Additional file 2: Supplementary figures for

Comparative genomics and transcriptomics of *Pichia pastoris*



Figure S1: Experimental timeline for RNA sequencing of *Komagatella* **strains**. Schematic timeline for collection of RNA samples during batch cultivation of strains in shake flasks on 3 different carbon sources. Three independent cultivations were sampled for each time point.



Figure S2: Phylogenetic comparison of *K. pastoris* and *K. phaffii* to other related

yeasts. Phylogeny was generated using a concatenated, gap-free alignment of ten orthologous proteins. Phylogenetic tree was calculated using neighbor-joining, distance-based, maximum likelihood and maximum parsimony methods; reliability was assessed using bootstrapping. Clades marked with an asterisk are supported by 100% of bootstrap replicates in all four methods.



b) pGAPDH





Figure S3: Gene conservation between *K. pastoris* and *K. phaffii.* a) Histogram denoting homology at the base pair level for all 1:1 orthologous genes. b) Alignment of the P_{GAPDH} promoter element between *K. pastoris* and *K. phaffii.* c) Alignment of the P_{AOX1} promoter element between *K. pastoris* and *K. phaffii.*





Figure S4: Copy number determination for major chromosomes in a) *K. pastoris*, b) *K. phaffii* wild-type and c) *K. phaffii* GS115 strains.

a) K. pastoris

Second Base								
	U	С	А	G				
	UUU DH.H	UCU 1.56	UAU 1.03	UGU 1.24	U			
	UUC 1 0.86	UCC 1.04	UAC 1 20.97	UGC 20.76	С			
Ŭ	UUA T 1.07	UCA 1.25	UAA Stigil	UGA SOND	A			
	UUG 1.71	UCG 0.57	UAG State	neel 1100	G			
	CUU 1.02	CCU 1.35	CAU 1.23	CGU 0.84	U			
	CUC _ 0.51	CCC_ 0.73	CAC -0.77	CGC 0.32	с			
e C	CUA 0.76	CCA 1.51	CAA _ 1.20	CGA 0.72	A =			
Bas	CUG 0.93	CCG 0.41	CAG 0.80	CGG 0.32	G lind			
ts:	AUU 1.38	ACU 1.46	AAU 1.07	AGU 0.95	Das			
Ξ.	AUC 0.91	ACC 0.96	AAC A -0.93	AGC 0.63	c ®			
A	AUA 0.71	ACA 11.09	AAA _ 1.04	AGA 7 2.72	A			
	AUG Met / 1:00	ACG 0.49	AAG 20.96	AGG 1.09	G			
	GUU 1.55	GCU 1.62	GAU 1.23	GGU 1.43	U			
-	GUC T 7 0.87	GCC -0.97	GAC 0.27	GGC - 10.63	С			
G	GUA V 0.72	GCA 1.09	GAA 1.18	GGA 1.45	A			
	GUG 0.87	GCG 0.32	GAG 0.82	GGG 0.49	G			

b) K. phaffii

Second Base								
	U	С	A	G				
	UUU DH1.16	UCU 1.56	UAU 1.03	UGU 1.24	U			
	UUC ¹ 0.84	UCC 1.04		UGC 10.76	C			
Ŭ	UUA T 1.07	UCA 1.25	UAA Seligili	UGA SOLO	А			
		UCG 0.57	UAG SEDED	neel 11 Do	G			
	CUU 1.03	CCU 1.36	CAU1.23	CGU 0.84	U			
200	CUC _ 0.51	CCC0.72	CAC 10.77	CGC 0.31	С			
c	CUA 1 9.74	CCA 1.52	CAA _ 1.21	CGAA 0.72	A _			
Basi	CUG 0.91	CCG 0.41	CAG 0.79	CGG 0.32	G			
ž	AUU 1.40	ACU 1.46	AAU 1.07	AGU 0.95	UB			
iii.	AUC0.89	ACC 0.95	AACA 0.93	AGC 0.63	c ⁶			
A	AUA 0.71	ACA 1.10	AAA _ 1.04	AGA 7 2.72	A			
	AUG Met / 1:80	ACG 0.49	AAG 0.96	AGG 1.09	G			
	GUU 1.56	GCU 1.60	GAU 1.23	GGU 1.43	U			
G	GUC 0.86	GCC	GAC 0.17	GGC0.63	С			
	GUA 0.72	GCAA 1.10	GAA 11.19	GGA 1.46	A			
	GUG 0.87	GCG 0.32	GAG 0.81	GGG 0.49	G			

Figure S5: Codon usage for a) *K. pastoris* and b) *K. phaffii* as determined from all coding sequences identified in genome annotation. The relative abundance observed for each codon is represented as a percentage of total codon usage for the corresponding amino acid.



Figure S6: Isoform expression in *K. pastoris* and *K. phaffii* as a function of cultivation conditions. Heat maps of gene expression (log2 fpkm) for isoforms of alternatively spliced genes that alter coding sequences in a) *K. pastoris* and b) *K. phaffii* detected in initial genome annotation. Alternatively spliced genes with sufficient homology to *S. cerevisiae* are named, otherwise gene identifiers from genome annotation are used. Isoform expression is shown as a function of batch growth in glycerol, glucose or methanol during a 48 h cultivation period.

a)	K. pastoris	K. phaffii		
	cinii ¹ 8 - "nukuntiiku, liikunnikulliiku, mirikki kiliiinkkultiiku	Glycerol 24	Chri 8- 9- 10- 10- 11- 11- 11- 11- 11- 11- 11- 11	Glycerol 24
Average log2 FPKM	12 - 8 - 14 Mai -]]]] 48 h	12 - 	48 h
	יין 12- 18- היולאניה באניי, בהרגיה המליונטו ביות, היה האור שלי בנו, באורה לאין הערבור.	Glucose	12 - 6 - מינה להביר הבושר (10 מינה 10 מינה 6 - מינה מינה 10	Glucose
	ייייייייייייייייייייייייייייייייייייי	48	י 12- פ- 10 - 10 - 10 - 10 - 10 - 10 - 10 - 10	ULU. 10 48
	12	Methanol	12 - 8 - 14 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Methanol 24
	יייייייייייייייייייייייייייייייייייי	1 1 48	12 - 9 - נווניוניי באור באור באור אונט אונט אונט גער אונט אונט גער אונט אונט גער אונט גער אונט גער אונט גער אונט גער 10 - גער	
	0 1.0 2.0 3.0	Mbp	0 1.2	2.4 Mbp
	Chr2	-	Chr2	5
Average log2 FPKM	12 - 12 הנקאות האו המנה בישר 10 לעלים על היי היא האו האו האו האו האו האו האו האו האו הא	Glycerol 24	וער אין	Glycerol 24
	יור אין	48 h	12 = 5 = ,1; /(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,	↓ 48 h
	12 - 8 - גומראשון או האנריה, אילו לאמניר, אילו לאמניר גייר אילה אולה אילה אילה אילו אילו אילו אילו אילו אילו אילו	Glucose 24	12 - 6 - J. Mai y. a. May Ministry, 11, Mar. 10,	Glucose 24
	12 - 8 - Jakashin 104, ndinadilika da international da in	48	12 = 8 = 11. huit, dan 14 Mil dahila, 1. m, 16 Mil dili pandan huit	48
	וויין 12 	Methanol 24	12 - 8 - "A. JALLIA, AMMAN, ANDAR, A. AMMAN, ALAMAN, ALAMAN, ALAMAN, ALAMAN, ALAMAN, ALAMAN, ALAMAN, ALAMAN, ALAMAN	Methanol 24
	$\overset{12}{\text{s}}\overset{=}{=} \underbrace{1} \mu_{1} \mu_{2} \mu_{3} \mu_{4} \underbrace{1} \mu_{4} \mu_{3} \mu_{3} \mu_{3} \mu_{4} \underbrace{1} \mu_{4} \mu_{4} \underbrace{1} \mu_{4} \mu_{4} \underbrace{1} \underbrace{1} \mu_{4} \underbrace{1} \underbrace{1} \underbrace{1} \mu_{4} \underbrace{1} \mu_{4$	48	12 - 6 - Juliu, WWW.Ninke Juliu, Juliu, Alin, Ali	48
	0 1.0 2.0 3.0	Mbp	0 1.2	2.4 Mbp
	12 8 - Lau de Delane, la dela dela 111. dela	Glycerol 24	Cnr3 ^{12 -} 8 - 11101,11100,000,000,000,000,000,000,000	Glycerol 24
5	12 8 - 11 100 100 . 100 10 10 100 10 100	48 h	12 - 6 - 11 dai 1, 1 1 10 11, 1011 anii 101, 1010 anii 1010 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 - 1010 -	48 h
ige log2 FPKN	12 - 6 - Lou va val com jacon val val i j	Glucose 24	ייייייייייייייייייייייייייייייייייייי	Glucose 24
		48	יי 12 - 6 - גונולוסד, א לעלוד מידר העירט המעלידות, 11 גולעד, אלומונט אונג	48
Aver	ווייים 12-12-11-11-11-11-11-11-11-11-11-11-11-1	Methanol 24	12 - 6 - 11,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,	Methanol 24
	12 - 6 - 11 11 10 10 	48	6 - 1111 (11 1) (11 1 1 1 1 1 1 1 1 1 1	48
	0 1.0 2.0 3.0	Mbp	0 1.2	2.4 Mbp
	Chr4	Glycerol	Chr4	Chronzel
Average log2 FPKM	, ș = 1' în	24	12 - 5 - 1, p., n., p., p., p., p., p., p., p., p., p., p	24
	12 - 1, huunna , huunna , dalla, n,	48 h	12 - 0 - 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,	48 h
	12 - 6 - 1. Iliumi, 1. Iliumi, 1. Ini 1. I	Glucose 24	12 - 8 - ,,111,10,00,111,111111,10,11,111111,1,111111	Glucose 24
	12 - 11 11 11	48	12 - 1, o 1, o n di di n di di na di di tangan di sa di di tangan di sa di di tangan di sa di	48
	18 - 11 JULI 10 JULI 10 JULI 10 JULI 10 JULI	Methanol 24	12 - 6 - 11/11.0.0.00.00.01/10.0.00.000/01/01/01/01/00000000	Methanol 24
	12 - 6 - 11 11 1 11 11 11 11 11 11 11 11 11 11	48	12 5 - 11 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	48
	0 1.0 2.0 3.0	Mbp	0 1.2	2.4 Mbp

Figure S7: Chromosomal locations of highly expressed genes. Map of

chromosomal location (base pair identity) for the most highly expressed genes (top 10% expression) in a) *K. pastoris* and b) *K. phaffii*. Black lines indicate gene expression level at 24 h time points during batch cultivation in either glycerol, glucose or methanol. Red lines indicate locations of GC-rich autonomously replicating sequence (GC-ARS) motifs identified by BLAST.



Figure S8: Variance in gene expression during batch cultivation of *K. pastoris* **and** *K. phaffii.* Scatter plots of the average variance versus expression observed for all annotated genes across the 10 conditional averages generated from either a) *K. pastoris* or b) *K. phaffii* expression data. Genes with average log2 fpkm <1 and variance < 0.05 were excluded from further analyses. c) Elbow analysis of input cluster number to identify optimal expression data clustering by self- organizing maps (SOMs). Large circles denote the number of clusters for each expression data set where the additional variance captured by further clustering was < 1%.



Figure S9: Gene expression phenotypes in *K. pastoris* as a function of cultivation **conditions.** Self-organizing maps (SOMs) of genes changing expression similarly in *K. pastoris* during a 48 h batch cultivation in a) glycerol, b) glucose or c) methanol.



Figure S10: Gene expression phenotypes in *K. phaffii* as a function of cultivation **conditions.** Self-organizing maps (SOMs) of genes changing expression similarly in *K. phaffii* during a 48 h batch cultivation in a) glycerol, b) glucose or c) methanol.



Figure S11: Biological process enrichment as a function of cultivation in

glucose. Heat map representation of the enrichment of GO biological process terms for expression phenotypes observed in *K. pastoris* and *K. phaffii* during a 48 h batch cultivation in glucose as characterized by self-organizing maps (SOMs). Representative temporal trajectories of gene expression were generated for each SOM by averaging expression data at each time point for genes present within a given map. Color density relates to the number of genes assigned to a particular process as a percentage of the total number genes present in a particular expression phenotype or map.



Figure S12: Biological process enrichment as a function of cultivation in

glycerol. Heat map representation of the enrichment of GO biological process terms for expression phenotypes observed in *K. pastoris* and *K. phaffii* during a 48 h batch cultivation in glycerol as characterized by self-organizing maps (SOMs). Representative temporal trajectories of gene expression were generated for each SOM by averaging expression data at each time point for genes present within a given map. Color density relates to the number of genes assigned to a particular process as a percentage of the total number genes present in a particular expression phenotype or map.



Figure S13: Secretory pathway protein expression in *K. pastoris* and *K. phaffii*.

a) Row normalized single set Gene Set Enrichment Analysis (ssGSEA) projections for 170 proteins bearing a signal peptide as identified by Signalp. b) SDS-PAGE analysis of host-cell protein expression in supernantants during batch cultivation of *K. pastoris* and *K. phaffii* for 48h in glucose or glycerol-containing media.



Figure S14: Correlation of gene expression at two different cultivation densities. Scatter plots of gene expression between similar cultivation conditions for a) *K. pastoris*, b) wildtype *K. phaffii*, and c) *K. phaffii* GS115 grown at two different cell densities. Density A corresponds to cultures outgrown to $OD_{600} = 2.0$ prior to sampling and Density B corresponds to cultures outgrown to $OD_{600} = 20$ prior to sampling. Pearson correlation coefficients were calculated from expression vectors that were averages of three biological replicates for each cultivation condition and density.