8th of December – 4th meeting Debaryomyces hansenii

Important points of previous meeting:

- Try kraken with custom database – done
- ITS regions – started
- SNP calling – not done yet
- Know assemblers – done
- Things of previous meetings that were left behind for now:
- Annotation using Maker
- BLAST of the assemblies

New results and information:

Everything located on GitHub:
https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii

Kraken customDB study:
Assemblies:
https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii/tree/master/Work_files/mahesha_assemblies_workfolder/contamination_check

Raw reads:
https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii/tree/master/Work_files/rawdata_workfolder/contamination_check

Tables with results included in this file.
Also resume of the results.

ITS sequences: Found in almost all strains, better conditions depending of the assembler used.
Resume at the end of this document.

Information about strains

Alternative names for strains and its origin:

Strain	Alternative strain names	Origin	Year
1001	NCYC2572, CBS767, ATCC36239, CCRC21394, DBVPG6050, IFO0083, JCM1990, JCM2102, KCTC7645, MUCL30242,	Carlsberg laboratories, habitat associated with fermentation	1994

	NRRLY-7426, NRRLY-10976, UCD74-86		
1002	NCYC8, NCTC2059	Throat of pacient with angina	1925
1003	NCYC9, NCTC2048	Dutch cheese prepared in Russia	1924
1004	NCYC10, NCTC2056	Unknown	1925
1005	NCYC103, NCTC1681	Carlsberg laboratories, habitat associated with fermentation	1923
1006	NCYC459	Soil, New Zealand	1955
1007	NCYC475, CBS811, JCM1439, NRRLY-1454, UCD75-11	Fermenting Kentucky tobacco	1956
1008	NCYC792, NCMB1230 43	Sea Water	1974
1009	NCYC3045	Dried salted black olives from Thassos, Greece	2002
1010	NCYC3364	Capping machine of soft drinks factory in Brazil	2006
1011	CBS117	Rennet from New Zealand	1985
1012	CBS5140	Skin of a man, Hungary	
1013	CBS1101, IFO0027, IFO0093	Salt pork	1946
1014	CBS1792	Chilled beef from Brisbane, Australia	
1015	J63	Seawater, Sweden	
1016	J26	Seawater, Sweden	
1017	J16	-	-
1018	J52	-	-
1019	DBH9	CBS767 carrying a mutation in DhHIS4 gene	

Kraken

Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomics DNA sequences. Previous programs designed for this task have been relatively slow and computationally expensive, forcing researchers to use faster abundance estimation programs, which only classify small subsets of metagenomic data.Using exact alignment of k-mers, Kraken achieves classification accuracy comparable to the fastest BLAST program. In its fastest mode, Kraken classifies 100 base pair reads at a rate of over 4.1 million reads per minute, 909 times faster than Megablast and 11 times faster than the abundance estimation program MetaPhlAn.

k-mer to lowest common ancestor database

At the core of Kraken is a database that contains records consisting of a k-mer and the LCA of all organisms whose genomes contain that k-mer. This database, built using a user-specified library of genomes, allows a quick lookup of the most specific node in the taxonomic tree that is associated with a given k-mer. Sequences are classified by querying the database for each k-mer in a sequence, and then using the resulting set of LCA taxa to determine an appropriate label for the sequence (Figure 1). Sequences that have no kmers in the database are left unclassified by Kraken. By default, Kraken builds the database with k = 31, but this value is user-modifiable.

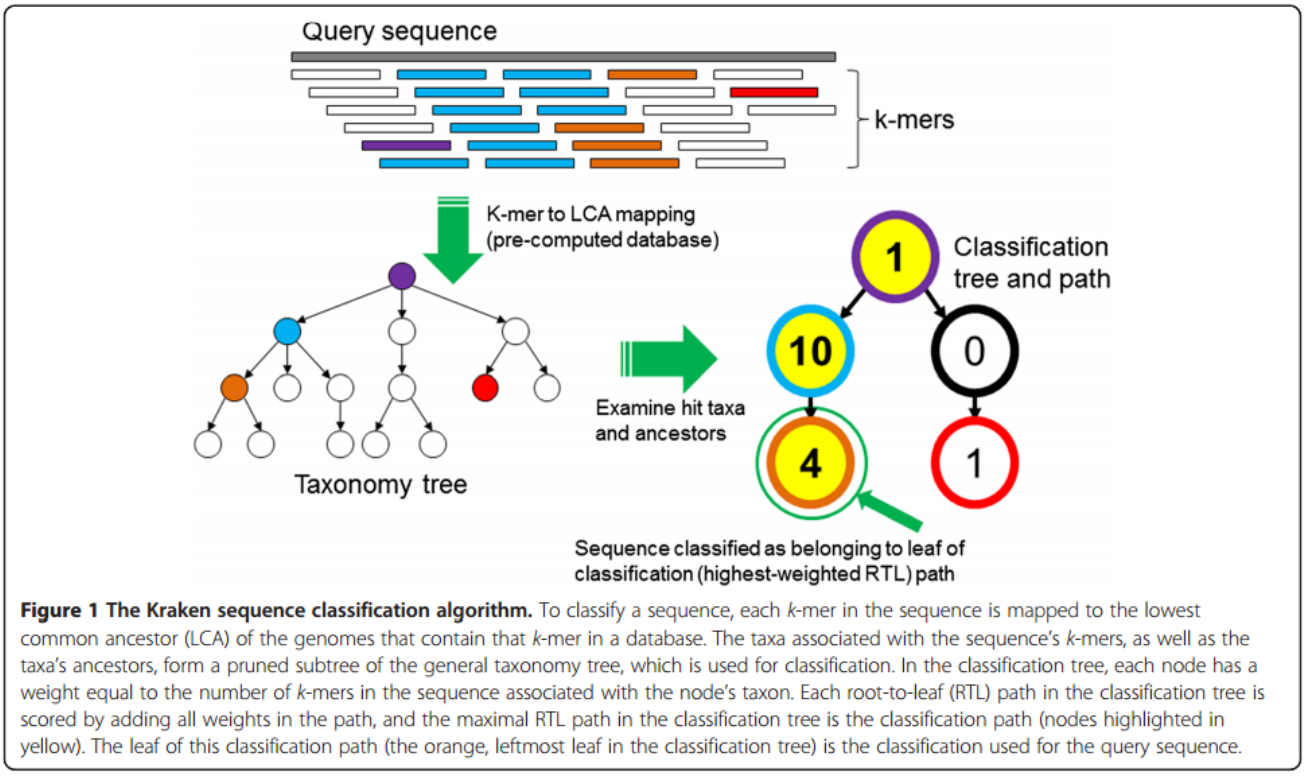

Figure 1 The Kraken sequence classification algorithm. To classify a sequence, each *k*-mer in the sequence is mapped to the lowest common ancestor (LCA) of the genomes that contain that *k*-mer in a database. The taxa associated with the sequence's *k*-mers, as well as the taxa's ancestors, form a pruned subtree of the general taxonomy tree, which is used for classification. In the classification tree, each node has a weight equal to the number of *k*-mers in the sequence associated with the node's taxon. Each root-to-leaf (RTL) path in the classification tree is scored by adding all weights in the path, and the maximal RTL path in the classification tree is the classification path (nodes highlighted in yellow). The leaf of this classification path (the orange, leftmost leaf in the classification tree) is the classification used for the query sequence.

Database structure and search algorithm

Because Kraken very frequently uses a k-mer as a database query immediately after querying an adjacent kmer, and because adjacent k-mers share a substantial amount of sequence, we utilize the minimizer concept to group similar k-mers together. To explain our application of this concept, we here define the canonical representation of a DNA sequence S as the lexicographically smaller of S and the reverse complement of S. To determine a k-mer's minimizer of length M, we consider the canonical representation of all M-mers in the k-mer, and select the lexicographically smallest of those M-mers as the k-mer's minimizer. In practice, adjacent k-mers will often have the same minimizer.

In Kraken's database, all k-mers with the same minimizer are stored consecutively, and are sorted in lexicographical order of their canonical representations. A query for a k-mer R can then be

processed by looking up in an index the positions in the database where the k-mers with R's minimizer would be stored, and then performing a binary search within that region (Figure 5). Because adjacent k-mers often have the same minimizer, the search range is often the same between two consecutive queries, and the search in the first query can often bring data into the CPU cache that will be used in the second query. By allowing memory accesses in subsequent queries to access data in the CPU cache instead of RAM, this strategy makes subsequent queries much faster than they would otherwise be.

The index containing the offsets of each group of k-mers in the database requires 8 × 4M bytes. By default Kraken uses 15-bp minimizers, but the user can modify this value; for example, in creating MiniKraken, we used 13-bp minimizers to ensure the total database size stayed under 4 GB.

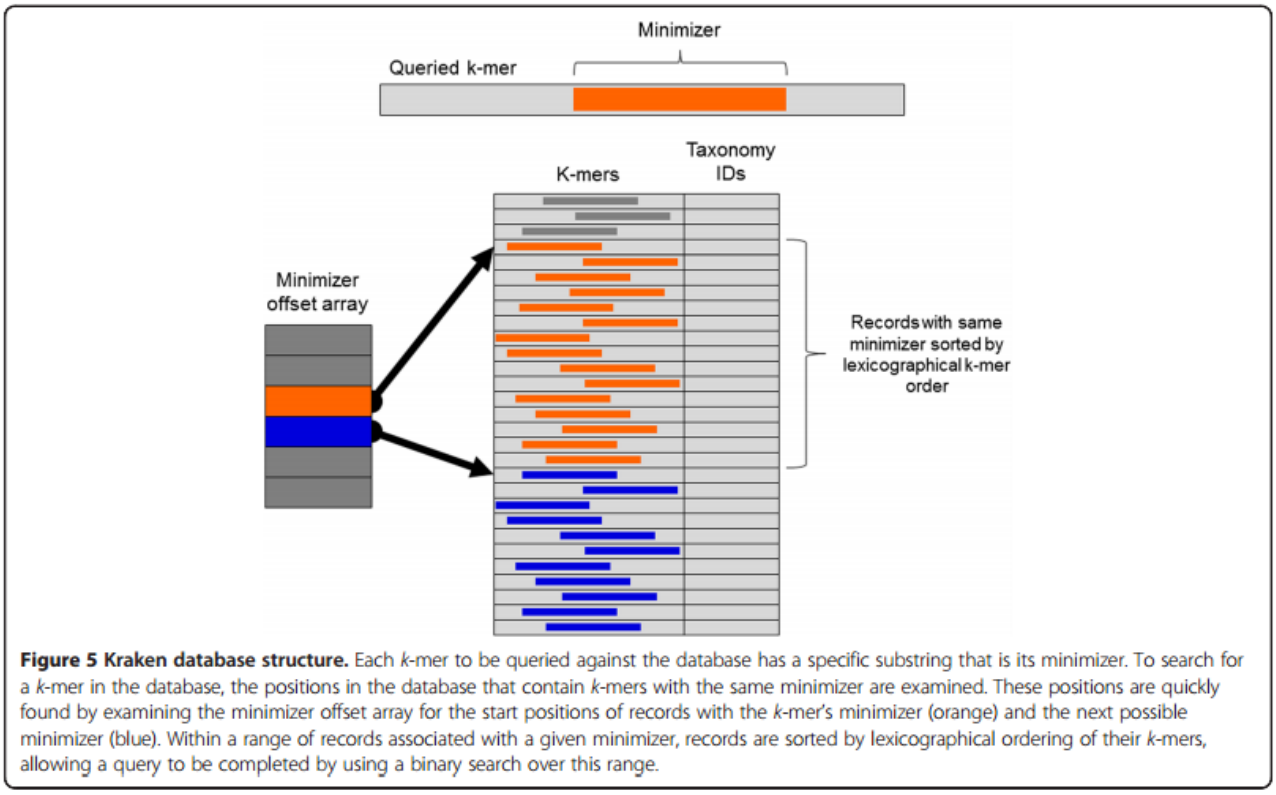

Figure 5 Kraken database structure. Each *k*-mer to be queried against the database has a specific substring that is its minimizer. To search for a *k*-mer in the database, the positions in the database that contain *k*-mers with the same minimizer are examined. These positions are quickly found by examining the minimizer offset array for the start positions of records with the *k*-mer's minimizer (orange) and the next possible minimizer (blue). Within a range of records associated with a given minimizer, records are sorted by lexicographical ordering of their *k*-mers, allowing a query to be completed by using a binary search over this range.

Raw reads results

- ✧ **Classified sequences:** Found in the kraken custom database, sequences matching bacterial, archaeal, plasmids, viral, human or fungi domains.
- ✧ **Unclassified sequences:** Not found in the custom database.

fastq	Total sequences	Classified sequences	Unclassified sequences
1001 AH 1	2462393	2029147 (82.41%)	433246 (17.59%)
1001 AH 2	2462393	2000937 (81.26%)	461456 (18.74%)
1001 BC 1	3480996	2858026 (82.10%)	622970 (17.90%)
1001 BC 2	3480996	2832253 (81.36%)	648743 (18.64%)
1002 AH 1	1160378	1140993 (98.33%)	19385 (1.67%)
1002 AH 2	1160378	1134927 (97.81%)	25451 (2.19%)
1002 BC 1	1853630	1822492 (98.32%)	31138 (1.68%)
1002 BC 2	1853630	1805406 (97.40%)	48224 (2.60%)
1003 AH 1	2193546	1817189 (82.84%)	376357 (17.16%)
1003 AH 2	2193546	1776305 (80.98%)	417241 (19.02%)
1003 BC 1	3349842	2766520 (82.59%)	583322 (17.41%)
1003 BC 2	3349842	2709763 (80.89%)	640079 (19.11%)

1004 AH 1	1558691	1433365 (91.96%)	125326 (8.04%)
1004 AH 2	1558691	1374655 (88.19%)	184036 (11.81%)
1004 BC 1	2449087	2248776 (91.82%)	200311 (8.18%)
1004 BC 2	2449087	2189404 (89.40%)	259683 (10.60%)
1005 AH 1	2088288	2057955 (98.55%)	30333 (1.45%)
1005 AH 2	2088288	2027995 (97.11%)	60293 (2.89%)
1005 BC 1	2922955	2879153 (98.50%)	43802 (1.50%)
1005 BC 2	2922955	2835228 (97.00%)	87727 (3.00%)
1006 AH 1	1929277	335761 (17.40%)	1593516 (82.60%)
1006 AH 2	1929277	324501 (16.82%)	1604776 (83.18%)
1006 BC 1	2847843	492928 (17.31%)	2354915 (82.69%)
1006 BC 2	2847843	479112 (16.82%)	2368731 (83.18%)
1007 AH 1	2305984	2267226 (98.32%)	38758 (1.68%)
1007 AH 2	2305984	2239638 (97.12%)	66346 (2.88%)
1007 BC 1	3473851	3413498 (98.26%)	60353 (1.74%)
1007 BC 2	3473851	3372019 (97.07%)	101832 (2.93%)
1008 AH 1	1312616	1281379 (97.62%)	31237 (2.38%)
1008 AH 2	1312616	1268340 (96.63%)	44276 (3.37%)
1008 BC 1	1940489	1893091 (97.56%)	47398 (2.44%)
1008 BC 2	1940489	1873108 (96.53%)	67381 (3.47%)
1009 AH 1	1944249	1419587 (73.01%)	524662 (26.99%)
1009 AH 2	1944249	1394897 (71.74%)	549352 (28.26%)
1009 BC 1	2894462	2107051 (72.80%)	787411 (27.20%)
1009 BC 2	2894462	2077020 (71.76%)	817442 (28.24%)
1010 AH 1	1319601	189904 (14.39%)	1129697 (85.61%)
1010 AH 2	1319601	185313 (14.04%)	1134288 (85.96%)
1010 BC 1	1909206	270625 (14.17%)	1638581 (85.83%)
1010 BC 2	1909206	264627 (13.86%)	1644579 (86.14%)
1011 AH 1	768699	450133 (58.56%)	318566 (41.44%)
1011 AH 2	768699	446288 (58.06%)	322411 (41.94%)
1011 BC 1	1178421	687776 (58.36%)	490645 (41.64%)
1011 BC 2	1178421	680994 (57.79%)	497427 (42.21%)
1012 AH 1	2073205	359992 (17.36%)	1713213 (82.64%)
1012 AH 2	2073205	350775 (16.92%)	1722430 (83.08%)
1012 BC 1	3224847	556618 (17.26%)	2668229 (82.74%)
1012 BC 2	3224847	544512 (16.88%)	2680335 (83.12%)
1013 AH 1	1828482	1765439 (96.55%)	63043 (3.45%)
1013 AH 2	1828482	1740777 (95.20%)	87705 (4.80%)
1013 BC 1	2736996	2640904 (96.49%)	96092 (3.51%)
1013 BC 2	2736996	2604369 (95.15%)	132627 (4.85%)
1014 AH 1	2206902	1806751 (81.87%)	400151 (18.13%)
1014 AH 2	2206902	1780380 (80.67%)	426522 (19.33%)
1014 BC 1	3528456	2882136 (81.68%)	646320 (18.32%)
1014 BC 2	3528456	2837996 (80.43%)	690460 (19.57%)
1015 AH 1	1923464	1859416 (96.67%)	64048 (3.33%)
1015 AH 2	1923464	1844644 (95.90%)	78820 (4.10%)
1015 BC 1	3181425	3074019 (96.62%)	107406 (3.38%)
1015 BC 2	3181425	3044236 (95.69%)	137189 (4.31%)
1016 AH 1	2646260	2151919 (81.32%)	494341 (18.68%)
1016 AH 2	2646260	2117272 (80.01%)	528988 (19.99%)
1016 BC 1	3919163	3177692 (81.08%)	741471 (18.92%)
1016 BC 2	3919163	3130716 (79.88%)	788447 (20.12%)

1017 AH 1	1889165	1853424 (98.11%)	35741 (1.89%)
1017 AH 2	1889165	1828433 (96.79%)	60732 (3.21%)
1017 BC 1	2738771	2685950 (98.07%)	52821 (1.93%)
1017 BC 2	2738771	2659430 (97.10%)	79341 (2.90%)
1018 AH 1	2628541	2580100 (98.16%)	48441 (1.84%)
1018 AH 2	2628541	2552444 (97.10%)	76097 (2.90%)
1018 BC 1	3662396	3593609 (98.12%)	68787 (1.88%)
1018 BC 2	3662396	3555962 (97.09%)	106434 (2.91%)
1019 AH 1	2000344	1964810 (98.22%)	35534 (1.78%)
1019 AH 2	2000344	1915216 (95.74%)	85128 (4.26%)
1019 BC 1	2664496	2615956 (98.18%)	48540 (1.82%)
1019 BC 2	2664496	2576230 (96.69%)	88266 (3.31%)

Table resume of the reports

Strain	Fungi	Bacteria	Viruses	Archaea	Others	Unclassified	Deha	Sc	Human
1001 AH 1	82.38%	0.01%	0.00%	0.00%	0.01%	17.59%	78.50%	0.01%	0.01%
1001 AH 2	81.22%	0.01%	0.00%	0.00%	0.01%	18.74%	77.38%	0.01%	0.01%
1001 BC 1	82.08%	0.01%	0.00%	0.00%	0.01%	17.90%	78.25%	0.01%	0.01%
1001 BC 2	81.33%	0.01%	0.00%	0.00%	0.02%	18.64%	77.53%	0.01%	0.02%
1002 AH 1	98.31%	0.01%	-	-	0.01%	1.97%	93.71%	0.01%	0.01%
1002 AH 2	97.78%	0.00%	-	-	0.01%	2.19%	93.20%	0.01%	0.01%
1002 BC 1	98.30%	0.01%	-	-	0.01%	1.68%	93.71%	0.01%	0.01%
1002 BC 2	97.37%	0.00%	-	-	0.02%	2.60%	92.82%	0.01%	0.02%
1003 AH 1	82.81%	0.02%	0.00%	0.00%	0.01%	17.16%	78.91%	0.00%	0.01%
1003 AH 2	80.93%	0.02%	0.00%	0.00%	0.01%	19.02%	77.08%	0.01%	0.01%
1003 BC 1	82.55%	0.02%	0.00%	0.00%	0.01%	17.41%	78.71%	0.00%	0.01
1003 BC 2	80.84%	0.02%	0.00%	0.00%	0.02%	19.11%	77.05%	0.01%	0.02%
1004 AH 1	91.91%	0.04%	0.00%	0.00%	0.00%	8.04%	89.12%	0.00%	0.00%
1004 AH 2	88.13%	0.04%	0.00%	0.00%	0.01%	11.81%	85.43%	0.00%	0.01%
1004 BC 1	91.77%	0.05%	0.00%	0.00%	0.01%	8.18%	89.03%	0.00%	0.01%
1004 BC 2	89.33%	0.04%	0.00%	0.00%	0.02%	10.60%	86.66%	0.00%	0.02%
1005 AH 1	98.48%	0.06%	0.00%	-	0.00%	1.45%	93.36%	0.00%	0.00%
1005	97.04%	0.06%	0.00%	-	0.01%	2.89%	91.97%	0.00%	0.01%

AH 2									
1005 BC 1	98.43%	0.06%	0.00%	-	0.00%	1.50%	93.46%	0.00%	0.00%
1005 BC 2	96.92%	0.06%	0.00%	-	0.01%	3.00%	92.09%	0.00%	0.01%
1006 AH 1	17.32%	0.05%	0.00%	-	0.02%	82.60%	12.74%	0.02%	0.02%
1006 AH 2	16.73%	0.04%	0.00%	-	0.03%	83.18%	12.21%	0.02%	0.03%
1006 BC 1	17.22%	0.05%	0.00%	-	0.02%	82.69%	12.69%	0.02%	0.02%
1006 BC 2	16.73%	0.04%	0.00%	-	0.03%	83.18%	12.30%	0.02%	0.03%
1007 AH 1	98.29%	0.02%	0.00%	-	0.00%	1.68%	91.13%	0.00%	0.00%
1007 AH 2	97.08%	0.02%	0.00%	0.00%	0.01%	2.88%	89.97%	0.00%	0.01%
1007 BC 1	98.23%	0.02%	0.00%	0.00%	0.00%	1.74%	91.22%	0.00%	0.00%
1007 BC 2	97.03%	0.02%	0.00%	0.00%	0.01%	2.93%	90.12%	0.00%	0.01%
1008 AH 1	97.60%	0.01%	-	-	0.01%	2.38%	91.44%	0.00%	0.01%
1008 AH 2	96.60%	0.00%	0.00%	0.00%	0.01%	3.37%	90.49%	0.00%	0.01%
1008 BC 1	97.54%	0.01%	-	0.00%	0.01%	2.44%	91.52%	0.00%	0.01%
1008 BC 2	96.50%	0.00%	-	-	0.02%	3.47%	90.59%	0.00%	0.02%
1009 AH 1	72.97%	0.01%	0.00%	0.00%	0.02%	26.99%	69.45%	0.01%	0.02%
1009 AH 2	71.69%	0.01%	0.00%	0.00%	0.03%	28.26%	68.21%	0.01%	0.03%
1009 BC 1	72.75%	0.02%	0.00%	0.00%	0.02%	27.20%	69.22%	0.01%	0.02%
1009 BC 2	71.70%	0.01%	0.00%	0.00%	0.03%	28.24%	68.23%	0.01%	0.03%
1010 AH 1	14.33%	0.03%	0.00%	0.00%	0.02%	85.61%	9.20%	0.03%	0.02%
1010 AH 2	13.97%	0.02%	0.00%	0.00%	0.02%	85.96%	8.94%	0.03%	0.02%
1010 BC 1	14.12%	0.03%	0.00%	0.00%	0.02%	85.83%	9.15%	0.03%	0.02%
1010 BC 2	13.79%	0.02%	0.00%	0.00%	0.03%	86.14%	8.91%	0.02%	0.03%
1011 AH 1	58.50%	0.02%	0.00%	0.00%	0.02%	41.44%	53.28%	0.01%	0.02%
1011 AH 2	58.00%	0.01%	0.00%	0.00%	0.03%	41.94%	52.80%	0.02%	0.03%
1011 BC 1	58.30%	0.02%	0.00%	0.00%	0.03%	41.64%	53.10%	0.02%	0.03%
1011	57.72%	0.01%	0.00%	0.00%	0.04%	42.21%	52.57%	0.02%	0.04%

BC 2									
1012 AH 1	17.31%	0.02%	0.00%	-	0.02%	82.64%	12.17%	0.02%	0.02%
1012 AH 2	16.85%	0.02%	0.00%	0.00%	0.02%	83.08%	11.82%	0.02%	0.02%
1012 BC 1	17.20%	0.02%	0.00%	-	0.02%	82.74%	12.13%	0.02%	0.02%
1012 BC 2	16.82%	0.02%	0.00%	0.00%	0.03%	83.12%	11.83%	0.02%	0.03%
1013 AH 1	96.50%	0.04%	0.00%	0.00%	0.00%	3.45%	90.88%	0.00%	0.00%
1013 AH 2	95.14%	0.04%	0.00%	0.00%	0.01%	4.80%	89.58%	0.00%	0.01%
1013 BC 1	96.44%	0.05%	0.00%	0.00%	0.00%	3.51%	90.90%	0.00%	0.00%
1013 BC 2	95.09%	0.04%	0.00%	0.00%	0.01%	4.85%	89.65%	0.00%	0.01%
1014 AH 1	81.84%	0.01%	0.00%	0.00%	0.01%	18.13%	78.14%	0.00%	0.01%
1014 AH 2	80.64%	0.01%	0.00%	-	0.01%	19.33%	76.97%	0.00%	0.01%
1014 BC 1	81.65%	0.01%	0.00%	0.00%	0.01%	18.32%	77.98%	0.00%	0.01%
1014 BC 2	80.39%	0.01%	0.00%	-	0.02%	19.57%	76.75%	0.00%	0.02%
1015 AH 1	96.65%	0.01%	-	-	0.01%	3.33%	90.17%	0.00%	0.01%
1015 AH 2	95.87%	0.01%	0.00%	0.00%	0.02%	4.10%	89.43%	0.00%	0.02%
1015 BC 1	96.60%	0.01%	0.00%	0.00%	0.01%	3.38%	90.32%	0.00%	0.01%
1015 BC 2	95.65%	0.01%	0.00%	0.00%	0.02%	4.31%	89.48%	0.00%	0.02%
1016 AH 1	81.29%	0.01%	0.00%	0.00%	0.01%	18.68%	76.99%	0.01%	0.01%
1016 AH 2	79.98%	0.01%	0.00%	-	0.01%	19.99%	75.72%	0.00%	0.01%
1016 BC 1	81.05%	0.01%	0.00%	0.00%	0.01%	18.92%	76.80%	0.01%	0.01%
1016 BC 2	79.84%	0.01%	0.00%	0.00%	0.02%	20.12%	75.64%	0.01%	0.02%
1017 AH 1	98.04%	0.03%	0.00%	-	0.02%	1.89%	0.05%	93.14%	0.02%
1017 AH 2	96.71%	0.03%	0.00%	0.00%	0.03%	3.21%	0.04%	91.83%	0.03%
1017 BC 1	97.99%	0.05%	0.00%	-	0.02%	1.93%	0.05%	93.03%	0.02%
1017 BC 2	97.02%	0.02%	0.00%	0.00%	0.04%	2.90%	0.04%	92.09%	0.04%
1018 AH 1	98.12%	0.01%	-	-	0.01%	1.84%	0.04%	94.68%	0.01%
1018	97.06%	0.01%	0.00%	-	0.02%	2.90%	0.03%	93.62%	0.02%

AH 2									
1018 BC 1	98.08%	0.01%	0.00%	-	0.02%	1.88%	0.04%	94.65%	0.02%
1018 BC 2	97.04%	0.01%	0.00%	-	0.03%	2.91%	0.03%	93.61%	0.03%
1019 AH 1	98.19%	0.03%	0.00%	-	0.00%	1.78%	95.54%	0.00%	0.00%
1019 AH 2	95.70%	0.03%	0.00%	0.00%	0.01%	4.26%	93.10%	0.00%	0.01%
1019 BC 1	98.14%	0.04%	0.00%	0.00%	0.00%	1.82%	95.51%	0.00%	0.00%
1019 BC 2	96.64%	0.03%	0.00%	-	0.01%	3.31%	94.04%	0.00%	0.01%

1001 CBS767 reference
1006, 1010 and 1012 Probably not *Debaryomyces hansenii*
1009, 1011 Genome ~double size
1017, 1018 *Saccharomyces cerevisiae*

Mahesh assemblies results

- ✧ **Classified sequences:** Found in the kraken custom database, sequences matching bacterial, archaeal, plasmids, viral, human or fungi domains.
- ✧ **Unclassified sequences:** Not found in the custom database.

fasta	Total sequences	Classified sequences	Unclassified sequence
1001	840	684 (81.43%)	156 (18.57%)
1002	804	735 (91.42%)	69 (8.58%)
1003	714	517 (72.41%)	197 (27.59%)
1004	15474	14975 (96.78%)	499 (3.22%)
1005	2070	1090 (52.66%)	980 (47.34%)
1006	593	218 (36.76%)	375 (63.24%)
1007	1421	1162 (81.77%)	259 (18.23%)
1008	1024	929 (90.72%)	95 (9.28%)
1009	1029	994 (96.60%)	35 (3.40%)
1010	396	297 (75.00%)	99 (25.00%)
1011	5889	5641 (95.79%)	248 (4.21%)
1012	454	200 (44.05%)	254 (55.95%)
1013	1483	996 (67.16%)	487 (32.84%)
1014	906	670 (73.95%)	236 (26.05%)
1015	1140	1029 (90.26%)	111 (9.74%)
1016	723	545 (75.38%)	178 (24.62%)
1017	1008	998 (99.01%)	10 (0.99%)
1018	1326	1298 (97.89%)	28 (2.11%)
1019	1023	715 (69.89%)	308 (30.11%)

Table resume of the reports

Strain	Fungi	Bacteria	Viruses	Archaea	Others	Unclassified	Deha	Sc
1001	80.36%	0.48%	-	-	-	18.57%	74.29%	-
1002	89.18%	0.25%	-	-	0.37%	8.58%	81.59%	-

1003	70.17%	1.26%	0.14%	-	-	27.59%	64.85%	-
1004	96.37%	0.25%	0.03%	-	0.01%	3.22%	95.83%	-
1005	45.51%	6.14%	0.43%	-	0.05%	47.34%	41.59%	-
1006	30.19%	4.22%	0.51%	-	-	63.24%	25.80%	-
1007	79.24%	1.69%	-	-	0.21%	18.23%	73.61%	-
1008	88.96%	0.39%	-	-	0.10%	9.28%	79.20%	-
1009	96.60%	-	-	-	-	3.40%	95.43%	-
1010	73.48%	0.51%	-	-	0.76%	25.00%	66.41%	-
1011	95.72%	-	-	0.02%	0.07%	4.19%	95.19%	00.02%
1012	40.53%	0.88%	-	-	0.22%	55.95%	33.48%	00.22%
1013	61.90%	3.78%	0.40%	-	0.07%	32.84%	55.02%	-
1014	72.41%	0.33%	-	-	0.11%	26.05%	66.67%	-
1015	88.51%	0.26%	-	-	0.35%	9.74%	77.82%	-
1016	73.44%	0.69%	-	-	0.28%	24.62%	65.15%	-
1017	98.91%	0.10%	-	-	-	0.99%	00.00%	98.12%
1018	97.89%	-	-	-	-	2.11%	00.00%	97.51%
1019	65.59%	3.03%	0.20%	-	-	30.11%	61.00%	00.20%

1001 CBS767 reference

1006, 1010 and 1012 Probably not *Debaryomyces hansenii*

1009, 1011 Genome ~double size

1017, 1018 *Saccharomyces cerevisiae*

Results more resumed explained in the meeting.

ITS sequences.

I have been able to get the ITS regions of all the strains except from 1018, one of the *Saccharomyces cerevisiae* strains. Another problem that I had is that all of them had the region quite extended except from some strains that casually corresponds to the two different assemblers used: Spades or MaSuRCA. The ones assembled with Spades have better regions normally containing 18s and 28s at the extremes, and the ones with MaSuRCA I was only able to find some really short sequences (1009, 1010, 1011, 1017, 1018).

When trying to get this region on the reference genome CBS767 I discovered that it doesn't exist! So there is no way to align them to a reference. I will try to find it for the other reference genome.

I did some alignments that I will show you, but it is quite impossible to make some phylogeny from that to be reliable because of the MaSuRCA assemblies.