2nd of November – 3rd meeting *Debaryomyces hansenii*

Important points of previous meetings:

- Put everything on github – done
- Map against *Saccharomyces cerevisiae* – done
- Check contaminants – done with kraken
- Variant calling – done using VarScan and mpileup (results not revised yet)
- Know what assembler used by Mahesha – not known
- Annotation using Maker – Not started
- BLAST of the assemblies – Running

New results and information:

- GitHub: https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii
    - Results included on README files of each folder, or specified where to find them on the README file.
- Mapping against genome reference of *D. hansenii*, CBS 767 and *S. cerevisiae*,S288c
- Contamination check using kraken on the raw-reads and on Mahesha assemblies, also done with the reference CBS 767, but as it is a whole sequence each chromosome, there is a 100% match of the sequences with classified sequences of the kraken standard database.
- Reports of contamination checks located on GitHub – In this document included a resume of contamination found in Mahesha assemblies.
- Strains information – alternative names for each one.
- Specific information of previous investigations using 1006 (NCYC 459), 1010 (NCYC 3364) and 1012 (CBS 5140) (that probably are not *Debaryomyces hansenii*).
- New mapping, this time against ATCC 6260, the genome reference of *Pichia guilliermondii*, also known as *Meyerozyma guilliermondii*, known to be impossible to differentiate from *D. hansenii* phenotypically, no match found.
- *Candida palmioleophila* also known to be mistaken with *D. hansenii*. No genome reference found.
- Blastx against nr running with 4 assemblies for now (1001, 1002, 1006 and 1017). No results yet.

Results

Mapping against CBS767

Overall alignment rate (%)

Strain	AH	BC
1001	79.02	78.83
1002	90.26	89.92
1003	77.46	77.38
1004	85.48	85.85
1005	87.30	87.29
1006	8.68	8.76
1007	87.61	87.48
1008	89.23	89.00
1009	68.99	68.94
1010	5.71	5.73
1011	51.41	51.11
1012	8.98	8.96
1013	86.48	86.45
1014	78.13	77.92
1015	88.65	88.47
1016	75.98	75.87
1017	0.08	0.09
1018	0.08	0.08
1019	87.51	87.94

Mapping against S288c

Overall alignment rate (%)

Strain	AH	BC
1001	1.53	1.49
1002	2.25	2.20
1003	1.77	1.73
1004	1.17	1.14
1005	1.42	1.39
1006	1.74	1.70
1007	2.45	2.39
1008	2.61	2.55
1009	1.70	1.69
1010	2.06	1.96
1011	2.55	2.52
1012	2.12	2.08
1013	2.47	2.42
1014	1.65	1.62
1015	2.48	2.40
1016	2.10	2.07
1017	95.75	95.59
1018	96.10	95.91
1019	1.09	1.08

1001 = CBS767 (*D. hansenii* reference)

1017 and 1018 known to be *S. cerevisiae*

1006, 1010 and 1012, probably not *D. hanenii*

Contamination check

General first result, percentaje of contaminated sequences.

- **Classified sequences:** Found in the kraken standard database, sequences matching bacterial, archaeal or viral domains.
- **Unclassified sequences:** Not found in the standard database. Not archaeal, bacterial or viral domain. Before future analysis we could assume that the rest would be sequences of our specie.

Mahesha assemblies:

fasta	Total sequences	Classified sequences	Unclassified sequences
1001	840	69 (8.21%)	771 (91.79%)
1002	804	80 (9.95%)	724 (90.05%)
1003	714	71 (9.94%)	643 (90.06%)
1004	15474	305 (1.97%)	15169 (98.03%)
1005	2070	194 (9.37%)	1876 (90.63%)
1006	593	88 (14.84%)	505 (85.16%)
1007	1421	76 (5.35%)	1345 (94.65%)
1008	1024	126 (12.30%)	898 (87.70%)
1009	1029	228 (22.16%)	801 (77.84%)
1010	396	84 (21.21%)	312 (78.79%)
1011	5889	268 (4.55%)	5621 (95.45%)
1012	454	74 (16.30%)	380 (83.70%)
1013	1483	133 (8.97%)	1350 (91.03%)
1014	906	88 (9.71%)	818 (90.29%)
1015	1140	96 (8.42%)	1044 (91.58%)
1016	723	72 (9.96%)	651 (90.04%)
1017	1008	158 (15.67%)	850 (84.33%)
1018	1326	138 (10.41%)	1188 (89.59%)
1019	1023	89 (8.70%)	934 (91.30%)

Raw-reads

fastq	Total sequences	Classified sequences	Unclassified sequences
1001 AH 1	2462393	4218 (0.17%)	2458175 (99.83%)
1001 AH 2	2462393	4490 (0.18%)	2457903 (99.82%)
1001 BC 1	3480996	6040 (0.17%)	3474956 (99.83%)
1001 BC 2	3480996	6252 (0.18%)	3474744 (99.82%)
1002 AH 1	1160378	2457 (0.21%)	1157921 (99.79%)
1002 AH 2	1160378	2479 (0.21%)	1157899 (99.79%)
1002 BC 1	1853630	3816 (0.21%)	1849814 (99.79%)
1002 BC 2	1853630	3927 (0.21%)	1849703 (99.79%)
1003 AH 1	2193546	4473 (0.20%)	2189073 (99.80%)
1003 AH 2	2193546	4551 (0.21%)	2188995 (99.79%)
1003 BC 1	3349842	6988 (0.21%)	3342854 (99.79%)
1003 BC 2	3349842	6964 (0.21%)	3342878 (99.79%)
1004 AH 1	1558691	3031 (0.19%)	1555660 (99.81%)
1004 AH 2	1558691	2975 (0.19%)	1555716 (99.81%)
1004 BC 1	2449087	4880 (0.20%)	2444207 (99.80%)
1004 BC 2	2449087	4581 (0.19%)	2444506 (99.81%)
1005 AH 1	2088288	4554 (0.22%)	2083734 (99.78%)
1005 AH 2	2088288	4600 (0.22%)	2083688 (99.78%)
1005 BC 1	2922955	6602 (0.23%)	2916353 (99.77%)
1005 BC 2	2922955	6244 (0.21%)	2916711 (99.79%)
1006 AH 1	1929277	4030 (0.21%)	1925247 (99.79%)
1006 AH 2	1929277	4074 (0.21%)	1925203 (99.79%)
1006 BC 1	2847843	5991 (0.21%)	2841852 (99.79%)

fastq	Total sequences	Classified sequences	Unclassified sequences
1006 BC 2	2847843	5937 (0.21%)	2841906 (99.79%)
1007 AH 1	2305984	5389 (0.23%)	2300595 (99.77%)
1007 AH 2	2305984	5370 (0.23%)	2300614 (99.77%)
1007 BC 1	3473851	8250 (0.24%)	3465601 (99.76%)
1007 BC 2	3473851	8072 (0.23%)	3465779 (99.77%)
1008 AH 1	1312616	3184 (0.24%)	1309432 (99.76%)
1008 AH 2	1312616	3037 (0.23%)	1309579 (99.77%)
1008 BC 1	1940489	4563 (0.24%)	1935926 (99.76%)
1008 BC 2	1940489	4432 (0.23%)	1936057 (99.77%)
1009 AH 1	1944249	3770 (0.19%)	1940479 (99.81%)
1009 AH 2	1944249	3908 (0.20%)	1940341 (99.80%)
1009 BC 1	2894462	5777 (0.20%)	2888685 (99.80%)
1009 BC 2	2894462	5932 (0.20%)	2888530 (99.80%)
1010 AH 1	1319601	2676 (0.20%)	1316925 (99.80%)
1010 AH 2	1319601	2714 (0.21%)	1316887 (99.79%)
1010 BC 1	1909206	3707 (0.19%)	1905499 (99.81%)
1010 BC 2	1909206	3663 (0.19%)	1905543 (99.81%)
1011 AH 1	768699	1806 (0.23%)	766893 (99.77%)
1011 AH 2	768699	1853 (0.24%)	766846 (99.76%)
1011 BC 1	1178421	2840 (0.24%)	1175581 (99.76%)
1011 BC 2	1178421	2867 (0.24%)	1175554 (99.76%)
1012 AH 1	2073205	4299 (0.21%)	2068906 (99.79%)
1012 AH 2	2073205	4274 (0.21%)	2068931 (99.79%)
1012 BC 1	3224847	6586 (0.20%)	3218261 (99.80%)
1012 BC 2	3224847	6744 (0.21%)	3218103 (99.79%)
1013 AH 1	1828482	4889 (0.27%)	1823593 (99.73%)
1013 AH 2	1828482	4751 (0.26%)	1823731 (99.74%)
1013 BC 1	2736996	7244 (0.26%)	2729752 (99.74%)
1013 BC 2	2736996	7079 (0.26%)	2729917 (99.74%)
1014 AH 1	2206902	4290 (0.19%)	2202612 (99.81%)
1014 AH 2	2206902	4180 (0.19%)	2202722 (99.81%)
1014 BC 1	3528456	6791 (0.19%)	3521665 (99.81%)
1014 BC 2	3528456	6712 (0.19%)	3521744 (99.81%)
1015 AH 1	1923464	4333 (0.23%)	1919131 (99.77%)
1015 AH 2	1923464	4365 (0.23%)	1919099 (99.77%)
1015 BC 1	3181425	7000 (0.22%)	3174425 (99.78%)
1015 BC 2	3181425	7160 (0.23%)	3174265 (99.77%)
1016 AH 1	2646260	5434 (0.21%)	2640826 (99.79%)
1016 AH 2	2646260	5487 (0.21%)	2640773 (99.79%)
1016 BC 1	3919163	8065 (0.21%)	3911098 (99.79%)
1016 BC 2	3919163	7991 (0.20%)	3911172 (99.80%)
1017 AH 1	1889165	8730 (0.46%)	1880435 (99.54%)
1017 AH 2	1889165	8552 (0.45%)	1880613 (99.55%)
1017 BC 1	2738771	13410 (0.49%)	2725361 (99.51%)
1017 BC 2	2738771	13166 (0.48%)	2725605 (99.52%)
1018 AH 1	2628541	8423 (0.32%)	2620118 (99.68%)
1018 AH 2	2628541	8583 (0.33%)	2619958 (99.67%)
1018 BC 1	3662396	12651 (0.35%)	3649745 (99.65%)

fastq	Total sequences	Classified sequences	Unclassified sequences
1018 BC 2	3662396	12665 (0.35%)	3649731 (99.65%)
1019 AH 1	2000344	3714 (0.19%)	1996630 (99.81%)
1019 AH 2	2000344	3698 (0.18%)	1996646 (99.82%)
1019 BC 1	2664496	5139 (0.19%)	2659357 (99.81%)
1019 BC 2	2664496	5031 (0.19%)	2659465 (99.81%)

1006, 1010 and 1012: same levels of contamination than the rest. Not the reason why they map worse to *Debaryomyces hansenii* CBS 767.

More specific results of contamination using Kraken on the Mahesha assemblies

1001

- Unclassified: 91.79 %
- Classified: 8.21 %
  - Cellular organisms: 6.90 %
    - Bacteria: 5.95 %
    - Archaea: 0.95 %
  - Viruses: 0.48 %

1002

- Unclassified: 90.05 %
- Classified: 9.95 %
  - Cellular organisms: 7.21 %
    - Bacteria: 6.47 %
    - Archaea: 0.37 %
  - Viruses: 1.24 %

1003

- Unclassified: 90.06 %
- Classified: 9.94 %
  - Cellular organisms: 8.26 %
    - Bacteria: 7.56 %
    - Archaea: 0.70 %
  - Viruses: 0.56 %

1004

- Unclassified: 90.63 %
- Classified: 9.37 %
  - Cellular organisms: 8.21 %
    - Bacteria: 8.12 %
    - Archaea: 0.05 %
  - Viruses: 0.72 %

1005

- Unclassified: 98.03 %
- Classified: 1.97 %
  - Cellular organisms: 1.49 %
    - Bacteria: 1.40 %
    - Archaea: 0.08 %
  - Viruses: 0.30 %

1006

- Unclassified: 85.16 %
- Classified: 14.84 %
  - Cellular organisms: 11.47 %
    - Bacteria: 10.12 %
    - Archaea: 1.18 %
  - Viruses: 1.69 %

1007

- Unclassified: 94.65 %
- Classified: 5.35 %
  - Cellular organisms: 4.15 %
    - Bacteria: 3.94 %
    - Archaea: 0.21 %
  - Viruses: 0.63 %

1008

- Unclassified: 87.70 %
- Classified: 12.30 %
  - Cellular organisms: 9.18 %
    - Bacteria: 7.91 %
    - Archaea: 1.17 %
  - Viruses: 1.95 %

1009

- Unclassified: 77.84 %
- Classified: 22.16 %
  - Cellular organisms: 17.98 %
    - Bacteria: 16.72 %
    - Archaea: 1.26 %
  - Viruses: 3.30 %

1010

- Unclassified: 78.79 %
- Classified: 21.21 %
  - Cellular organisms: 17.42 %
    - Bacteria: 16.92 %
    - Archaea: 0.51 %
  - Viruses: 2.78 %

1011

- Unclassified: 95.45 %
- Classified: 4.55 % - Cellular organisms: 3.50 % - Bacteria: 3.38 % - Archaea: 0.08 %
  - Viruses: 0.68 %

1012

- Unclassified: 83.70 %
- Classified: 16.30 %
  - Cellular organisms: 11.89 %
    - Bacteria: 10.35 %
    - Archaea: 0.88 %
  - Viruses: 1.98 %

1013

- Unclassified: 91.03 %
- Classified: 8.97 %
  - Cellular organisms: 7.08 %
    - Bacteria: 6.68 %
    - Archaea: 0.34 %
  - Viruses: 1.15 %

1014

- Unclassified: 90.29 %
- Classified: 9.71 %
  - Cellular organisms: 7.28 %
    - Bacteria: 6.84 %
    - Archaea: 0.44 %
  - Viruses: 0.99 %

1015

- Unclassified: 91.58 %
- Classified: 8.42 %
  - Cellular organisms: 6.05 %
    - Bacteria: 5.09 %
    - Archaea: 0.61 %
  - Viruses: 1.32 %

1016

- Unclassified: 90.04 %
- Classified: 9.96 %
  - Cellular organisms: 7.88 %
    - Bacteria: 6.78 %
    - Archaea: 1.11 %
  - Viruses: 0.83 %

1017

- Unclassified: 84.33 %
- Classified: 15.67 %
  - Cellular organisms: 11.41 %
    - Bacteria: 9.62 %
    - Archaea: 1.09 %
  - Viruses: 3.57 %

1018

- Unclassified: 89.59 %
- Classified: 10.41 %
  - Cellular organisms: 7.69 %
    - Bacteria: 6.18 %
    - Archaea: 0.98 %
  - Viruses: 2.11 %

1019

- Unclassified: 91.30 %
- Classified: 8.70 %
  - Cellular organisms: 6.84 %
    - Bacteria: 6.65 %
    - Archaea: 0.10 %
  - Viruses: 0.98 %

Extracted from reports located in https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii/tree/master/Work_files/mahesha_assemblies_workfolder/contamination_check/kraken

Reports of Kraken using raw-reads: https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii/tree/master/Work_files/rawdata_workfolder/contamination_check/kraken_results

Report of Kraken using the reference genome CBS 767: https://github.com/The-Bioinformatics-Group/Debaryomyces_hansenii/tree/master/Work_files/reference/kraken_results

More information about each strain:

Strain	Alternative strain names
1001	NCYC 2572, CBS767, ATCC 36239, CCRC 21394, DBVPG 6050, IFO 0083, JCM 1990, JCM 2102, KCTC 7645, MUCL 30242, NRRL Y-7426, NRRL Y-10976, UCD 74-86
1002	NCYC 8, NCTC 2059
1003	NCYC 9, NCTC 2048
1004	NCYC 10, NCTC 2056
1005	NCYC 103, NCTC 1681
1006	NCYC 459
1007	NCYC 475, CBS 811, JCM 1439, NRRL Y-1454, UCD 75-11
1008	NCYC 792, NCMB 1230 strain 43
1009	NCYC 3045
1010	NCYC 3364
1011	CBS 117
1012	CBS 5140
1013	CBS 1101, IFO 0027, IFO 0093
1014	CBS 1792
1015	J63
1016	J26
1017	J16
1018	J52
1019	DBH9

NCYC 459 (1006) in publications (All of them are phenotypic studies):

Named as *Debaryomyces subglobulosus*:

Anderson, F.B. and Harris, G. 1963. The production of Riboflavin and D-Arabitol by *Debaryomyces subglobulosus*. Journal of genetic Microbiology 33: 137-156.

Wase, D.A.J. and Hough, J.S. 1966. Continuous culture of yeast on phenol. Microbiology 42: 13-23.

Johnson, B. Nelson, S.J., Brown, C.M. 1972. Influence of glucose concentration on the physiology and lipid composition of some yeasts. Antoine van Leeuwenhoek 38: 129-136.

Johnson, B. and Brown, C.M. 1972. A possible relationship between the fatty acid composition of yeasts and the 'petite' mutation. Antoine van Leeuwenhoek 38: 137-144.

Nelson, G. and Young, T.W. 1987. The addition of proteases to the fermenter to control chill-haze formation. Journal of the Institute of Brewing 93, 2: 116-120.

Named as *Debaryomyces hansenii*:

Nelson, G. and Young, T.W. 1986. Yeast extracellular proteolytic enzymes for chill-proofing beer. Journal of the Institute of Brewing 92, 6: 599-603.

Gadd, G.M. and Edwards, S.W. 1986. Heavy-metal-induced flavin production by *Debaryomyces hansenii* and possible connexions with iron metabolism. Transactions of the British Mycological Society 84, 4: 533-542.

Reed, R.H., Chudek, J.A., Foster, R., Gadd, G.M. 1987. Osmotic significance of glycerol accumulation in exponentially growing yeasts. Applied and Enviromental Microbiology 53, 9: 2119-2123.

Van Dyke, M.I., Lee, H., Trevors, J.T. 1989. Germanium toxicity in selected bacterial and yeast strains. Journal of Industrial Microbiology 4: 299-306.

Kierans, M., Staines, A.M., Bennett, H., Gadd, G.M. 1991. Silver tolerance and accumulation in yeasts.

Biology of Metals 4: 100-106.
Meikle, A.J., Chudek, J.A., Reed, R.H., Gadd, G.M. 1991. Natural abundance of 13C-nuclear magnetic resonance spectroscopic analysis of acyclic polyool and trehalose accumulation by several yeast species in response to salt stress. FEMS Microbiology Letters 82, 2: 163-168.
Gharieb, M.M., Wilkinson, S.C., Gadd, G.M. 1994. Reduction of selenium oxyanions by unicellular, polymorphic and filamentous fungi: cellular location of reduces selenium and implications for tolerance. Journal of Industrial Microbiology 14: 300-311.

NCYC 3364 (1010): No publications using this strain. Registered as: *Debaryomyces hansenii var. fabryii*

CBS 5140 (1012): Recieved as *Debaryomyces subglobulosus*, changed to *Debaryomyces hansenii*.
Publications:

Named as *Debaryomyces hansenii*:

Tunblad-Johansson, I. And Adler, L. 1987. Effects of sodium chloride concentration on phospholipid fatty acid composition of yeasts differing in osmotolerance. FEMS Microbiology Letters 43, 3: 275-278.

Named as *Candida haemulonii*:

Chowdhary, A., Anil Kumar, V., Sharma, C., Prakash, A., Agarwal, K., Babu, R., Dinesh, K.R., Karim, S., Singh, S.K., Hagen, F., Meis, J.F. 2014. Multidrug-resistant endemic clonal strain of *Candida auris* in India. European Journal of Clinical Microbiology and Infectious Diseases 33, 6: 919-926.

Newest results

Because of articles explaining the difficulties on differentiating between *Debaryomyces hansenii* and *Picchia guillermondii*, mapping against ATCC 6260 results:

Overall alignment rate (%)

Strain	AH	BC
1001	2.16	2.12
1002	3.22	3.20
1003	2.73	2.69
1004	2.05	2.05
1005	3.12	3.09
1006	2.54	2.51
1007	3.99	3.93
1008	3.61	3.54
1009	2.48	2.47
1010	2.88	2.77

Strain	AH	BC
1011	3.83	3.80
1012	2.94	2.91
1013	3.92	3.88
1014	2.35	2.33
1015	3.54	3.46
1016	3.06	3.03
1017	4.12	4.18
1018	2.81	2.80
1019	2.61	2.62