

Supplementary material

Genotypic and phenotypic diversity among *Komagataella* species reveals a hidden pathway for xylose utilization

Lina Heisteringer^{1,#,+,*}, Juliane C. Dohm^{2*}, Barbara G. Paes^{1,3*}, Daniel Koizar¹, Christina Troyer⁴, Özge Ata^{1,5}, Teresa Steininger-Mairinger⁴, Diethard Mattanovich^{1,5}

¹ University of Natural Resources and Life Sciences Vienna (BOKU), Department of Biotechnology, Institute of Microbiology and Microbial Biotechnology, 1190 Vienna, Austria

² University of Natural Resources and Life Sciences Vienna (BOKU), Department of Biotechnology, Institute of Computational Biology, 1190 Vienna, Austria

³ University of Brasilia (UnB), Institute of Biological Sciences, Department of Cell Biology, Brasilia, Brazil

⁴ University of Natural Resources and Life Sciences Vienna (BOKU), Department of Chemistry, Institute of Analytical Chemistry, 1190 Vienna, Austria

⁵ Austrian Centre of Industrial Biotechnology (acib GmbH), 1190 Vienna, Austria

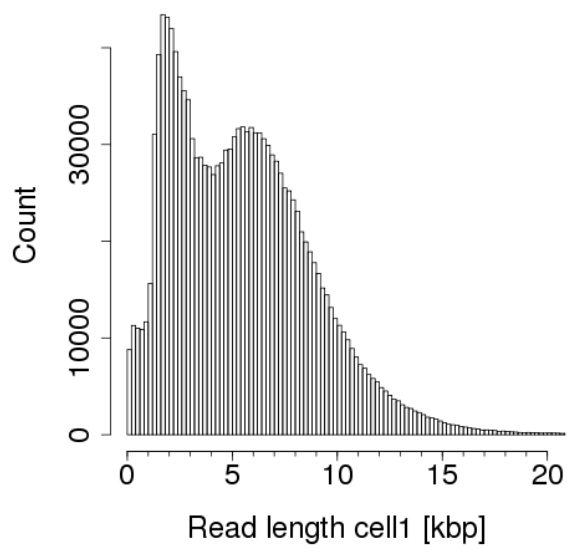


Figure S1: Length distribution of PacBio subreads from one SMRT cell. The other two cells showed similar length profiles.

Table S1: Pairwise distances between genomes of eleven species as determined by Mash dist [1].

	<i>C. matrit.</i>	<i>K. mondav.</i>	<i>K. ulmi</i>	<i>K. populi</i>	<i>K. kurtzm.</i>	<i>K. phaffii</i> (2)	<i>K. pastoris</i> (2)	<i>K. pastoris</i> (3)	<i>K. pseudopast.</i>	Ref_ <i>K.pastoris</i>	Ref_ <i>K. phaffii</i>
<i>C. matritensis</i>	0	0.295981	0.263022	0.263022	0.263022	0.263022	0.263022	0.263022	0.263022	0.263022	0.263022
<i>K. mondav.</i>	0.295981	0	0.149777	0.0554283	0.149777	0.145727	0.14383	0.145727	0.052714	0.145727	0.151946
<i>K. ulmi</i>	0.263022	0.149777	0	0.174408	0.0774628	0.0782587	0.0374354	0.0374354	0.167688	0.0374354	0.0786628
<i>K. populi</i>	0.263022	0.0554283	0.174408	0	0.159146	0.151946	0.174408	0.174408	0.0259767	0.182269	0.161822
<i>K. kurtzm.</i>	0.263022	0.149777	0.0774628	0.159146	0	0.0181917	0.0747985	0.0751681	0.159146	0.0751681	0.0182524
<i>K. phaffii</i> (2)	0.263022	0.145727	0.0782587	0.151946	0.0181917	0	0.0759181	0.0762985	0.156619	0.0770708	0.00420918
<i>K. pastoris</i> (2)	0.263022	0.14383	0.0374354	0.174408	0.0747985	0.0759181	0	0.00238081	0.159146	0.0025468	0.0774628
<i>K. pastoris</i> (3)	0.263022	0.145727	0.0374354	0.174408	0.0751681	0.0762985	0.00238081	0	0.159146	0.00274254	0.0778587
<i>K. pseudopast.</i>	0.263022	0.052714	0.167688	0.0259767	0.159146	0.156619	0.159146	0.159146	0	0.164662	0.167688
Ref_ <i>K. pastoris</i>	0.263022	0.145727	0.0374354	0.182269	0.0751681	0.0770708	0.0025468	0.00274254	0.164662	0	0.0770708
Ref_ <i>K. phaffii</i>	0.263022	0.151946	0.0786628	0.161822	0.0182524	0.00420918	0.0774628	0.0778587	0.167688	0.0770708	0

C. matrit. = *Citeromyces matritensis*, *K. mondav.* = *K. mondaviorum*, *K. pseudopast.* = *K. pseudopastoris*, *K. phaffii* (2) = *K. phaffii* UWOPS 03-328y3, *K. pastoris* (2) = *K. pastoris* DSMZ 70877, *K. pastoris* (3) = *K. pastoris* CBS 9178, Ref_*K. pastoris* = *K. pastoris* CBS 704, Ref_*K. phaffii* = *K. phaffii* CBS 7435

Table S2: Orthologs of *Scheffersomyces stipitis* xylose reductase (XR) in different *Komagataella* species. Start and Stop refer to the nucleotide position of each gene in the genome assembly. Seq id gives the % sequence identity of the translated ORF to the *S. stipitis* protein.

Strain	Contig	Start	Stop	DNA strand	Seq id [%]	Protein length	Annotation in <i>K. phaffii</i> CBS 7435
<i>K. monodavium</i> CBS 15017	contig5	680262	679303	minus	68.15	319	Gre3
	contig2	289047	289977	plus	36.79	309	Ypr1
	contig1	2254062	2254907	plus	32.39	281	PP7435_Ch2-0714
	contig1	1034828	1033929	minus	30.65	299	PP7435_Ch4-0551
<i>K. ulmi</i> CBS 12361	contig3	1026994	1026035	minus	67.87	320	Gre3
	contig1	1253391	1254320	plus	36.16	309	Ypr1
	contig2	1355033	1355878	plus	33.33	281	PP7435_Ch2-0714
	contig4	640886	639987	minus	29.17	299	PP7435_Ch4-0551
<i>K. populi</i> CBS 12362	contig4	818460	819419	plus	67.62	319	Gre3
	contig5	474100	473171	minus	36.8	309	Ypr1
	contig2	845514	844669	minus	32.7	281	PP7435_Ch2-0714
	contig1	1800090	1800989	plus	29.97	299	PP7435_Ch4-0551
<i>K. kurtzmanii</i> CBS 12817	contig2	1412680	1411721	minus	67.54	319	Gre3
	contig3	25157	26086	plus	36.36	309	Ypr1
	contig1	1344402	1345247	plus	33.65	281	PP7435_Ch2-0714
	contig4	382089	382988	plus	29.49	299	PP7435_Ch4-0551
<i>K. phaffii</i> UWOPS 03-328y3	contig2	1446336	1445377	minus	67.54	319	Gre3
	contig1	1235528	1236457	plus	36.99	309	Ypr1
	contig3	1065727	1064882	minus	33.96	281	PP7435_Ch2-0714
	contig4	345894	344995	minus	29.11	299	PP7435_Ch4-0551
<i>K. pastoris</i> DSMZ 70877	contig3	825331	826290	plus	68.53	319	Gre3
	contig1	2099022	2098093	minus	36.79	309	Ypr1
	contig2	1369684	1370529	plus	33.33	281	PP7435_Ch2-0714
	contig4	939156	940055	plus	28.80	299	PP7435_Ch4-0551
<i>K. pastoris</i> CBS 9178	contig3	1027767	1026808	minus	68.2	319	Gre3
	contig1	2079669	2078740	minus	36.79	309	Ypr1
	contig2	1336610	1335765	minus	33.33	281	PP7435_Ch2-0714
	contig4	939592	940491	plus	28.80	299	PP7435_Ch4-0551
<i>K. pseudopastoris</i> CBS 9187	contig2	1279287	1278328	minus	67.83	319	Gre3
	contig2	293297	294226	plus	37.12	309	Ypr1
	contig4	836018	835173	minus	32.39	281	PP7435_Ch2-0714
	contig1	993828	992929	minus	29.97	299	PP7435_Ch4-0551

Table S3: Orthologs of *S. stipitis* xylitol dehydrogenase (XDH) in different *Komagataella* species. Start and Stop refer to the nucleotide position of each gene in the genome assembly. Seq id gives the % sequence identity of the translated ORF to the *S. stipitis* protein.

Strain	Contig	Start	Stop	DNA strand	Seq id [%]	Protein length	Annotation in <i>K. phaffii</i> CBS 7435
<i>K. mondavium</i> CBS 15017	contig2	169306	170352	plus	54.25	348	Sor1
<i>K. ulmi</i> CBS 12361	contig1	1115051	1116097	plus	54.8	348	Sor1
<i>K. populi</i> CBS 12362	contig1	2634432	2633386	minus	53.7	348	Sor1
<i>K. kurtzmanii</i> CBS 12817	contig5	1095113	1096159	plus	54.52	348	Sor1
<i>K. phaffii</i> UWOPS 03-328y3	contig1	1097155	1098201	plus	54.8	348	Sor1
<i>K. pastoris</i> DSMZ 70877	contig1	2237414	2236368	minus	54.52	409	Sor1
<i>K. pastoris</i> CBS 9178	contig1	2218021	2216975	minus	54.52	409	Sor1
<i>K. pseudopastoris</i> CBS 9187	contig1	158944	159990	plus	53.97	348	Sor1

Table S4: Orthologs of *S. stipitis* xylulokinase (XKS) in different *Komagataella* species. Start and Stop refer to the nucleotide position of each gene in the genome assembly. Seq id gives the % sequence identity of the translated ORF to the *S. stipitis* protein.

Strain	Contig	Start	Stop	DNA strand	Seq id [%]	Protein length	Annotation in <i>K. phaffii</i> CBS 7435
<i>K. mondavium</i> CBS 15017	contig2	170770	172593	plus	49.76	607	Xks1
<i>K. ulmi</i> CBS 12361	contig1	1116531	1118348	plus	49.44	605	Xks1
<i>K. populi</i> CBS 12362	contig1	2632967	2631144	minus	50.08	607	Xks1
<i>K. kurtzmanii</i> CBS 12817	contig5	1096552	1098405	plus	49.28	617	Xks1
<i>K. phaffii</i> UWOPS 03-328y3	contig1	1098596	1100449	plus	49.76	617	Xks1
<i>K. pastoris</i> DSMZ 70877	contig1	2235937	2234120	minus	50.08	605	Xks1
<i>K. pastoris</i> CBS 9178	contig1	2216544	2214727	minus	50.24	605	Xks1
<i>K. pseudopastoris</i> CBS 9187	contig1	160409	162232	plus	49.84	607	Xks1

<i>S.stipitis</i> _CBS6054_XKS	INRLFVSDMNSFEV-KDKWLEYANGVGMLAKMESELKH*--	623
<i>K.mondavium</i> _CBS15017	LTKNFHWKENVKPVEADSSLWLQYAEVGVILSEIEQTLDRDS*	607
<i>K.populi</i> _CBS12362	LSKNFHWKENVKPVEADSSLWLQYAEVGVILSEIEQTLDRDH*	607
<i>K.pseudopastoris</i> _CBS9187	LTKYFHWKENVKPVEADSSLWLQYAEVGVILSEIEQTLDRDL*	607
<i>K.kurtzmanii</i> _CBS12817	LRKNFHWKENVKQVEADSSLWLQYVDGVGILSEIEQTLEK*--	617
<i>K.phaffii</i> _UWOPS03-328y3	LRKNFHWKENVKPVEADSSLWLQYVDGVGILSEIEQTLEK*--	617
<i>K.phaffii</i> _CBS7435	LRKNFHWKENVKPVEADSSLWLQYVDGVGILSEIEQTLEK*--	617
<i>K.ulmi</i> _CBS12361	LSKNFHWKENVKPVEADSSLWLQYVDGVGILSEIEQTLEK*--	605
<i>K.pastoris</i> _DSMZ70877	LSKNFHWKENVKPVEADSSLWLQYVDGVGILSEIEQTLEK*--	605
<i>K.pastoris</i> _CBS9178	LSKNFHWKENVKPVEADSSLWLQYVDGVGILSEIEQTLEK*--	605
<i>K.pastoris</i> _CBS704	LSKNFHWKENVKPVEADSSLWLQYVDGVGILSEIEQTLEK*--	605
	: : *. :.: : *. : **.*.:***.*.:.:*. *.:	

Figure S2: Multiple sequence alignments of the *S. stipitis* CBS 6054 xylose pathway enzymes A) xylose reductase (XR), B) xylulose dehydrogenase (XDH) and C) xylulokinase (XKS) with the closest orthologs from all seven *Komagataella* species. Alignments were generated using Clustal Omega [2]. Identical amino acids are marked with an asterisk, a colon indicates residues with highly similar properties and a point indicates residues with weakly similar properties.

<i>K.phaffii</i> _PP7435_Chr4-0551	-----MALAIPQIGFGTGTWYKYGRESEIDQTLVDALVSALQVGITHLDG	46
<i>K.phaffii</i> _PP7435_Chr2-0714	-----MSRYLQLNNGNKIPHIGFGTWEL----G-----RSQAANVVYHALKAGFRLIDT	45
<i>K.phaffii</i> _Gre3	-----MATLLKLNNGLKLPQVGLGVWKI----P-----NELTAETVYNAIKQGYRLFDG	45
<i>K.phaffii</i> _Ypr1	MAVPSNSTKTFKLNNGLSIPAVGLGTWQS----T-----DEEAYNAVIAALKAGYRHIDT	51
	. :* :*:.*. . :.: *.: * :*	
<i>K.phaffii</i> _PP7435_Chr4-0551	AECYGTALETRDAIKKA-----NIPREKLWITEKYAGDSSHKSskaanplaalk-EI	99
<i>K.phaffii</i> _PP7435_Chr2-0714	AYLYRNEKEVGEGERWLEDDEPENNRSDVFYTTKLWDSQFGYENAKR-----SIQRAI	99
<i>K.phaffii</i> _Gre3	AEDYGNEKEVGQVRRADIEG--LVKREDLFIVSKLWNNYHHPDNVVGK-----ALDRTL	97
<i>K.phaffii</i> _Ypr1	AYCYGNEEPIGKAI----KDS--GVARKDIFITTKLWGTDHTRT--EE-----GLDRSL	97
	* * . :.: *.: : * : . :.: :	
<i>K.phaffii</i> _PP7435_Chr4-0551	LRLEELEYVDLYLLHSPYITKET-----HGFTLEEAWGYLEEAQS	139
<i>K.phaffii</i> _PP7435_Chr2-0714	DQVPGLEYIDLLLHSPNGGPK-----VRKETYQAMQEAVD	135
<i>K.phaffii</i> _Gre3	S-DLGLDYLDLFYIHFPFAFKFVPLEEKYPFAFYCGDGNNFHYEDVPLLDTYRALERLVD	156
<i>K.phaffii</i> _Ypr1	K-LLGLDYVDLFLMHPVPMNPNNGHDKFPTLP--DGKRDILFDWNFVDTYREMQLVA	153
	*.:** :* * :.: :.	
<i>K.phaffii</i> _PP7435_Chr4-0551	LGLAKNIGVSNFTVADLEKILKVA--KVKPQVNQIEYNAFLQDQTPGVVEFSQKNGILIE	197
<i>K.phaffii</i> _PP7435_Chr2-0714	QGIVKQLGVSSWGEAHIKELFSWDGLKYKPVVNQVELSPWCMRE--RLVDFCHKNEILTE	193
<i>K.phaffii</i> _Gre3	AGRIKSLGVSNFNGALLQDLLRGA--RIKPVALQIEHHPYLVQQ--KLIEYAQSEDIVVV	212
<i>K.phaffii</i> _Ypr1	SGKTKAIGVSNFSITNLKLLADPEITIKPVVNQVEIHGYLPQQ--RLLEYAKENDIVLE	211
	* * :***.: : :.: : ** . *:* : : : : : : *	
<i>K.phaffii</i> _PP7435_Chr4-0551	AYSPLAPLYKGDKS----IKEVKEF-LDYVYELGKKYKGSdTQVLLKWWLQKGIIPITTS	252
<i>K.phaffii</i> _PP7435_Chr2-0714	AYSPLARGGR-----FNEQVVKSIKNGKVTAPQVLLRWSIQKGIPIPKT	239
<i>K.phaffii</i> _Gre3	AYSSFGPQSFLKVNKALTAVSLFEHDVICKIAQAHNRSAGEVLLRWATQRLAIIPKS	272
<i>K.phaffii</i> _Ypr1	AYSPLGSTGA-----PLLKDELVDLAKKNGISESTLLISWAVWRGIVVLPKS	259
	*** :. : : : : : : : * : * : * : : :	
<i>K.phaffii</i> _PP7435_Chr4-0551	TKTERVKEFLKIDDFQLEPSE-VEKITTLGAKAPV-VRQYWVPEYSKFD*--	299
<i>K.phaffii</i> _PP7435_Chr2-0714	QTIGRLPENIDVFGFELNDQE---IGQLDH-PNAHDPTDWDVNVNP*----	281
<i>K.phaffii</i> _Gre3	SKPERLSSNLHINSFDLTK-EDLETISSLDLGLRFNDPWDWDKIPIFA*---	319
<i>K.phaffii</i> _Ypr1	VTPSRIADNLKIIELCEEDGKKLNELASIRGEKRL-VSPPWDPIVVFND*--	309
	. * : . :.: : : : : *	

Figure S3: Multiple sequence alignments of the *K. phaffii* CBS 7435 Gre3 homologs. Alignments were generated using Clustal Omega [2]. Identical amino acids are marked with an asterisk, a colon indicates residues with highly similar properties and a point indicates residues with weakly similar properties.

Table S5: Relative abundance of isotopologues of selected metabolites after cultivation on ^{12}C or $1\text{-}^{13}\text{C}$ D-xylose, average and standard deviation for n=3 biological replicates.

	^{12}C Xylose		$1\text{-}^{13}\text{C}$ D-xylose	
Isotopologue	average	s	average	s
2PG M0	92.6	4.2	69.9	9.1
2PG M1	7.0	3.9	24.0	6.0
2PG M2	0.3	0.5	6.1	3.5
2PG M3	0.0	0.0	0.0	0.0
3PG M0	93.9	5.3	69.7	8.7
3PG M1	5.6	4.7	23.4	6.7
3PG M2	0.5	0.7	6.9	2.1
3PG M3	0.0	0.0	0.0	0.0
6PGA M0	93.6	3.0	91.2	2.1
6PGA M1	3.1	4.4	3.3	4.7
6PGA M2	0.7	1.1	2.0	1.9
6PGA M3	1.9	1.5	3.0	2.1
6PGA M4	0.6	0.4	0.5	0.7
6PGA M5	0.0	0.0	0.0	0.0
6PGA M6	0.0	0.0	0.0	0.0
F6P M0	100.0	0.0	21.6	4.2
F6P M1	0.0	0.0	37.6	4.4
F6P M2	0.0	0.0	25.0	4.2
F6P M3	0.0	0.0	11.9	5.4
F6P M4	0.0	0.0	3.6	1.6
F6P M5	0.0	0.0	0.0	0.0
F6P M6	0.0	0.0	0.4	0.3
G6P M0	98.2	1.0	19.9	2.6
G6P M1	0.4	0.6	39.0	2.9
G6P M2	0.0	0.0	25.2	3.6
G6P M3	1.1	0.6	14.1	0.6
G6P M4	0.2	0.2	1.8	0.7
G6P M5	0.0	0.0	0.0	0.0
G6P M6	0.0	0.0	0.1	0.0
Glc M0	94.7	0.8	36.1	11.3
Glc M1	4.1	0.9	31.5	4.1
Glc M2	0.0	0.0	21.2	4.7
Glc M3	0.5	0.5	10.2	3.7
Glc M4	0.7	1.0	1.0	1.0
Glc M5	0.0	0.0	0.0	0.0
Glc M6	0.0	0.0	0.0	0.1
R5P M0	95.9	0.2	19.7	0.1
R5P M1	4.0	0.2	67.6	1.1
R5P M2	0.0	0.0	11.5	0.5
R5P M3	0.0	0.0	0.8	1.1
R5P M4	0.0	0.0	0.1	0.1
R5P M5	0.0	0.0	0.3	0.2
S7P M0	85.3	9.4	11.2	1.8
S7P M1	7.9	6.1	31.8	1.1
S7P M2	2.0	1.4	40.0	3.9

S7P M3	3.1	1.3	9.8	1.8
S7P M4	1.7	2.4	2.9	1.5
S7P M5	0.0	0.0	2.7	1.0
S7P M6	0.0	0.0	1.6	2.0
S7P M7	0.0	0.0	0.0	0.0

Table S6: List of metabolites and corresponding GC retention times (RT), chemical ionization adducts/fragments evaluated for ^{13}C isotopologue distribution analysis, m/z values of isotopologues extracted and respective mass extraction windows.

Name of derivatized metabolite *	RT (min)	Adduct/ Fragment Evaluated**	Isotopologue* **	m/z	m/z extraction window left / right (ppm)
2PG 4TMS	17.15	$[\text{MH}]^+$	M0	475.1583	50 / 50
			M1	476.1617	50 / 50
			M2	477.1650	50 / 50
			M3	478.1684	50 / 50
3PG 4TMS	17.56	$[\text{MH}]^+$	M0	475.1583	50 / 50
			M1	476.1617	50 / 50
			M2	477.1650	50 / 50
			M3	478.1684	50 / 50
Glc 5TMS 1EtOx	20.55	$[\text{M}-\text{CH}_3]^+$	M0	568.2792	50 / 50
			M1	569.2826	50 / 50
			M2	570.2859	50 / 50
			M3	571.2893	50 / 50
			M4	572.2926	50 / 50
			M5	573.2960	50 / 50
			M6	574.2993	50 / 50
R5P 5TMS 1EtOx	24.35	$[\text{M}-\text{CH}_3]^+$	M0	618.2350	50 / 50
			M1	619.2383	50 / 50
			M2	620.2417	50 / 50
			M3	621.2450	50 / 50
			M4	622.2484	50 / 50
			M5	623.2517	50 / 50
F6P 6TMS 1EtOx	28.59	$[\text{M}-\text{CH}_3]^+$	M0	720.2851	50 / 50
			M1	721.2884	50 / 50
			M2	722.2918	50 / 50
			M3	723.2951	50 / 50
			M4	724.2985	50 / 50
			M5	725.3018	50 / 50
			M6	726.3052	50 / 50
G6P 6TMS 1EtOx	28.91	$[\text{M}-\text{CH}_3]^+$	M0	720.2851	50 / 50
			M1	721.2884	50 / 50
			M2	722.2918	50 / 50
			M3	723.2951	50 / 50
			M4	724.2985	50 / 50
			M5	725.3018	50 / 50
			M6	726.3052	50 / 50

6PGA 7TMS	30.02	$[M-CH_3]^+$	M0	765.2773	50 / 50
			M1	766.2807	50 / 50
			M2	767.2840	50 / 50
			M3	768.2874	50 / 50
			M4	769.2907	50 / 50
			M5	770.2941	50 / 50
			M6	771.2974	50 / 50
S7P 7TMS 1EtOx	32.31	$[M-CH_3]^+$	M0	822.3351	50 / 50
			M1	823.3385	50 / 50
			M2	824.3419	50 / 50
			M3	825.3452	50 / 50
			M4	826.3486	50 / 50
			M5	827.3519	50 / 50
			M6	828.3553	50 / 50
			M7	829.3586	50 / 50

* names of derivatized metabolites are given specifying the number of trimethylsilyl (TMS) and ethoxyamino groups (EtOx)

** Positive chemical ionization fragments used for data evaluation: $[M+H]^+$: protonated molecular ion, $[M-CH_3]^+$: abstraction of CH_3 , $[M+C_2H_5]^+$: addition of C_2H_5

*** number denotes the number of ^{13}C atoms in the backbone

Additional references

1. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17(1):132.
2. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539.