'everi', 'clone', 'nucleotid', 'notabl', 'albic
an', 'sc', 'genom_sequenc', 'avail', 'cgd', 'version', 'contain', 'orf', 'sequen
c', 'ambigu', 'least_one', 'allel', 'ambigu', 'includ', 'stretch', 'ns', 'iupac',
'code', 'nucleotid', 'orfeom', 'gener', 'sequenc', 'data', 'resolv', 'sequenc', 'a
mbigu', 'orf', 'supplement_tabl', 'author_funder_right_reserv', 'reus_allow_withou
t_permiss', 'copyright_holder_preprint_peer', 'review_doi_biorxiv_preprint', 'stud
i', 'also', 'detect', 'recombin', 'haplotyp', 'clone', 'orf', 'supplement_tabl',
'display', 'characterist', 'refer', 'haplotyp', 'taken_togeth', 'studi', 'provid',
'almost', 'comprehens', 'extrem_high', 'qualiti', 'albican_orfeom', 'candidaorfd
b', 'creat', 'integr', 'data', 'albican_orfeom', 'project', 'avail', 'candidaorfd
b', 'enabl', 'scientif_commun', 'search', 'avail', 'qualiti', 'clone', 'data', 'pe
rtin', 'clone', 'process', 'store', 'databas', 'inform', 'includ', 'name', 'sequen
c', 'primer', 'use', 'amplifi', 'orf', 'albican', 'sc', 'genom', 'dna', 'coordin',
'primer', 'well', 'storag', 'plate', 'sequenc', 'inform_avail', 'clone', 'snp', 'd
ip', 'locat', 'orf', 'influenc', 'correspond_amino_acid', 'sequenc', 'shown', 'nuc
leotid_sequenc', 'clone', 'orf', 'correspond_amino_acid', 'sequenc', 'display', 'h
ighlight', 'snp', 'dip', 'rel', 'closest', 'haplotyp', 'sequenc', 'refer', 'sequen
c', 'data', 'extract', 'cgd', 'databas_queri', 'orf', 'number', 'orf', 'xxxx', 'xx
xx', 'assembl', 'name', 'ci', 'xxxxxw', 'ci', 'xxxxxc', 'gene', 'name', 'albican_o
rfeom', 'describ', 'develop', 'pdonr', 'allow', 'subsequ', 'transfer', 'clone', 'o
rf', 'gateway', 'adapt', 'destin_vector', 'destin_vector', 'orf', 'express', 'hos
t', 'coli', 'cerevisia', 'avail', 'destin_vector', 'orf', 'express', 'albican', 'r
eport', 'therefor', 'set', 'establish', 'collect', 'gateway', 'adapt', 'destin_vec
tor', 'author_funder_right_reserv', 'reus_allow_without_permiss', 'copyright_holde
r_preprint_peer', 'review_doi_biorxiv_preprint', 'constitut', 'condit', 'express',
'untag', 'tag', 'orf', 'albican', 'primari', 'collect', 'albican', 'destin_vecto
r', 'deriv', 'cip', 'provid', 'choic', 'two', 'promot', 'either', 'constitut', 'pr
omot', 'tdh', 'induc', 'promot', 'tet', 'option', 'termin', 'fusion', 'variou', 'e
pitop_tag', 'xha', 'tap', 'tabl', 'integr', 'destin_vector', 'deriv', 'rp', 'loc

u', 'albican', 'genom', 'promot', 'stui', 'scei', 'linear', 'cipderiv', 'vector',
'carri', 'either', 'auxotroph_marker', 'ura', 'nat', 'marker', 'confer_resist', 'n
ourseothricin', 'use', 'clinic', 'isol', 'howev', 'tet', 'bear', 'plasmid', 'canno
t', 'carri', 'nat', 'marker', 'sinc', 'transactiv', 'need', 'induct', 'tet', 'bor
n', 'nat', 'carri', 'pnimx', 'plasmid_vector', 'harbor', 'call', 'spacer', 'sequen
c', 'sp', 'origin', 'coli', 'kanamycin_resist', 'gene', 'barcod', 'insert', 'subse
qu', 'illumina', 'base', 'barcod', 'sequenc', 'second', 'set', 'plasmid', 'also',
'gener', 'constitut', 'promot', 'act', 'induc', 'promot', 'pck', 'option', 'termi
n', 'fusion', 'gfp', 'tabl', 'destin_vector', 'design', 'standard_nomenclatur', 'p
ca', 'destijkl', 'wherebi', 'stand', 'transform', 'marker', 'ura', 'nat', 'stand',
'promot', 'tet', 'pck', 'tdh', 'act', 'stand', 'termin_tag', 'tag_xha', 'gfp', 'ta
p', 'stand', 'termin_tag', 'tag_xha', 'gfp', 'tap', 'tabl', 'destin_vector', 'avai
l', 'addgen', 'order', 'valid', 'primari', 'set', 'destin_vector', 'could', 'driv
e', 'constitut', 'condit', 'express', 'untag', 'tag', 'orf', 'ume', 'orf', 'transf
er', 'use', 'lr_clonas', 'ume', 'encod', 'transcript_factor', 'whose', 'overexpres
s', 'shown', 'forc', 'filament', 'albican', 'valid', 'criteria', 'destin_vector',
'success', 'lr_reaction', 'ii', 'integr', 'author_funder_right_reserv', 'reus_allo
w_without_permiss', 'copyright_holder_preprint_peer', 'review_doi_biorxiv_preprin
t', 'albican', 'genom', 'rp', 'locu', 'iii', 'filament', 'phenotyp', 'iv', 'detec
t', 'xha', 'tap_tag', 'ume', 'protein', 'western_blot', 'lr_reaction', 'success',
'use', 'transfer', 'ume', 'pdonr', 'ume', 'plasmid', 'destin_vector', 'result', 'e
xpress_plasmid', 'success', 'integr', 'albican', 'rp', 'locu', 'strain', 'contai
n', 'tet', 'induc', 'overnight', 'ad', 'ml', 'atc', 'cultur_medium', 'strain', 'co
ntain', 'tdh', 'grown_overnight', 'ypd', 'describ', 'chauvel', 'overexpress', 'um
e', 'led', 'phenotyp', 'observ', 'strain', 'regardless', 'tag_xha', 'tap_tag', 'um
e', 'detect', 'relev', 'strain', 'appropri', 'growth', 'condit', 'taken_togeth',
'data', 'present', 'collect', 'avail', 'destin_vector', 'valid', 'applic', 'albica
n_orfeom', 'avail', 'orfeom', 'collect', 'prerequisit', 'larg', 'systemat', 'two_h
ybrid', 'screen', 'albican_candida', 'ch', 'system', 'develop', 'target', 'one',
'one', 'protein', 'protein', 'interact', 'detect', 'henc', 'need', 'engin', 'syste
m', 'high_throughput', 'screen', 'base', 'rapid', 'conveni', 'recombin', 'clone',
'system', 'gener', 'prey_bait', 'array', 'protein', 'use', 'mate', 'approach', 'ci
rcumv', 'low', 'effici', 'transform', 'fungu', 'former', 'obtain', 'adapt', 'prey_
bait', 'recipi', 'vector', 'gateway', 'mediat', 'transfer', 'orfeom', 'librari',
'yield', 'pchb_gc', 'pchp_gc', 'respect', 'latter', 'involv', 'combin', 'process',
'white_opaqu', 'switch', 'prepar', 'mate', 'compat', 'strain', 'pleiomorph', 'fung
u', 'albican', 'character', 'parasexu', 'cycl', 'author_funder_right_reserv', 'reu
s_allow_without_permiss', 'copyright_holder_preprint_peer', 'review_doi_biorxiv_pr
eprint', 'possibl', 'mate', 'diploid', 'gener', 'tetraploid', 'absenc', 'meiosi',
'revers', 'diploid', 'state', 'chromosom', 'loss', 'review', 'essenti_prerequisi
t', 'mate', 'albican', 'phenotyp', 'switch', 'white_opaqu', 'stabl', 'revers', 'sw
itch', 'regul', 'environment_stimuli', 'low_temperatur', 'carbon_dioxid', 'nutrien
t', 'epigenet', 'switch', 'tightli_link', 'mate', 'sinc', 'complex', 'encod', 'mat
e', 'type', 'like', 'mtl', 'loci', 'act', 'repressor', 'opaqu', 'switch', 'contex
t', 'two_hybrid', 'strain', 'sch', 'origin', 'diploid', 'sn', 'strain', 'delet',
'mtla', 'mtl_locu', 'construct', 'sch', 'scha', 'strain', 'effici', 'switch', 'opa
qu', 'hemizygot', 'strain', 'achiev', 'express', 'wor', 'major', 'regul', 'white_o
paqu', 'switch', 'proof_principl', 'whole', 'procedur', 'protein', 'protein', 'det
ect', 'shown_figur', 'map_kinas', 'kinas', 'hst', 'use', 'bait', 'link', 'dna', 'b
ind_domain', 'lexa', 'part', 'pchb_gc', 'vector', 'similarli', 'cek', 'kinas', 'ex
press', 'prey', 'protein', 'fuse', 'activ', 'domain', 'vp', 'compon', 'pchp_gc',
'vector', 'pchb', 'hst', 'pchp', 'cek', 'transform', 'opaqu', 'sch', 'scha', 'stra
in', 'respect', 'mate', 'result', 'transform', 'achiev', 'select', 'leucin', 'argi
nin', 'prototrophi', 'upon', 'interact', 'two', 'protein', 'reconstitut', 'transcr
ipt', 'modul', 'promot', 'express', 'marker', 'allow', 'mate', 'deriv', 'strain',
'grow', 'histidin', 'free_medium', 'wherea', 'origin', 'diploid', 'strain', 'coul

d', 'mechan', 'chromosom', 'loss', 'cross', 'tetraploid', 'strain', 'return_back',
'diploid', 'state', 'show', 'plasmid', 'stabli_express', 'diploid', 'offspr', 'str
ain', 'strain', 'conserv', 'abil_grow', 'absenc', 'arginin', 'leucin', 'indic', 'p
resenc', 'prey_bait', 'vector', 'respect', 'addit', 'strain', 'abl', 'grow', 'abse
nc', 'histidin', 'indic', 'target', 'protein', 'protein', 'interact', 'author_fund
er_right_reserv', 'reus_allow_without_permiss', 'copyright_holder_preprint_peer',
'review_doi_biorxiv_preprint', 'discuss', 'partial', 'albican', 'gateway', 'adap
t', 'orfeom', 'orfeomev', 'orf', 'correspond', 'albican', 'overexpress', 'strain',
'alreadi', 'success', 'use', 'investig', 'morphogenesi', 'biofilm_format', 'geno
m', 'dynam', 'albican', 'studi', 'gener', 'albican_orfeomev', 'within', 'given',
'high', 'standard', 'success', 'rate', 'achiev', 'albican_orfeomev', 'quit', 'rema
rk', 'eukaryot', 'orfeom', 'usual', 'clone', 'full_length', 'cdna_librari', 'accou
nt', 'presenc', 'intron', 'therefor', 'splice_variant', 'howev', 'albican', 'gen
e', 'intron', 'render', 'gener', 'albican_orfeom', 'straightforward', 'pluricellul
ar', 'eukaryot', 'inde', 'orf', 'simpli', 'amplifi', 'genom', 'dna', 'refer', 'str
ain', 'sc', 'therefor', 'orf', 'albican_orfeomev', 'harbor', 'intron', 'matter',
'taken_consider', 'express', 'orf', 'prokaryot', 'host', 'possibl', 'eukaryot', 'h
ost', 'albican', 'second', 'aspect', 'taken_consider', 'use', 'orfeom', 'host', 'a
uthor_funder_right_reserv', 'reus_allow_without_permiss', 'copyright_holder_prepri
nt_peer', 'review_doi_biorxiv_preprint', 'albican', 'unusu', 'codon_usag', 'spec
i', 'inde', 'albican', 'cug', 'decod', 'serin', 'instead', 'leucin', 'major', 'tim
e', 'henc', 'improp', 'translat', 'host', 'standard', 'genet_code', 'may', 'distor
t', 'protein', 'structur', 'function', 'sanger_sequenc', 'end', 'perform', 'confir
m', 'orf', 'ident', 'exclud', 'clone', 'contain', 'primer', 'recombin', 'error',
'previou_studi', 'report', 'mutat_rate', 'primer', 'similar', 'rate', 'observ', 'w
ork', 'unlik', 'gateway', 'recombin', 'clone', 'project', 'see', 'major', 'diffe
r', 'clone', 'effici', 'orf', 'kb', 'decreas', 'success', 'rate', 'observ', 'orf',
'kb_size', 'bia', 'could_attribut', 'increas', 'difficulti', 'amplifi', 'orf', 'pc
r', 'reduc', 'effici', 'gateway_bp', 'clonas', 'reaction', 'long', 'pcr_product',
'error_rate', 'pcr', 'polymeras', 'increas', 'longer', 'product', 'addit', 'also',
'correct', 'sequenc', 'orf', 'sequenc', 'ambigu', 'tract', 'iupac', 'code', 'nucle
otid', 'accord', 'candida', 'genom', 'databas', 'identifi', 'orf', 'display', 'rec
ombin', 'haplotyp', 'recombin', 'haplotyp', 'could_explain', 'templat_switch', 'pc
r', 'cycl', 'incorrectli', 'phase', 'snp', 'refer', 'sequenc', 'could', 'address',
'perform', 'long', 'read', 'sequenc', 'albican', 'sc', 'genom', 'howev', 'mani',
'challeng', 'remain', 'address', 'heterozyg', 'gene', 'one', 'allel', 'present',
'orfeom', 'function', 'differ', 'report', 'allel', 'heterozyg', 'gene', 'would',
'relev', 'also', 'clone', 'valid', 'second', 'allel', 'instanc', 'addit', 'desig
n', 'base', 'annot', 'haploid', 'set', 'assembl', 'albican', 'genom', 'consequ',
'gene', 'mtl_locu', 'includ', 'orfeom', 'despit', 'effort', 'collect', 'still', 'm
iss', 'clone', 'next', 'version', 'albican_orfeom', 'could', 'author_funder_right_
reserv', 'reus_allow_without_permiss', 'copyright_holder_preprint_peer', 'review_d
oi_biorxiv_preprint', 'extend', 'gene', 'coverag', 'ad', 'miss', 'orf', 'allel_var
iant', 'strain', 'specif', 'variat', 'orfeom', 'essenti', 'bridg_gap', 'genom_anno
t', 'system', 'biolog', 'allow', 'larg_scale', 'gene', 'protein', 'develop', 'coll
ect', 'albican', 'constitut', 'condit', 'overexpress', 'strain', 'one', 'applic',
'albican_orfeom', 'success', 'explor', 'respect', 'albican_orfeom', 'project', 'cu
rrent', 'complet', 'second_third', 'phase', 'consist', 'transfer', 'clone', 'orf',
'barcod', 'destin_vector', 'gener', 'collect', 'albican', 'barcod', 'overexpress',
'mutant', 'mutant', 'carri', 'one', 'clone', 'orf', 'control', 'induc', 'tet', 'pr
omot', 'candidaorfdb', 'alreadi', 'provid', 'inform_avail', 'overexpress', 'plasmi
d', 'strain', 'studi', 'pave_way', 'second', 'applic', 'albican_orfeom', 'develo
p', 'tool', 'matrix', 'approach', 'protein', 'protein', 'interact', 'detect', 'vi
a', 'mate', 'albican', 'data', 'show', 'mate', 'diploid', 'albican', 'express', 'b
ait', 'fuse', 'lexa', 'dna', 'bind_domain', 'prey', 'fuse', 'vp', 'activ', 'domai
n', 'respect', 'allow', 'protein', 'protein', 'interact', 'test', 'cerevisia', 'la

```
rg_scale', 'screen', 'perform', 'matrix', 'design', 'wherebi', 'haploid', 'strai
n', 'express', 'bait_prey', 'mate', 'protein', 'protein', 'interact', 'score', 're
sult', 'diploid', 'toolkit', 'enabl', 'implement', 'matrix', 'design', 'larg_scal
e', 'screen', 'albican', 'aim', 'collect', 'prey', 'clone', 'alreadi', 'gener', 'm
ention', 'albican', 'codon_usag', 'unusu', 'limit', 'develop', 'applic', 'albican_
orfeom', 'speci', 'use', 'standard', 'decod', 'cerevisia', 'develop', 'tool', 'lar
g_scale', 'screen', 'albican', 'circumv_limit', 'open', 'path', 'albican', 'intera
ctom', 'author_funder_right_reserv', 'reus_allow_without_permiss', 'copyright_hold
er_preprint_peer', 'review_doi_biorxiv_preprint', 'defin', 'albican', 'interacto
m', 'undoubtedli', 'impact', 'understand', 'albican', 'pathogenesi', 'human', 'sum
mari', 'high_level', 'gene', 'coverag', 'coupl', 'versatil', 'gateway', 'recombi
n', 'clone', 'albican', 'adapt', 'function', 'genom', 'tool', 'full', 'access', 'c
lone', 'make', 'albican_orfeomev', 'uniqu', 'valuabl_resourc', 'scientif_commun',
'greatli_facilit', 'futur', 'function', 'studi', 'albican', 'grate', 'member', 'gr
oup', 'support', 'throughout', 'develop', 'albican_orfeom', 'author_funder_right_r
eserv', 'reus_allow_without_permiss', 'copyright_holder_preprint_peer', 'review',
'ter_ter', 'ter_ter', 'promot', 'tag', 'posit', 'promot', 'tag', 'posit', 'author_
funder_right_reserv', 'reus_allow_without_permiss', 'copyright_holder_preprint_pee
r', 'review_doi_biorxiv_preprint']"
```

In [ ]:
```python
from sklearn.feature_extraction.text import CountVectorizer

train, test = train_test_splitter(processed_articles.processed_trigrams, 0.90)

vectorizer = CountVectorizer(min_df = 50, max_df = 0.8, max_features = 50000)
tf = vectorizer.fit_transform([t[4] for t in train]) ## Vectorize training set
tf_feature_names = vectorizer.get_feature_names_out() ## Pull out words for use in

# Transform test data for perplexity eval

tf_test = vectorizer.transform([t[4] for t in test])
```

Importing the Gensim library for topic modeling

In [ ]:
```python
import gensim.corpora as corpora
from gensim.models import CoherenceModel
import logging
```

In [ ]:
```python
# creating a dictionary and a corpus for the LDA model

id2word = corpora.Dictionary([t[4].split() for t in train])
corpus = [id2word.doc2bow(text.split()) for text in [t[4] for t in train]]
```

Modeling -LDA

In [ ]:
```python
# setting the lists and dictionaries for the models
num_topics_list = [8, 9, 10]
models = {}
perplexity = {}
coherence = {}

# nice little loop to build the models
```

```
for num_topics in num_topics_list:
    lda_model = gensim.models.LdaModel(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics,
                                       random_state=100,
                                       update_every=1,
                                       chunksize=100,
                                       passes=10,
                                       alpha='auto',
                                       per_word_topics=True)
    models[num_topics] = lda_model
    # evaluate using perplexity
    perplexity[num_topics] = lda_model.log_perplexity(corpus)

    # evaluate using coherence
    coherence_model_lda = CoherenceModel(model=lda_model, texts=[t[4].split() for t
    coherence[num_topics] = coherence_model_lda.get_coherence()
```

That was a 17 minute run time for the cell above

In [ ]:
```
# Print the evaluation results
for num_topics in num_topics_list:
    print(f'Number of topics: {num_topics}')
    print(f'Perplexity: {perplexity[num_topics]}')
    print(f'Coherence: {coherence[num_topics]}')
    print('')
```

```
Number of topics: 8
Perplexity: -9.311266383217047
Coherence: 0.4842853560855015

Number of topics: 9
Perplexity: -9.621981029842669
Coherence: 0.4329586510443721

Number of topics: 10
Perplexity: -10.019501432435934
Coherence: 0.45750258984138614
```

We are going to go with the 10 topic model as it is the best model based on the perplexity and coherence scores

In [ ]:
```
# printing the top 3 most common words in each topic for the 10 topic model

lda_model = models[10]
topics = lda_model.print_topics(num_words=3)

# prints
print("Most common words in each topic (model with 10 topics):")
```

```
for topic_num, topic_words in topics:
    print(f"Topic {topic_num}: {topic_words}")
```

```
Most common words in each topic (model with 10 topics):
Topic 0: 0.028*"'use'," + 0.012*"'concentr'," + 0.009*"'compound',"
Topic 1: 0.039*"'protein'," + 0.011*"'cell'," + 0.010*"'structur',"
Topic 2: 0.028*"'patient'," + 0.017*"'infect'," + 0.012*"'studi',"
Topic 3: 0.024*"'use'," + 0.022*"'sequenc'," + 0.017*"'detect',"
Topic 4: 0.008*"'develop'," + 0.008*"'provid'," + 0.008*"'health',"
Topic 5: 0.010*"'use'," + 0.009*"'case'," + 0.008*"'studi',"
Topic 6: 0.047*"'de'," + 0.021*"'cat'," + 0.013*"'dog',"
Topic 7: 0.022*"'may'," + 0.017*"'diseas'," + 0.015*"'anim',"
Topic 8: 0.048*"'cell'," + 0.018*"'infect'," + 0.012*"'activ',"
Topic 9: 0.046*"'viru'," + 0.037*"'infect'," + 0.032*"'vaccin',"
```

Topic breakdowns by the most common words and What I think the topics are about

- `Topic 0` : This looks like it is talking about some kind of using some kind of compund and looks like its talking about development of drugs
- `Topic 1` : this looks like it us talking about the structure of viruses with the words like protein, cell and structure
- `Topic 2` : This looks like the articles are talking about the patinets and the way they are infected during a study.
- `Topic 3` : To me it looks like the sequence of events when someone is infected by a virus as it has the word sequencing in there.
- `Topic 4` : It looks like they are talking about the action plans to make a vaccine for the virus, and provide it, for healthcare.
- `Topic 5` : Maybe how to use studies to help with the virus and using specific cases to help tailor the vaccine?
- `Topic 6` : This is maybe talking about the way that the dogs and cats are being infected by the virus or testing subjects?
- `Topic 7` : This is talking about the lock down because it icludes the word 'May'
- `Topic 8` : This could be talking about what the virus is doing to the cells and how it is affecting the person when the infection sets it?
- `Topic 9` : Maybe this is talking about the way that the virus is being spread and how it is being spread, and the vaccines?

Summary of all the topics 2-3 sentences

I think that all of the topics are looking at different things and the way that the effect people and the way that the virus is being spread. I think that the topics are all very different and are all looking at different thing, and explore multiple topics, but they all are dealing with Viruses, and the way that they are being spread, and the way that they are being treated historically speaking that is.