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

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

Introduction

Saccharomyces cerevisiae has been a highly relevant organism throughout human history. *S. cerevisiae* was domesticated thousands of years ago and was the first eukaryotic model organisms established in the scientific community (Botstein 1997). This fungus, generally known as "Baker's yeast", has been utilized in numerous applications— from brewing beer to studying the aging process— assisting in society's scientific and cultural development (Legras 2007, Nicolass 2013). Despite its clear relevance, its evolutionary history prior to domestication (and the underlying evolutionary mechanisms) are still largely unknown (Liti 2009). Previous attempts to directly study purely natural *S. cerevisiae* populations have proven difficult (Sniegowski 2002), therefore attention has grown towards *S. cerevisiae's* non-domesticated sister species: *Saccharomyces paradoxus* (Sniegowski 2002).

Several factors are necessary to confidently determine under which mechanisms a population has evolved, including an accurate and large enough sample size, a confident assessment of the population's genetic variation, and a clear understanding of which environmental factors interact to determine the organism's fitness.

This study investigates one such mechanism: local temporal adaptation and temperature as a likely influential parameter. Temperature was an obvious choice, given that it has played a significant role in directing the domestication of *S. cerevisiae* (Salvadó 2011, Paget 2014). Furthermore, throughout the *Saccharomyces* genus, different species display different temperature-dependent phenotypes (i.e. optimal growth rates) (Paget 2014). Naturally, temperature heavily influences ecosystem dynamics and therefore is a sound influential candidate factor (Enquist 2003).

In terms of spatial local adaptation, previous studies have shown that *S. paradoxus* can become geographically isolated (matings between North American isolates and European isolates showed partial reproductive isolation compared to *S. cerevisiae* (Sniegowski 2002) -- genomic analyses have shown that there is significantly greater genetic variation in *S. paradoxus* compared to *S. cerevisiae* (Liti 2009).

Therefore, genetic variation necessary for potential niche adaptation (sourced chiefly from mutations (Tsai 2008)) may sufficiently exist. More recently, Portuguese *Saccharomyces cerevisiae*, *paradoxus,* and *kudriavzevii* isolates were shown to be living in sympatry, isolated at different temperatures from the same bark sample (Salvadó 2011). Fitness assays proved that local adaptation to specific temperature ranges was the probable cause for this sympatric relationship (Salvadó 2011). Perhaps local adaptation could occur in a similar manner temporally, coinciding with seasons.

In 2016, Kraemer et. al alluded to the lack of investigation into local adaptation in natural free-living microbial populations. That same year Kowallik et. al discovered a seasonally consistent population abundance of *S. paradoxus* in a northern German forest. Continuing investigation aims to solve the question of whether this is due to local temporal adaptation (i.e. whether different genotypes correlate to particular seasons) or selection for genetic generalization (i.e. phenotypic plasticity).

In this study, a pipeline was established for comparing performances of multiple genome alignment software to confidently analyze genome-wide genetic variation found in *S. paradoxus*. The aim was to use the best possible software (and filtering method) to analyze a dataset of *S. paradoxus* isolates from various geographical regions-- identifying genes that held phenotypically relevant variation (i.e. non-synonymous single nucleotide polymorphisms (SNP) that specifically relate to a temperature-dependent phenotype (Cherry 2012). Non-synonymous SNPs have the potential to highly impact the organism's phenotype (Tang 2018). The main phenotype investigated was heat-sensitive growth, defined as lower fitness at temperatures higher than optimal growth temperatures (Cherry 2012). The isolates composing the first dataset were used with the assumption that due to drastic differences in geographical regions, ideal temperature ranges for optimal growth would also vary.

A similar project will eventually be applied to a sequenced assembly dataset encompassing seasonal isolates collected from a northern German forest. Alleles identified in this study will be tested for their effects on fitness patterns, testing for local adaptation, using a mutant collection of the German isolates. Fitness assays have the power to rule out other sources of genetic variation using characteristic fitness patterns that correspond to different evolutionary mechanisms (Kraemer 2016). In addition to local adaptation, genetic drift, linkage, and gene flow can generate variation within a population (Kraemer 2016).

To confidently analyse genome-wide genetic variation within a dataset, an accurate alignment of the isolate genomes must be made. Unfortunately, many alignment programs still produce many false positives (or variants) which significantly impact the alignments and the downstream analysis (Markova-Raina 2011). Therefore a performance comparison of multiple genome alignment programs was necessary to find the best software that produces the most confident alignment and variant dataset.

This process is especially important in whole genome studies when analyzing thousands of genes (Markova-Raina 2011). The amount of errors that are significant also depend on the biological question trying to be answered (Markova-Raina 2011).

Alignment problems such as false positives can arise from elements that produce discrepancies in genomic synteny such as duplications, indels, transposable elements, or repeat rich regions (Markova-Raina 2011). Organisms possess different amounts of such elements, therefore, it is important to understand the specific biology of the organisms being compared. For the datasets used in this study, a whole-genome duplication (WGD) event led to the rise of the *Saccharomyces* genus (Kellis 2004) meaning duplications most likely will cause a significant amount of false variants. Overall, accurate variant detection is necessary to make correct conclusions about an organism's evolution (Yi 2014). Incorrect conclusions can significantly influence evolutionary pattern or candidate gene inferences and the direction of future studies.

Materials & Methods

In this study, Threaded Blockset Aligner (TBA) (Blanchette 2004) and Mugsy (Angiuoli and Salsber 2011) were evaluated. Two versions of the Mugsy aligner were included: Mugsy-parallel (MugsyP), the modified version which compares multiple pairwise alignments at once and the original version that works serially (MugsyS). All three tools first generated pairwise genome alignments that were then used as the base for generating the multiple genome alignments.

A general pipeline was established for downloading and preparing assemblies, aligning them, filtering the alignments, generating relevant statistics, and calling single nucleotide polymorphic (SNP) variants. Scripts and commands for the pipeline can be found in the supplementary material in order of operation. After calling the variants, the two datasets underwent different analyses.

Datasets

Two sets of isolates were used, the first was for determining the best alignment program to analyze *Saccharomyces* genomes and the second was for creating the candidate gene survey. The first set included assemblies from the UK, Russia and the US (Table 1). The second set included assemblies from the UK and Italy with the exception of the two annotated reference genomes (*S. paradoxus* CBS432 from Russia and *S. cerevisiae* S288c from the US) (Table 2). All isolates and references were found in Bergstrom et al 2014 and Liti et al 2009.

The software performance comparison was conducted first. The file names and individual chromosome designations were renamed for ease of work using. The dataset assemblies were then inspected for short contigs using the program Quast, that assesses assembly quality. An optional step could be applied here that splits the assembly at uncalled nucleotide ("N") regions if longer than 30 consecutive base pairs. Short contigs (with lengths < 1 kbp) were filtered out in all assemblies to reduce processing time.

Multiple Genome Alignments with TBA

We identified the *beta-tubulin* gene using a reference from *S. paradoxus* BCRC23154 in all isolates using *blastn* to create a guide phylogenetic tree for the alignment software (Huang 2009 & Altschul 1990). Blast output format 7 gave the necessary fasta format for the next step. The program Seaview (Gouy et al 2010) was then used to generate the tree and export it in Newick format, necessary for the genome aligner MultiZ.

Using the sequences specified in Newick format, the *all_bz* program generated a series of blastz commands for all possible pairwise comparisons. The list was written into a bash script that would execute the pairwise alignments. Headers in each fasta file were reformatted into a readable context for the alignment program MultiZ ("Strain ID:ChrID:Start:Strand:Length"). Pairwise alignments were made for TBA locally. Pairwise alignment files (ending in ".orig.maf" or ".sing.maf") along with the specified Newick phylogenetic tree were then used to create the multiple genome alignment (MGA) using TBA. Once the MGA was created it was projected against the reference genome (Sp_CBS432) to orient annotations.

Multiple Genome Alignments with Mugsy

The same formatted fasta files were used, however ":" were removed from the headers. The guide tree was reformatted specifying strain names separated by spaces. The serial version was executed first, specifying the options "-fullsearch -duplications 1". "Fullsearch" specified that both sequences in the pairwise comparisons should be treated as the reference and "duplications 1" created two separate alignments, one with duplications and one without (Angiuoli 2011). The parallel version followed. Everything was the same as in the serial commands except for the name of the program "mugsy_parallel" and the number of parallel comparisons (option "-nproc #"). Eight comparisons were run in parallel (-nproc 8).

Alignment Filtering

The next step in the pipeline was filtering. One alignment set was left unfiltered, another just filtered, and the third was locally re-aligned and subsequently filtered. The program package Maffilter (Dutheil 2014) was used for filtering and the multiple alignment program Mafft (Katoh 2002) via Maffilter was used for re-aligning the MGA block by block. The Maffilter options used for filtering included: *MinBlockSize*()-- only kept blocks with all 14 individuals, *XFull Gap*()-- removed gap-only columns from blocks, *AlnFilter2*()-- masked columns within a window that contained more than 5 gaps, *MinBlockLength*() --kept blocks with more than 50 nucleotides, *MaskFilter*()-- split blocks by removing regions that contained too many masked columns (2 masked columns), and *AlnFilter*()-- split blocks by removing ambiguous regions. Bpp files were used to specify the filtering options, the online Maffilter manual 1.2.1 describes the bpp file format.

QMS quality statistics were generated after filtering. The final alignments were designated unfiltered ("UF"), filtered ("F"), and re-aligned and filtered ("RF"). Two separate comparisons were made, the first between filtering types and the second between MGA software. Relevant sequence statistics included total block length frequencies, N50, L50, Gap %, and total length of the alignments. Maffilter was used to generate the site frequency spectrums using *SequenceStatistics*().

The final common step for both datasets was calling genome-wide nucleotide variants using Maffilter with *VcfOutput()*. This option produced files in variant calling format (VCF) and specified all variants found within the alignment including gap sites, ambiguous bases ("N"), and importantly the "passed" variants (meaning at that specific site all bases within each genome were successfully determined when assembled). For software comparisons, variants were called for all alignments to visualize shared and unique SNPs for confidence assessments; the program Upset created these visualizations in R. Only "passed" variants were assessed in the Upset program. For filtering of the VCF files, two linux commands were used to write the "PASS" variants to new files.

Evaluation of Alignments

Quality statistics were evaluated using a scoring system: the best values were assigned +1, middle values +0, and the worst values -1. The total length, N50, L50, and the total block frequencies correlated positively with the accuracy of the alignment. The gap % correlated negatively. For future analyses, scores can be weighed corresponding to the significance of the statistics.

Population Genomic Analyses using highest quality Alignment

Sequence diversity statistics and principal component analysis (of the dataset population structure) were calculated and graphed for the best alignment. Maffilter with option *DiversityStatisics*() under *SequenceStatistics*() was used to calculate the nucleotide diversity estimators: Tajima's D, Tajima's Pi, and Watterson Theta. The PCA plot was constructed using three scripts, that converted the VCF file into a MAP formatted file.

For the second dataset, the same steps were used as described above with the exception of only using MugsyP-RF to generate the MGA. The genome-wide variant dataset (in VCF format) was used to generate a SnpEFF survey (Cingolani 2012) of predicted annotated genes that contained high-impact SNPs. SnpEff variants were annotated using the pre-packaged reference genome *S. cerevisiae* S288c from GenBank assembly GCA_000146045.2, "R64-1-1.86" --included in the database with the latest snpEff version 4.3Q, 2018-01-04. SnpEFF genes containing the annotated variation were matched to the temperature-dependent genes found in the yeastgenome.org database under "temperature sensitive growth" and "heat sensitivity" (Cherry 2012). Finally, matched genes were selected for the survey based on how many high impact SNPs they contained. Genes that contained greater than or equal to 10 SNPs listed under the SNPeff column "variants_impact_HIGH" were added to the survey.

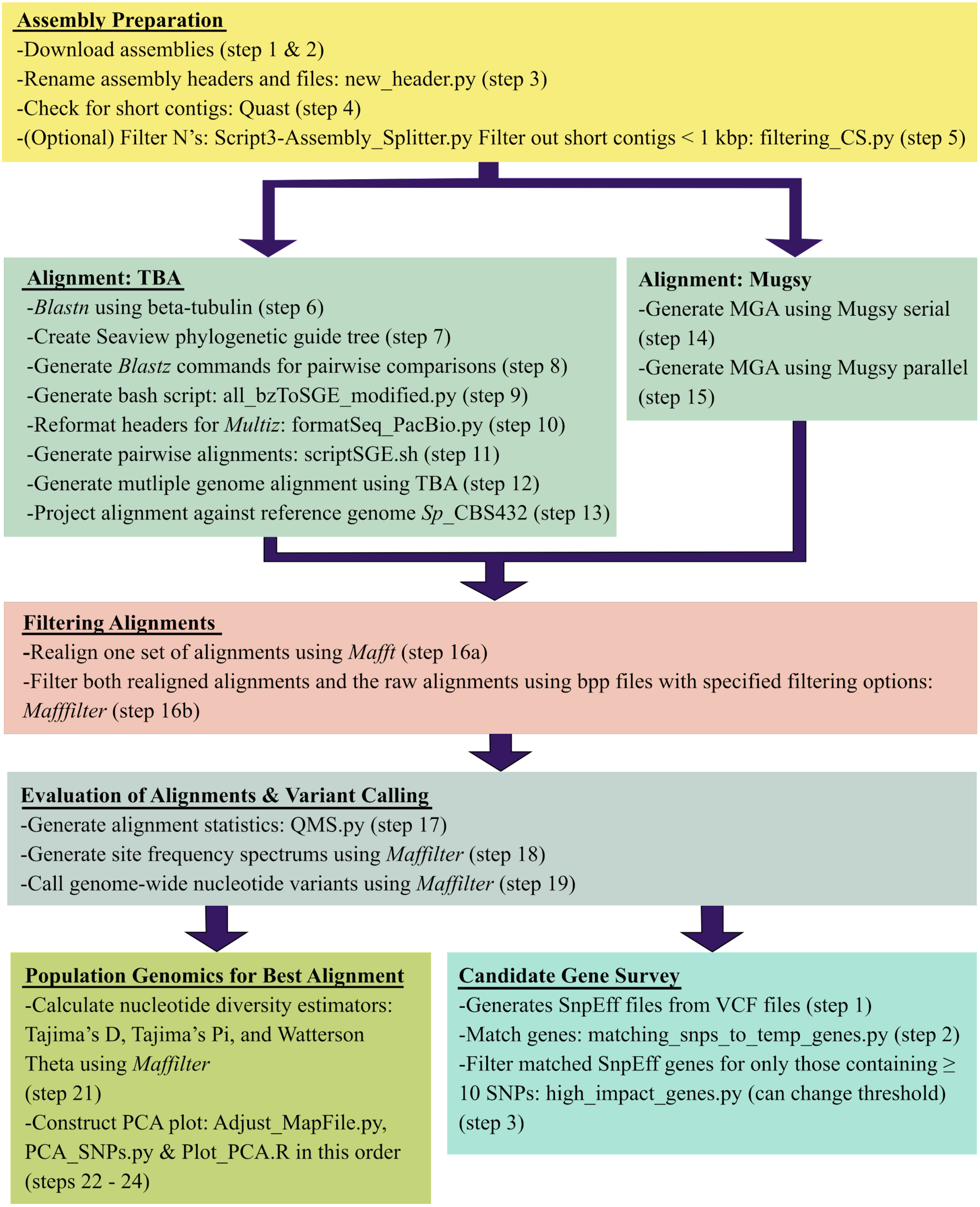
Table 1. Strains used in the software comparison dataset.

S. paradoxus strain	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
N-44	Far Eastern	*Quercus* spp.	Russia
YPS138	North American	*Q. velutina*	USA
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
CBS432 (annotated reference)	Europe	*Quercus* spp.	Moscow, Russia

Table 2. Strains used for surveying candidate temperature-related genes.

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

*All strains descriptions can be found in the Liti et al 2009 supplementary material



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)

Figure 1. Pipeline diagram for both datasets. Corresponds to commands listed in Supplementary Material.

Results

MGA Software Comparison

The consensus tree used for the first dataset isolates can be found in Figure 2. Comparisons of the MGA software and filtering options led to the conclusion that Mugsy parallel or serial (there was no significant difference between the two) with local realignment and filtering produced the most accurate SNP dataset. The statistics generated from the filtering options and the MGA software can be found in Table 3 and Table 4, respectively. The statics found in grey did not hold any merit due to their biases or lack of significance. The total length, N50, L50, and total block lengths correlated positively with the accuracy of the alignment. The larger the region of the alignment or block, the better confidence or tactness the alignment had. The larger the regions compared, the higher the accuracy of the nucleotide locations within the sequences and alignment. Correspondingly, the lower the gap percentage was the more intact the alignment was as well. N50 is the length of the smallest block included in the summed block lengths that make up 50 % of the total alignment, therefore an N50 that is larger suggests higher accuracy. L50 refers to the number of summed blocks that gives N50: again, the greater the number of larger blocks the better and the larger the total sequence covered the more analyzed. The RF option consistently produced the most accurate alignments (Table 3).

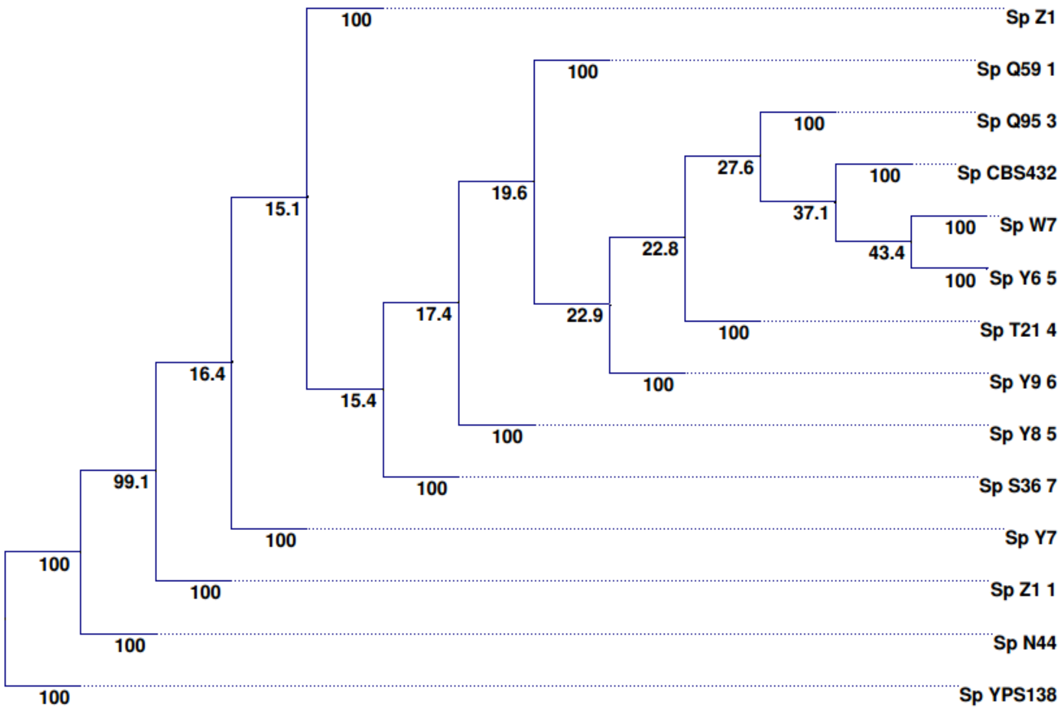


Figure 2. Consensus tree generated using genome-wide SNPs from MugsyP-RF. Sub-trees constructed using UPGMA

Table 3. QMS results for Filtering

	Total Length	Blocks	Gap%	N50	L50	Total Block Length: 1kb	Total Block Length: 5kb	Total Block Length: 10kb	Score
Tba-UF	12229423	10061	1.28	3309	1101	11057280	7830332	4360009	NA
Tba-F	8458355	40541	0.36	163	18831	236325	0	0	-7
Tba-RF	11028652	6397	0.3	1906	1940	9794899	3795678	946740	7
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

*Note: when using TBA-- realigning is a must. Green box = best score

Table 4. QMS results for MGA software

	Total Length	Blocks	Gap%	N50	L50	Total Block Length: 1kb	Total Block Length: 5kb	Total Block Length: 10kb	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
MugsyS-RF	11172450	6263	0.76	1886	1980	9883914	3773364	991599	0
MugsyP-F	11190803	6576	0.99	1792	2085	9818100	3536276	920109	-1

*MugsyP-RF and MugsyS-RF are very comparable = no significant differences.

Comparing MugsyP's filtering methods, filtering removed about 19,887 variants from the unfiltered variants, realigning and filtering restored 11,084 variants. When comparing software, the difference between TBA-RF and MugsyP-RF was 12,713 variants and between MugsyP-RF and MugsyS-RF was about 539 variants. Little difference was seen between MugsyP and MugsyS (the only difference came from seeding the alignment). TBA was the worst among all software and therefore most likely called the most false positives. Importantly, when considering filtering options while using TBA, it is essential to realign locally and filter before assessing variant datasets.

Upset figures (Figures 3 and 4) showed a large number of variants (~92%) that were shared between all filtering options and between all MGA software (~93% among the three RF MGAs and ~89% among the three RF MGAs plus MugsyP-F). We can more confidently suggest that most of the "passed" variants were called accurately and are in fact not false positives-- the more alignment programs compared the greater the confidence. In Figure 3, there were no unique variants found in MugsyP-F compared to

MugsyP-UF because all variants in the filtered alignment were a subset of the unfiltered.

Population genomic analysis included nucleotide diversity statistics Tajima's D, Tajima's Pi (π_T), and Waterson theta ($\boldsymbol{\theta}_W$). Even though MugsyP-RF was ultimately the best, MugsyP-F was used for further analysis (shown in green box versus red box, Table 2). Initially, MugsyP-F was mistaken as the best software-filtering combination. Overall, Tajima's D values were very low genome-wide (around -2), with an exception of the mitochondrial DNA average being slightly above -2.

As a follow up, the nucleotide diversity of multiple subsets excluding Russian and/or American isolates were analyzed (see Figure 5). The removal of both increased the average Tajima's D value to 0. The site frequency spectrums for MugsyP-F in Figure 6 show that rare alleles (1 SNP/nucleotide site) were more frequently found among the whole dataset compared to just the UK dataset.

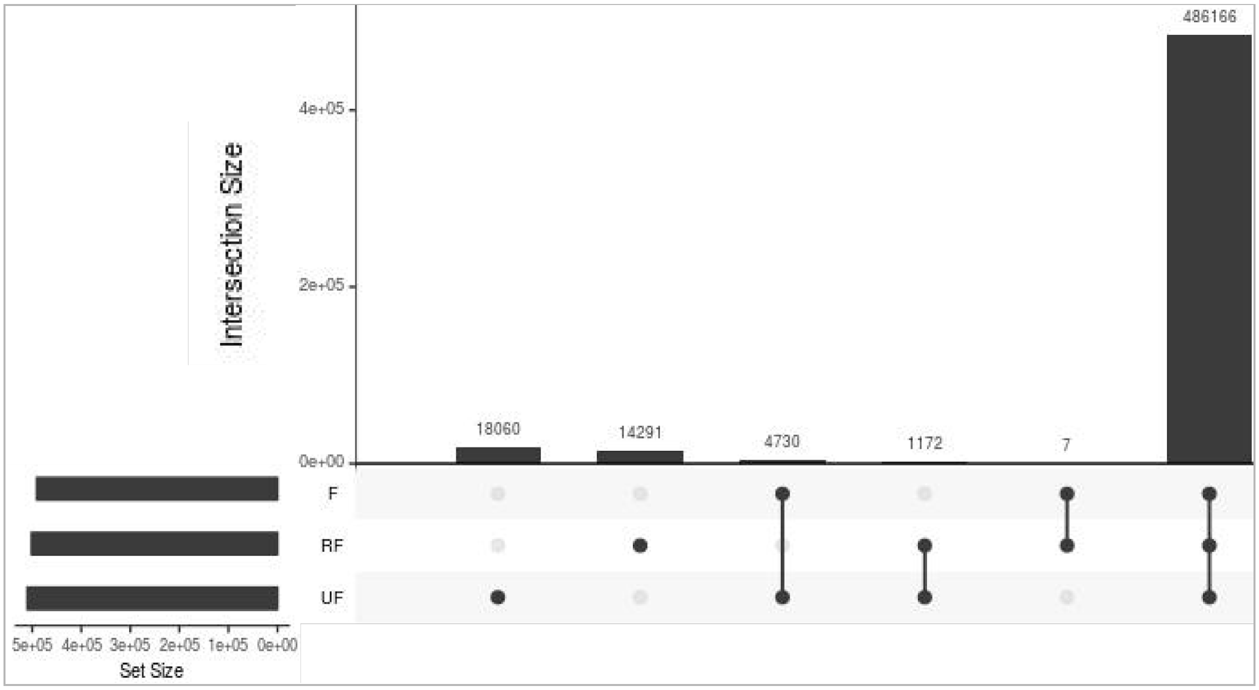


Figure 3. Upset comparison of all filtering options of the MugsyP alignment. Upset visualization of common and unique SNPs. The y-axis refers to the number of SNPs corresponding to the combination of filtering types seen as dot plots below the x-axis.

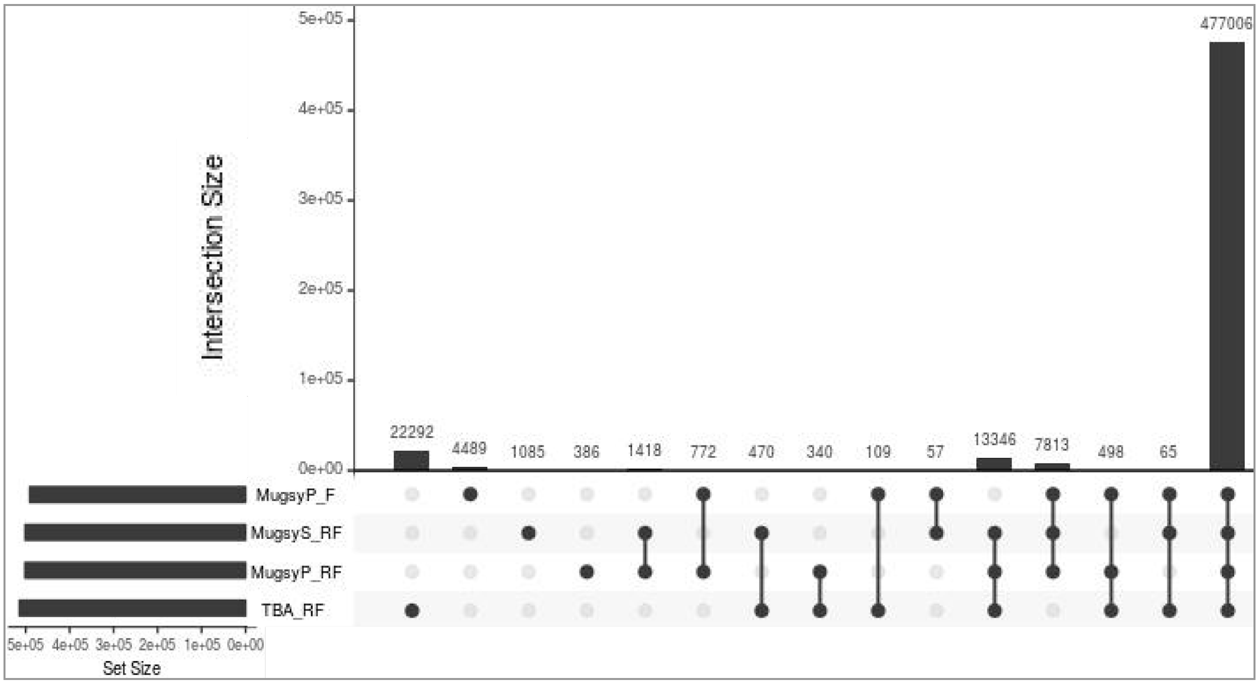


Figure 4. Upset comparison of MGA software all RF alignments plus MugsyP-F. Upset visualization of common and unique SNPs. The y-axis refers to the number of SNPs corresponding to the combination of filtering types seen as dot plots below the x-axis.

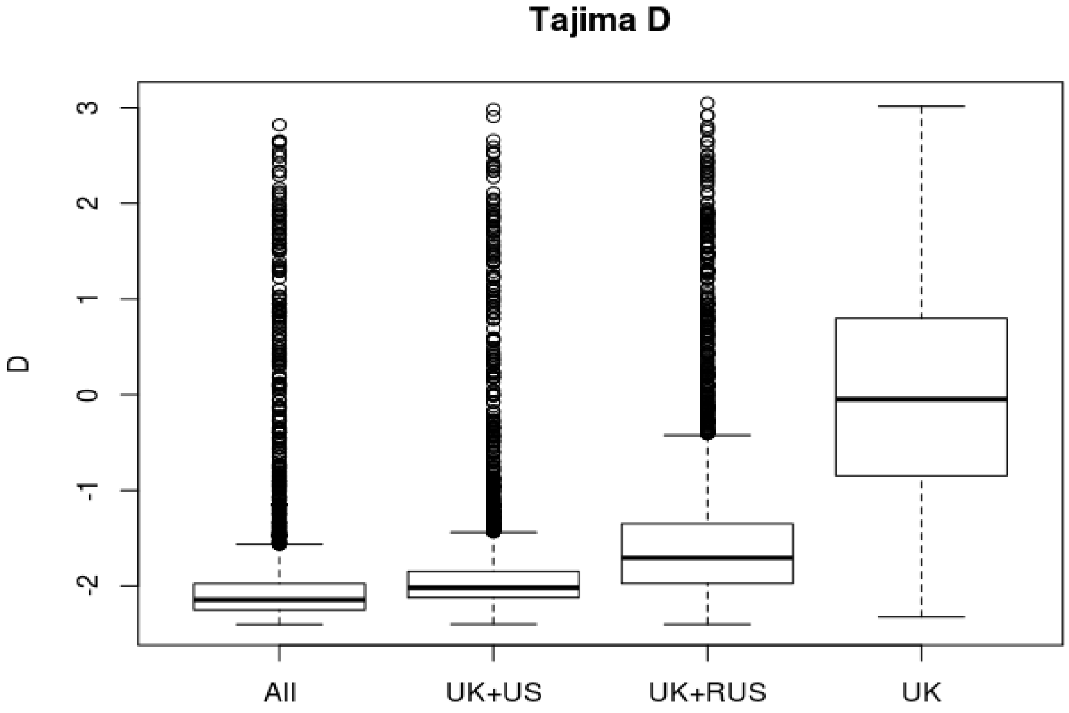


Figure 5. Tajima's D values per assembly dataset. Tajima's D values of UK subset have an average value of 0 compared to the full dataset of -2.

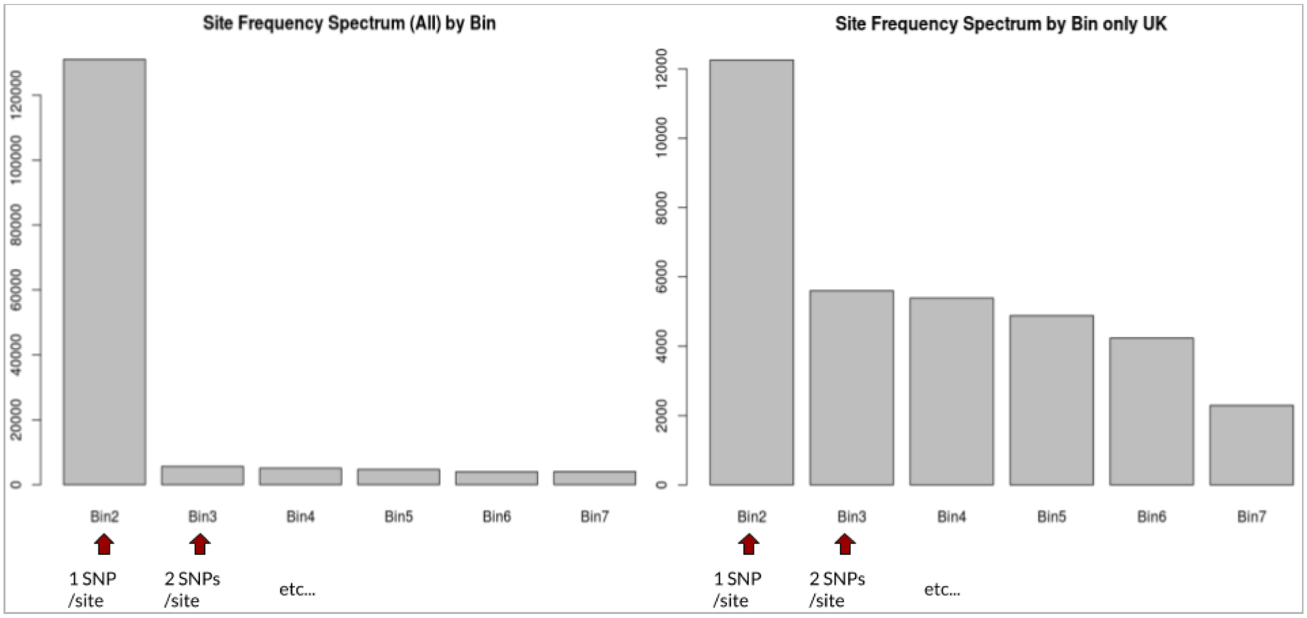


Figure 6. Site Frequency Spectrums for MugsyP-F including the entire assembly dataset (left) and only the UK assemblies (right). Bins refer to the number of SNPs found per site, for example, Bin7 refers to 6 SNPs per site. There is a more even distribution in the UK versus all isolates.

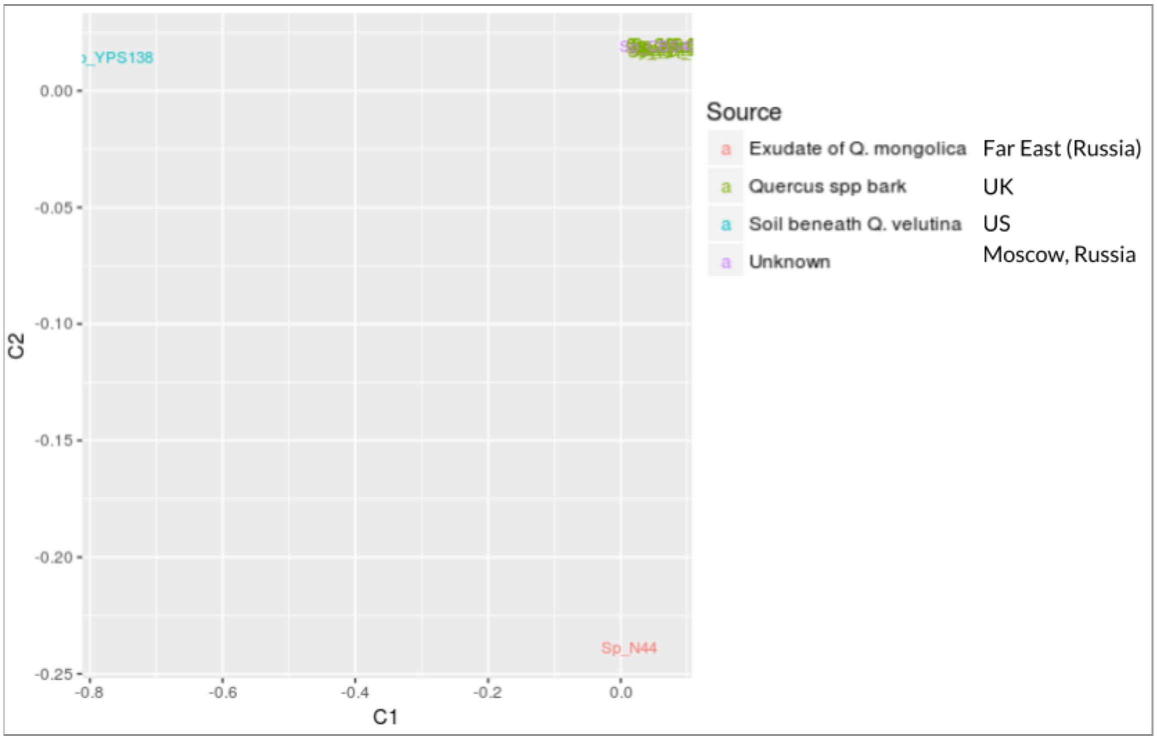
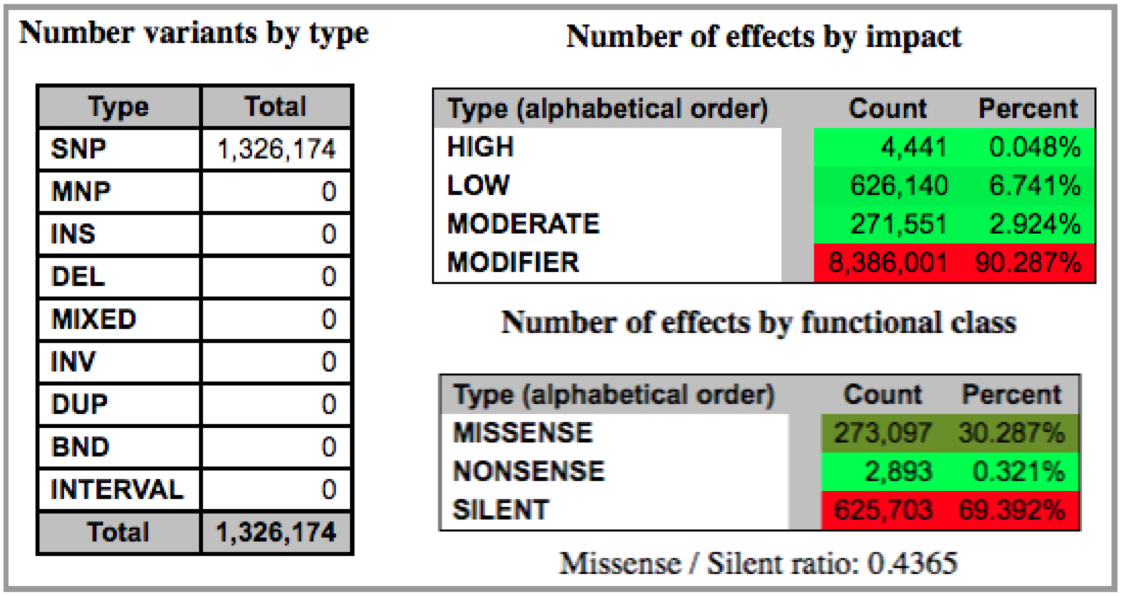


Figure 7. PCA plot of dataset population structure. *S. paradoxus* YPS138 and N44 (US and Far East Russian isolates) are drastically different from the rest.

Candidate Gene Survey

Following SnpEff analysis and filtering of the second dataset, 15 genes containing 10 or more high impact SNPs (see Table 5) were identified. According to SnpEff, a total of more than 1 million SNPs were present within the alignment, out of those 0.048% made high impacting effects and 30.608% were of either missense or nonsense function (potentially codon changing, i.e. non-synonymous) (19,483 SNPs) (Figure 8). After matching SnpEff annotated genes to the temperature-related database (with a total of 1,816 genes) and filtering, 320 SNPs (1.64%) of the 19,483 variants remained within the relevant temperature related genes (Figure 8).

The identified genes interact with or are involved in a multitude of phenotypes (see Table 5). Our results suggest that these cellular functions and components are sensitive to higher than optimal temperatures. Starting with cell membrane formation (ARP2), cytoskeletal and microtubule formation (CIN2 and TUB3), and extracellular plasma membrane binding proteins (ECM 33 and 9). Perhaps some variants may help the cell cope with higher temperatures (DDI3) or are directly involved in expression of other genes (HPC2, MUD1, IWR1, TFC3, and STO1). Others are related to different kinds of pathways: KIN28 is a protein kinase, PMI40 encodes Mannose-6-phosphate isomerase-- required for mannose glycoside synthesis, and RFA2 encodes a subunit for Replication Protein A-- a protein involved in DNA replication, repair, and recombination (see Table 5). The gene that possessed the greatest amount of high impact variants was TFC3, a subunit of the RNA polymerase III transcription initiation factor complex (see Figure 10).



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

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	4,441	0.048%
LOW	626,140	6.741%
MODERATE	271,551	2.924%
MODIFIER	8,386,001	90.287%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	273,097	30.287%
NONSENSE	2,893	0.321%
SILENT	625,703	69.392%

Missense / Silent ratio: 0.4365

Figure 8. Relevant SnpEff variant analysis results.

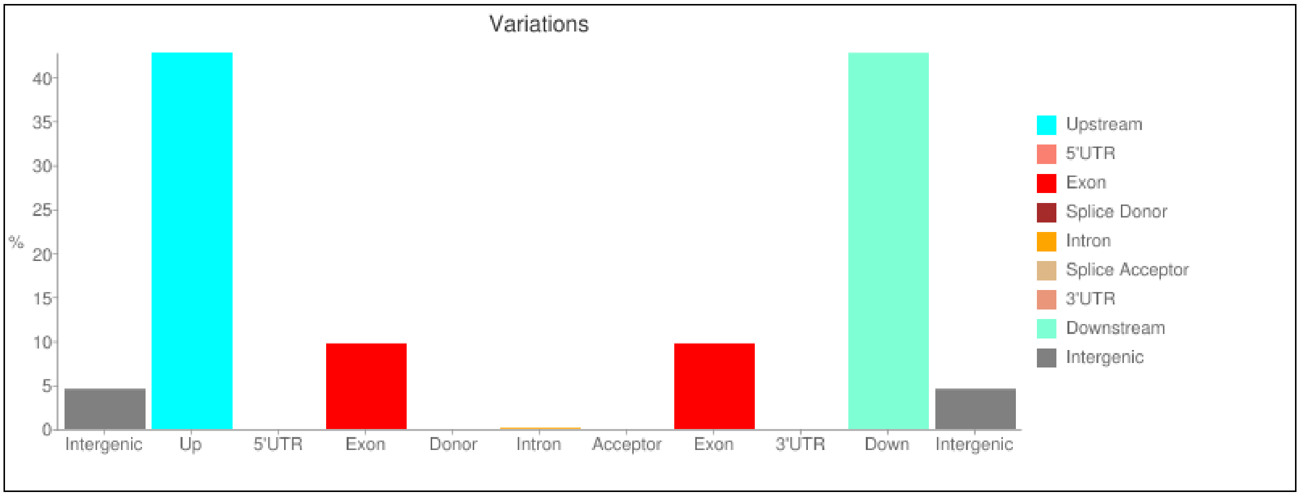


Figure 9. Regions where variants were found in SnpEff. Phenotypically relevant variation may be found not only in exons but in the regulatory regions of genes.

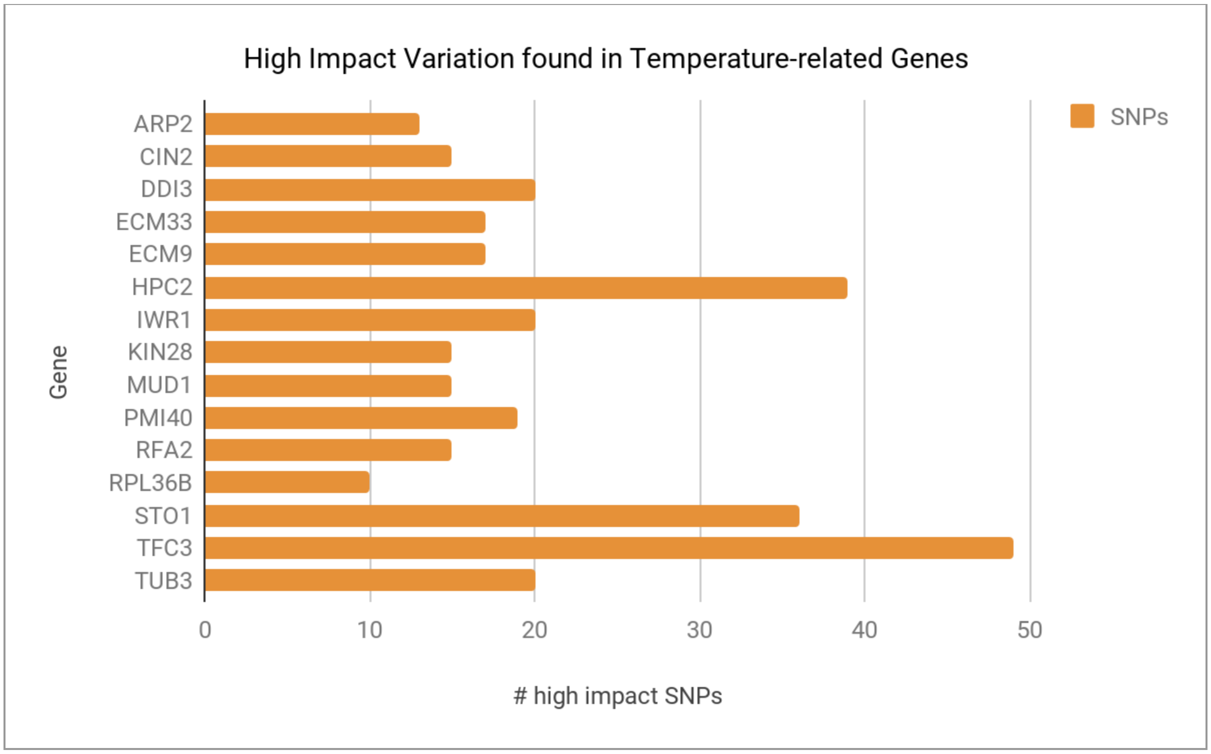


Figure 10. The number of high impact SNPs found in their corresponding genes possessing a temperature-dependent phenotype.

Table 5. Survey for candidate genes that possess a temperature-dependent phenotype

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
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

Discussion & Conclusion

False positives called from a multiple genome alignment do not always possess equal power, their importance depends on the question being investigated (Markova-Raina 2011). For example, the severity of an error (i.e. mis-alignment) varies between creating a phylogeny and finding evidence for positive selection at a specific site (Markova-Raina 2011). Severity of an error positively relates with the specificity needed for the analysis. It is therefore more critical to have a more accurate alignment when searching for positive selection because positive selection occurs not across the entire genome, but at particular genes (at single nucleotides or regions within those genes).

In this study, we searched for variation in genes related to temperature (i.e. a stable frequency of polymorphic genes that largely contribute to fitness) within the grouped dataset spanning multiple geographically distinct groups.

In this ongoing study, we are hoping to prove the hypothesis that temperature positively selects for alleles that have the largest effects on phenotype under higher temperatures and that due to seasonal changes, maintains polymorphisms at these loci (Kawecki 2004). Ultimately, we are searching for evidence of local temporal adaptation to temperature. If the population is seasonally undergoing selection, these loci will show greater differentiation within the entire population (Kawecki 2004). Moreover, alleles with stronger fitness effects are more likely involved in local adaptation because their chances are lower at being lost by genetic drift due to their initially higher frequencies (Kawecki 2004).

Nucleotide Diversity of both Datasets

A proven sampling bias was observed in the first dataset. Tajima's D values corresponded to this explanation. After the removal of two isolates from highly differentiated locations, the average Tajima's D value increased, indicating removal of an excess of rare alleles. When the nucleotide diversity estimator is negative, a higher number of rare alleles exist within the population (Tajima 1989). The only two isolates introducing most of these rare alleles were from the two poorly sampled populations. To increase the confidence in our findings, our sampling of these populations should be made equivalent to the other sample sizes. Additionally, the principal component analysis showed a distinct population structure (Figure 7), where there are three distinct groupings-- the two most distantly related isolates grouped separately alone. Thus, the average Tajima's D values per chromosome were based on artifacts and not biological mechanisms. Among the second dataset only one isolate was from a warmer region (Italy), this isolate would also introduce a disproportionate number of rare alleles.

If a negative Tajima's D value was of true significance, it could suggest a population expansion after a bottleneck event, a selective sweep that brought an allele to fixation, linkage to a swept allele, or purifying selection when deleterious alleles are removed from the population. Alleles that survived the bottleneck, that were fixed in the population, or were being selectively maintained within the population would be at relatively high frequencies, therefore any mutations that occurred would be rarer (Tajima et al 1989). When Tajima's D is greater than 0 or more positive than another population's value, polymorphisms are maintained at lower frequencies, meaning that there is a higher diversity of polymorphisms kept within the population. If overall there are more polymorphisms, an allele with a lower frequency (i.e. a rare allele) would still be relatively frequent when comparing all frequencies of different alleles. This could indicate balancing selection or population shrinkage (Tajima et al 1989). If Tajima's D is equal to 0, then there is no mechanism producing a dominant effect on variation in either direction (Tajima et al 1989).

A positive Tajima's D value that supports balancing selection within a population could mean that local adaptation is occurring-- fitnesses could vary within the population depending on local niches or conditions and possess corresponding genotypes (Kawecki 2004). This could also lead to the divergence of the population potentially (Kawecki 2004). In the case of seasonally adapted *S. paradoxus*, we should expect to see balancing selection of certain alleles when looking at the population as a whole. In contrast, when looking at each population individually by season, we should expect the opposite effect: a selective sweep at those allele positions.

A local population in this context, belongs to the same species and is connected by dispersal and gene flow (Kawecki 2004). When explaining geographical local adaptation, gene flow is more easily restricted when a population becomes diverged enough. To support this, Tsai et al found that in the UK *S. paradoxus* population, members of the same clone showed spatial aggregation, in that, genotypes were differentially dispersed due to localized expansion of the clones. Local adaptation is known to be reduced by gene flow, genetic drift, and generalization (naturally selected for by high variability of conditions) (Kawecki 2009). In terms of temporal local adaptation, a population has to be segregated enough by temporally changing conditions to acquire different genotypes. The conditions have to be changing but stable enough to apply sufficient selective pressure on the population (Kawecki 2004).

Patterns of local adaptation have to be tested for-- this mechanism is not alone in causing frequency changes in genetic variation within a population. There are other sources of change of genetic variation including genetic drift, linkage, and gene flow (Kraemer 2016). Other values can be used to measure such sources in combination with Tajima's D, such as linkage disequilibrium, F_{ST} & F_{IS} inbreeding coefficients, and effective population size (Goode 2011, Springer 2008, Husemann 2018). Calculating Tajima's D and its significance is one method for finding evidence for local adaptation (Pavlidis 2017), another is to test for specific fitness patterns using fitness assays (Kraemer 2016).

Local adaptation can be described using the equation: **Fitness = genotype + environment + (genotype x environment)** (Kraemer 2016). The last variable is the interaction term meaning, "genotype-by-environment [or] forces of natural selection often vary in space" (Kawecki 2004). In other words, divergent selection can produce traits (and genotypes) that provide a fitness advantage in that specific environment (Kawecki 2004). The genotype-by-environment interaction is what indicates whether local adaption is the responsible force influencing the variation seen within the population. Kawecki et al describes three fitness patterns that are characteristic of local adaptation:

1) A significant genotype-by-environment interaction.
2) Antagonistic pleiotropy (trade-offs)
3) The phenomenon of a local organism that possesses a particular genotype being fitter in its home habitat compared to others from a different niche (having a distinct advantage over the others).

Testing the fitness of the genes proposed in this survey (Table 5) is the next step. The genes will be tested for their fitness effects using a mutant collection of the German *S. paradoxus* isolates. There is a decent amount of genes that contain many high impact SNPs (Figure 10).

However, there are still issues to consider. Perhaps the variation that is truly determining the adapted phenotype is in a region outside of the expressed exon(s) in the regulatory parts of the gene (Figure 9). It is also conceivable that not enough genetic variation exists for local adaptation to occur (Kawecki 2004), therefore, it would be beneficial if this pipeline was repeated with a larger, more even dataset-- significant variation would be detected more accurately with a larger sample size per region or with the sequenced German collection itself. It is also important to keep in mind that temperature may only be one component of the interaction that determines the fitness patterns (Kraemer 2016). Lastly, temperature changes may also be too variable for any stable shift in allele variation.

Theoretically, it is possible for a population of free-living microbes to adapt to patterns of environmental change within one geographic area-- local adaptation is multi-dimensional. There are temperature adapted *Saccharomyces,* however not of one species population (meaning gene flow is already reduced) (Paget 2012).

Although temporal local adaptation has yet to be observed in a free-living microbial population of one species, there are too few investigations to make any sound conclusions (Kramer 2016). The fundamental questions addressed in this ongoing research are important to keep in mind. Firstly, does temporal local adaptation exist in populations of free-living microbes? Secondly, does temporal local adaptation occur within the wild *Saccharomyces paradoxus* population collected from a forest in Northern Germany? Seasonal variation in Germany may be sufficient for temporal local adaptation to temperature (www.dwd.de/EN/climate_environment/climateatlas/climateatlas_node.html). Thirdly, do annual temperature patterns select for certain yeast phenotypes and genotypes? Finally, can significant genetic variation be identified within the German isolates? Perhaps identification of other kinds of variation or use of other methods can help answer the above questions such as GWAS or QTL analyses.

References

Angiuoli, SV and Salzberg, SL. Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 2011 27(3):334-4

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

Botstein, D, Steven A. Chervitz, and J. Michael Cherry. "Yeast as a Model Organism." *Science (New York, N.Y.)* 277.5330 (1997): 1259–1260.

Brudno, Do, Cooper, Kim, Davydov. NISC Comparative Sequencing Program, E. D. Green, A. Sidow, S. Batzoglou, LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13, 721–731 (2003).

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res: 40:D700-5. [PMID: 22110037]

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM."A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3". 2012. Apr-Jun;6(2):80-92. PMID: 22728672

Darling AE, Mau B, Perna NT (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE 5(6): e11147.https://doi.org/10.1371/journal.pone.0011147

Dutheil JY, Gaillard S, Stukenbrock EH. 2014. BMC Genomics. Jan 22;15:53. MafFilter: a highly flexible and extensible multiple genome alignment files processor. For filtering option descriptions: http://biopp.univ-montp2.fr/manual/html/maffilter/v1.2.1/

Goode E.L. (2011) Linkage Disequilibrium. In: Schwab M. (eds) Encyclopedia of Cancer. Springer, Berlin, Heidelberg.

Gouy M., Guindon S. & Gascuel O. (2010). SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2):221-224.

Inbreeding Coefficient. In: Encyclopedia of Genetics, Genomics, Proteomics and Informatics. (2008) Springer, Dordrecht.

Huang, CH., Lee, FL. & Tai, CJ. The beta-tubulin gene as a molecular phylogenetic marker for classification and discrimination of the Saccharomyces sensu stricto complex.
Antonie van Leeuwenhoek (2009) 95: 135. https://doi.org/10.1007/s10482-008-9296-1

Husemann, M et al. “Effective Population Size in Ecology and Evolution.” *Heredity* 117.4 (2016): 191–192. *PMC*. Web. 22 Apr. 2018.

Katoh, Misawa, Kuma, Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059-3066

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.

Pavlidis P, Alachiotis N. A survey of methods and tools to detect recent and strong positive selection. J Biol Res [Internet]. 2017 Apr;24(1):7. Available from: https://doi.org/10.1186/s40709-017-0064-0

Tajima, F. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123.3 (1989): 585–595.

Tang, Haiming, and Paul D. Thomas. “Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation.” *Genetics* 203.2 (2016): 635–647. *PMC*. Web. 22 Apr. 2018.

Supplementary Material

Commands for pipeline

1.
   ```
   wget ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest/para_assemblies.tgz
   ```
   (Liti et al. 2009)

2.
   ```
   wget http://www.moseslab.csb.utoronto.ca/sgrp/data/SGRP2-assemblies_Jun25.tar.gz
   ```
   (Bergström et al. 2014)

3.
   ```
   python new_header.py
   ```
   Raw assembly files were renamed manually and given new scaffold headers.

4.
   ```
   python quast.py Assembly1.fa Assembly 2.fa Assembly 3.fa …
   ```
   QUAST is an assembly quality assessment tool-- with many short assembly scaffolds, filtering was required to reduce processing time.

5.
   ```
   python filtering_CS.py
   ```
   Filters out scaffolds shorter than 1000 bps.

Steps 6 through 13 are for running TBA

6.
   ```
   python ./blast_script.py -h
   ```

Blasts all sequences using beta-tubulin gene for creation of a phylogenetic tree.

7.
   ```
   python ./Create_Seaview_Input.py -h
   ```
   Seaview generates the tree and allows for conversion of the tree into Newick format, necessary for the genome aligner MultiZ. (“,” are replaced with a space)

8.
   ```
   all_bz - ‘Phylogenetic_Tree_in_newick_format’ >& all_bz.log
   ```
   List of all pairwise blastz commands are created with all_bz to be executed.

9.
   ```
   ./all_bzToSGE_modified.py
   ```
   Creates a bash script that will execute the pairwise alignments.

10.
    ```
    ./formatSeq_PacBio.py -h
    ```
    Re-formats the fasta assemblies to a format readable by the aligning program MultiZ within TBA.

11. **qsub scriptSGE.sh** & **qstat**
    Generates pairwise alignments for TBA. Qstat checks the state of the alignment.

12.
    ```
    time tba ‘Phylogenic_Tree’ *.*.maf tba_Alignment_name.maf>&
    tba_Alignment_name.log
    ```
    Runs TBA to create the multiple genome alignment (MGA).

13.
    ```
    maf_project tba_Alignment_name.maf Reference_ID>&
    tba_Alignment_name_reference_name.maf
    ```
    Projects the MGA against the reference strain.

14.
    ```
    source ~/Programs/mugsy_x86-64-v1r2.2/mugsyenv.sh
    time mugsy --directory Output_Directory -fullsearch -duplications 1 -prefix
    Output_prefix Files_separates_by_spaces
    ```
    Runs Mugsy-serial

15.
    ```
    source ~/Programs/mugsy_x86-64-v1r2.2/mugsyenv.sh
    time mugsy_parallel --directory Output_Directory -nproc 8 -fullsearch -duplications
    1  -prefix Output_prefix Files_separates_by_spaces
    ```
    Runs Mugsy-parallel

16. For maffilter bpp file format see Maffilter manual (Dutheil 2014)

    a. ```
       maffilter param=Option_file_with_mafft_settings.bpp
       DATA=MGA_maf_file
       ```
       See MafFilter_Realign.bpp and MafFilter_Realign_ploen.bpp

    b. ```
       maffilter param=Option_file_with_Filter_Settings.bpp DATA=
       MGA_maf_file
       ```
       Filters MGA using specified filtering options, see filter_settings_for16_maffilter.bpp, filter_settings_for14_maffilter.bpp, and filter_settings_forploen_maffilter.bpp

17.
    ```
    QMS.py -bs 10 -f mafFile1,mafFile2,mafFile3
    ```
    Generates alignment statistics relevant for accuracy evaluation.

18.
    ```
    maffilter param=option_file_with_sequencestatitics.bpp DATA=MGA_maf_file
    ```
    Generates site frequency spectrums.

19.
    ```
    maffilter param=option_file_with_VcfOutput.bpp DATA=MGA_maf_file
    ```

20.
    ```
    grep “#” VCF_file.vcf > Filtered_vcf_file.vcf
    grep “#” -v VCF_file.vcf | grep “PASS” >> Filtered_vcf_file.vcf
    ```
    Writes variants only with filter flag “PASS” to a new vcf file for further analysis.

21.
    ```
    maffilter param=option_file_with_population_statistics.bpp DATA=MGA_maf_file
    ```

22.
    ```
    python PCA_SNPs.py
    ```
    Vcf file as input, outputs converted Ped file to Map file

23.
    ```
    python Adjust_MapFile.py
    ```
    Map file input

24.
    ```
    R Plot_PCA.R
    ```

Steps for processing second dataset

1.
   ```
   java -Xmx4g -jar /directory/to/snpEff.jar -c /directory/to/snpEff_config_file
   R64-1-1.86 /directory/to/vcf_file > name_of_annotated_vcf_file.ann.vcf
   ```
   Generates SnpEff files.

2.
   ```
   python path/to/matching_snps_to_temp_genes.py
   ```
   Matches SnpEFF output genes to heat sensitive genes (from "temperature sensitive growth"observed phenotype annotation database; https://www.yeastgenome.org/observable/APO:0000092).

3.
   ```
   python path/to/high_impact_genes.py
   ```
   Filters the matched snpEff genes that contain 10 or more SNPs (can change SNP threshold).

***Scripts** can be found in folder: “scripts_Corinn_biol608”