Surveying candidate genes for temperature adaptation in wild *Saccharomyces paradoxus* populations using highest performing Multiple Genome Alignment software

Christian-Albrechts-Universität zu Kiel

MAX-PLANCK-GESELLSCHAFT

Corinn Small
Biol 607 & 608 MAMBE
Kiel University & Max Planck Institute for Evolutionary Biology
Supervisors: Christoph Eschenbrenner, Primrose Boynton and Prof. Dr. Eva Stuckenbrock
December 2017 - March 2018

Saccharomyces cerevisiae has been a highly relevant organism throughout human history.

- Domesticated thousands of years ago (McGovern *et al.* 2004)

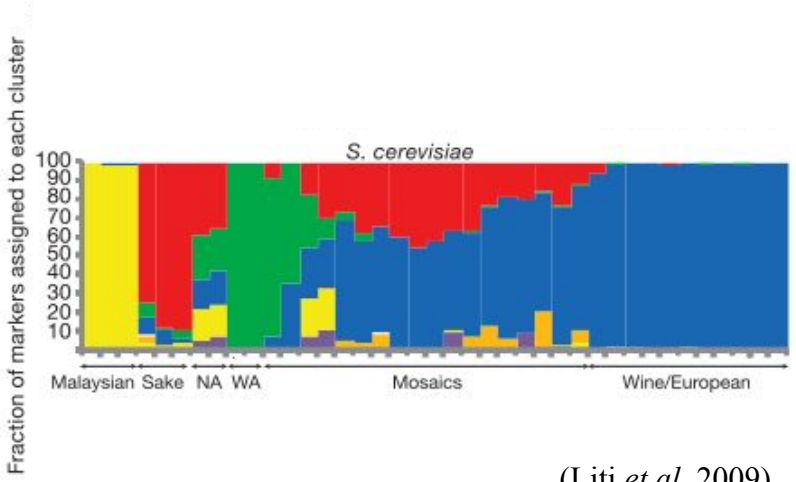
- 1st eukaryotic model organisms established in scientific community (Botstein *et al.* 1997)

- Utilized in numerous applications— from brewing beer to studying the aging process (Legras *et al.* 2007 & Nicolass *et al.* 2013)

However, there are barriers to studying the natural history of *Saccharomyces cerevisiae*

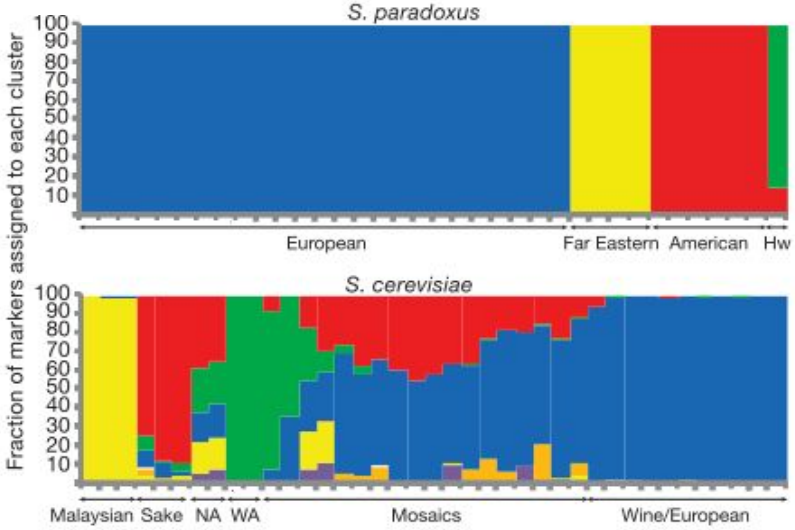
- Purely natural populations of *S. cerevisiae* have been difficult to find (Wang *et al.* 2012).

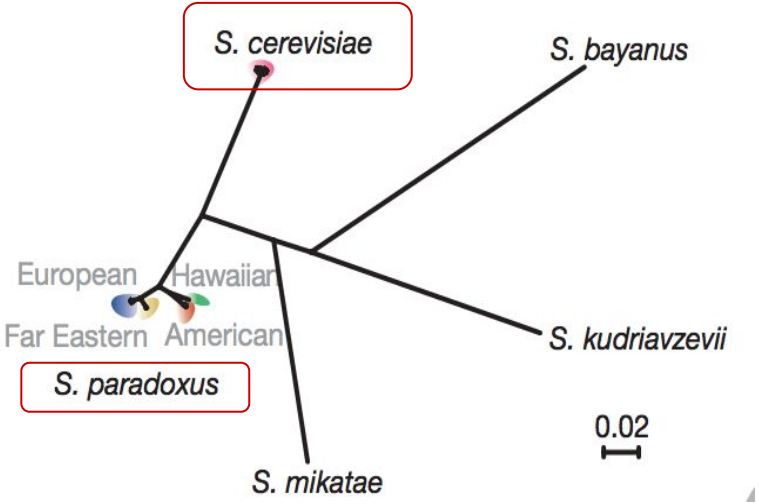
- Populations found in nature
  - may contain mixed genetics (not consistent with geography) (Liti *et al*. 2009).
  - or may have been influenced by other domestications
    - Origin? → comigration with grape varieties (Marsit & Dequin 2015)



(Liti *et al*. 2009).

S. paradoxus = wild sister species helpful for investigating the natural evolution & ecology of *Saccharomyces cerevisiae*

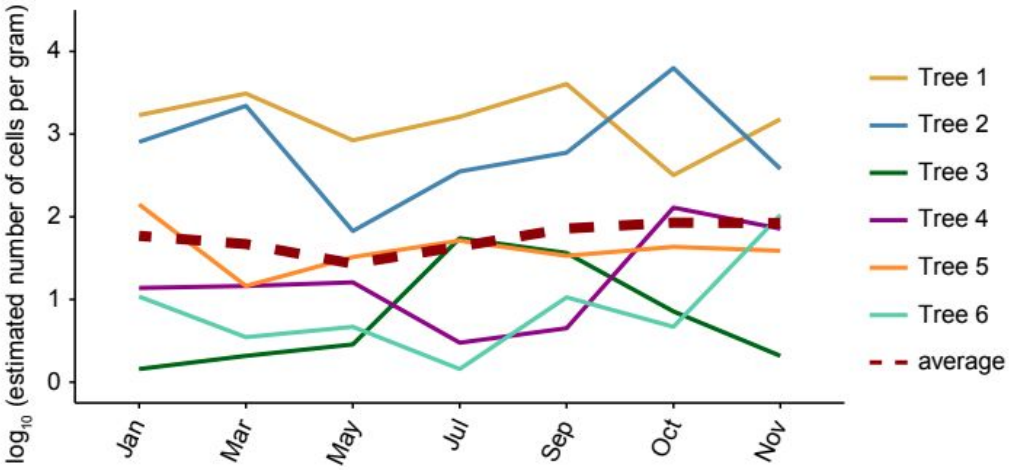




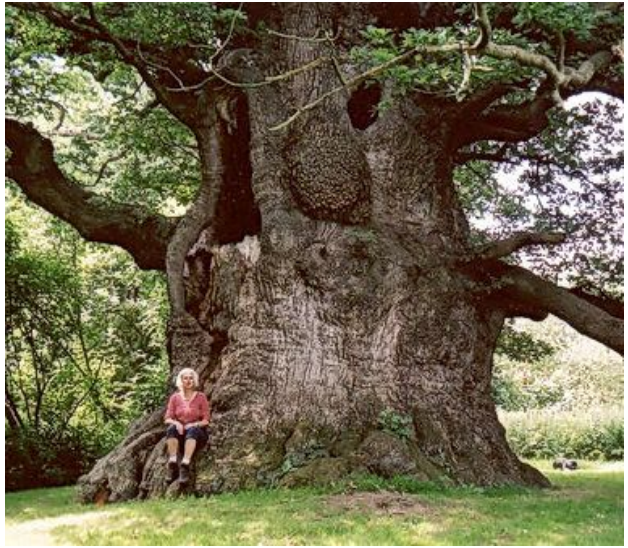
Liti *et al.* 2009

From a population of *S. paradoxus*... What genomic diversity can we observe?

- Northern Germany
- Free living in soil under oak trees
- Constant abundance throughout the year (Kowallik & Greig 2016)



(Kowallik & Greig 2016)



http://www.cathedralgrove.eu/text/08-Tree-Websites.htm

Is this population undergoing positive selection due to seasonal changes in temperature?

Overall hypothesis: due to seasonal changes, temperature positively selects for alleles involved in temperature adaptation and therefore maintains polymorphisms at these loci.

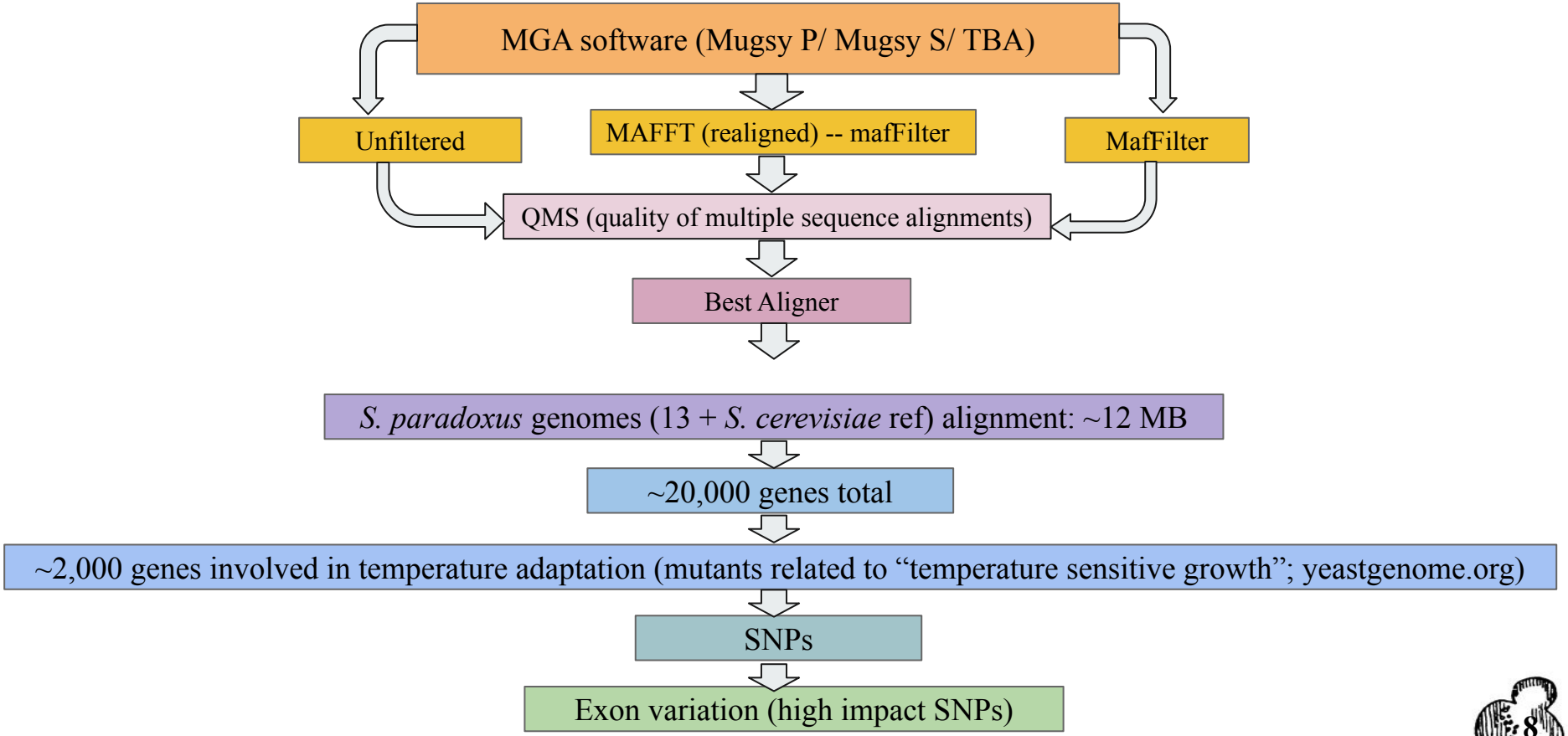
Null hypothesis: due to rate of temperature fluctuations, the population is not undergoing positive selection, but is rather phenotypically plastic.

Question: Is there evidence for <u>local temporal adaptation</u> to temperature?

Aim:

To identify candidate genes involved in temperature adaptation based on evolutionary predictions of an *S. paradoxus* genome dataset

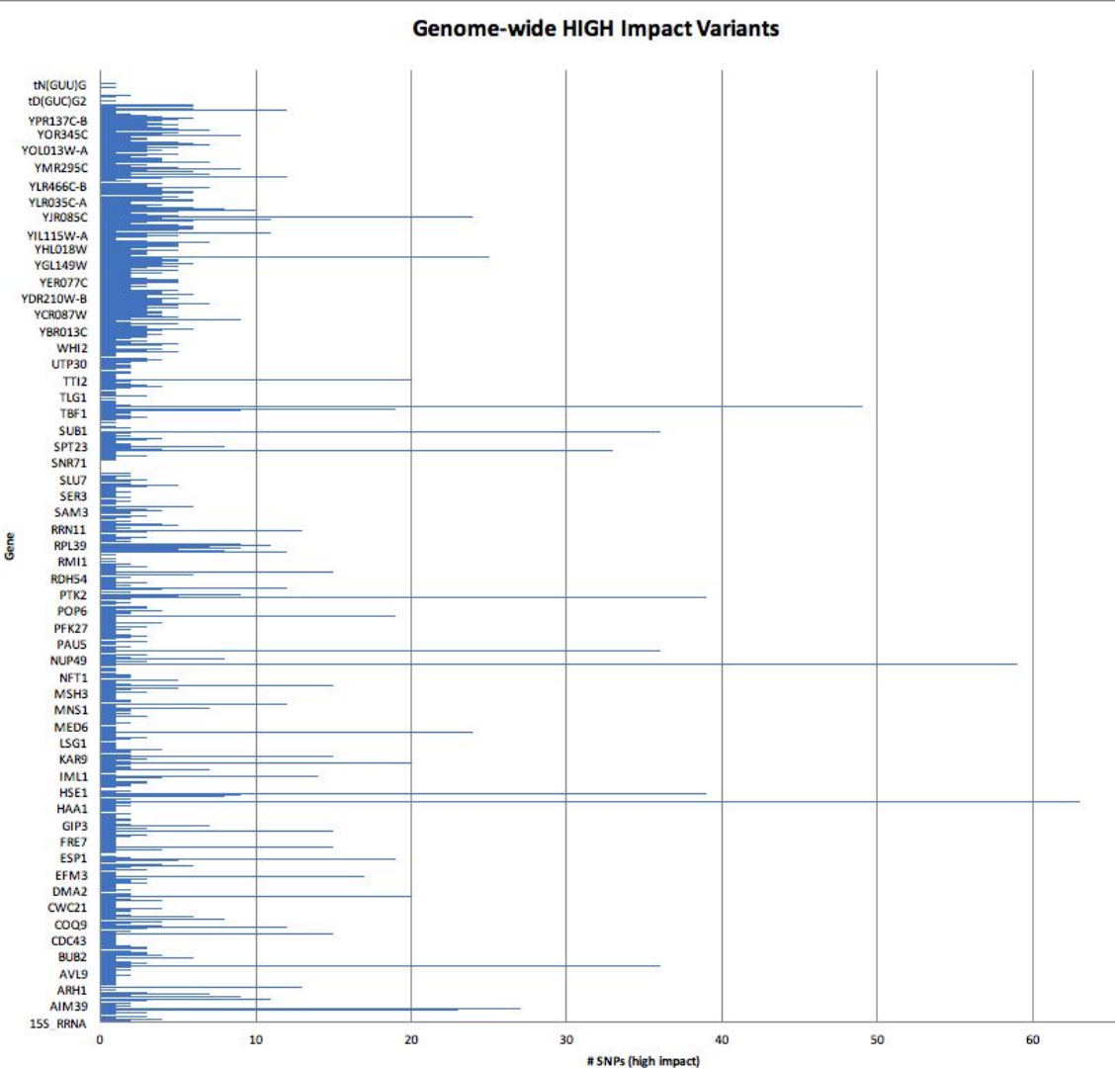
Generated SNP dataset from MugsyP-RF to investigate variation found in exons



Strains sampled from populations from various temperature ranges

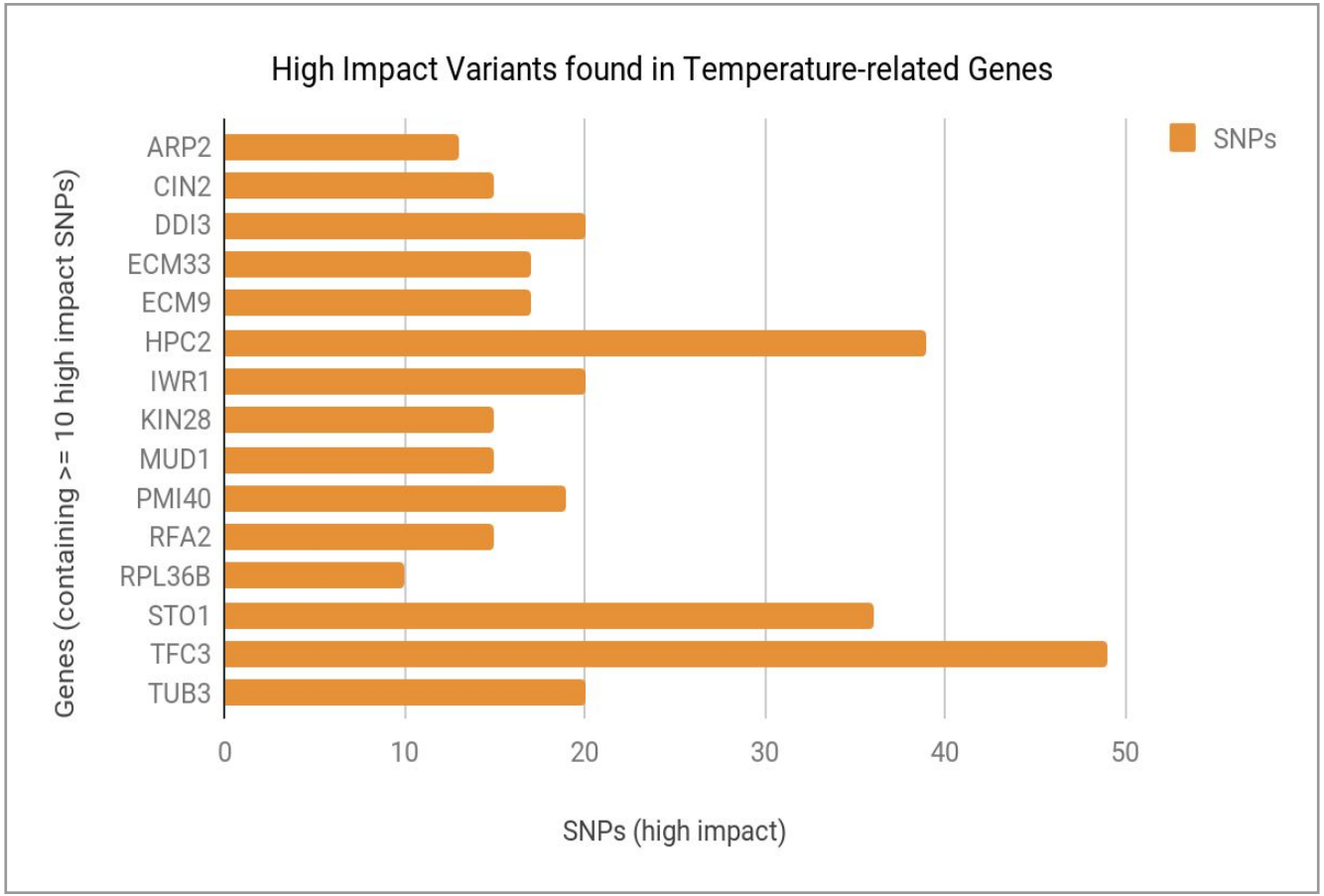
S. paradoxus strains	Region	Source	Country
Y8.5	Europe	*Quercus* spp.	UK
Z1.1	Europe	*Quercus* spp.	UK
Y9.6	Europe	*Quercus* spp.	UK
Z1	Europe	*Quercus* spp.	UK
Q59.1	Europe	*Quercus* spp.	UK
S36.7	Europe	*Quercus* spp.	UK
Y6.5	Europe	*Quercus* spp.	UK
Y7	Europe	*Quercus* spp.	UK
Q95.3	Europe	*Quercus* spp.	UK
T21.4	Europe	*Quercus* spp.	UK
W7	Europe	*Quercus* spp.	UK
DBVPG4650	Europe	Fossilized guano	Italy
Annotated reference strains			
S. cerevisiae S288c	North America	Rotting fig/lab strain	US
S. paradoxus CBS432	European Russia	*Quercus* spp.	Moscow, Russia

Liti *et al* 2009 supplementary material



From list of temperature-related genes:

Looked at those that had high dn/ds ratios (≥ 10 high impact SNPs)

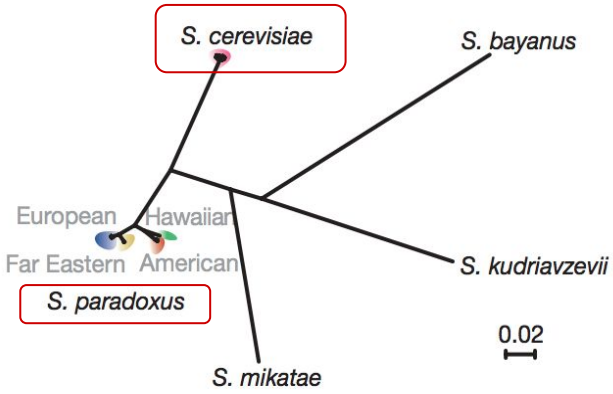

High Impact Variants found in Temperature-related Genes
SNPs
Genes (containing >= 10 high impact SNPs)
ARP2
CIN2
DDI3
ECM33
ECM9
HPC2
IWR1
KIN28
MUD1
PMI40
RFA2
RPL36B
STO1
TFC3
TUB3
0
10
20
30
40
50
SNPs (high impact)

Comparisons to related research from Liti *et al* 2009 may help to assess alignment quality

Number variants by type

Type	Total
SNP	1,326,174
MNP	0
INS	0
DEL	0
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	**1,326,174**

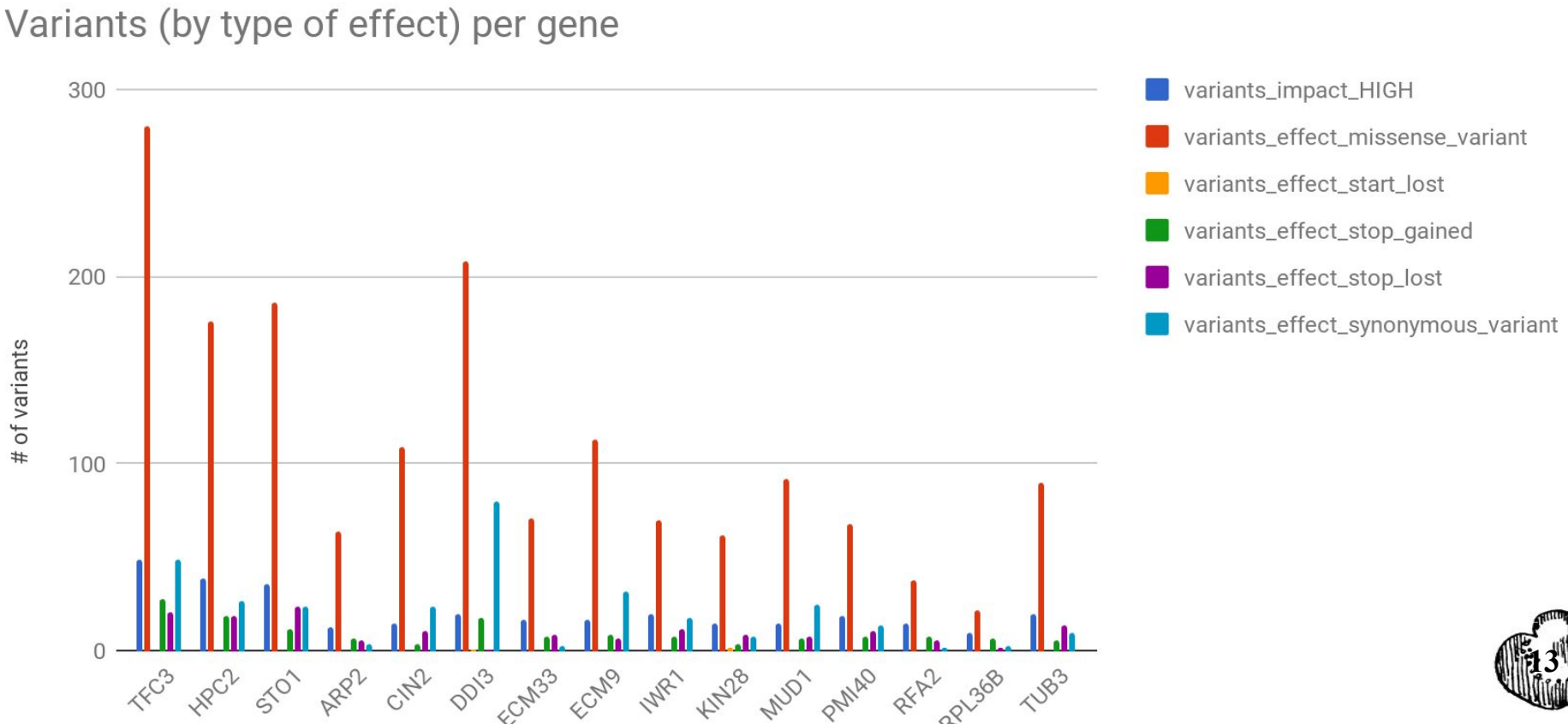
Using dataset of 13 *S. paradoxus* + 1 *S. cerevisiae*



623,287 SNPs total found in *S. paradoxus* alignment from Liti *et al* 2009

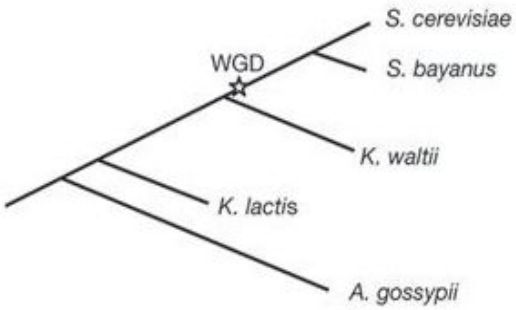
Source or location[a]		Strains
S. paradoxus		35
	England	18
	Continental Europe/Siberia	6
	Far East Russia/Japan	4
	North & South America	6
	Hawaii	1

Large amount of missense variation could be caused by several factors impacting the alignment other than positive selection



False variation can arise from...

- Mis-alignment /mis-annotation
- Duplications/ psuedogenization
- Indels
- Transposable elements
- Repeat rich regions



Kellis *et al.* 2004

```
score=700
Sp_CBS432.CP020258.1            7566 86 + 71482 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_W7.scaffold1420_size3763      464 86 +  3763 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Y6_5.scaffold344_size800      465 86 +   800 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Q95_3.scaffold307_size775     440 86 -   775 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_T21_4.scaffold419_size555     223 86 +   555 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Y9_6.scaffold505_size729        0 62 +   729 -TATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Q59_1.scaffold245_size843     511 86 -   843 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_N44.scaffold368_size5437     3828 74 +  5437 ATATAAT-TA-----CC-ATAATAA-ATT---AATTTT
Sp_Y8_5.scaffold386_size1105       0 69 +  1105 -TATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_S36_7.scaffold413_size1167    414 86 -  1167 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Z1.scaffold647_size1173       168 86 +  1173 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_YPS138.scaffold420_size4720   252 83 -  4720 -TATTATATATTTTTTTATTATAATATTTTAAATAAAT
Sp_Y7.scaffold254_size1697       434 86 -  1697 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Z1_1.scaffold212_size475      143 86 -   475 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
```

Survey of candidate genes & their functions

Gene Name	Gene ID	# Snps (>=10)	Function
ARP2	YDL029W	13	Actin-related; actin nucleation center required for the motility and integrity of actin patches; involved in endocytosis and membrane growth and polarity; https://www.yeastgenome.org/reference/S000055626
CIN2	YPL241C	15	GTPase-activating protein (GAP) for Cin4p; tubulin folding factor C involved in beta-tubulin (Tub2p) folding; mutants display increased chromosome loss; https://www.yeastgenome.org/locus/S000006162
DDI3	YNL335W	20	DNA Damage Inducible; Cyanamide hydratase, detoxifies cyanamide (a dehydration agent (reacts with H_2O); https://www.yeastgenome.org/locus/S000005279
ECM33	YBR078W	17	Extra-cellular mutant: Glycosylphosphatidylinositol (GPI) - associated protein, anchors to plasma membrane; https://www.yeastgenome.org/reference/S000072533
ECM9	YKR004C	17	Extra-cellular mutant: unknown function, non-essential; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1208169/pdf/ge1472435.pdf
HPC2	YBR215W	39	HIR (a nucleosome assembly) complex subunit; involved in regulation of histone gene transcription; https://www.yeastgenome.org/locus/S000000419#reference

Gene Name	Gene ID	# Snps (>=10)	Function
IWR1	YDL115C	20	RNA polymerase II transport factor, nucleo-cytoplasmic shuttling protein; https://www.yeastgenome.org/locus/S000002273
KIN28	YDL108W	15	Protein kinase; https://www.yeastgenome.org/locus/S000002266
MUD1	YBR119W	15	Involved in nuclear mRNA splicing; https://www.yeastgenome.org/locus/S000000323
PMI40	YER003C	19	Mannose-6-phosphate isomerase, required for early steps in mannose glycoside synthesis; https://www.yeastgenome.org/reference/S000042091
RFA2	YNL312W	15	Replication Protein A subunit- involved in DNA replication, repair, recombination, https://www.yeastgenome.org/locus/S000005256
RPL36B	YPL249C+AC0-A	10	Ribosomal 60S subunit protein L36B; httphttps://www.yeastgenome.org/locus/S000006438s://www.yeastgenome.org/locus/S000006438
STO1	YMR125W	36	Large subunit of the nuclear mRNA cap-binding protein complex; interacts with Npl3p to carry nuclear poly(A)+ mRNA to cytoplasm; https://www.yeastgenome.org/locus/S000004732
TFC3	YAL001C	49	RNA polymerase III transcription initiation factor complex subunit; https://www.yeastgenome.org/locus/S000000001
TUB3	YML124C	20	Alpha-tubulin associates with Tub2p forms tubulin dimer, polymerizes to form microtubules; https://www.yeastgenome.org/locus/S000004593

REFERENCES

- Samuel V. Angiuoli, Steven L. Salzberg; Mugsy: fast multiple alignment of closely related whole genomes, *Bioinformatics*, Volume 27, Issue 3, 1 February 2011, Pages 334–342, https://doi.org/10.1093/bioinformatics/btq665
- Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. Nature [Internet]. 2009;458(7236):337–41. Available from: http://dx.doi.org/10.1038/nature07743
- Kowallik, Vienna & Greig, Duncan. (2016). A systematic forest survey showing an association of Saccharomyces with oak leaf litter. Environmental Microbiology Reports. 8. 10.1111/1758-2229.12446.
- Fermented beverages of pre- and proto-historic China. Patrick E. McGovern, Juzhong Zhang, Jigen Tang, Zhiqing Zhang, Gretchen R. Hall, Robert A. Moreau, Alberto Nuñez, Eric D. Butrym, Michael P.Richards, Chen-shan Wang, Guangsheng Cheng, Zhijun Zhao, ChangsuiWang Proceedings of the National Academy of Sciences Dec 2004, 101 (51) 17593-17598; DOI:10.1073/pnas.0407921102
- Botstein, D, Steven A. Chervitz, and J. Michael Cherry. “Yeast as a Model Organism.” *Science (New York, N.Y.)* 277.5330 (1997): 1259–1260.
- Legras, J. , Merdinoglu, D. , Cornuet, J. And Karst, F. 2007. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. Molecular Ecology, 16: 2091-2102. doi:10.1111/j.1365-294X.2007.03266.x
- Nicolaas A Buijs, Verena Siewers, Jens Nielsen. 2013. Advanced biofuel production by the yeast Saccharomyces cerevisiae. Current Opinion in Chemical Biology, 17: 3: 480-488. ISSN 1367-5931. https://doi.org/10.1016/j.cbpa.2013.03.036.
- Manolis Kellis, Nick Patterson, Bruce Birren, Bonnie Berger, and Eric S. Lander.Journal of Computational Biology.Mar 2004.ahead of printhttp://doi.org/10.1089/1066527041410319
- Wang, Q. M., Liu, W. Q., Liti, G., Wang, S. A. & Bai, F. Y. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. Mol. Ecol. 21, 5404–5417 (2012).
- Souhir Marsit, Sylvie Dequin; Diversity and adaptive evolution of *Saccharomyces* wine yeast: a review, *FEMS Yeast Research*, Volume 15, Issue 7, 1 November 2015, fov067, https://doi.org/10.1093/femsyr/fov067
- Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu & Douglas M. Ruden (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff, Fly, 6:2, 80-92,DOI: 10.4161/fly.19695

Thank you to…

Prof. Dr. Eva H. Stukenbrock
Dr. Primrose Boynton
Dr. Alice Feurtey

Questions?

Supplementary Material

General Pipeline for Alignment Comparison and Survey Generation

Assembly Preparation

-Download assemblies (step 1 & 2)

-Rename assembly headers and files: new_header.py (step 3)

-Check for short contigs: Quast (step 4)

-(Optional) Filter N's: Script3-Assembly_Splitter.py Filter out short contigs < 1 kbp: filtering_CS.py (step 5)

Alignment: TBA

-*Blastn* using beta-tubulin (step 6)

-Create Seaview phylogenetic guide tree (step 7)

-Generate *Blastz* commands for pairwise comparisons (step 8)

-Generate bash script: all_bzToSGE_modified.py (step 9)

-Reformat headers for *Multiz*: formatSeq_PacBio.py (step 10)

-Generate pairwise alignments: scriptSGE.sh (step 11)

-Generate mutliple genome alignment using TBA (step 12)

-Project alignment against reference genome *Sp*_CBS432 (step 13)

Alignment: Mugsy

-Generate MGA using Mugsy serial (step 14)

-Generate MGA using Mugsy parallel (step 15)

Filtering Alignments

-Realign one set of alignments using *Mafft* (step 16a)

-Filter both realigned alignments and the raw alignments using bpp files with specified filtering options: *Mafffilter* (step 16b)

Evaluation of Alignments & Variant Calling

-Generate alignment statistics: QMS.py (step 17)

-Generate site frequency spectrums using *Maffilter* (step 18)

-Call genome-wide nucleotide variants using *Maffilter* (step 19)

Population Genomics for Best Alignment

-Calculate nucleotide diversity estimators: Tajima's D, Tajima's Pi, and Watterson Theta using *Maffilter* (step 21)

-Construct PCA plot: Adjust_MapFile.py, PCA_SNPs.py & Plot_PCA.R in this order (steps 22 - 24)

Candidate Gene Survey

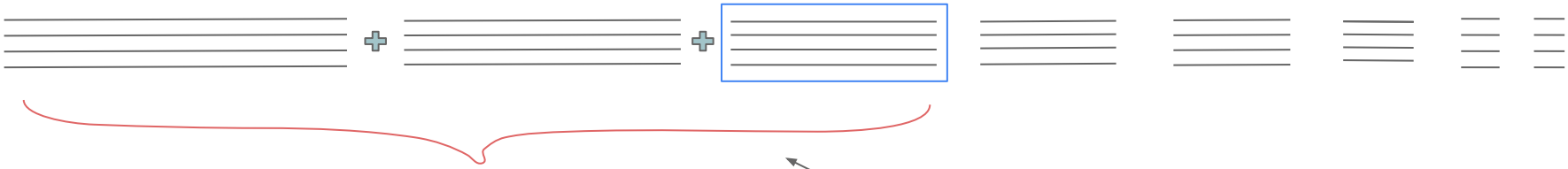
-Generates SnpEff files from VCF files (step 1)

-Match genes: matching_snps_to_temp_genes.py (step 2)

-Filter matched SnpEff genes for only those containing ≥ 10 SNPs: high_impact_genes.py (can change threshold) (step 3)

Additional stats refer to the quality of the alignment

N50 & L50



N50 -> length of the smallest block included in the summed block lengths that make up 50% of total alignment

Larger N50 & L50 => the more bases are contiguous

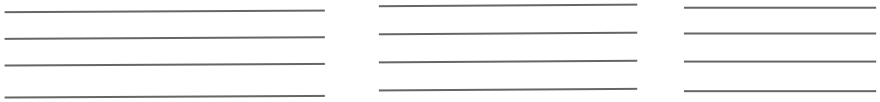
L50 refers to the number of summed blocks that give N50

L50 = 3

Additional stats refer to the quality of the alignment

Block Length and Gap% reflect homology

Longer



Shorter

More homology

Less homology

Less introduced gaps

More introduced gaps

(Angiuoli & Salzberg 2011)

QMS statistics scoring shows that re-alignment is necessary before filtering.

First, best type of filtering per software was chosen

Key	-1 (worst)	+1 (best)	NA (not included in the analysis)

	↑ Total Length	Blocks	↓ Gap%	↑ N50	↑ L50	↑ Total Block Length ≥ 1kb	Total Block Length: ≥ 5kb	Total Block Length: ≥10kb	Evaluation score
Tba UF	12229423	10061	1.28	3309	1101	11057280	7830332	4360009	NA
Tba F	8458355	40541	0.36	163	18831	236325	0	0	-6
Tba RF *	11028652	6397	0.3	1906	1940	9794899	3795678	946740	6
MugsyP UF	12023677	7036	2.59	3396	1049	10935200	7801887	4536046	NA
MugsyP F	11190803	6576	0.99	1792	2085	9818100	3536276	920109	-3
MugsyP RF	11176614	6301	0.77	1880	1982	9886060	3796347	1001818	3
MugsyS UF	12004021	6965	2.5	3424	1046	10931937	7785588	4501238	NA
MugsyS F	11187565	6529	0.98	1797	2080	9814145	3525269	920242	-3
MugsyS RF	11172450	6263	0.76	1886	1980	9883914	3773364	991599	3

UF= unfiltered, F = filtered, RF = realigned & filtered

Mugsy + local realignment + MafFiltering = Highest quality alignment

Key	-1	0 (intermediate)	1	NA

	↑ Total Length	Blocks	↓ Gap%	↑ N50	↑ L50	↑ Total Block Length: 1kb	Total Block Length: 5kb	Total Block Length: 10kb	Evaluation score
Tba RF	11028652	6397	0.3	1906	1940	9794899	3795678	946740	-2
MugsyP RF	11176614	6301	0.77	1880	1982	9886060	3796347	1001818	3
MugsyP F	11190803	6576	0.99	1792	2085	9818100	3536276	920109	-1
MugsyS RF	11172450	6263	0.76	1886	1980	9883914	3773364	991599	0

* Mugsy P-RF and Mugsy S-RF are very comparable= no significant differences

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
downstream_gene_variant	3,970,211	42.735%
initiator_codon_variant	31	0%
intergenic_region	424,054	4.564%
intron_variant	13,780	0.148%
missense_variant	271,551	2.923%
non_coding_transcript_exon_variant	5,118	0.055%
splice_acceptor_variant	27	0%
splice_donor_variant	7	0%
splice_region_variant	2,170	0.023%
start_lost	444	0.005%
stop_gained	2,893	0.031%
stop_lost	1,071	0.012%
stop_retained_variant	1,408	0.015%
synonymous_variant	624,295	6.72%
upstream_gene_variant	3,973,277	42.768%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	3,970,211	42.745%
EXON	905,500	9.749%
INTERGENIC	424,054	4.566%
INTRON	13,379	0.144%
SPLICE_SITE_ACCEPTOR	27	0%
SPLICE_SITE_DONOR	6	0%
SPLICE_SITE_REGION	1,679	0.018%
UPSTREAM	3,973,277	42.778%

Type of Exonic effect	Predicted Effect Count	Percent of total effects
Initiator codon variant	31	~0%
Start lost	444	0.005%
Stop gained	2,893	0.031%
Stop lost	1,071	0.012%
Missense variant	271,551	2.92%
Synonymous variant	624,295	6.72%
Total exonic effects	900,285	9.74%

Variation found in upstream regions could point to plasticity rather than adaptation