

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede amministrativa: Università degli Studi di Padova
Dipartimento di Agronomia Animali Alimenti Risorse Naturali e Ambiente

DOTTORATO DI RICERCA IN
Viticoltura, Enologia e Marketing delle Imprese Vitivinicole
CICLO XXIV

**A genomic and transcriptomic approach
to characterize oenological *Saccharomyces cerevisiae* strains.**

--

Caratterizzazione genomica e trascrittomica di ceppi naturali di
Saccharomyces cerevisiae di importanza enologica.

Coordinatore: Prof. Viviana Corich

Supervisore: Prof. Viviana Corich

Co-Supervisore: Dott. Stefano Campanaro

Dottoranda: Laura Treu

*Nature has a great simplicity
and therefore a great beauty*

Richard Feynmann

*In un universo subitamente spogliato di illusioni e di luci
l'uomo si sente un estraneo.
Persuaso dell'origine esclusivamente umana di tutto ciò che è umano,
cieco che desidera vedere e che sa che la notte non ha fine,
egli è sempre in cammino.*

da "il mito di Sisifo"
Albert Camus

*It is sometimes an appropriate response to reality
to go insane*

Philip K. Dick

ABSTRACT

Genus *Saccharomyces* includes a large number of microorganisms that are important for industrial applications such as the production of fermented beverages, biofuel and baking. Natural selection combined with domestication applied selective pressures to the genome of this yeast producing large numbers of different strains with specialized phenotypes.

During the last decades thousand of strains have been phenotypically characterized but correlation between phenotype and genotype is not yet completely unveiled. Genome sequence analysis is a crucial step to obtain a general description of gene content and highlight differences between strains. In this study the homozygous derivatives of four ecotypical *Saccharomyces cerevisiae* strains isolated from Raboso and Prosecco fermented grape bunch have been successfully sequenced using next generation sequencing, and a variety of tools have been used and developed to solve the complex task of genome finishing.

A detailed overview of gene expression in different winemaking and laboratory strains has also been performed using SOLiD RNA-seq. Samples growth in synthetic wine media on controlled bioreactors have been collected during fermentation process. Our results revealed a transcriptional fingerprint characterizing oenological strains adaptation to stressful environment. A comparison between differences in promoter sequences between strains and their downstream effect on gene expression have been performed and the results show a higher influence of tandem repeat variability respect to mutations on transcription factor binding sites. Finally using statistical analysis we correlate the genetic traits of strains with their metabolic properties and we obtained a global overview of fermentation performances in the different genetic groups.

Table of Contents

1. INTRODUCTION.....	5
YEAST BETWEEN BIOLOGY AND INDUSTRIES	5
<i>Yeast in winemaking.....</i>	6
<i>Wine Yeast Ecology</i>	8
YEAST METABOLISM	9
<i>Technological Characters</i>	11
FROM GENOTYPE TO PHENOTYPE.....	13
NEXT GENERATION SEQUENCING TECHNOLOGY	14
<i>Phylogenetic Relationship.....</i>	15
TRANSCRIPTIONAL PROFILE.....	15
<i>Regulatory Elements.....</i>	16
PROJECT OUTLINE	17
REFERENCES.....	18
2. STRAIN SELECTION	23
INTRODUCTION	23
<i>Qualitative Trait and Aromas.....</i>	23
<i>Oenological Yeasts Collection</i>	24
<i>Yeast Improvement Strategy.....</i>	25
MATERIALS AND METHODS	27
<i>Sporulation and Tetrad Dissection.....</i>	27
<i>Pulsed Field Gel Electrophoreses</i>	28
<i>Fermentation Ability and Ethanol Resistance</i>	28
<i>Growth Curve.....</i>	30
<i>Sulphite Stress Resistance.....</i>	30
<i>Compounds of Technological Interest</i>	30
<i>Chemical Analysis on Fermented Must.....</i>	31
RESULTS AND DISCUSSION.....	32
<i>Natural Isolates Selection</i>	32
<i>Strains Genetic Stability.....</i>	32
<i>Chromosomes Pattern.....</i>	33
<i>Derivative Lines Selection</i>	35
<i>Oenological Trait Evaluation.....</i>	36
<i>Fermentation Profiles</i>	40
REFERENCES	45

3. GENOME SEQUENCES	47
INTRODUCTION	47
<i>Genetic Characteristics</i>	47
<i>Chromosomal Rearrangements and SNPs</i>	48
<i>The Finishing Task</i>	48
<i>Gene Prediction</i>	50
MATERIALS AND METHODS. MOLECULAR BIOLOGY	51
<i>DNA Purification</i>	51
<i>DNA concentration and quality</i>	52
<i>Amplification by polymerase chain reaction (PCR)</i>	52
<i>Genomic DNA Sequencing</i>	54
<i>Cesium Chloride Centrifugation</i>	55
MATERIALS AND METHODS. BIOINFORMATICS	56
<i>Sequence Assembly</i>	56
<i>GapResolution and Finishing Process</i>	57
<i>Genomes Alignment and Visualization</i>	59
<i>Gene Prediction and Annotation</i>	61
<i>Comparison of Intergenic Regions</i>	61
<i>Neighbor Joining Tree and SNPs</i>	63
RESULTS AND DISCUSSION	64
<i>Sequence Assemblies</i>	64
<i>Gap Filling Results</i>	65
<i>SNPs Distribution and Phylogenesis</i>	67
<i>Structural Variations</i>	70
<i>Genomes Annotation</i>	73
<i>Transcription Factor Binding Sites</i>	74
<i>Tandem Repeats</i>	75
REFERENCES	76

4. TRANSCRIPTIONAL PROFILES.....	79
INTRODUCTION	79
<i>RNA Sequencing</i>	79
<i>Transcription Factors</i>	80
MATERIALS AND METHODS. MOLECULAR BIOLOGY	81
<i>Total RNA extraction</i>	81
<i>rRNA Subtraction</i>	81
<i>mRNA deCAPping</i>	82
<i>SOLiD Libraries preparation</i>	82
<i>Sequencing with the SOLiD system</i>	84
MATERIALS AND METHODS. BIOINFORMATICS	85
<i>Reads Alignment and Differential Expression</i>	85
<i>Hierarchical Clustering using TMEV</i>	86
<i>Gene Ontology</i>	87
RESULTS AND DISCUSSION.....	88
<i>RNA-seq Results</i>	88
<i>Gene Expression Level Results</i>	89
<i>Specific protein coding genes absent in S288c</i>	91
<i>Influence Of Structural Variations On The Expression Of Flanking Genes</i>	92
<i>GO Classes Enriched in Oenological strains</i>	92
<i>Genes Involved in Ethanol Tolerance</i>	93
<i>Transcription Factor Binding Sites</i>	97
<i>Differential expression linked to differences in TR lenght</i>	98
<i>Global analysis of the influence of different factors on gene expression.</i>	98
REFERENCES.....	101
5. DISCUSSION AND CONCLUSIONS	105
REFERENCES	109
ACKNOWLEDGEMENTS	111

1. INTRODUCTION

S. cerevisiae has a long history of association with human activity. This microorganism is used for lots of industrial processes, such as baking, brewing, wine and bioethanol production. Natural selection combined with artificial domestication applied selective forces and constraints to its genome producing a large number of different strains with specialized phenotypes. For this *S. cerevisiae* is a model to study how divergent selective pressures can modify the genomic content of species and how these differences can influence the phenotype. The physiological characterization of different yeast strains is quite common especially in those industries where strains are used. Genomic characterization, on the contrary, is becoming widespread just recently thanks to Next Generation sequencing technologies that allow to sequence genomes in short times and at affordable prices. In 1996 *S. cerevisiae* S288c was the first eukaryotic organism completely sequenced (1). Thus this strain is a model organism, its characteristics are completely different respect to strains used in technological applications so other strains with different evolutionary histories were selected and sequenced. For example, the comparison of the genome of S288c strain with the genomes of other yeasts of the four other yeast of the hemiascomycete phylum allowed the reconstruction of the evolutionary path leading to the differentiation of these species. Differences among genomes were used to infer events leading to speciation, such as intron loss, gene duplication and diversification, the appearance of new centromeres and MAT cassettes and whole-genome duplication (2). On the other hand, the comparison of low coverage depth genomes of seventy isolates of the baker's yeast *S. cerevisiae* and its closest relative, *S. paradoxus* was useful to examine variation in gene content, single nucleotide polymorphisms, nucleotide insertions and deletions, copy numbers and transposable elements and to identify new hypothetical open reading frames present in more than one strain or specific to a single lineage (3). All these studies show the potentiality offered by yeast, with its eukaryotic but simple genome, to understand molecular mechanisms underlying genome evolution.

YEAST BETWEEN BIOLOGY AND INDUSTRIES

S. cerevisiae is a single-celled fungus used both for biological research and industrial processes. In research field *S. cerevisiae* is a very common model organism thanks to its characteristics: it is small (5-30 μm) and can be easily cultured in liquid and solid media. It has a short life cycle of 90 min and its generation time is short (doubling time 1.25–2 hours at 30°C). It is stable in both the haploid and diploid state and under favourable conditions it propagates indefinitely by mitotic divisions forming large clonal populations. Under stress conditions it can undergo sporulation, entering meiosis and producing haploid spores with two different mating types α and a , which can mate between them or with spores from a different progenitor leading to the exchange of genetic material.

This model system is used for understanding fundamental cellular processes, metabolic pathways and for performing molecular analysis on many disease-associated genes. These associations are possible because *S. cerevisiae* is an eukaryote and it shares the complex internal cell structure of plants and animals. Its genome is simpler than those of the higher

eukaryotes but nearly 50% of human genes implicated in heritable diseases have yeast homologs.

Furthermore its relatively high rate of recombination between homologous DNA sequences allows the insertion of DNA sequences at specific locations within the genome and the generation of knockout strains (4). The genome of *S. cerevisiae* S288c was completely sequenced through a worldwide collaboration in 1996 (1). The haploid genome is 12 Mb long, is packaged into 16 chromosomes and is quite compact: approximately 70% of its DNA is composed by coding sequences and it is predicted to encode nearly 6,200 genes. The genes of higher eukaryotes typically contain introns; however, only 263 of yeast genes do (5).

The simple genome of yeast is an interesting subject also for bioinformatics, in fact it is often used to test several programs. Various yeast are used for technological processes that range from the ancient arts of bread, wine and beer making, to the modern application of heterologous protein production. Modern yeast technology represents a vast industrial sector worth about US\$ 70 billion *per annum*. Consumers of yeasts and yeast-based products demand continually improved quality and economics (6).

In modern industries *S. cerevisiae* is used in baking, brewing, wine and sake fermentation, and bioethanol production. Despite their diverse roles, the different *S. cerevisiae* industrial strains all share the general ability to grow and live under the influences of lots of environmental stressors like low pH, poor nutrient availability, high ethanol concentrations and fluctuating temperatures. All industrial strains evolved due to different selective pressures and are able to adapt to their specific environment better than the others. Clear differences can be found between industrial and non-industrial strains of *S. cerevisiae*, however there are numerous subtle differences also between strains used in the same industrial process (7). Yeast is also widely used as a probiotic because 50 percent of its mass is composed by proteins and is a rich source of B vitamins, niacin, and folic acid (SGD, 2008) and, with today's ever-growing energy needs, yeast has broadened its scope from food into fuel production, as the industry keeps striving to increase the maximum yield from feedstock and microorganisms.

Yeast in winemaking

In 1863, Louis Pasteur revealed the presence of microbial activity during wine fermentation and he proved that yeast is the primary catalyst of this process. Wine fermentation is a complex ecological and biochemical process involving the sequential development of different yeast species. Yeasts are predominant in the ancient and complex process of winemaking. Winemakers have long noted that different strains of wine yeasts, even when used to ferment the same juice under identical conditions, can yield very different wines in terms of sensory characteristics, presumably as a result of variations in the strains' fermentative properties. Previous studies have demonstrated genetic diversity among both commercial and wild *S. cerevisiae* wine yeast strains, and it has been hypothesized that this genetic diversity may, at least in part, be a root cause of their differing fermentative and sensory qualities. The aroma and flavour profile of wine is the result of an almost infinite number of variations in production, whether in the vineyard or the winery. In addition to the obvious, such as the grapes selected, the winemaker employs a variety of techniques and tools to produce wines with specific flavour profiles. One of these tools is the choice of microorganism to use in the fermentation process.

During alcoholic fermentation, the wine yeast *S. cerevisiae* brings forth the major changes between grape must and wine: modifying aroma, flavour, mouth-feel, colour and chemical complexity.

Thus flavour-active yeasts and bacterial strains can produce desirable sensory results by helping to extract compounds from the solids in grape must, by modifying grape-derived molecules and by producing flavour-active metabolites (8). In spontaneous fermentations, there is a progressive growth pattern of indigenous yeasts, with the final stages invariably being dominated by the alcohol-tolerant strains of *S. cerevisiae*. This species is universally known as the 'wine yeast' and is widely preferred for initiating wine fermentations. The primary role of wine yeast is to catalyze the rapid, complete and efficient conversion of grape sugars to ethanol, carbon dioxide and other minor, but important, metabolites without the development of off-flavours. However, due to the demanding nature of modern winemaking practices and sophisticated wine markets, there is an ever-growing quest for specialized wine yeast strains possessing a wide range of optimized, improved or novel oenological properties (9)(10).

The microflora of grapes varies according to the grape variety, temperature, climatic influences, soil and viticulture practices. Must is complete in nutrient content but its low pH and high sugar content induce selective pressure on the microorganisms so only few yeast and bacterial species can survive and proliferate. Sulphur dioxide, added as an antioxidant and antimicrobial preservative, together with the increasing levels of ethanol produced during fermentation select the remaining microorganisms further on leaving only *S. cerevisiae* as unique responsible for alcoholic fermentation. Originally, wine was made by using the natural indigenous microflora for spontaneous fermentation. This process was performed by the alcohol-tolerant strains of *S. cerevisiae* but other yeasts, such as species of *Brettanomyces*, *Schizosaccharomyces*, and *Zygosaccharomyces*, might be present during the fermentation and some of them were capable of adversely affect sensory quality.

From 1890 the practice of inoculating must with pure yeast starter cultures started to diffuse and commercial active dried wine yeast were produced. The diffusion of commercial starter strains is quite controversial because they are thought to induce a standardization of the wine organoleptic characters (8). On the other hand non-commercial yeast strains associated with specific vineyards are thought to give a distinctive style and quality to wine. However the outcome of spontaneous fermentation depends not only on the yeasts, but also on grape chemistry and processing protocol.

For these reasons the identification and characterization of new ecotypical starter strains to be used exclusively in the area of isolation and selected to develop the desired organoleptic traits is becoming common. Characteristics of yeast that are important to determine the quality of the wine and that are used to select starter strains, are the fermentation rate, the alcohol tolerance, the resistance to the sulphur dioxide and the production of chemical compounds conditioning the aroma (11). Most of the selected strains are *S. cerevisiae* strains adapted to the specific wine-producing region and they can be quite different in their fermentation performance and their contribution to the final bouquet and quality of wine.

Wine Yeast Ecology

The diversity of yeast species on grapes has been investigated in vineyards worldwide (12,13) and numerous reviews have covered this topic (14). With respect to the vineyard and winery niche habitats, some of these yeasts are considered as “autochthonous” (essential) and others as “allochthonous” (transient or fortuitous) members of the communities found in these environments.

Their successful coexistence depends on the sum of all physical, chemical and biotic factors that pertain to vineyards and wineries. ‘Generalist’ yeasts are endowed with a broad niche and occupy many habitats, whereas ‘specialist’ yeasts occur in unique habitats (15). The microflora of grapes vary according to the grape variety; temperature, rainfall and other climatic influences; soil, fertilization, irrigation and viticultural practices (e.g. vine canopy management); development stage at which grapes are examined; physical damage caused by mould, insects and birds; and fungicides applied to vineyards. It is also important to note that harvesting equipment, including mechanical harvesters, picking baskets and other infrequently cleaned delivery containers can also represent sites for yeast accumulation and microbiological activity before grapes reach the winery (16). Using aggressive washing and analytical techniques, a concentration of 3×10^5 yeast cells cm^{-2} of the berry surface has been estimated. Other studies suggest a range of 10^4 – 10^6 cells cm^{-2} (14). The factors impacting which genera and species are found have also been evaluated. The methodologies have differed, but there is a striking similarity of the main genera and species found. There are three principal genera found on grapes: *Hanseniaspora uvarum* (anamorph: *Kloeckera apiculata*), *Metschnikowia pulcherrima* (anamorph: *Candida pulcherrima*), and *Candida stellata*. In some reports, *Hanseniaspora* is the dominant genre and in others it is *Candida* (17,18).



Figure 1.1 Prosecco wine grapes and image of yeast cells taken by electron microscopy.

Other yeasts can be commonly found, although they are not as universal. *Saccharomyces* can be detected, but is present on grape surfaces at very low levels and has been undetectable in some studies (19). A key factor determining the species present on the surface of grape appears to be the amount of damage to the fruit. The leakage of sugar substrates either through physical damage mediated by insects, birds, or invasive fungal species, or as a consequence of berry aging and shrivel on the vine due to dehydration, enriches for the ascomycetes (20,21). The presence of other yeast *genera* depends also upon regional and climatic influences, the grape variety, disease pressure and vineyard practices.

The major species identified using viable isolates and total DNA extraction were the same, but a greater number and diversity of yeasts were detected in the direct DNA isolation studies. In general, the number of yeast cells present on grapes increases with ripening, and the numbers are higher by one or two orders of magnitude nearer the peduncle. Seasonal variation has also been observed with warmer and dryer years yielding increased yeast populations (10). It was thought that the higher levels of *Saccharomyces* seen in some vineyards may be due to the practice of placing yeast lees from the fermentation in the vineyard as a source of vine fertilization. To test this hypothesis, the effect of deliberate inoculation of vineyards with *Saccharomyces* on the presence of *Saccharomyces* at the time of harvest has been investigated (22). The winery residents and vineyard *inocula* did not become established in the berry flora in spite of high inoculation levels. Puncturing the grapes to induce berry seepage and damage did not improve the chances of colonization by the *Saccharomyces* inoculums. Microbial flora often also coat winery walls, outer barrel surfaces, hoses, and drains, particularly during barrel ageing, as this is typically done under conditions of humidity to prevent evaporative loss of wine volume. Sanitation practices vary widely, as does the practice of supplementation with nutrients. All of these factors impact winery flora.

YEAST METABOLISM

Industrial cultivation of wine yeasts can have a profound effect on the microbiological quality, fermentation rate, production of hydrogen sulphide, ethanol yield and tolerance, resistance to sulphur dioxide as well as tolerance to drying and rehydration. The primary selection criteria applied to most strain development programs relate to the overall objective of achieving a better than 98% conversion of grape sugar to alcohol and carbon dioxide, at a controlled rate and without the development of off-flavours. The growth and fermentation properties of wine yeasts have, however, yet to be genetically defined.

What makes the genetic definition of these attributes even more complex is the fact that lag phase, rate and efficiency of sugar conversion, resistance to inhibitory substances and total time of fermentation are strongly affected by the physiological condition of the yeast, as well as by the physicochemical and nutrient properties of grape must. Generally, sugar catabolism and fermentation proceed at a rate greater than desired, and are usually controlled by lowering the fermentation temperature (23).

In *S. cerevisiae*, glucose and fructose, the main sugars present in grape must, are metabolized to pyruvate via the glycolytic pathway. Pyruvate is decarboxylated to acetaldehyde, which is then reduced to ethanol. The rate of fermentation and the amount of alcohol produced per unit of sugar during the transformation of grape must into wine is of considerable commercial importance. During wine yeast glycolysis, one molecule of glucose or fructose yields two molecules each of ethanol and carbon dioxide. However, the theoretical conversion of 180 g sugar into 92 g ethanol (51.1%) and 88 g carbon dioxide (48.9%) could only be expected in the absence of any yeast growth, production of other metabolites and loss of ethanol as vapour (24). In a model fermentation, about 95% of the sugar is converted into ethanol and carbon dioxide, 1% into cellular material and 4% into other products such as glycerol.

The first step to ensure efficient utilization of grape sugar by wine yeasts is to replace any mutant alleles of genes encoding the key glycolytic enzymes, namely hexokinase (HXX), glucokinase (GLK), phosphoglucose isomerase (PGI), phosphofructokinase (PFK), aldolase (FBA), triosephosphate isomerase (TPI), glyceraldehyde-3-phosphate dehydrogenase (TDH), phosphoglycerate kinase (PGK), phosphoglycerate mutase (PGM), enolase (ENO), pyruvate kinase (PYK), pyruvate decarboxylase (PDC) and alcohol dehydrogenase (ADH). The genes encoding PGI, TPI, PGM and PYK appear to be present in single copy in a haploid genome, while multiple forms exist for TDH (three isozymes), ENO (two isozymes) and GLK (three isozymes) (14).

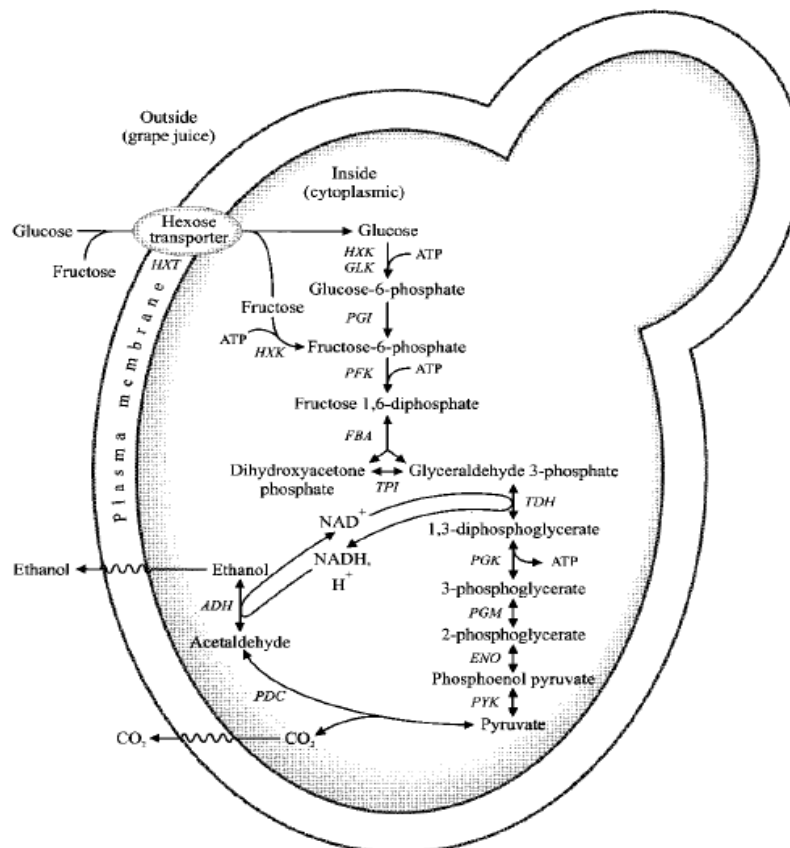


Figure 1.2 Glycolytic pathway in wine yeast (10)

The assumption that an increase in the dosage of genes encoding these glycolytic enzymes would result in an increase in the efficiency of conversion of grape sugar to alcohol has been disproved; it has been demonstrated that overproduction of the enzymes has no effect on the rate of ethanol formation (25). This indicates that the step of sugar uptake represents the major control site for the rate of glycolytic flux under anaerobic conditions, whereas the remaining enzymatic steps do not appear to be rate limiting. In other words, the rate of alcohol production by wine yeast is primarily limited by the rate of glucose and fructose uptake. Therefore, in winemaking, the loss of hexose transport towards the end of fermentation may result in reduced alcohol yields (15). Sugars enter yeast cells in one of three ways: simple net diffusion, facilitated (carrier-mediated) diffusion and active (energy-dependent) transport. In grape must fermentations where sugar concentrations above 1 M are common, free diffusion may account for a very small proportion of sugar uptake into yeast cells.

However, since the plasma membranes of yeast cells are not freely permeable to highly polar sugar molecules, various complex mechanisms are required for efficient translocation of glucose, fructose and other minor grape sugars into the cell. The hexose transporter family of *S. cerevisiae* consists of more than 20 proteins comprising high, intermediate and low affinity transporters and at least two glucose sensors. Many factors affect both the abundance and intrinsic affinities for hexoses of these transporters present in the plasma membrane of wine yeast cells, among them glucose concentration, stage of growth, presence or absence of molecular oxygen, growth rate, rate of flux through the glycolytic pathway and nutrient availability (particularly nitrogen) (24). Although the precise mechanisms and regulation of grape sugar transport of wine yeast are still unclear, some aspects about glucose and fructose uptake can be noted. Glucose uptake is rapid down a concentration gradient, reaching an equilibrium and is therefore not accumulative (26). Several specific, energy-dependent glucose carriers mediate the process of facilitated diffusion of glucose and proton symport is not involved. Phosphorylation by the HXK1- and HXK2-encoded hexokinases and the GLK1-encoded glucokinase is linked to high-affinity glucose uptake. Glucose transporters, encoded by HXT1-HXT18 and SNF3, are stereospecific for certain hexoses and will translocate glucose, fructose and mannose.

Some members of this multigene permease family affect glucose, galactose, glucose and mannose, or glucose, fructose and galactose uptake, but thus far none has been described as specifically affecting fructose uptake (15). It appears that in *S. cerevisiae*, fructose is transported via facilitated diffusion rather than active transport, whereas related species (*S. bayanus* and *S. pastorianus*) within the *Saccharomyces sensu stricto* group do possess fructose-proton symporters. Based on the spectacular increase in the amount of information on sugar sensing and their entry into yeast cells that has come to the fore over the last few years, several laboratories have identified this main point of control of glycolytic flux as one of the key targets for the improvement of wine yeasts. For example, in some instances, certain members of the HXT permease gene family are being overexpressed in an effort to enhance sugar uptake, thereby improving the fermentative performance of wine yeast strains. However, more in-depth details are required about the complex regulation of glucose and fructose uptake as well as glycolysis as it occurs in grape juice (especially in the presence of high sugar levels during the early phase of fermentation and during the final stages of sugar depletion coupled to nutrient limitation) before it will be possible to devise novel strategies to improve wine yeast's fermentation performance and to prevent sluggish or stuck fermentations.

Technological Characters

With the importance of *S. cerevisiae*'s role in winemaking now firmly established, there is an ever-growing demand for new and improved wine yeast strains. In addition to the primary role of wine yeast to catalyze the efficient and complete conversion of grape sugars to alcohol without the development of off-flavours, starter culture strains of *S. cerevisiae* must now possess a range of other properties, such as those listed in Table 2. The importance of these additional yeast characteristics differs with the type and style of wine to be made and the technical requirements of the winery. The need is for *S. cerevisiae* strains that are better adapted to the different wine-producing regions of the world with their respective grape varieties, viticultural practices and winemaking techniques (9).

Table 1.1 Desirable characteristics of wine yeast (9).(10)

Fermentation properties	Technological properties
Rapid initiation of fermentation	High genetic stability
High fermentation efficiency	High sulphite tolerance
High ethanol tolerance	Low sulphite binding activity
High osmotolerance	Low foam formation
Low temperature optimum	Flocculation properties
Moderate biomass production	Compacts sediment
Flavour characteristics	Resistance to desiccation
Low sulphide/DMS/thiol formation	Zymocidal (killer) properties
Low volatile acidity production	Genetic marking
Low higher alcohol production	Proteolytic activity
Liberation of glycosylated flavour precursors	Low nitrogen demand
High glycerol production	Metabolic properties with health implications
Hydrolytic activity	Low sulphite formation
Enhanced autolysis	Low biogenic amine formation
Modified esterase activity	Low ethyl carbamate (urea) potential

Some of the requirements listed above are complex and difficult to define genetically without a better understanding of the biochemistry and physiology involved. To date, no wine yeast in commercial use has all the characteristics listed, and it is well established that wine yeasts vary in their winemaking abilities. While some degree of variation can be achieved by altering the fermentation conditions, a major source of variation is the genetic constitution of the wine yeasts. One of the most important characters is the fermentation efficiency, together with the rapid initiation of the process itself, in the presence of antiseptics and in a temperature range between 18 and 28°C. This trait is stable, strain specific and positive selected in all commercial starters. The winemaker is confronted with the dilemma that, while ethyl alcohol is the major desired metabolic product of grape juice fermentation, it is also a potent chemical stress factor that is often the underlying cause of sluggish or stuck fermentations. Apart from the inhibitory effect of excessive sugar content on yeast growth and vinification fermentation, the production of excessive amounts of ethanol, coming from harvest of over-ripe grapes, is known to inhibit the uptake of solutes (e.g. sugars and amino acids) and to inhibit yeast growth rate, viability and fermentation capacity (27,28). Test of ethanol production in synthetic wine must with 300 g/l of glucose added is commonly performed on commercial strains together with ethanol stress resistance. Ethanol is highly toxic to yeast metabolism and growth and the cell membrane is the primary target for its action. A number of molecular pathways have evolved which ensure that the yeast cell can implement a response to these injuries, and the molecular and physiological response of an organism to changes in the environment is referred to as “stress response” The regulation of the stress response includes sensor systems and signal transduction pathways which result in the activation of the so-called stress response gene Hsp12 protects membranes against desiccation and ethanol-induced stress (29). Sulphur dioxide (SO₂) is the most widely used and controversial additive in organic winemaking. Sulphites are naturally produced by the yeasts during the wine processing, but the addition of SO₂ is traditionally considered as an efficient method to protect and preserve the wine at different stages of its elaboration. Sulphitation is allowed by all the standards for organic wine processing, but with restrictions compared to the wine regulation.

It improved fermentation processes by inhibiting the growth of undesirable bacteria and yeasts, furthermore it inactivates certain enzymes during the wine making process (30). In fact, it is used to control the microflora of a fermentation while *Saccharomyces* in general are quite resistant to it. Susceptibility to sulphite varies widely. The resulting differences in yeast population would be expected to yield wines with different flavour characteristics. Membrane transport of sulphite in wine yeasts is by simple diffusion of liberated sulphur dioxide rather than being carrier mediated (15).

SO₂ dissociates within the cell to SO₃²⁻ and HSO₃⁻ and the resulting decline in intracellular pH forms the basis of the inhibitory action. Although *S. cerevisiae* tolerates much higher levels of sulphite than most unwanted yeasts and bacteria, excessive SO₂ dosages may cause sluggish or stuck fermentations. Wine yeasts vary widely in their tolerance to sulphite, and the underlying mechanism of tolerance as well as the genetic basis for resistance differs between strains and is not completely clarified. Once these have been better defined, it may be advantageous to engineer wine yeast starter strains with elevated SO₂ tolerance. This, however, should not replace efforts to lower the levels of chemical preservatives in wine.

FROM GENOTYPE TO PHENOTYPE

The correlation between different phenotypes with importance in enology and specific molecular patterns would simplify the characterization of the indigenous yeast populations in wine yeast selection programs. Recently, a close correlation between molecular polymorphism and specific phenotypic traits was reported in non-*Saccharomyces* wild yeast strains (31). However, the results obtained from genotype-phenotype relationships studies in wild wine *S. cerevisiae* populations are controversial (32,33). In these studies, the degree of correlation was estimated taking into account the total number of isolates as a whole. In these studies, the degree of correlation was estimated taking into account the total number of isolates as a whole. In other works, when this statistical method is applied very low correlation coefficients are obtained. The use of more powerful statistical tools as the Generalized Procrustes Analysis (GPA) for the simultaneous analysis of molecular and physiological traits (34) allow to weigh the relationships for each isolate in particular, denoting a better degree of agreement between molecular and physiological data for most of the population analysed. Application of the GPA in studies on the genetic and/or phenotypic variability in the microbiological field evidence the possibility to quantify the relationship between molecular and phenotypic characteristics in wine yeasts (35).

The NCBI Genome Project Database reports 46 genome sequencing projects on different strains of *S. cerevisiae*. Only the genome of *S. cerevisiae* S288c is completed, among the other projects, 27 genomes are assembled with coverage depths varying from 2.6 to 20x and 18 are still in progress. The sequenced strains include lab, pathogenic, baking, wine, natural fermentation, sake, probiotic and plant isolates. Most of the sequencing projects led to the comparison of the genomes of different strains to correlate genomic traits to specific phenotypes and to infer phylogenetic relationships and evolutionary histories. Analysis of closely related strains have been performed too, for example genome of six commercial strains of *S. cerevisiae* used in wine fermentation and brewing were compared to find characteristics typical of these industrial classes of yeast (7).

Regularly updated information concerning the genomic and functional analysis of yeasts is available on a number of extensive databases. These include the Génolevures project web site (36), the Stanford Genome Database (SGD), the Munich information Center for Protein Sequences Comprehensive yeast Genome database (MIPS CYGD) and the Yeast Proteome Database (YPD).

Furthermore genome-wide transcriptional profiling has important applications in evolutionary biology for assaying the extent of heterozygosity for alleles showing quantitative variation in gene expression in natural populations. These studies have, in turn, stimulated renewed interest in the interactions among metabolic pathways and the control of metabolic flux. Most experiments thus far have dealt with comparisons of patterns of gene expression of organisms with the same genotype grown under different conditions or at different stages of the cell cycle. Genetic variability of wine yeasts has been demonstrated using various analysis tools at the molecular level (37). The aCGH analysis has established that major differences between laboratory strains of *S. cerevisiae* are found in subtelomeric regions (38) and that *S. cerevisiae* wine strains show a gene copy number variation that differentiate them from laboratory strains and strains of clinical origin. Differences were found in genes related to the fermentative process such as membrane transporters, ethanol metabolism and metal resistance (39,40). With the objective of studying genomic and phenotypic changes between similar yeasts isolated from different origins, several genomic and phenotypic comparison of strains has been carried out. Various kinetic and fermentative parameters were evaluated and significant phenotypic differences were detected between strains, some of which may be explained by differences at the genomic level.

NEXT GENERATION SEQUENCING TECHNOLOGY

In the last decade the incredible development of high-throughput and low-cost sequencing platforms have allowed to increase rapidly the number of sequenced genomes and stimulates the creation of new protocols to use these technologies to study other aspects of the cell, such as transcriptional profiles, chromatin structures, non-coding RNAs. In fact, Next Generation Sequencing (NGS) technologies have a great impact both at economical and at research level, with increasing of data production and cost reduction. This new kind of techniques allow the sequencing of thousands of genomes from humans to microbes and they open entirely new areas of biological inquiry, including the investigation of ancient genomes, of human disease, the characterization of ecological diversity, and the identification of unknown etiological agents. The application field could be divided into three main arguments: genomic tasks (genome assembly, SNPs and structural variations), transcriptome analysis (gene prediction and annotation, alternative splicing discovering) and epigenetic problems.

Three commercial platforms are currently well established on the market, the Roche 454 Genome Sequencer, the Illumina Genome Analyzer, and the Life Technologies SOLiD System, but other technologies are also available or under development. All these high-throughput sequencing systems use new sequencing chemistries replacing Sanger's technology and do not require electrophoresis and individual amplification of the templates. They are based on the parallelization of the sequencing process to produce thousands of sequences at once and lower costs and time required for DNA sequencing (41).

Before the coming of these technologies, big consortiums of laboratories were required to sequence just one genome. Today, on the contrary, also small labs can cope with sequencing projects. Thanks to these powerful technologies it is now possible to sequence lots of genomes and get several information by the comparison of them. As said, the sequencing of yeast strains used in winemaking, can be a powerful approach to identify the still unknown genes involved in fermentation and development of typical aroma. Moreover the transcriptional profile (complete set of transcripts in a cell for a specific physiological condition) of a strain, can be used to identify the differentially expressed genes with respect to other strains and to see how differences in the genome are mirrored by gene expression, and more generally by the phenotype.

Phylogenetic Relationship

During its long history of association with human activity, the genomic makeup of the yeast *S. cerevisiae* is thought to have been shaped through the action of multiple independent rounds of wild yeast domestication combined with thousands of generations of artificial selection. As the evolutionary constraints that were applied to the *S. cerevisiae* genome during these domestication events were ultimately dependent on the desired function of the yeast (e.g baking, brewing, wine or bioethanol production), this multitude of selective schemes have produced large numbers of *S. cerevisiae* strains, with highly specialized phenotypes that suit specific applications (42,43). As a result, the study of industrial strains of *S. cerevisiae* provides an excellent model of how reproductive isolation and divergent selective pressures can shape the genomic content of a species. There have been several attempts to characterize the genomes of industrial strains of *S. cerevisiae* which have uncovered differences that included single nucleotide polymorphisms (SNPs), strain-specific ORFs and localized variations in genomic copy number. However, the type and scope of genomic variation documented by these studies were limited either by technology constraints (e.g CGH arrays relying on the laboratory strain as a “reference” genome), or by the resources required for the production of high quality genomic assemblies which has limited the scope and number of whole-genome sequences available for comparison. In addition, to limit genomic complexity to a manageable level, previously published whole-genome sequencing studies on industrial strains used haploid representations of diploid, and often heterozygous, commercial and environmental strains (3,44-46).

TRANSCRIPTIONAL PROFILE

The phenotype of each organism is defined by a combination of its gene content and gene regulation. Variability in gene expression results from adaptive evolution of regulatory sequence and reflects changes in genomic sequences that influence the expression (47). Genome-wide transcriptional analysis has been employed to investigate yeast responses to a variety of stresses that arise during fermentation, such as glucose and ammonia limitations, salt stress, nitrogen concentration, unfolded protein stress, alterations in growth temperature, ethanol exposure, and hypoxia (48-52). Most of these studies have focused on the response of yeast to one or several specific stresses; however, the impact of the combination of these stresses, such as occurs during industrial fermentation, is likely to be far more complex.

Microarray analysis of brewer's yeast subjected to batch fermentation in a 3-l bioreactor has been carried out, and Varela *et al.* quantified gene expression profiles of industrial yeast under winemaking conditions in a 50-l bioreactor (53). Finally genome-wide expression analysis has been used to study responses of yeast to stresses that occur during wine fermentation with a 1-l working volume. It was in fact previously described a genome-wide transcriptional response of lager yeast during full scale batch brewery fermentation (55,56), but no transcriptional data obtained for multiple strain growth in parallel in synthetic wine must have already been reported.

Regulatory Elements

Genomic elements that mostly control it are the promoter regions, so they represent ideal candidates for driving gene expression divergence. Eukaryotic promoters are structures difficult to characterize because they can have regulatory elements lying quite far from the transcription start site. In general, the nucleus of the promoter includes the transcription start site plus an additional sequence that can be a TATA box upstream, or a downstream promoter element. This region is bound by the basal transcription apparatus, but the efficiency and the specificity of binding depend on the presence of transcription factor binding sites and possibly on chromatin accessibility. In 2004 Harbison *et al.* constructed a map of yeast's transcription regulatory code by identifying the sequence elements bound by regulators (54). 3337 regions along the genome of *S. cerevisiae* have been annotated. These elements were identified merging information from genome-wide location data, phylogenetically conserved sequences, and prior knowledge. A genome wide analysis on how differences in both TF binding sites and tandem repeats variation affect gene expression have not been performed yet.

PROJECT OUTLINE

The aim of this research was the selection and characterization of four representatives from a wine yeast collection of *S. cerevisiae* strains isolated in Veneto vineyards to be used as starter of fermentation in the production of Prosecco di Valdobbiadene DOCG and DOC Piave wines. These strains, together with S288c and EC1118 as controls, were used to correlate genome structure and transcriptional profile with metabolites production in synthetic wine must. PFGEs have been performed both on natural isolates and their homozygous lines derived from ascii dissection to detect large genomic differences. Oenological properties have been tested and it was assessed that the main phenotypic characters of natural isolates were maintained in derivative lines. A mixed approach of paired-end and shotgun 454-FLX sequencing was applied to obtain high-quality assemblies of genomes.

By genome comparisons we were able to highlight genetic differences among strains including the presence of genomic rearrangements and the variability on Ty elements distribution and frequency. To explain the genetic basis of oenological traits and to understand among the high number of polymorphic sequences those involved in wine adaptation, we also performed transcription profiling with RNA-seq using SOLiD technology.

RNA-seq of the four ecotypical strains plus S288c and EC1118 was performed on RNA extracted during fermentation process under winemaking conditions in controlled bioreactors, collecting samples growth in synthetic wine media. The molecular adaptation and metabolites production of wine yeasts in presence of high sugar content, low pH, and high ethanol concentration during mid-exponential and early-stationary phases was investigated. RNA-seq was also used to facilitate gene annotation, to evaluate splicing sites, to identify hundreds of new non-protein coding transcripts localized in intergenic regions and antisense transcripts overlapping protein coding genes.

REFERENCES

- (1) Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. *Science* 1996 Oct 25;274(5287):546, 563-7.
- (2) Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, et al. Genome evolution in yeasts. *Nature* 2004 Jul 1;430(6995):35-44.
- (3) Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. *Nature* 2009 Mar 19;458(7236):337-341.
- (4) Suter B, Auerbach D, Stagljar I. Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *BioTechniques* 2006 May;40(5):625-644.
- (5) Kumar A, Snyder M. Emerging technologies in yeast genomics. *Nat Rev Genet* 2001 Apr;2(4):302-312.
- (6) Winde JHd. Functional genetics of industrial yeasts. Berlin: Springer; 2003.
- (7) Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, et al. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. *PLoS Genet* 2011 Feb 3;7(2):e1001287.
- (8) Swiegers JH, Kievit RL, Siebert T, Lattey KA, Bramley BR, Francis IL, et al. The influence of yeast on the aroma of Sauvignon Blanc wine. *Food Microbiol* 2009 Apr;26(2):204-211.
- (9) Pretorius IS. Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. *Yeast* 2000 Jun 15;16(8):675-729.
- (10) Rementeria A, Rodriguez JA, Cadaval A, Amenabar R, Muguruza JR, Hernando FL, et al. Yeast associated with spontaneous fermentations of white wines from the "Txakoli de Bizkaia" region (Basque Country, North Spain). *Int J Food Microbiol* 2003 Sep 1;86(1-2):201-207.
- (11) Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, et al. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci U S A* 2009 Sep 22;106(38):16333-16338.
- (12) Nisiotou AA, Spiropoulos AE, Nychas GJ. Yeast community structures and dynamics in healthy and Botrytis-affected grape must fermentations. *Appl Environ Microbiol* 2007 Nov;73(21):6705-6713.
- (13) Barnett JA. A quick procedure for anaerobic fermentation tests in the identification of yeasts. *Arch Mikrobiol* 1972;84(3):266-269.
- (14) Fleet GH. Wine microbiology and biotechnology. London: Taylor & Francis; 2002.
- (15) Walker GM. Yeast physiology and biotechnology. Chichester, West Sussex: Wiley; 1998.

- (16) Fugelsang KC. Wine microbiology. New York, N.Y.: Chapman and Hall; 1997.
- (17) Beltran G, Torija MJ, Novo M, Ferrer N, Poblet M, Guillamon JM, et al. Analysis of yeast populations during alcoholic fermentation: a six year follow-up study. *Syst Appl Microbiol* 2002 Aug;25(2):287-293.
- (18) Torija MJ, Rozes N, Poblet M, Guillamon JM, Mas A. Yeast population dynamics in spontaneous fermentations: comparison between two different wine-producing areas over a period of three years. *Antonie Van Leeuwenhoek* 2001 Sep;79(3-4):345-352.
- (19) Combina M, Elia A, Mercado L, Catania C, Ganga A, Martinez C. Dynamics of indigenous yeast populations during spontaneous fermentation of wines from Mendoza, Argentina. *Int J Food Microbiol* 2005 Apr 1;99(3):237-243.
- (20) Mortimer R, Polsinelli M. On the origins of wine yeast. *Res Microbiol* 1999 Apr;150(3):199-204.
- (21) Prakitchaiwattana CJ, Fleet GH, Heard GM. Application and evaluation of denaturing gradient gel electrophoresis to analyse the yeast ecology of wine grapes. *FEMS Yeast Res* 2004 Sep;4(8):865-877.
- (22) Comitini F, Ciani M. Survival of inoculated *Saccharomyces cerevisiae* strain on wine grapes during two vintages. *Lett Appl Microbiol* 2006 Mar;42(3):248-253.
- (23) Romano P, Fiore C, Paraggio M, Caruso M, Capece A. Function of yeast species and strains in wine flavour. *Int J Food Microbiol* 2003 Sep 1;86(1-2):169-180.
- (24) Boulton RB. Principles and practices of winemaking. New York, N.Y.: Chapman and Hall; 1996.
- (25) Schaaff I, Heinisch J, Zimmermann FK. Overproduction of glycolytic enzymes in yeast. *Yeast* 1989 Jul-Aug;5(4):285-290.
- (26) Reifemberger E, Boles E, Ciriacy M. Kinetic characterization of individual hexose transporters of *Saccharomyces cerevisiae* and their relation to the triggering mechanisms of glucose repression. *Eur J Biochem* 1997 Apr 15;245(2):324-333.
- (27) Alexandre H, Heintz D, Chassagne D, Guilloux-Benatier M, Charpentier C, Feuillat M. Protease A activity and nitrogen fractions released during alcoholic fermentation and autolysis in enological conditions. *J Ind Microbiol Biotechnol* 2001 Apr;26(4):235-240.
- (28) Piper PW. The heat shock and ethanol stress responses of yeast exhibit extensive similarity and functional overlap. *FEMS Microbiol Lett* 1995 Dec 15;134(2-3):121-127.
- (29) Querol A, Fernandez-Espinar MT, del Olmo M, Barrio E. Adaptive evolution of wine yeast. *Int J Food Microbiol* 2003 Sep 1;86(1-2):3-10.
- (30) Romano P, Suzzi G. Acetoin production in *Saccharomyces cerevisiae* wine yeasts. *FEMS Microbiol Lett* 1993 Mar 15;108(1):23-26.

- (31) Rodriguez ME, Lopes CA, van Broock M, Valles S, Ramon D, Caballero AC. Screening and typing of Patagonian wine yeasts for glycosidase activities. *J Appl Microbiol* 2004;96(1):84-95.
- (32) Nadal D, Colomer B, Pina B. Molecular polymorphism distribution in phenotypically distinct populations of wine yeast strains. *Appl Environ Microbiol* 1996 Jun;62(6):1944-1950.
- (33) Comi G, Maifreni M, Manzano M, Lagazio C, Cocolin L. Mitochondrial DNA restriction enzyme analysis and evaluation of the enological characteristics of *Saccharomyces cerevisiae* strains isolated from grapes of the wine-producing area of Collio (Italy). *Int J Food Microbiol* 2000 Jun 30;58(1-2):117-121.
- (34) Gower JC. Generalized Procrustes Analysis. *Psychometrika* 1975;40:33-51.
- (35) Lopes CA, Rodriguez ME, Querol A, Bramardi S, Caballero AC. Relationship between molecular and enological features of Patagonian wine yeasts: relevance in selection protocols. *World Journal of Microbiology & Biotechnology* 2006;22:827-833.
- (36) Souciet JL, Genolevures Consortium GDR CNRS 2354. Ten years of the Genolevures Consortium: a brief history. *C R Biol* 2011 Aug-Sep;334(8-9):580-584.
- (37) Schuller D, Valero E, Dequin S, Casal M. Survey of molecular methods for the typing of wine yeast strains. *FEMS Microbiol Lett* 2004 Feb 9;231(1):19-26.
- (38) Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, et al. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* 2003 Jan;163(1):79-89.
- (39) Dunn B, Levine RP, Sherlock G. Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC Genomics* 2005 Apr 16;6:53.
- (40) Carreto L, Eiriz MF, Gomes AC, Pereira PM, Schuller D, Santos MA. Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics* 2008 Nov 4;9:524.
- (41) Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J. The next-generation sequencing technology and application. *Protein Cell* 2010 Jun;1(6):520-536.
- (42) Querol A, Belloch C, Fernandez-Espinar MT, Barrio E. Molecular evolution in yeast of biotechnological interest. *Int Microbiol* 2003 Sep;6(3):201-205.
- (43) Fay JC, Benavides JA. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet* 2005 Jul;1(1):66-71.
- (44) Borneman AR, Forgan AH, Pretorius IS, Chambers PJ. Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. *FEMS Yeast Res* 2008 Nov;8(7):1185-1195.

- (45) Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, et al. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet* 2008 Aug 29;4(8):e1000183.
- (46) Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OV, Missawa SK, et al. Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. *Genome Res* 2009 Dec;19(12):2258-2270.
- (47) Tirosh I, Weinberger A, Bezael D, Kaganovich M, Barkai N. On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol* 2008;4:159.
- (48) Kolkman A, Daran-Lapujade P, Fullaondo A, Olsthoorn MM, Pronk JT, Slijper M, et al. Proteome analysis of yeast response to various nutrient limitations. *Mol Syst Biol* 2006;2:2006.0026.
- (49) Melamed D, Pnueli L, Arava Y. Yeast translational response to high salinity: global analysis reveals regulation at multiple levels. *RNA* 2008 Jul;14(7):1337-1351.
- (50) Mendes-Ferreira A, del Olmo M, Garcia-Martinez J, Jimenez-Marti E, Mendes-Faia A, Perez-Ortin JE, et al. Transcriptional response of *Saccharomyces cerevisiae* to different nitrogen concentrations during alcoholic fermentation. *Appl Environ Microbiol* 2007 May;73(9):3049-3060.
- (51) Payne T, Hanfrey C, Bishop AL, Michael AJ, Avery SV, Archer DB. Transcript-specific translational regulation in the unfolded protein response of *Saccharomyces cerevisiae*. *FEBS Lett* 2008 Feb 20;582(4):503-509.
- (52) Pizarro FJ, Jewett MC, Nielsen J, Agosin E. Growth temperature exerts differential physiological and transcriptional responses in laboratory and wine strains of *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 2008 Oct;74(20):6358-6368.
- (53) Varela C, Cardenas J, Melo F, Agosin E. Quantitative analysis of wine yeast gene expression profiles under winemaking conditions. *Yeast* 2005 Apr 15;22(5):369-383.
- (54) Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004 Sep 2;431(7004):99-104.
- (55) Marks VD, Ho Sui SJ, Erasmus D, van der Merwe GK, Brumm J, Wasserman WW, et al. Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response. *FEMS Yeast Res* 2008 Feb;8(1):35-52.
- (56) Rossignol T, Dulau L, Julien A, Blondin B. Genome-wide monitoring of wine yeast gene expression during alcoholic fermentation. *Yeast* 2003 Dec;20(16):1369-1385.

2. STRAIN SELECTION

INTRODUCTION

The origins of non-Saccharomyces are grape skin and winery equipment (1). However, the origin of *S. cerevisiae* is the subject of some debate; the most significant finding was that *S. cerevisiae* is practically absent from grapes and vineyard soils (2). In contrast, some authors propose that this species is a “natural” organism present in plant fruits (3,4). Finally, other authors postulate that *S. cerevisiae* is a domesticated species originating from its closest relative *S. paradoxus*, a wild species found all around the world associated with insects, tree exudates and fermenting plant extracts (5). The occurrence of *S. cerevisiae* in vineyards would then be the consequence of back transportation from the cellars to the vineyards by insects (5). Although the origin of *S. cerevisiae* is a matter of controversy, its original genome has been subjected to strong selective pressures since its first unintended use in controlled fermentation processes, and this phenomenon could be related with the origin of the Saccharomyces in wine fermentation and the adaptation to this special environment. Intensive research has focused on elucidating the molecular mechanisms involved in stress response, as the genomic characteristics of the industrial wine yeast which have been selected over billions of generations.

Qualitative Trait and Aromas

Originally, all wine was made by taking advantage of natural microflora for spontaneous fermentation, no deliberate inoculation was made to start the process. Various yeasts found on the surface of grape skins and the indigenous microbiota associated with winery surfaces participate in these natural wine fermentations. Today, several companies producer of yeasts serving wine industries market a wide variety of dehydrated cultures of various *S. cerevisiae* strains. In guided fermentations, the actively growing starter culture dominates the native yeast species present in grape must. It is clear that the genetic and physiological characteristics of the wine yeast strain have a significant effect on the amount of volatile thiols released. It was shown that the VL3 yeast strain released more volatile thiols than strains VL1 and 522d (isolated from vineyards in France). Furthermore, *S. bayanus* strains appeared to release more 4MMP than the VL3 strain (6). Yeast strains do affect wine aroma and could influence the preference for particular wines. The fermentation product profiles of wines made with different yeast strains varied widely depending on the yeast strain used for fermentation. Chemical analyses of the wines indicated that the flavour compounds present in these wines made with different yeasts were significantly different and unique for each strain. Furthermore, some yeast strains could be very efficient fermenters while not necessarily producing the best flavour profiles. Other yeast strains might show desirable aroma-enhancing capabilities but might have tendencies to produce volatile acidity (7). In large-scale wine production, however, where rapid and reliable fermentations are essential for consistent wine flavour and predictable quality, the use of selected pure yeast inocula of known ability is preferred. These large wineries will be the main beneficiaries of programs aimed at selecting yeast strains with even more reliable performance, reducing processing inputs, and facilitating the production of affordable high-quality wines.

Alcoholic beverages contain mainly saturated, straight chain fatty acids. The volatile acid content of wine usually lies between 400 and 1000 mg/l, normally more than 90% of volatile acid consists of acetic acid (8). Although acetic and lactic acid bacteria can be associated with high levels of short chain fatty acid, acetic, propanoic and butanoic acids are by-products of alcoholic fermentation (9). Fermentation purity is expressed as of the ratio between volatile acidity (as g acetic acid/l) and ethanol (% volume) produced at the end of the fermentation process. High values of this ratio denote the ability to form few undesirable by-products in the course of fermentation. Wines cannot be commercialized if volatile acidity exceeds one tenth of the ethanol content. Another fermentation by-product affecting wine quality is glycerol.

In a model fermentation, about 95% of the sugar is converted into ethanol and carbon dioxide, 1% into cellular material and 4% into other products such as glycerol. Due to its non volatile nature, glycerol has no direct impact on the aromatic characteristics of wine. However, this triol imparts certain other sensory qualities; it has a slightly sweet taste, and owing to its viscous nature, also contributes to the smoothness, consistency and overall body of wine. Wine yeast strains producing a consistent amount of glycerol would therefore be of considerable value in improving the organoleptic quality of wine (10).

Oenological Yeasts Collection

Some critics of the practice of guided fermentations (using starter cultures) dislike the fact that the commercial wine strains, despite being numerous, possess very ordinary characteristics. Commercial yeast strains produce wines with average qualities and do not enhance the aromatic traits that characterise many yeasts isolated from specific geographical areas. Studies on the improvement and the selection of wine yeasts to overcome this problem have recently been carried out. In the last few years, there has been an increasing use of new local selected yeasts for controlled must fermentation in countries with a wine-making tradition. Though there are commercial yeasts to accomplish must fermentation, the use of local selected yeasts is believed to be much more effective (8,11). Local yeasts are presumed to be more competitive because they are better acclimated to the environmental conditions. Therefore, they would be better able to dominate the fermentation and become the most important biological agent responsible for the winemaking. Selection of the appropriate local yeasts assures the maintenance of the typical sensory properties of the wines produced in any given region (12). During the last years the microbiology research group of Prof. V. Corichi in the Department of Agricultural Biotechnology of University of Padua isolated approximately 600 yeast strains collected in the vineyards of the “Prosecco di Conegliano-Valdobbiadene” VQPRD District and of the “Raboso DOC Piave” District in Venetian region (Fig. 2.1).



Figure 2.1 Sampling areas of “Prosecco di Conegliano-Valdobbiadene” (yellow) and “Raboso DOC Piave” (red) vineyards in Veneto region.

Combination of molecular genetic analysis (microsatellites, PCR-RFLP of MET2, the ITS1-ITS2 region and the NTS region) and physiological examination (SO₂ resistance, ethanol production and tolerance, killer activity, fermentation vigour and production of metabolites) of yeasts isolated from spontaneously fermenting wines in two wine regions revealed very high diversity in the *S. cerevisiae* population. Selection process included sampling in soil, vineyard, grapes, must and cellar walls to be sure to collect the highest number of strains.

Yeast Improvement Strategy

Traditionally the genetic manipulation strategies of wine yeasts to produce better new strains exploits different strategies, which included the selection of natural and induced mutants by sexual recombination methods (13). Hybridisation of laboratory heterothallic strains was the first method used for yeast improvement. The wild strains are mostly homothallic and heterozygous [14]; for this reason conjugation by micromanipulator or mixing sporulated cultures is possible among germinating spores before autodiploidization. The sexual recombination can be performed with gametes obtained by single-spore cultures or with spores obtained directly from parental strains. The recombination among a small number of parental strains allows to collect a complex progeny, which is then submitted to Chapter 1 26 selective processes. This method is based on random events, and it is very similar to the new combinatorial approaches that were used for the determination of the optimal genetic configuration in industrial microbes (14). To rationalize the latter strategy, the first requirement is to try to establish the importance of the genetic determinism of the oenological parameters of yeast. Specifically, crosses and progeny analysis could theoretically be used to improve genotypes, thereby accumulating general and specific properties in a strain. The availability of relevant and reliable phenotypic tests to screen a large population of yeast strains in laboratory conditions is the prerequisite condition to appreciate the contribution of genetics in different characters (15). In particular, hybridization can be carried out to support different methods depending on yeast strains characteristics. Intra-species hybridization (mating) involves the mating of haploids of opposite mating-types to yield a heterozygous diploid.

Recombinant progeny are recovered by sporulating the diploid, collecting individual haploid ascospores and repeating the mating/sporulation cycle as required. Thus, in theory, crossbreeding can permit the selection of desirable characteristics and the elimination of undesirable ones. Elimination or inclusion of a specific property could thus be achieved relatively quickly by hybridization, when the trait has simple genetic basis, for example it is coded by one or two genes (16). Unfortunately, many desirable wine yeast characteristics are determined by several genes or are the result of numerous controlling system interacting each other. Wine yeast strains that fail to express a mating-type can be forced to mate (rare-mating) with haploid MATa and MATa strains.

MATERIALS AND METHODS

Common media and growth conditions are listed together with list of abbreviation and standard solution in the Appendix I section. Strains were routinely grown on YPD medium at 28 °C for 12 to 24h under agitation.

Sporulation and Tetrad Dissection

Yeasts were inoculated into 50 ml tubes containing 10 ml of liquid YPD and incubated with rotator shaking in a New Brunswick incubator at 30°C until the stationary phase was reached (about 10^8 cells per ml). Presporulation medium tubes were inoculated with a stationary-phase culture grown in YPD to reach an initial A660 of 0.05. After inoculation, the tubes were incubated at 30°C with shaking until either middle exponential phase (about 1×10^7 to 5×10^7 cells per ml) or stationary phase (about 5×10^8 cells per ml). The cells were then centrifuged, washed twice with distilled water, inoculated into liquid PRE5 medium and incubated at 30°C with shaking for at least 24 h. Then, the cells were newly centrifuged, washed twice with distilled water, transferred to solid SPO2 medium and incubated at 30°C for at least 4 days. The percent of asci formed, as well as the number of ascospores per ascus, was determined by counts of cells under the optical microscope Olympus BX60 (17).

Sporulated cultures usually consist of unsporulated vegetative cells, four-spored asci, three-spored asci, etc. Dissection of asci requires the identification of four-spored asci and the relocation of each of the four ascospores to separate positions where they will form isolated spore colonies. The procedure requires the digestion of the ascus wall with Zymolyase, without dissociating the four spores from the ascus. Sporulated cells from sporulation medium are harvested and then suspended in 50 ml of a stock solution of Zymolyase T100 (50 mg/ml in 1 M sorbitol), and the suspension is incubated at 30°C for approximately 10 minutes. The exact time of incubation is strain dependent and the progress of the digestion can be followed by removing a sample of the digest to a glass slide and examining it under phase contrast at 100x magnification. The sample is ready for dissection when the spores in most of the asci are visible as discrete spheres, arranged in a diamond shape.

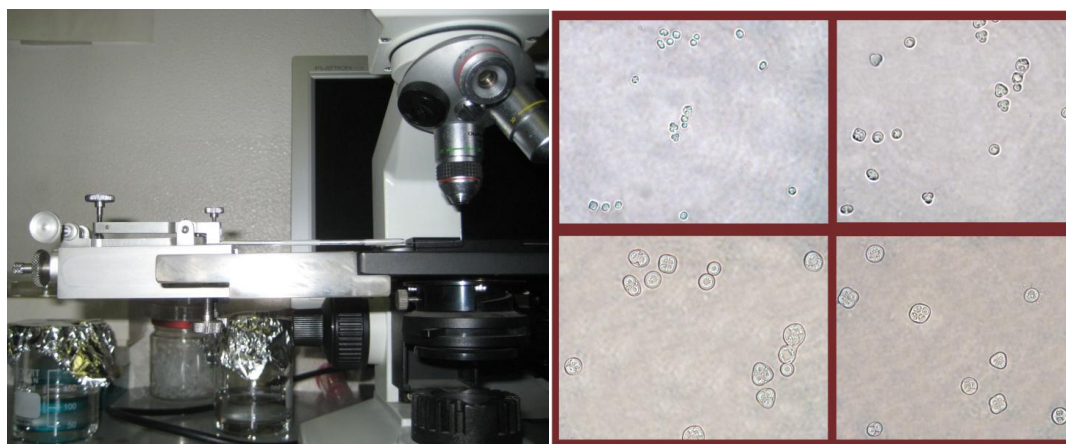


Figure 2.2 Micromanipulator and typical digested asci at 40x and 100x magnification.

The culture is suspended by gently rotating the tube; an aliquot is transferred with a wire loop to the surface of a petri plate or agar slab. It is important not to agitate the spores once they have been treated.

If the treated spores are vortexed or shaken, the integrity of the ascus cannot be assured since the contents of one ascus may disperse and reassemble with the contents of another. Micromanipulation can be implemented directly on the surfaces of ordinary petri dishes filled with nutrient medium or in special chambers on thin agar slabs. A cluster of four spores is picked up on the microneedle by positioning microneedle tip next to the four-spored cluster on the surface of the agar.

Once the four spores have been transferred to the first position, it is necessary to separate at least one spore from the rest so that it can be left behind. After picking up the four spores from an ascus, it is often convenient to set the stage micrometer so that each group of four spore colonies falls on cardinal points such as 15, 20, 25, etc. This makes it easier to keep track of progress and prevents the spore colonies from growing too close together. Likewise, positions on the y axis can be marked on the stage micrometer so that the four spore colonies from each ascus are evenly spaced.

Pulsed Field Gel Electrophoreses

Protoplasts generation and PFGE run condition were previously described by Vaughan-Martini et al. (18). Cells were grown to middle exponential phase (about 1×10^7 cells per ml), collected by centrifugation at 8000 rpm for 5 minutes at 4°C. The cells were then washed twice with cold distilled water and EDTA 50mM pH8.0 and then gently resuspended in 120 µl of fresh protoplast forming medium SPG with 25 mg/ml of Zymolyase. After 2h incubation at 30°C with shaking cells were transferred at 37°C for 10 min. Equal volume of low melting point agarose solution (10mM Tris-HCl pH 7.5, 0.125M EDTA, 2% Low Melting Point Agarose) kept at 50°C was added to cellular suspension. Mixture was immediately poured into plug molds (disposable plug mold, Biorad) and left solidifying at 4°C for 20 min. Formed plugs were immersed in solubilisation buffer LET and incubated 3h with shaking at 30°C. After a washing step with cold 50mM EDTA pH 8.0 plugs were left O/N at 50°C and 400rpm in 600 µl of NDS buffer with 2mg/ml of Proteinase K. Plugs were finally rinsed several times in 1ml cold 50mM EDTA pH 8.0 for all the day long and stored indefinitely at 4°C in 500mM EDTA pH 9.0.

Electrophoresis Parameters

A 120ml gel 1.2% agarose for Pulsed field (SIGMA) was prepared with 0.5X TBE buffer, the corresponding running buffer was 0.5X TBE kept 9°C constant by PFGE chiller. The running program was 5.1 V/cm voltage gradient, 34 h run time with 60 sec initial switch and 120 sec final switch. Reference ladders with DNA size standards routinely used for pulsed field runs were commercial chromosomal preparations from *S.cerevisiae* purchased from Biorad. Post running staining was done in 0.5X TBE with EtBr in standard concentration for 30 min and rinsed O/N at 4°C in TBE before image capturing.

Fermentation Ability and Ethanol Resistance

Fermentation ability was tested in MNS media (19) in small-scale winemaking trials using 100 ml bottles. Yeasts were inoculated into 50-ml tubes containing 12 ml of liquid MNS and incubated with rotatory shaking in a New Brunswick incubator at 30°C until the stationary phase was reached (about 10^8 cells per ml).

An inoculum of 5 ml was added to 95 ml of MNS in the 100 ml bottles to reach an initial A_{600} of 0.05. Fermentations were performed under isothermal conditions 25°C and bottles were sealed with sterile rubber caps and clamped with aluminium rings to maintain anaerobiosis. Caps were then punctured with a needle to allow fermentation gases release. Glucose fermented was determined by the measurement of bottles weight loss every 24 h with a precision balance (Sartorius, BL210S) and the rate of CO_2 production was calculated using a polynomial smoothing.

Fermentation in Controlled Bioreactors

Yeast cultures were grown in 100 ml YPD medium at 25 °C in agitation for 18 hours. Each culture has been centrifuged and the pellet was resuspended into the volume of synthetic must MS300 required to obtain an OD_{600} of 0.5 of the 1:10 diluted solution (5×10^6 cells/ml). 100 ml of this preinoculum has been added to 900 ml of MS300, a synthetic medium that mimics the composition of a white wine must.

Fermentation was performed at 25°C in 1 l bioreactors (Multifors, Infors HT) constantly monitoring the temperature, the pH, and the CO_2 flux in a range of 1-20 ml/min (red-y mod. GSM-A95A-BN00).

The fermentations have been performed for each strain in three replicates and samples from each replica have been taken at specific time points during the fermentation. The first samples were taken at the beginning of the fermentation when the CO_2 produced was 6 g/l, second samples were taken at 45 g/l and the third at 80 g/l. Yeast cells were immediately centrifuged, washed with water and the pellet was immediately frozen by immersion in EtOH previously refrigerated at -80°C in order to maintain unaltered the transcriptional profile. All corresponding supernatants were conserved for chemical analysis.



Figure 2.3 Bioreactors used to perform yeast fermentations

Ethanol Resistance

Starting from a cellular concentration normalized at 5×10^6 cells/ml for all strain inocula, 5 serial dilutions 1:10 have been performed using a microtitre plate with 96 wells. The four higher dilutions for every strain have been spotted in YPD agar medium added with different ethanol quantities to reach respectively final 8%, 9%, 10% and 11% concentration. All inocula have been executed in three independent replicates. Petri dishes were incubated at 25°C and growth was registered after 24h, 48h and 120h.

Growth Curve

Preinocula has been prepared with standard method in YPD and MNS and incubated at 30°C for 12h. The absorbance was measured at OD_{595nm} with an automated system using the Beckman coulter DTX 880 multimode detector and with shaking on the Heidolph titramax 1000 device at 450 rpm. 20 µl or 50 µl of cultures were inoculated into 24 wells plate containing respectively 1.5 ml of liquid YPD and 3 ml of MNS incubated with a rotatory shaking of 3.0 mm orbiting radius at 30°C in a Heidolph inkubator 1000. Blank wells were used as negative control and for each sample four experimental replicates have been realized. Measurement of OD_{595nm} was registered every 20 min.

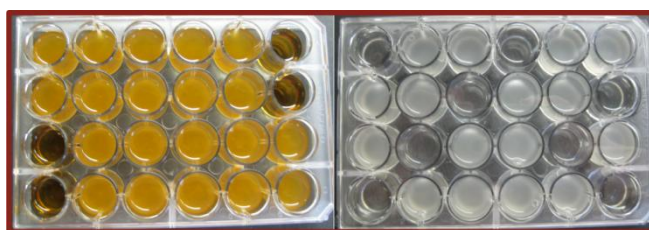


Figure 2.4 Example of 24 well plate with YPD and MNS liquid media at the end of growth curve.

Sulphite Stress Resistance

For the determination of resistance to sulphite compound has been used the method described above. It has been added to MNS medium different concentrations of sulphur dioxide, diluted from a stock solution with SO₂ concentration 10 g/l in 50 ml tubes (1 g of sodium metabisulphite NaHSO₃, SO₂ = 0.81g). 2 ml of MNS sulphated were aliquoted in the wells plates to reach respectively the final concentrations of 25 mg/l, 50 mg/l, 75 mg/l and 100 mg/l and inoculated with 50µl of cultures. To prevent evaporation of sulphur dioxide a double layer of parafilm has been applied on top of the wells under the plate cover and also sealed around the perimeter of the plate.

Compounds of Technological Interest

Ethanol Production

It is interesting to evaluate the maximum alcohol content that a yeast can produce in optimal conditions of development and in the presence of 300 g/l of sugar. For this test synthetic must have prepared modifying MNS media recipe, increasing glucose content (300 g/l), tartaric acid (to 6g/L), malic acid (6 g/l), hydrolyzed casein (1 g/l), ammonium sulphate and ammonium phosphate (both 0.9g/l). The medium was aliquoted into 100 ml flasks and pasteurized at 100°C for 5 minutes. The procedure and condition were previously described by Delfini (19). Yeasts were grown in 100 ml of YPD at 25 ° C for 12 h and inoculated to normalize the final OD for all strains and replicas. Then the flasks were incubated at a constant temperature of 25°C and glucose fermented was determined by the measurement of flasks weight loss every 12 h with a precision balance (Gibertini EU-7500DR C), with a sensitivity of 0.01g. The amount of ethanol produced at the end of fermentation was determined with HPLC by measuring the amount of residual sugar and using the conversion factor for sugar/alcohol of 0.61 (19).

Hydrogen Sulphide and Sulphur Dioxide

Selective solid media have been used to determine yeast production of sulphite compound. Natural strains and commercial controls have been incubated for 72 h at 25°C and then changing in colour have been evaluated. It has been used Biggy Agar for hydrogen sulphide production and Fucsina Agar for sulphur dioxide. The following table reports the chromatic scales used for result consideration:

Table 2.1 Chromatic scales used for sulphite compound production evaluation

Colour	White	Beige	Light Brown	Dark Brown
H ₂ S production	None	Low	Medium	High
Colour	Dark Pink	Pink	Light Pink	White
SO ₂ production	Low	Medium	High	Very High

Total and free sulfur dioxide were quantified at the end of synthetic must fermentation using iodometric titration.

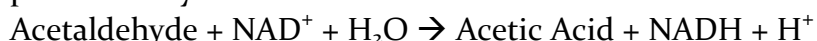
Chemical Analysis on Fermented Must

Samples of synthetic must fermented by the different strains were analyzed with HPLC technique to verify the exact amount of ethanol, glycerol, residual glucose, malic acid, succinic acid, citric acid and acetic acid. Components separation was carried out using a Waters 1525 binary HPLC pump with an Aminex ion exclusion column to HPX_87H 300 mm x 7.8 mm.

A Waters 2414 Refractive Index Detector was set at 600nm wavelength for the determination of ethanol, glycerol and glucose, while for the detection of the peaks related to organic acids we used a Waters 2487 Dual Absorbance detector set at 210nm wavelength. A calibration has been done for each individual compound and it was used to calculate the corresponding g/L in each sample.

Acetaldehyde

Acetaldehyde enzymatic determination was carried out using the kit R-BIOPHARM purchased by Roche. The chemical reaction used is:



The determination of acetaldehyde is controlled by measuring the amount of NADH produced at OD_{340nm}.

RESULTS AND DISCUSSION

Natural Isolates Selection

Genetic and physiological characteristics of the isolated strains were used to evaluate the presence of phenotypical traits interesting for winemaking and to select those strains that better represent the populations among the 600 isolates. Starting from the genetic and phenotypic characterization of strains a variety of statistical analysis has been performed (Principal Component Analysis, multivariate, ANOVA), both separately and together on isolates of Prosecco and Raboso. The aim was to assess the distribution of yeast populations in the different characteristics taken into account and obtain selection of strains representative for the technological properties of interest. The chart below shows the distribution of strains according to the PCA, which facilitates the choice allowing an immediate vision of yeasts distribution. Axes report the variability due to fermentation rate (glucose consumption, days of fermentation), rapid sedimentation, adhesion, production hydrogen sulphide and other technological features of interest. F1 indicates variability due to glucose degradation and fermentation rate. F2 axis shows variability due to clearness time, adhesiveness in Raboso graph and variability due to clearness time and H₂S production in Prosecco graph

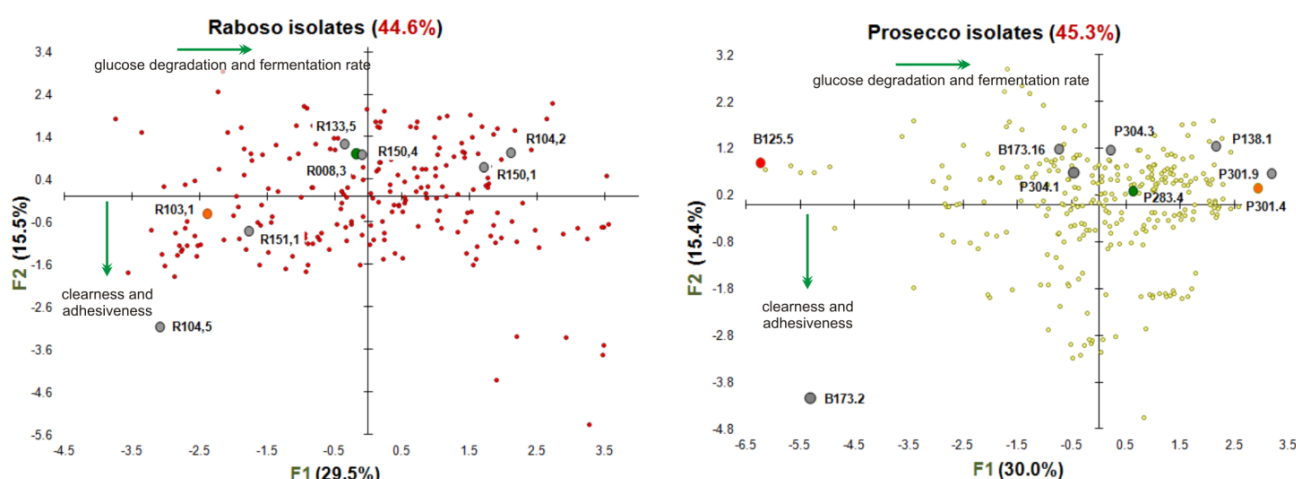


Figure 2.5 Principal Component Analysis performed on Raboso and Prosecco natural isolates to separate them in groups corresponding to their phenotypic characteristics.

Among all isolates only 17 strains (highlighted in the graph) have been chosen to deepen the analysis, together with a commercial and a laboratory strain as a comparison.

Strains Genetic Stability

Yeast is especially suited for meiotic mapping because the four spores in an ascus are the products of a single meiotic event, and the genetic analysis of these tetrads provides a sensitive means for determining linkage relationships of genes present in the heterozygous condition. The separation of the four ascospores from individual asci by micromanipulation is required for meiotic genetic analyses and for the construction of strains with specific markers.

The 17 strain selection were induced to sporulate and a minimum of 10 asci for each strain were dissected in order to evaluate spore viability and obtain single haploid spores. Result are reported in the following table.

Table 2.2 Chromatic scales used for sulphite compound production evaluation

mt DNA profile	Strain	Spore efficiency	Asci dissected	Spore Viability				
				4/4	3/4	2/4	1/4	0/4
2	P301.Z	high	10	80%	20%	-	-	-
28	P304.1	low	11	27%	-	46%	-	27%
18	P283.4	medium	12	92%	8%	-	-	-
20	P304.3	low	14	7%	57%	21%	14%	-
11	B173.16	high	13	46%	23%	23%	8%	-
101	R008.3	high	13	100%	-	-	-	-
230	R151.1	high	10	70%	30%	-	-	-
186	R150.4	medium	15	7%	53%	27%	7%	7%
155	R150.1	medium	10	80%	10%	-	10%	-
185	R133.5	medium	12	100%	-	-	-	-
189	R103.1	high	11	100%	-	-	-	-
142	R104.2	high	10	100%	-	-	-	-
139	R104.5	low	10	-	-	30%	50%	20%
10	B125.5	high	11	64%	18%	9%	-	9%
1	P301.4	high	13	85%	-	8%	8%	-
37	B173.2	low	10	100%	-	-	-	-
22	P138.1	low	10	-	10%	40%	30%	20%

Most of the examined strains exhibited a high sporulation efficiency producing asci with 4, 3 or 2 spores. A small percentage (<10%) of strains showed all four ascospores viable while a consistent fraction have 3 or less vital spores, with less than 10% presenting a strong reduction in spore vitality probably due to chromosomal aneuploidy.

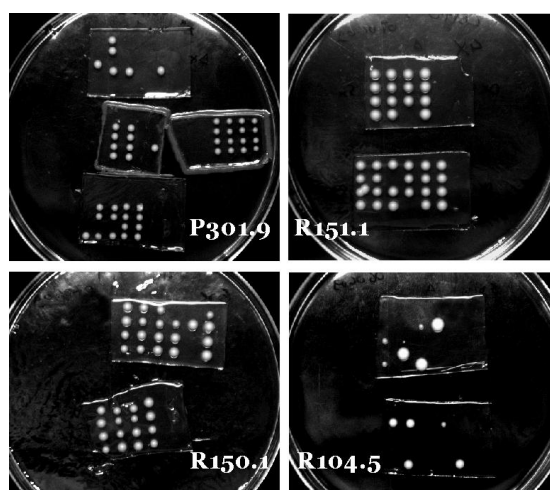


Figure 2.6 Spore colonies derived from asci separated on the surface of petri dishes.

Chromosomes Pattern

Genome stability and large genomic differences have been compared analyzing the chromosomes pattern produced. Several PFGE were performed both on natural isolates and on at least four of all homozygous lines derivated from diploids sporulation and spores autodiploidization. PFGE results revealed extensive genomic differences even between strains isolated in the same VQPRD District.

Analysis of their meiotic products (four ascospores) is important in order to identify strains having extensive chromosomal reorganization that occurs with a very high frequency during meiosis. Analysis of the derivatives homozygotes obtained from dissection of tetrads was used as a screening for genomic stability of strains, to detect errors in chromosome segregation and translocation of portions. Karyotypic differences and genetic stability are some of the fundamental criteria on which we based candidate strains selection.

Below is a portion of the dendrogram of similarity obtained from the comparison of all PFGE performed. It is representative of the relationship between heterozygous parental strains and the four homozygous derived from a single tetrad. It can be clearly seen in the fig. 2.7 that strain B125.5 has a high chromosomal stability with an almost perfect correspondence between parental and derivatives. It also present a karyotype profile quite different from that of the reference strain S288c. On the contrary, strain P138.1 is very unstable from a genetic perspective, it shows in fact enormous differences in the comparison between profiles, even among the four homozygotes. This strain had already a low viability in spore dissection, showing the existing strong relationship between chromosomal instability and poor spores viability.

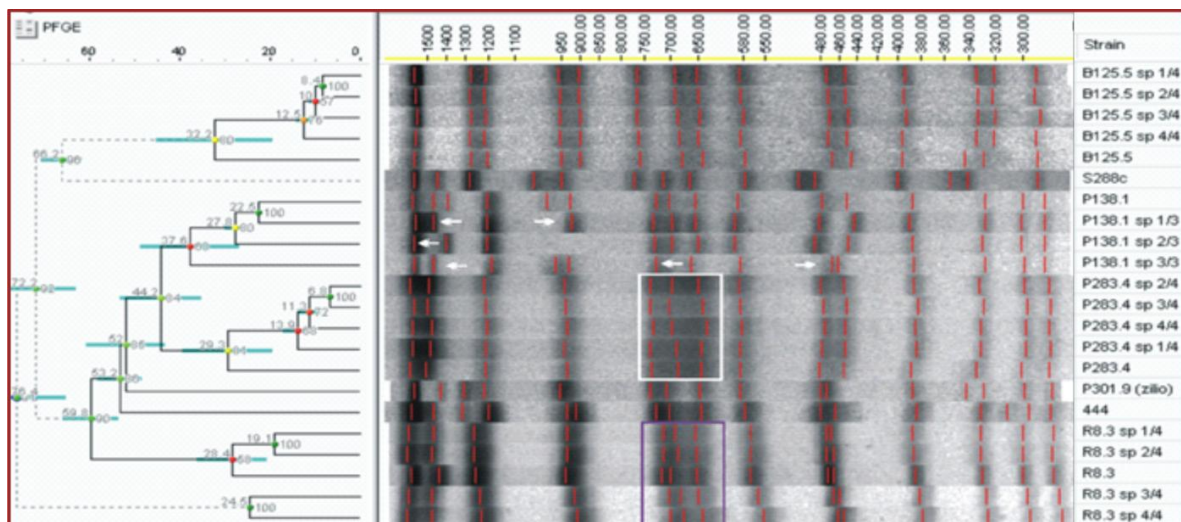


Figure 2.7 Phylogenetic tree built using stable and unstable strain as prove of chromosome recombination.

As indicated by the arrows chromosomes size is highly variable and bands not present in the parental strain appear in the derivatives, it happened also in other cases not reported in the figure. The origin of these chromosome changes is not clear, it could result from an "illegal crossing over." This phenomenon takes place, for example, both in strain P301.9 and R150.1 but while two of the derivative of P301.9 have a profile identical to that of the parental, in the case of R150.1 none of the homozugous seems to have a corresponding karyotype. Focusing on two bands corresponding to chromosome 13 in the heterozygous strain, are inherited independently in the derivatives, as they were two copies of the same chromosome but with different size. To support this speculation it has been seen that he two derivatives that inherit the smaller copy showed a reduced bands intensity, while the other two gave a higher signal. Finally P283.4 and Roo8.3 strains, those with the best performance of fermentation and spore viability, appear to be quite stable, apart from some slight variations in the chromosomes of intermediate size, marked by squares in the figure, which can be attributed to normal variations in telomere length.

Finally those strains with a high frequency of viable spores and with a chromosome structure correspondent between parental and derivative diploids were chosen for the next steps of the project.

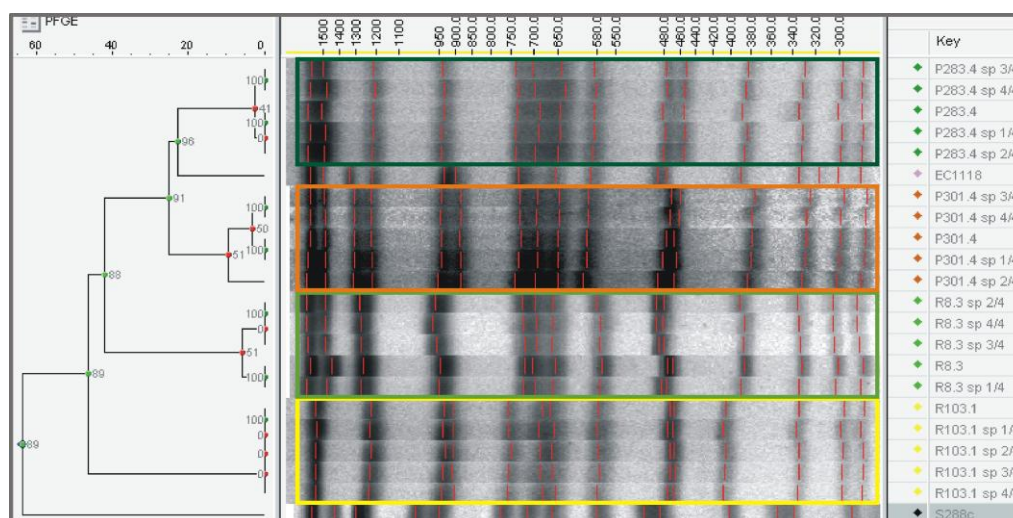


Figure 2.8 Phylogenetic tree of selected strains showing chromosomal correspondence between parental and derivatives.

The selected strains are two from Prosecco vineyards (P283.4 and P301.4) and two from Raboso ones (Roo8.3 and R103.1), hereafter called for simplicity P283, P301, Roo8 and R103. Viable spores from these strains gave homozygous derivative cultures that have been also chosen to facilitate the sequencing processes and assembly. It's important that homozygous lines maintain the same physiological characteristics of the parental strains to be sure that they are still representative of the yeast populations. This point will be discuss in the next paragraph.

Derivative Lines Selection

A first valuation of the most important oenological trat correspondence have been performer on the four natural isolates and on all homozygous derivative lines obtained, 24 for each strain. It was also carried out total DNA extraction of all strains and the enzymatic restriction to verify the corrispondence of mitochondrial DNA profile that confirmed the absence of contamination. All the following test were performed on both parental strains and derivatives searching for those with technological performances more similar to those of the parental.

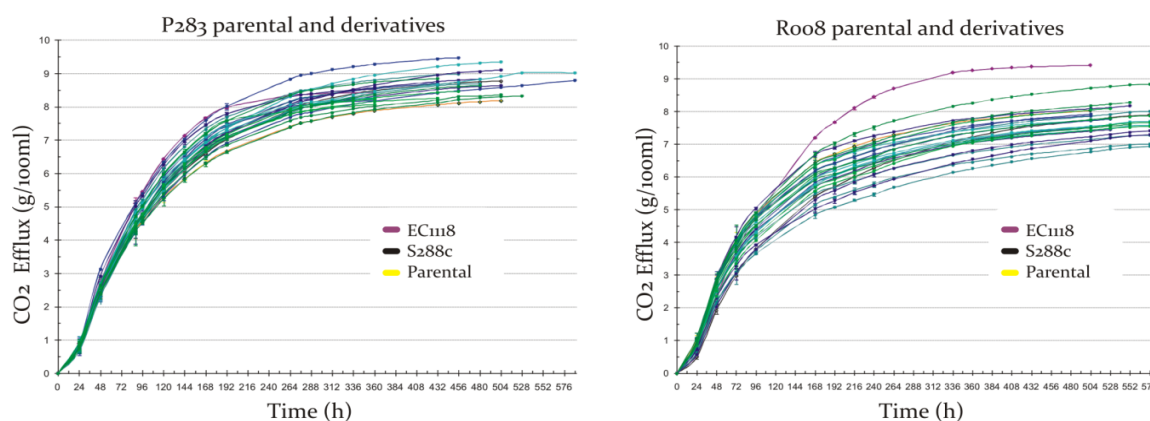


Figure 2.9 Fermentation cumulative curve of two selected strain and their homozygous derivative lines.

It was possible to evaluate the distribution of characters of interest among the variability in the fermentation performance of the first generation and to compare it with the fermentation kinetics of the parental strain. The validity of the assay used was confirmed by the reproducibility of the results obtained. Homozygous derivatives from P283 and P301 strains, isolated from Prosecco wine, showed a greater amount of glucose consumption in less time than the parental, but P283 derivatives showed a lower variation respect to P301 ones. Raboso isolate Roo8 produced a first generation less powerful than the parental, while R103 strain is positioned exactly in the middle of the derivatives distribution. It was also possible to identify among all derivatives line those, in terms of fermentation performances, the more similar to the parental strain that will be used in the sequencing process.

Ethanol Stress Resistance

A further test was performed by phenotypic matching the growth of heterozygotes and their derivatives in YPD media with the addition of known concentrations of ethanol (8%, 9%, 10% and 11%). However no detectable differences emerged and it seems that the resistance trait was transferred equally to all the spores. This result can be explained by a low heterozygosity of gene composition responsible for resistance to ethanol in the strains.

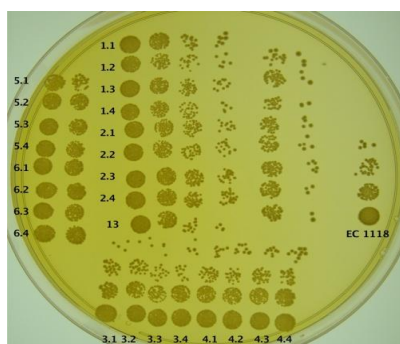


Figure 2.10 Strain R103 and its derivative growth on YPD solid media 10% ethanol with EC1118 strain as comparison

Despite we cannot evidence clear difference in ethanol resistance even among different strains, in paragraph "Genes involved in ethanol tolerance" (Chapter 4) it will be described a marked expression variation between strains in genes responsible for ethanol resistance. Since this differences are more evident in the first step of fermentation curve, corresponding to nearly 0.5% of ethanol produced, we are planning for the future to analyze strains growth curve in presence of low ethanol concentrations.

Oenological Trait Evaluation

With the first screening we were able to identify the candidates for each strain that are more similar to the parent. This strains were afterward re-tested for a variety of other technological trait of interest. Below are reported the fermentation curves R103 strain, representative for the other, in which we compared the homozygotes fermentation curve with the parental one obtained using respectively the 1l bioreactors with MS300 media and the small-scale method with MNS media. Only minor differences are visible.

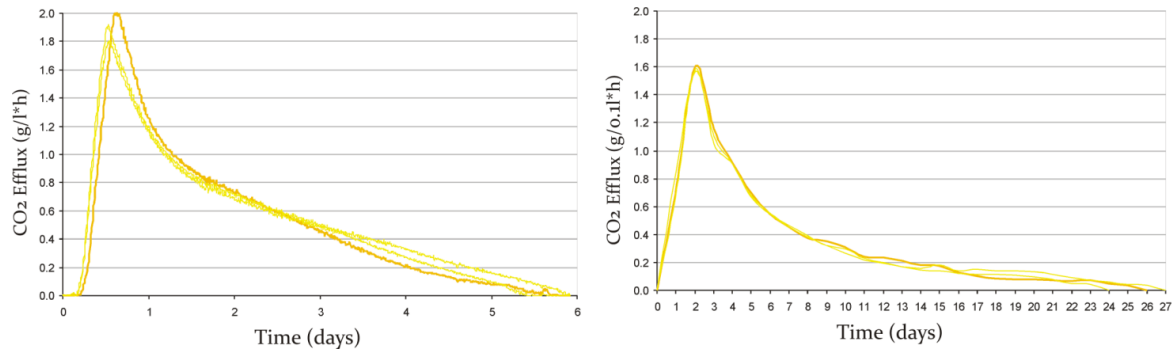


Figure 2.11 Strain R103 monitored during fermentation process in bioreactors (left) and in small-scale (right).

Growth Curves

Analyzing strains growth curve in YPD media it was possible to notice a similar pattern between our strains and the commercial starter. First of all it is important to note that natural isolates growth course was correspondent to that of their homozygous derivative lines. R103 strain was slightly slower in the exponential phase compared to the others, but in stationary phase reaches the highest values of final cell concentration.

Roo8 strain instead has a slightly higher rate of exponential growth. Yeasts growth curve was also monitored in MNS to evaluate the differences of performance in two different environments and under fermentation conditions. The major differences between the two media are pH values (3.2 in MNS and 5.0-5.5 in YPD) and osmotic pressure exerted by glucose concentration (200 g/l in MNS and 20 g/l in YPD), which leads to a completely different metabolic response in yeast (Crabtree effect more intense).

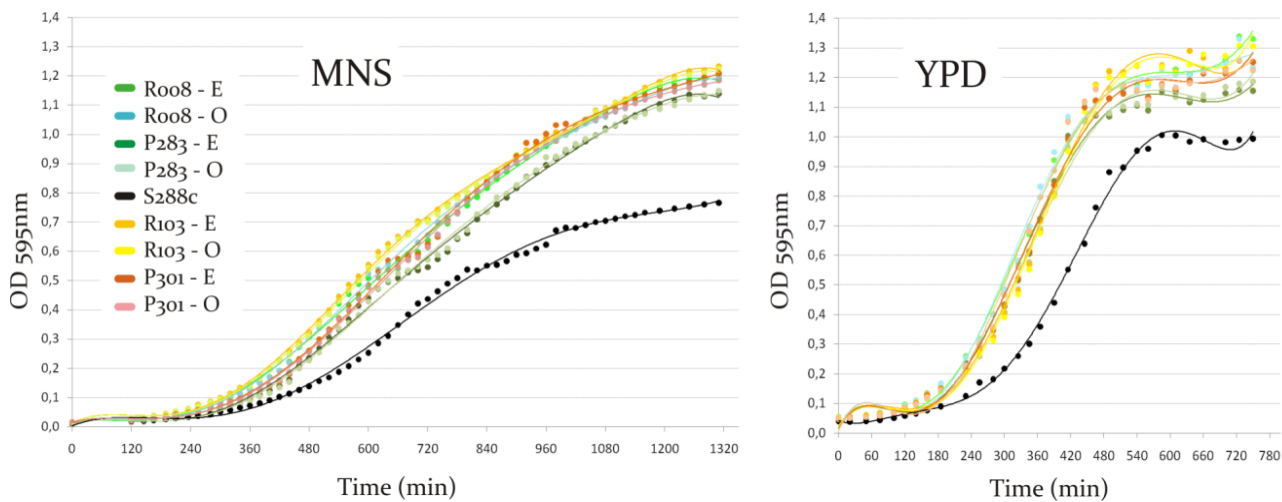


Figure 2.12 Growth curves of strains in MNS media (left) and YPD media (right)

The first thing we saw is that the growth rate is slower in MNS compared to YPD, the stationary phase is almost double time long and inflection points are less pronounced. Analyzing individual strains it was found that P283 growth was slower than the others, while R103 that was slightly slower to enter in exponential growth in YPD media is the fastest in MNS. Roo8 strain is faster in the early exponential phase, but decreases slightly in the later stages. The stationary phase of all strains came in a range between 1.5 and 1.4, higher than 0.1 OD relative to growth in YPD. It is clear that despite the stressful environment natural yeasts are able to overcome very well the conditions imposed.

SO₂ Resistance

Sulfur dioxide is commonly added to must in pre-fermentation phase for its antiseptic, antioxidant and anti-fermentative properties. We tested our strains resistance to different SO₂ concentrations.

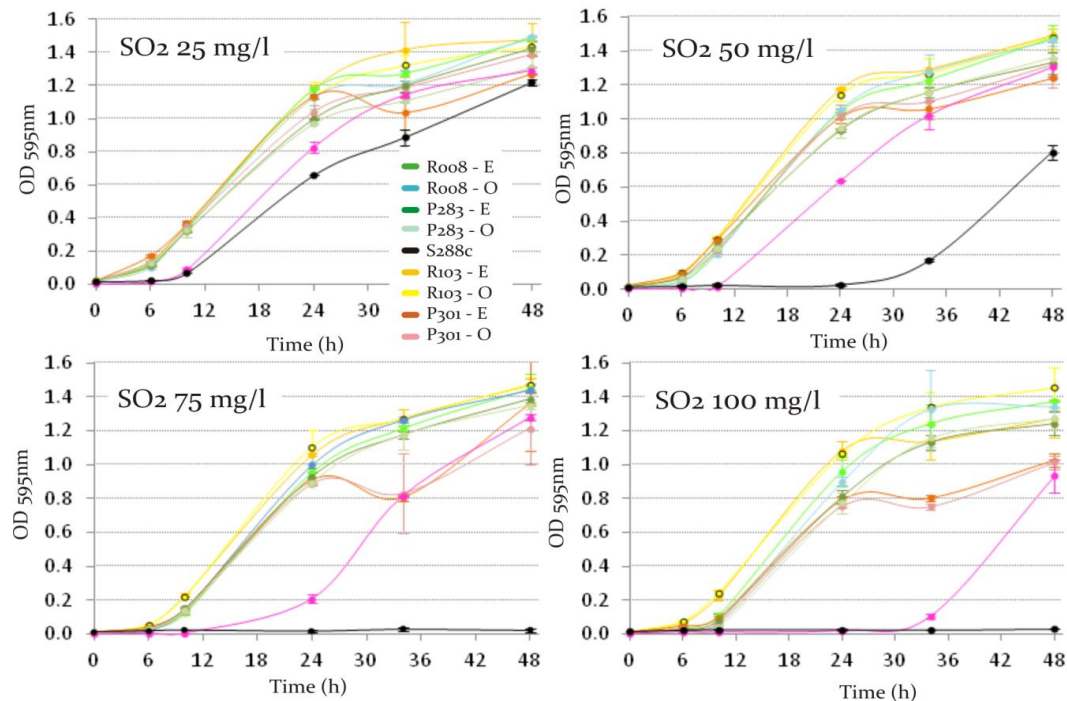


Figure 2.13 Yeasts growth in the presence of higher SO₂ concentration, respectively 25, 50, 75 and 100 mg/l.

Results indicate that, compared to a standard growth curve, the addition of SO₂ leads to a decrease in fitness in all strains. It is noteworthy to highlight that even with the variation of concentration of sulfur dioxide homozygous derivatives tend to maintain a pattern corresponding to that of the parent. In general, our natural isolates and their derivative lines are more resistant to sulfur dioxide than S288c and EC1118, especially at high concentrations. Among our strains the two R103 strains (-E and -O) have the best performances, while P301 are the more sensitive. It is important to highlight the differences seen at 50 and 75 mg/l where a substantial inhibition of strain S288c appear, with an OD less than 0.1 in 24h, while at 75 mg/l the resistance of EC1118 strain is lower than those of our strains.

It is also interesting to notice the behavior of P301 strains (-E and -O) that shows an initial sensitivity at 75 mg/l, confirmed at 100mg/l where their growth has slowed further. In general the Prosecco isolates seem to be more affected by the high concentration of sulfur when compared to those of Raboso. At 24 hours and 100 mg/l SO₂ EC1118 strain is strongly inhibited and S288c can be defined as non-viable. The data obtained can be seen that the ecotype strains have a higher resistance than the commercial strain EC1118 and the laboratory strain S288c. We can conclude that isolated strains have acquired, over the generations, genetic mutation useful to survive in adverse conditions in the stressful environment in which they lived.

Sulphur Compounds Production

From the intensity and colour change of single colonies in the two selective media used an suggestion of sulphite reductase enzyme activity and of the SO₂ production was obtained.

Hydrogen sulphide is a compound that in wine gives negative aroma like "reduced" or "rotten eggs" while SO_2 often generate unwanted high concentration of acetaldehyde. From data obtained we note that natural yeasts ability to produce H_2S and SO_2 are similar to each other, with the only exception of a low production in strains R103 (-E and -O), which is the best in this test. The correspondence of phenotype between parental strains and their derivatives was confirmed for all strains except Roo8, that showed a minimal difference. S288c strain has a high aptitude for the production of this compound, understandable for a laboratory strain. SO_2 production is linked to the production of H_2S , generally by an inverse correlation. This behaviour in fact is confirmed by our data. S288c and Roo8 strains showed and opposite trend in SO_2 production respect to H_2S , while the other strains were again medium producers.

Same conclusions can be deduced from iodometric titration data for determining total and free SO_2 production at the end of small-scale fermentation process.

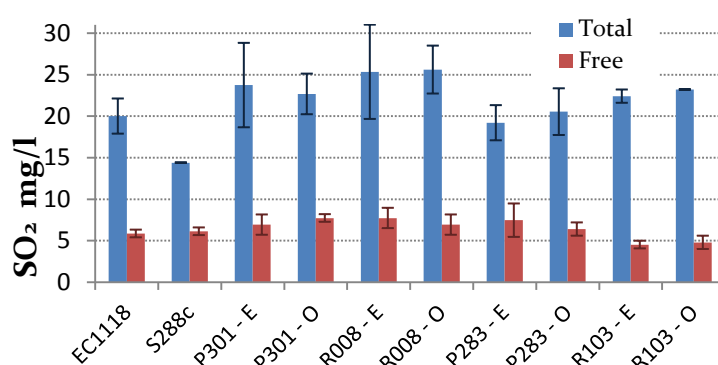


Figure 2.14 SO_2 presence at the end of fermentation process using small-scale method

Ethanol production

The medium used for this test has been deliberately modified to supply a sufficient amount of glucose for the production of higher ethanol amount and was also enhanced the availability of nitrogen to facilitate yeast metabolism.

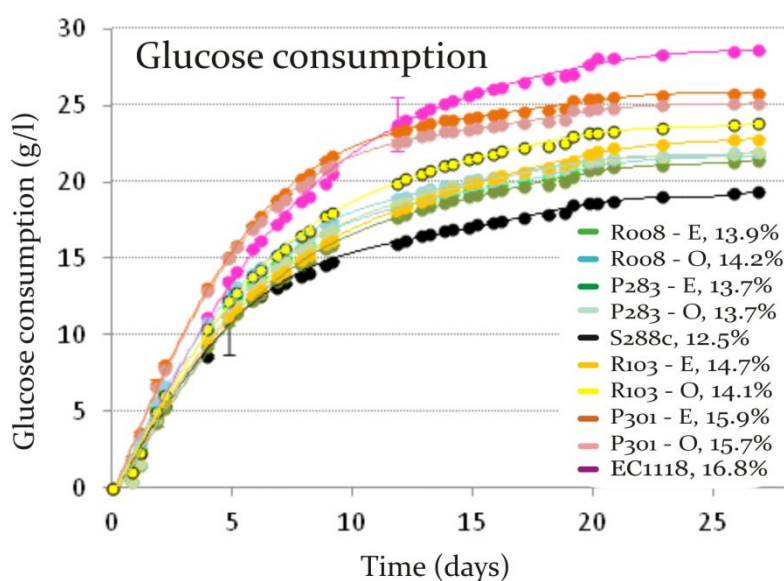


Figure 2.15 In y axes are reported the cumulative glucose consumption in g/l during fermentation (x axes). In the labels are reported final ethanol concentration produced (% v/v).

Ethanol production was determined by measuring with HPLC the quantity of residual sugar at the end of fermentation process and using the conversion factor sugar/alcohol of 0.61 (19). Strains P283 (-E and -O) had exactly the same values, while others have a negligible difference between parental and derivatives. Results show that strains P301 (-E and -O) produced the highest quantity of ethanol, near to 16% v/v, and also R103 strains (-E and -O) have a good alcohol production attitude. Strains P283 and R008 are less efficient. Finally even if EC1118 glucose consumption is more consistent, strain P301 is more efficient in the first ten days of fermentation.

Fermentation Profiles

Alcoholic fermentation in a synthetic white must containing 200 g/l of glucose under strict anaerobiosis conditions was monitored. The fermentation profiles for the four homozygotes strains plus S288c and EC1118 as references was determined. Usually fermentation rate (dCO_2/dt) reaches its maximal value around 12h, before entering stationary phase, and gradually declines thereafter until the end when sugar reserves are exhausted. Final development of CO_2 reaches the maximal expected value around 76 g/L when the fermentation is concluded and usually when less than 2 g/l of residual sugar are present. The accumulation of ethanol follows the same time course, as well as the cumulative release of CO_2 . Samples were taken along the whole process. The first at the beginning of the fermentation, when the cumulated CO_2 produced in the synthetic must reached 6 g/l, the second at 45 g/l and the third at the end of fermentation, when ethanol concentration had nearly reached 10% (v/v).

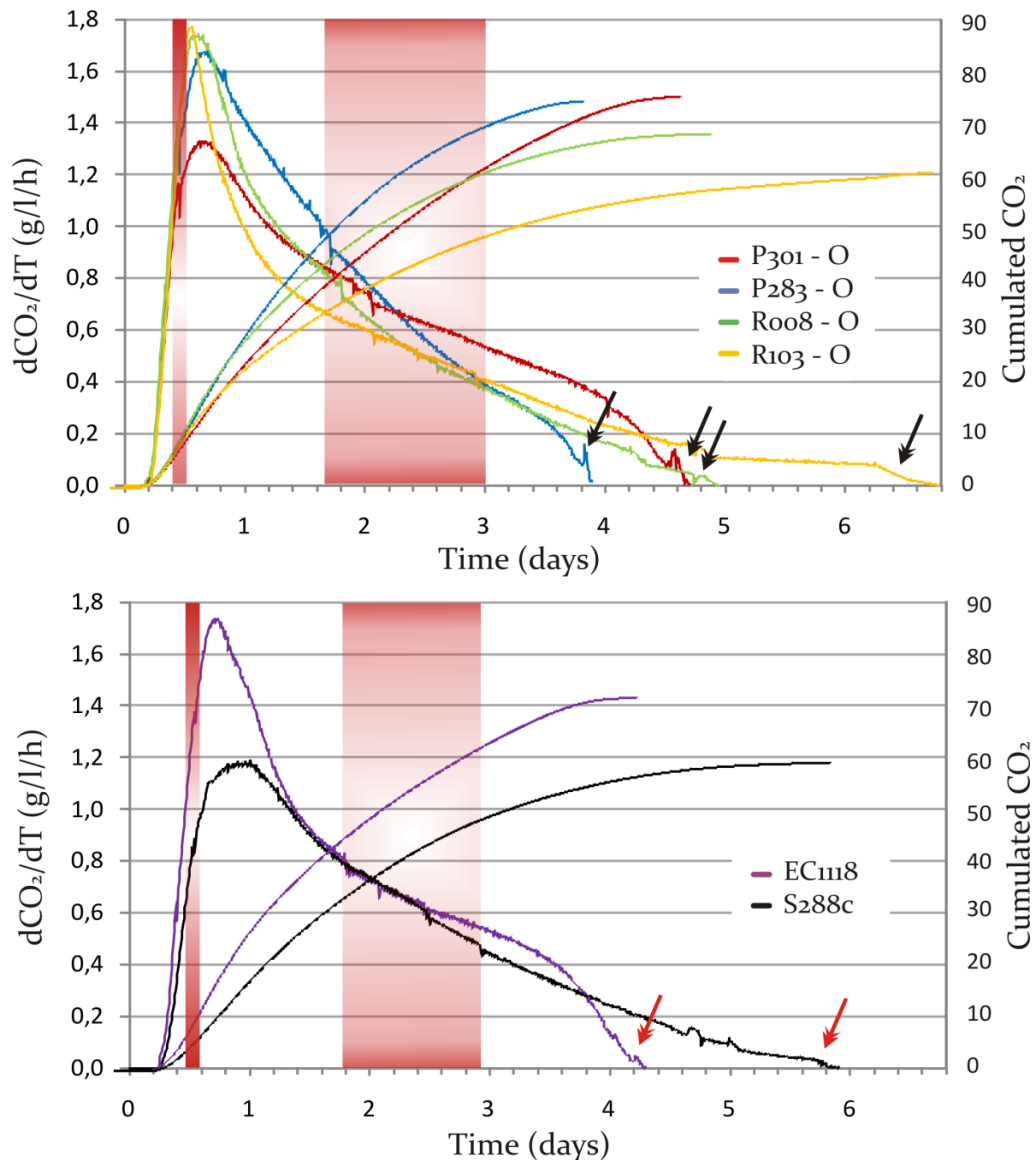


Figure 2.16 Fermentation kinetics of the four natural strains plus the two references. Both cumulated CO_2 and CO_2 per hour produced are displayed and represent the mean values between three independent replicates performed for each strain. Concentrations of 6g/l and 45g/l of CO_2 production are reached at different time points, red boxes represent temporary intervals of sampling and black or red arrows indicate samplings at the end of fermentation.

These concentrations are not reached contemporaneously by the different strains, because the amount of CO_2 produced depends on the rapidity of the specific strain. In winemaking, those strains that are able to complete the fermentation quickly and thus consuming all the glucose and releasing CO_2 in solution in shorter times, are preferred.

Fig. 2.16 highlights the typical differences between wine and non-wine strains during fermentation. EC1118 (20) is a commercial patented strain which shows good performances during fermentation. In fact if we compare the cumulative productions of CO_2 of EC1118 and of S288c, it is possible to see that the oenological strain reaches the concentration of 45 g/l faster than the laboratory strain. EC1118 concludes the fermentation more than one day before S288c. Furthermore it displays a high peak of production of CO_2 and a sudden closure of the fermentative process. P283 - O and P301 - O are characterized by a faster fermentation and concluded respectively with final 75 and 76 CO_2 g/l and no residual glucose. Strain R103 instead is characterized by low fermentation rate and was not able to conclude the fermentation together with the laboratory strain S288c.

This oenological strain has been chosen on purpose as “negative” control for the gene expression comparison. In the following table are reported mean values of the three independent experiment performed for each strain, describing fermentation progress: the fermentation duration (Tf), the maximum rate (Vm), the rate at 50% of fermentation (V50) and the total CO₂ released (Total CO₂).

Table 2.3 Principal parameters describing fermentation profiles of strains

Strain	V _m (g/l/h)	V ₅₀	T _f (h)	Total CO ₂
P283	1.7	1.2	94	74.7
R008	1.7	1.0	120	68.4
R103	1.8	0.7	163	61.0
P301	1.3	0.8	113	75.6
EC1118	1.7	0.9	103	72.1
S288c	1.2	0.8	142	59.6

Fermented must and yeast cells sampling points are marked in Fig. 2.16 by red boxes and arrows. We chose the first two time points for the third phase of the project, RNA-seq, because they represent two very different steps during fermentation so they should allow us to correlate differences in performance with corresponding gene expression. During the first time point (6 g/l) cells are actively reproducing and are increasing their ability to produce CO₂. In the second step (45 g/l) cells are in the stationary phase and they have passed the peak of high production of CO₂ but undergo ethanol stress. The final sampling was taken to complete all chemical analyses and, if needed, for real time PCR confirmation of RNA-seq results.

Fermented Must Evaluation

A variety of chemical analysis have been performed on fermented must sampled in the three time point during fermentation process. Glucose consumption rate and ethanol production were evaluated using HPLC. Results indicated that final concentration of all strains is around 10% v/v except the two worst fermentative strains, S288c and R103 – O, that left some residual glucose, respectively 8.7 and 9.2 g/l. Due to volatile property of ethanol, data reported present some biases.

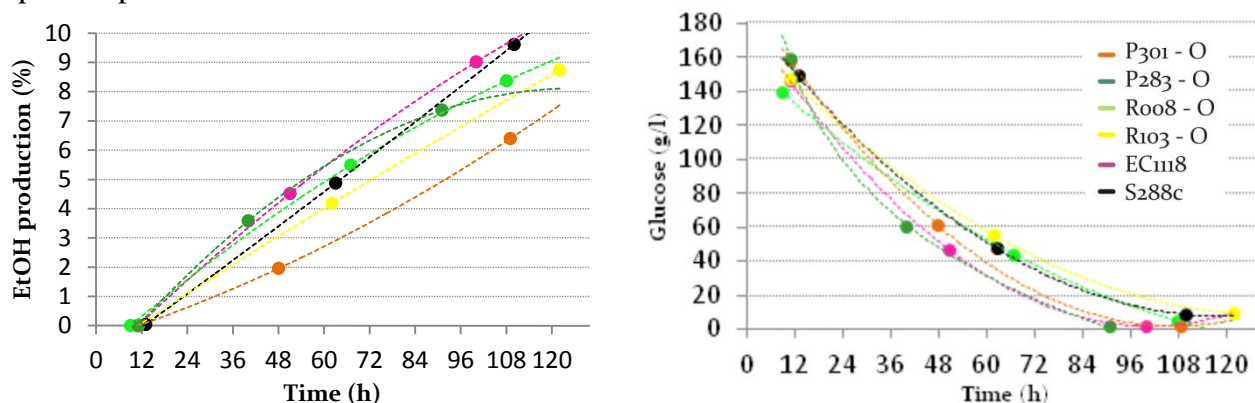


Figure 2.17 Ethanol production and corresponding glucose utilization of strains in the three fermentation time point.

The histogram below shows the variations of chemicals composition (glycerol, maleic acid, succinic acid, citric acid and acetic acid) observed during the exponential phase of glucose consumption (6g/l), in the middle of the curve of fermentation (45g/l) and at the end of fermentation (80g/l). Normally *S. cerevisiae* glycerol production is between 3.5 and 6.0 g/l, data obtained are positioned in a range between 3.82 and 4.35 g/l for all strains except for S288C, which is lower, 3.4 g/l. Raboso isolates have a lower production compared to that of Prosecco ones, that are similar to EC1118. The malic acid present in mature grapes is variable, depending on the growth zone, between 4 and 6.5 g/l and in MS300 media is 6g/l. Usual succinic acid production in *S. cerevisiae* is between 0.5 and 1.5 g/l, in EC1118 we found 1.99 g/l and slightly lower production in the two Prosecco isolates while it is much higher in Raboso ones. Citric acid is present at low concentrations in wine and is responsible for a pleasant flavour. The amount of citric acid present in MS300 is 6g/l.

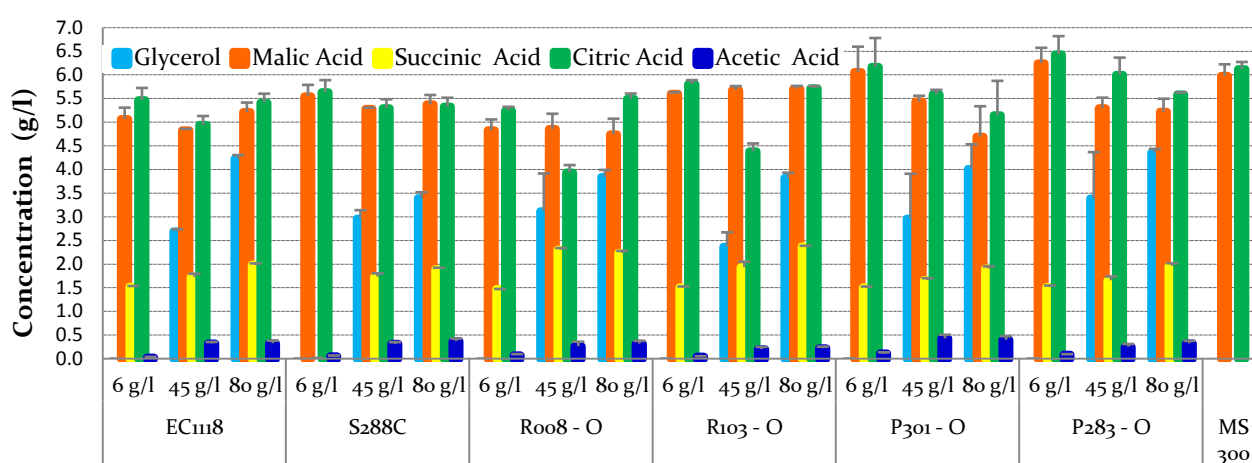


Figure 2.18 Chemical compound determined with HPLC

The trend in citric acid assimilation is common between strains. In particular Roo8-O and R103-O consume more citric acid than EC1118, while in S288c in almost zero. Rather different behavior was found in Prosecco isolates which continue even after mid-fermentation to assimilate citric acid, in particular strain P301.4 assimilates more citric acid than all other strains and is also the highest producer of acetic acid. In high concentration, greater than 0.6 g/l, the volatile acidity of this compound affects the quality of the wine and enhances the astringency. Acetic acid is formed during fermentation as a result of secondary reaction of acetaldehyde oxidation. In Italy, the value of volatile acidity permitted by law, respectively, 1.5 g/l for white wines and 1.7 g/l for red wines.

All concentrations reported produced by our strains are more than acceptable. The highest value was determined in strain P301-O, while the lowest in the strain R103-O.

Acetaldehyde

This compound is usually produced during fermentation and its formation is generally common to all strains at concentrations rather small (between 7 and 16 mg/l). A high level of acetaldehyde is undesirable because it is associated with the smell of rowan, which gives the wine an aroma of faded and removes the fruity freshness and vivacity. In addition, the acetaldehyde combines with sulphuric acid and decreases its effect of antioxidant and anti-fermentative.

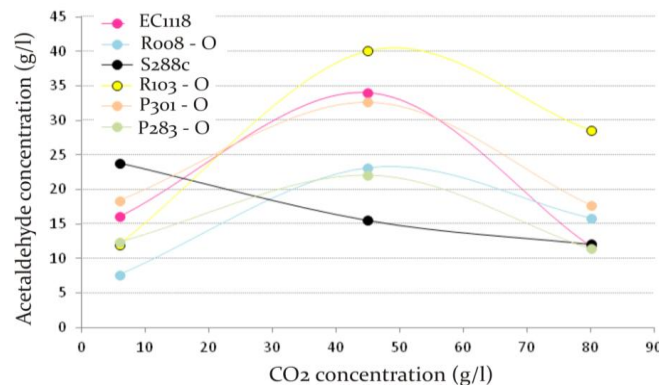


Figure 2.19 Acetaldehyde concentration during fermentation

In S288c strain the acetaldehyde concentration decreases steadily until the end of fermentation (0.13 g/l), while a common trend in wine yeasts shows a peak production at 45 g/l and then fall down again at end fermentation to values of 10-15 g/l. Yeast free acetaldehyde is useful to counteract the toxic effects of sulfur dioxide. Analyzing the production of natural isolates and comparing it with that of EC1118 they are quite similar except for R103-O strain. This Raboso isolate present the highest production during fermentation has the highest final concentration and it has an interesting correspondence between resistance to sulfur compound and the production of acetaldehyde.

REFERENCES

- (1) Heard GM, Fleet GH. Growth of Natural Yeast Flora during the Fermentation of Inoculated Wines. *Appl Environ Microbiol* 1985 Sep;50(3):727-728.
- (2) Vaughan-Martini A, Martini A. Isolation, purification, and analysis of nuclear DNA in yeast taxonomy. *Methods Mol Biol* 1996;53:89-102.
- (3) Mortimer R, Polsinelli M. On the origins of wine yeast. *Res Microbiol* 1999 Apr;150(3):199-204.
- (4) Sniegowski PD, Dombrowski PG, Fingerman E. *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* 2002 Jan;1(4):299-306.
- (5) Naumov GI, Naumova ES, Sancho ED, Korhola MP. Polymeric SUC genes in natural populations of *Saccharomyces cerevisiae*. *FEMS Microbiol Lett* 1996 Jan 1;135(1):31-35.
- (6) Murat ML, Tominaga T, Dubourdieu D. Assessing the aromatic potential of Cabernet Sauvignon and Merlot musts used to produce rose wine by assaying the cysteinylated precursor of 3-mercaptohexan-1-ol. *J Agric Food Chem* 2001 Nov;49(11):5412-5417.
- (7) Swiegers JH, Kievit RL, Siebert T, Lattey KA, Bramley BR, Francis IL, et al. The influence of yeast on the aroma of Sauvignon Blanc wine. *Food Microbiol* 2009 Apr;26(2):204-211.
- (8) Fleet GH. *Wine microbiology and biotechnology*. London: Taylor & Francis; 2002.
- (9) Ribéreau-Gayon P. *The handbook of enology*. Chichester, England: Wiley; 2000.
- (10) Michnick S, Roustan JL, Remize F, Barre P, Dequin S. Modulation of glycerol and ethanol yields during alcoholic fermentation in *Saccharomyces cerevisiae* strains overexpressed or disrupted for GPD1 encoding glycerol 3-phosphate dehydrogenase. *Yeast* 1997 Jul;13(9):783-793.
- (11) Querol A, Barrio E, Huerta T, Ramon D. Molecular monitoring of wine fermentations conducted by active dry yeast strains. *Appl Environ Microbiol* 1992 Sep;58(9):2948-2953.
- (12) Regodon JA, Perez F, Valdes ME, deMiguel C, Ramirez M. A simple and effective procedure for selection of wine yeast strains. *Food Microbiol* 1997;14:247-254.
- (13) Giudici P, Solieri L, Pulvirenti AM, Cassanelli S. Strategies and perspectives for genetic improvement of wine yeasts. 66, 622-628. *Appl Microbiol Biotechnol* 2005;66:622-628.
- (14) Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WP, del Cardayre SB. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 2002 Feb 7;415(6872):644-646.

- (15) Marullo P, Bely M, Masneuf-Pomarede I, Aigle M, Dubourdieu D. Inheritable nature of enological quantitative traits is demonstrated by meiotic segregation of industrial wine yeast strains. *FEMS Yeast Res* 2004 May;4(7):711-719.
- (16) van der Westhuizen TJ, Pretorius IS. The value of electrophoretic fingerprinting and karyotyping in wine yeast breeding programmes. *Antonie Van Leeuwenhoek* 1992 May;61(4):249-257.
- (17) Codon AC, Gasent-Ramirez JM, Benitez T. Factors which affect the frequency of sporulation and tetrad formation in *Saccharomyces cerevisiae* baker's yeasts. *Appl Environ Microbiol* 1995 Feb;61(2):630-638.
- (18) Vaughan-Martini A, Martini A, Cardinali G. Electrophoretic karyotyping as a taxonomic tool in the genus *Saccharomyces*. *Antonie Van Leeuwenhoek* 1993 Feb;63(2):145-156.
- (19) Delfini C. *Scienza e tecnica di microbiologia enologica*. Asti: Edizioni Il Lievito; 1995.
- (20) Walker GM. *Yeast physiology and biotechnology*. Chichester, West Sussex: Wiley; 1998.

3. GENOME SEQUENCES

INTRODUCTION

The yeast genome is quite small and highly packed, with about 6000 genes distributed over 16 chromosomes. *S. cerevisiae* also has two small cytoplasmic genomes: mitochondrial DNA and 2 μ plasmid. The nuclear genome structure is intimately linked to yeast genetic properties, which reciprocally influence its life style. The first strain sequenced, S288c, is a commonly used laboratory strain that was obtained in 1950s by mating a strain isolated from a rotten fig (EM93) with a commercial strain (1). While experimental condition may have left a significant footprint on the evolution of S288c (2), since 1996 its genome sequence has been the only reference sequence available for *S. cerevisiae*. Today the genomes of several other yeast strains have been sequenced, including that of RM11-1a, a haploid derivative of a natural vineyard isolate (www.broadinstitute.org/annotation/genome/saccharomyces_cerevisiae/Home.html), the clinical isolate YJM789 (3), and the diploid, heterozygous wine yeast strain EC1118 widely used as starter in the wine industry (4). The sequence divergence between these strains and the reference has been estimated at 0.5-1%, similar to that between humans and chimpanzees.

Genetic Characteristics

S. cerevisiae strains are mostly diploid in natural condition and display vegetative reproduction through multi-polar budding. Under specific nutritional condition cells may sporulate to form four haploid spores of different mating types, a or α . One peculiarity of wine strains is that many are homotallic, and descendants of these haploid spores mate with their own progeny to form a diploid. Homotallicism is frequent in wine yeast, with about 70% of strains known to be homotallic (5), but the ecological significance of this property remains unclear. Upon sporulation and the self-mating of homotallic spores, homozygous diploids are generated. This process makes it possible to eliminate recessive mutations deleterious for the strains or to ensure that recessive mutations increasing strain fitness are expressed. Genome renewal is therefore likely to play a role in adaptation of yeasts to stressful wine environment. Little is known about the sexual activity of yeasts in wine environments. The frequency at which yeasts sporulate and mate in such environment is unknown. The ability of wine yeast to sporulate is highly heterogeneous and varies from 0% to 100% on laboratory media. Early genetic studies on wine yeasts indicated that most strains were diploid though some were polyploid or aneuploid (6). An estimation of DNA content of a large set of commercial "fermentation" strains recently showed that most of these strains had a DNA content close to 2n (7). Unlike other industrial yeasts (baker's yeast and brewing yeast strains), which have ploidy levels exceeding 2n, most of the *S. cerevisiae* strains used in wine-making seem to be diploid. *S. cerevisiae* has a small (75 kb), circular mitochondrial DNA genome that encodes a small set of proteins involved principally in respiration. Mitochondrial DNA is not essential for yeast survival but it was observed that the ethanol resistance can depend on it and that the ethanol tolerance of a laboratory strain could be enhanced by introducing mitochondria from a flor yeast (8).

Chromosomal Rearrangements and SNPs

The existence of gross chromosomal rearrangements, such as translocations, deletions and insertion, was rapidly suspected based on the high level of chromosome polymorphism found in wine yeasts. Analysis of wine yeast chromosomes by Pulsed Field Gel Electrophoresis (PFGE) demonstrates major chromosome length polymorphism between wine yeast strains. Such variation in chromosome size clearly resulted from gross chromosomal rearrangements (GCR). Recombination between repeated Ty elements interspersed throughout the genome is shown to be a major cause of chromosomal translocation (9). Other types of repeated sequences may also serve as substrates for ectopic recombinations leading to chromosomal rearrangements (10). Some gene copy-number changes are specific to wine yeasts and have been identified as a possible wine yeast signature (11). The differences between wine strains are moderated and mostly concern genes encoding membrane transporters. The gene amplified in wine yeasts are mostly located at the end of chromosomes confirming the plasticity of sub-telomeric regions and their role in adaptation to industrial environments (12). The effects on yeast fitness of most of these rearrangements remain unclear, although no differences in fermentation properties are found between different structural variants (13). The best studied case of contribution to adaptation is that of a translocation between chromosome VIII and XVI, which has a direct impact on sulfite resistance (14).

With their small and compact *S. cerevisiae* and hemiascomycetes represent a powerful model for comparative genomics and studies of genome evolution. As a result, more than 18 hemiascomycetes species are either completely or partially sequenced. The availability of the sequence data has presented an unprecedented opportunity to evaluate DNA sequence variation and genome evolution in a phylum spanning a broad evolutionary range. This wealth of data on interspecific sequence differences stands in contrast to our limited knowledge of sequence variation within *S. cerevisiae*. Several work recently tried to cover this gap of understanding (15,16).

The Finishing Task

The process of *finishing* a genome is aimed to move it from a draft stage, the result of sequencing and initial assembly, to a complete genome. This process is very challenging and time consuming but indispensable because only with a small number of scaffolds and gaps in the assembly it is possible to reach a good level genomic and SNPs comparison. Furthermore only a complete genome sequence allows a reliable gene finding and annotation.

A good strategy to sequence a genome is based on two kind of genomic libraries, a shotgun library, prepared fragmenting the DNA randomly into numerous small segments and a paired-end library, created breaking the DNA into large fragments (usually between 3 and 20 kb) and processing them into molecules having only the two end sequences of the fragments. Once sequenced, in fact, the two libraries provide two kind of information, the sequences themselves for increasing the coverage, and the relative pairwise distance and orientation for scaffolding reads positions along the genome. All these information are analyzed using bioinformatics programs to create final assemblies.

Overlaps between reads are used to order and merge them into structures called *contigs* representing sequences of the genome in which the order of bases is known, without gaps. Paired-end sequences are used to assemble *contigs* into longer sequences called *scaffolds*. In fact, knowing the fixed length of the fragments used to produce paired-end libraries it is possible to infer the pairwise distance and orientation between the two ends sequences (17)(18). This information is used to order and orient *contigs* with respect to each other analyzing where the two ends of the same pair map and to infer the length of the gaps between them.

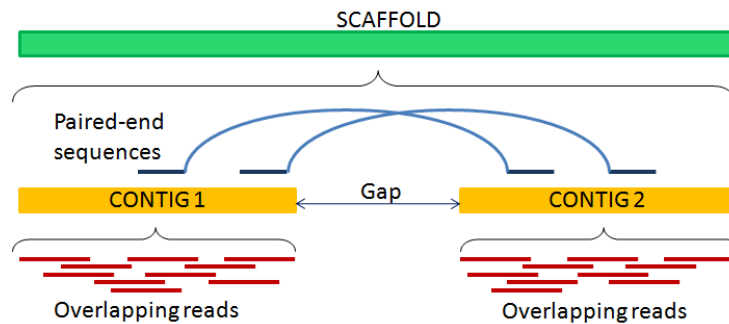


Figure 3.1 Cartoon describing general mechanisms used to assemble reads into contigs and scaffolds.

Once *scaffolds* are created, the remaining gaps can be filled with bioinformatics approaches or sequencing specific missing regions and this is the *finishing* step. Gaps creation in the contigs assembly step are due to problems in overlap graphs creations caused mainly by low coverage regions, sequencing errors and repeated sequences. True overlaps between overlapping sequences can be missed in regions with low coverage because there are not enough reads to confirm the connection between two sequences. Reads with sequencing errors can induce the path of the overlap graph to diverge into two different paths, or can induce two paths to converge into a single one, because of the co-presence of correct and wrong sequences. Repeats can increase graph complexity, leading to tangles that are difficult to resolve. Multiple copies of a repeat can collapse into a single unitig so regions with similar repeats can have reads joining several contigs to a single unitig containing the repeat. Other kinds of repeats such as homopolymers and short tandem repeats are generally low quality and, depending on their length, they tend to have paths converging to themselves because reads are highly repetitive and their ends overlap with other reads with the same repeat or with themselves (19)(20). These problems are usually solved during branch-point analysis on the overlap graph identifying critical regions of the path and breaking contigs at these points to avoid misassemblies and losing possible true overlaps.

Gene Prediction

The gene prediction, or annotation, is the problem of identifying stretches of sequence (genes) in genomic DNA that are biologically functional, and to define their internal structure. Existing approaches to solve this problem fall into two groups with respect to the technique they utilize: *intrinsic* or *ab initio* methods and extrinsic or similarity-based ones. The first class uses only the information contained in the input genomic sequence: it searches for typical patterns that generally characterize coding boundaries, and other signals inside and outside gene regions. The second type applies the information coming from external sources as EST, proteins, or other known references.

As the entire genomes of many different species are sequenced, a promising direction in current research on gene finding is a comparative genomics approach. This is based on the principle that the forces of natural selection cause genes and other functional elements to undergo mutation at a slower rate than the rest of the genome, since mutations in functional elements are more likely to negatively impact the organism than mutations elsewhere. Genes can thus be detected by comparing the genomes of related species to detect this evolutionary pressure for conservation. This approach was first applied to the mouse and human genomes, using programs such as SLAM, SGP and Twinscan/N-SCAN. Comparative gene finding can also be used to project high quality annotations from one genome to another. Notable examples include Projector, GeneWise and GeneMapper (21). Such techniques now play a central role in the annotation of all genomes.

MATERIALS AND METHODS. MOLECULAR BIOLOGY

Common experimental procedures can be found together with list of abbreviation and standard solution in the Appendix I section.

DNA Purification

Nucleic acids were purified as previously described in Barnett J.A (22), with minor variations. 100 ml of an overnight culture of yeast in their late exponential growth phase were harvested by centrifuging, to obtain a pellet of 10 g roughly. Cells were transferred to a 50ml tube and resuspended in 20ml extraction buffer (0.1M TrisHCl pH 8.5, 0.1 M EDTA pH 8.0, 0.2 M NaCl, 2.5%(w/v) SDS, 1mg/ml ProteinaseK (added right before use)). Mixture was then incubated on a rocking platform for 15 50 minutes at room temperature. Mixture was centrifuged at 9,500xg at room temperature for 10 min and supernatant was transferred to a new tube. 1/2 volume of phenol was added to supernatant, vortexed and mixed for 15 minutes. 1/2 volume of chloroform:isoamyl alcohol (24:1) was added only at this point, vortexed and mixed for an additional 20 minutes. Upper aqueous phase was recovered after centrifuging at 500xg for 515 min at room temperature and re-extracted with an equal volume of chloroform:isoamyl alcohol. 0.6 volumes of isopropanol were added to the recovered upper phase, mixed thoroughly and stored at -20°C for 90 minutes to precipitate nucleic acids. DNA were recovered by centrifuging at 12,000xg for 30 minutes at 4°C, supernatant was discarded and pellet rinsed in 5ml of 70% ethanol. After centrifuging pellet was drained from residual ethanol under the hood and resuspended in 200µl of cold mQ water. Centrifuge tube wall were rinsed with another 200µl of mQ water and combined with the resuspended pellet. Isolated nucleic acids were extracted again using 1/2 volume of cold phenol vortexing for 1 minute and 1/2 volume of cold chloroform:isoamyl alcohol, vortexing again. Mixture was then centrifuged at 12,000xg for 5 min (4°C) in Phase Lock Assemblies (PRIME). Upper phase was decanted and extracted with an equal volume of cold chloroform:isoamyl alcohol, gently mixed and recentrifuged in a Phase lock Assembly. Upper phase was transferred to a new tube and 1/10 of the volume of 3M sodium acetate was added together with 2 volumes of ethanol, mixed well and store at -20°C for 60 minutes. After centrifuging at 12,000xg for 30 minutes at 4°C, pellet was overlayed with 200µl of 70% ethanol and centrifuged again for 15 minutes. Pellet was air dried in sterile transfer hood for 10 min and resuspended in 400µl of cold mQ water. 200 µl of 8M LiCl were added, mixed thoroughly and solution placed at 4°C overnight. After centrifugation at 12,000xg for 30 minutes at 4°C supernatant, which contained DNA and tRNA, was collected. 1/10 of the volume of 3M sodium acetate and 2 volumes of ethanol were added to DNA solutions for precipitation. Samples were kept at -20°C overnight, centrifuged at 16,000xg for 30 minutes at 4°C and washed with 100µl of 70% ethanol. Obtained pellets were air dried in sterile transfer hood for 15 min and resuspended in 50µl of cold mQ water. The preparations were routinely assessed for quality and concentration using respectively Nanodrop and Qbit and stored either at 4°C or -20°C.

DNA concentration and quality

The concentration and quality of nucleic acid preparations were determined with a NanoDrop instrument (Nanodrop1000, Thermo Scientific) at a wavelength of 260 nm (A₂₆₀). An A₂₆₀ of 1.0 is equivalent to a concentration of approximately 50 µg ml⁻¹ of double stranded DNA, 33 µg ml⁻¹ of single stranded DNA or 40 µg ml⁻¹ RNA (23). The degree of contamination in the preparations could be estimated by measuring the A₂₆₀/A₂₈₀ ratio and A₂₆₀/A₂₃₀ ratio. Values above 1.95 for the measured A₂₆₀/A₂₈₀ and A₂₆₀/A₂₃₀ suggested a clean sample, whereas lower values indicated the presence of contaminants. The concentration and quality of DNA preparations were also visually estimated after agarose gel electrophoresis in the presence of Ethidium Bromide under UV illumination. The signal for the DNA with the unknown concentration was compared to the intensity of a marker DNA with a known DNA concentration. Moreover, in order to quantify the RNA contamination in DNA samples and parallel the DNA contamination in RNA samples, solutions were also examined using Qbit fluorometric quantization kits (Qbit 1.0 fluorometer, Invitrogen), which allow the registering of different signals from the two nucleic acids using specific fluorescent probes. Samples were prepared for dsDNA broad range assay and RNA assay following the manufacturer instruction. Fluorometric assay yielded a quantification of each of the nucleic acids in the samples and could be compared with the data obtained using the spectrophotometer.

Amplification by polymerase chain reaction (PCR)

The thermo stable DNA polymerases used in this study were: GOTAQ (Promega) and PHUSION (New England Biolabs). GOTAQ DNA polymerase was used for routine screening. PHUSION DNA polymerase was used to amplify DNA fragments for high fidelity cloning and sequencing and produced blunt-ended PCR products.

Either the Mastercycler gradient (Eppendorf) or the X T gradient (Biometra) PCR machine was used to amplify a desired DNA fragment using different DNA templates and the primers listed in tables specific for each experiment. A typical 25 µl reaction mixture, in which 0,2µl of GOTAQ DNA polymerase were used, contained: 5 µl of 5x reaction buffer supplied by the manufacturer (0.5 M KCl, 0.1 M Tris/HCl pH 8.3, 7.5 mM MgCl₂), 0,5 µl of 10 mM dNTP mixture (Invitrogen; final [0.2 mM] for each nucleotide: dATP, dCTP, dGTP and dTTP), 1 µl of 5 µM forward primer (Invitrogen; final [0,1 µM]), 1 µl of 5 µM reverse primer (Invitrogen; final [0,1 µM]) and 1-5 µl template DNA. This reaction mixture was made up to 25 µl with sterile mQ water, mixed and briefly centrifuged. When possible Green Buffer, containing already the loading dyes (Xylene cianolo e tartrazina) for the 58 subsequent electrophoresis, was used. The lid of the PCR machine was heated during the program to prevent sample evaporation and condensation in the lid of the tube. A standard PCR program consisted of an initial denaturation step at 94 °C for 2 min and 35 subsequent cycles of 94 °C for 30 sec (denaturation), from 46 to 60 °C for 30 sec (primer annealing) and 72 °C for 1 to 6 min (primer extension; 1 min per 1 kb). The final extension step was performed at 72 °C for 10 min. The reaction mixture and the PCR program were varied when the standard procedure did not yield an optimum amplification.

Semi-quantitative PCR

The semi-quantitative PCR is technique that allows a quantitative comparison between two different templates and to estimate the number of gene copies, normalized with respect to a reference gene. Quantification of mitochondrial DNA copies was obtained by using as nuclear DNA references two genes coding for fructose 1,6-bisphosphate aldolase (FBA) and actin (Act1) that are known to be present in single copy in the genome. The genes used as reference for the mitochondrial genome are Cox 2 and Cox 3, respectively, coding for the subunit II and III of Cytochrom C oxidase.

Primers Construction

PCR primers construction is an essential aspect for the success of the reaction. Suitable primers for semi-quantitative PCR, in particular, should be very similar among them and should give an amplify of similar length. They were designed using the software Primer 3 (<http://frodo.wi.mit.edu/primer3/>) and for further confirmation they were analyzed with the program Oligo Melting (http://promix.cribi.unipd.it/cgi-bin/ProMix/melting/oligo_melting.exe), which provides the melting temperature and GC content. Once you have selected your sequences of choice, these are produced lyophilized, re-dissolved in water at a concentration 10 mM and stored at -20°C

Table 3.1 Sequences of primer used for semi-quantitative PCR

Primer	Sequence
ACT1-fw	AATGCAAACCGCTGCTCAATCTTCTTCA
ACT1-rv	AATACCGGCAGATTCCAAACCCAAAACAG
FBA-fw	CTCCATTGCTGCTGCTTTTCGGTAACTGT
FBA-rv	GAACCACCGTGGAAGACCAAGAACAATG
Cox2-fw	GCTGCTGATGTTATTCATGATTTTGCTATTCC
Cox2-rv	GGCATATTTGCATGACCTGTCCACAC
Cox3-fw	TCCAACATGATGTCCAGCTGTAAATG
Cox3-rv	TGCTGCATTCACTATCTCTGATGGTGTT

Quick DNA extraction

Strains have been inoculated in liquid YPD at different glucose concentrations at 28°C for 12-18h until 1.8 OD_{600nm} was reached and cells concentration was normalized among strains. 200 µl of culture were centrifugated down at 1500 g and in a refrigerated centrifuge Eppendorf 4515R Centrifuge for 5 minutes and the supernatant was removed. After two wash with 2 ml of water the pellets were re-suspended in 30 µl of diluted Zimolyase (10µl of 26.7 mg/l and 20 µl H₂O), incubated at 28°C for 20 min and at 96°C for 10 min. After the lyses the pellet was centrifuged and the supernatant containing the DNA was diluted 1:10 with water. Proceed with the semi-quantitative PCR using this DNA as template. Mix and PCR cycles used are reported below:

Mix PCR	1X
Buffer 1X	1 µl
dNTPs 10 mM	0.1 µl
PRIMER F 10 µM	1 µl
PRIMER R 10 µM	1 µl
Taq 5 U/µl	0.1 µl
microlysate	2 µl
MgCl ₂	0.3 µl
water	4.5 µl
TOT:	10 µl

PCR cycle			
Step 1	1X	initial denaturation	9°C for 3 min
Step 2	30X	denaturation	95°C for 30 sec
		annealing	60°C for 30 sec
		extension	72°C for 30 sec
Step 3	1X	final extension	72°C for 5 min

PCR reactions were prepared in the same way in independent replicas for each sample in order to remove the same PCR from thermocycler at different cycles of amplification. In this way it is possible to see the exact moment at which the amplification product starts to become visible in the gel. Each PCR is then analyzed by electrophoresis on 2% agarose gel to visualize the results.

Genomic DNA Sequencing

Genomes were sequenced using the Roche 454 Genome Sequencer FLX system. This platform generates more than 1.25×10^6 (0.5 Gb) individual reads per run with read length of 400 bases. Although the per-base cost of sequencing of this system is higher than that of other next-generation platforms, it was chosen because the length of the reads produced is longer. Long reads are useful because they are easier to assemble in *de novo* assembly and in repetitive regions of the sequenced genome (24)(25). The protocol to prepare the libraries is based on the fragmentation of the DNA and on the emulsion PCR. Libraries have been prepared and sequenced by pyrosequencing (26)(27) at the Ramaciotti Centre for Gene Function Analysis (a not-for-profit facility located at the University of New South Wales, Sydney).

DNA Fragmentation

For the construction of the paired-end libraries, fragments of approximately 8 kb are necessary. To obtain fragments so long we have used, according to Roche protocol 454, the DNA Shearing Device HydroShear® (Gene Machines). This machine uses a syringe pump, which allows the control of pressure with which the DNA in solution is forced through a membrane with a very small hole. The sharp contraction of the diameter of the fluid path force the solution to accelerate to maintain its volumetric flow rate. The acceleration of the solution creates drag forces that increase until DNA is fragmented. The size of the fragments is determined by the speed of fluid flow, by the applied pressure and by the size of the hole. We have uploaded 7 g of DNA in a volume of 300µl that were subjected to 20 cycles with "Speed Code" set to 16 in order to obtain fragments of 8Kb length. Samples were loaded on 1% agarose gel and compared with the Marker™ DNA Ladders GeneRuler. The DNA also was quantified and the amount of DNA required by the company was freeze at -80°C.

Cesium Chloride Centrifugation

Separation of genomic from mitochondrial DNA was obtained by caesium chloride gradient with Hoechst dye labelling. For the nuclear genome isolation we used the following procedure. The volume of the sample was adjusted to 4.347 ml with TE buffer and added to a 50-ml tube. 4.565 g CsCl. 150 μ l of Hoechst dye from a 10mg/ml stock solution was also added and solution mixed well. Solution was transferred to the gradient tubes using a sterile siringe and sealed following the manufacturer instructions. Centrifugation was carried for 20 h at 55000 rpm in a Beckman Coulter Optima 4E-80K Ultracentrifuge with Vti 65-2 rotor at 17°C Bands were visualized with long wave UV light and removed using an 18 gauge needle attached to a 1 ml syringe.

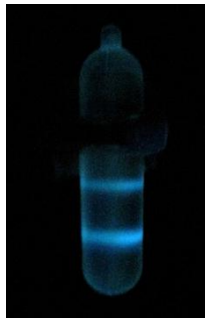


Figure 3.2 CsCl gradient separation of nuclear and mitochondrial DNA

Collected DNA was then diluted in three volumes of sterile mQ water and mix thoroughly, overlayed with 8.5 volumes of cold 100% ethanol and stored overnight at 4°C. Samples were centrifuged for 20 min at 20.000xg in corex tubes, pellet was washed twice first in 100% ethanol followed by 70% ethanol wash and air dried prior to resuspend in 50-100 μ l of water. Purification of nuclear DNA from mitochondrial DNA was verified using semi-quantitative PCR reaction with specific primers.

MATERIALS AND METHODS. BIOINFORMATICS

Sequence Assembly

The genomes of four strains of *S. cerevisiae* isolated from Prosecco and Raboso were sequenced with a combination of shotgun and 8kb paired-end libraries using the 454 GS FLX Titanium series chemistry (26). Raw data resulting from a 454 GS-FLX sequencing run consists of a series of digital images representing light emission during pyrosequencing reaction that takes place on the flow-cell into very small wells (44 μm). Images are analyzed and normalized to subtract background and to extract the raw signals that successively undergo normalization, correction, and quality filtering to generate Standard Flowgram Format (or SFF) files with base calls with associated quality scores for each individual read. Quality scores compute the probability that an individual basecall is correct. The depth of coverage is the average number of reads representing a given nucleotide in the reconstructed sequence produced by an assembly software. Sequence redundancy has been calculated identifying all the paired-end sequences with the same start and end coordinates of the two ends mapped to the reference genome of *S. cerevisiae* S288c and considered for coverage calculation.

High quality reads are used as input for different applications: Mira assembly software, Newbler Assembly, a software package for *de novo* DNA sequence assembly, and the Newbler Reference Mapper. The first two software generate a consensus sequence of the whole DNA sample, by assembly of the reads into contigs and then uses paired-end sequences info to order and orient the resulting contigs into scaffolds. The GS Reference Mapper application instead generates the consensus DNA sequence mapping of the reads on a reference sequence. Newbler is a software designed specifically for assembling sequence data generated by the 454 GS-series of pyrosequencing platforms. Mira is an open source multi-pass DNA sequence data assembler/mapper for whole genome and EST projects (28). The first step of the *de novo* assembly process is a complete all-against-all reads comparison to identify all possible overlaps between fragments. The set of all pair-wise overlaps between reads is used to merge these reads into unitigs. The second step is a contig optimization process that generates larger contigs from the unitigs. This step is based on an all-against-all unitig comparison to detect overlaps between unitig that can merge. In the end a quality controls are performed: contigs are broken in region where there are less than 4 spanning reads, and only contigs larger than 100 bases are output. All the assembled contigs are given as output by the program in a multi-Fasta file. Once contigs are assembled the scaffolding process starts. Two contigs can be inferred to be adjacent in the genome if one end of a paired-end sequence is assembled within the first contig, and the other end is assembled within the second contig. This step allows to create scaffolds defining contigs' order and orientation and the sizes of gaps between couples of contigs. Paired-end sequences with both ends mapping in the same contig are useful to validate contig assemblies. A multi-Fasta file with the sequences of all scaffolds, the structure of each scaffold with ordered contigs and the estimated gaps are given in output files.

In this project Newbler has been tested using different parameters but the default ones gave the best assembly results. To compute global assemblies the program was launched with and without paired-end sequences.

It was used also to compute local reassemblies giving as input only shotgun and paired-end reads mapping in the contigs that we wanted to assemble. The GS Reference Mapper works differently. The reference sequence can be used to guide the assembly of a genome using a process called comparative assembly. Reads are mapped to the reference genome and their placement is used to infer the structure of the sequenced genome. In this process care must be taken to avoid obscuring differences between the two genomes, anyway paired-ends provide a powerful tool for identifying large-scale misassemblies (19)(20). This application was used together with the *de novo* assembly to launch local reassemblies in regions not formerly assembled by Newbler during the *finishing* step.

GapResolution and Finishing Process

GapResolution is a software package used to help automate the process of closing intrascaffold gaps in Newbler assemblies. This software is not yet published but it can be obtained from Lawrence Berkeley National Laboratory, U.S. Dept. of Energy (<http://www.jgi.doe.gov/software>). The program considers all the gaps in the assembly and for each gap it identifies all the paired-ends reads with one end assembled in one of the two contigs flanking the gap (contig 1 or 2 in fig. 3.3) and the other end mapping somewhere else. If a defined number of these ends reside in a contig that is outside the scaffold (contig X in fig.3.3), that contig is assumed to be localized in the gap. It uses reads of the contigs adjacent to the gaps and reads of the identified contig to perform a local reassembly using Newbler to close the gap. This program should close gaps containing contigs or repeats collapsed on a contig. To check if the gap is closed short anchor sequences are created in un-repeated sequences in the two contigs flanking the gap (anchor 1 and 2 in fig.3.3). Left and right anchors are aligned to the Fasta file of all the contigs obtained from the local reassembly. If the anchors reside on the same contig and the distance is within the gap size (+/- standard deviation), the program gives as output a fake read representing the consensus sequence of the gap region.

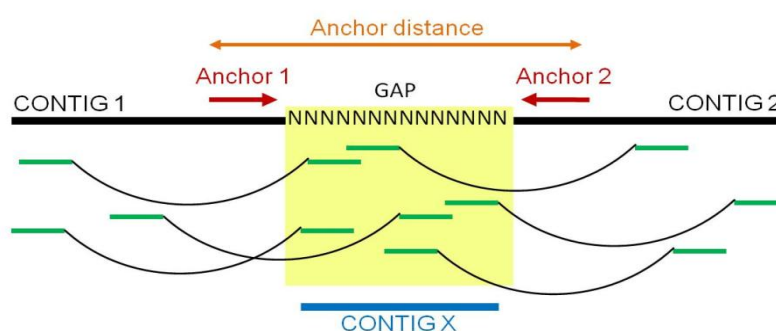


Figure 3.3 Cartoon explaining the mechanisms used by GapResolution to close gaps.

GapResolution outputs are fakes directory containing the Fasta and quality files of the consensus of closed gaps, and a directory for each gap analyzed containing resulting files from the local reassemblies and other files such as the anchor sequences. The program stitchClosed then takes all the fakes and uses the coordinates of the anchor sequences to replace the gap with the fake sequence in the output file of Newbler containing all the scaffolds. The resulting output of the stitcher includes a fasta file with fakes inserted in the scaffolds and a quality files with the quality of each base call of the assembly.

Poor results obtained using the program GapResolution in gap closure induce us to implement the program to try to solve more gaps. GapResolution is written using the programming language Perl. This programming language was used to modify the script to make it do different processes. The program was implemented to launch local reassemblies using both the *de novo* assembly and the reference assembly, and to create anchor sequences in forward and in reverse for all the gaps, augmenting the probability to find them. Another implementation was a new script to create a output file that should help to analyze all the reassemblies and create fakes manually for gaps where the program couldn't find anchors because of mismatches, and for where the reassembly closed the gap only partially. In this case the coordinates of contigs in the comparative assembly were used to estimate the number of "N" to completely fill the gap. Beside GapResolution and its implementation we wrote several other scripts using the programming language Perl. In fact GapResolution was able to close only the intrascaffold gaps identified by the program. To solve gaps between scaffolds, that GapResolution didn't consider, and also those that the program didn't close, we wrote a script based on the same general mechanism of GapResolution. This script takes as input the list of all the contigs that we want to use for local reassembly, and it launches both the *de novo* assembly by Newbler and the reference assembly after having recovered all the reads mapping in the considered contigs. The anchors and the results from the local reassemblies were used to create fake sequences as explained in the paragraph presenting GapResolution working mechanism. Fake sequences created were used to replace gaps in the scaffold automatically with the program stitchClosed.

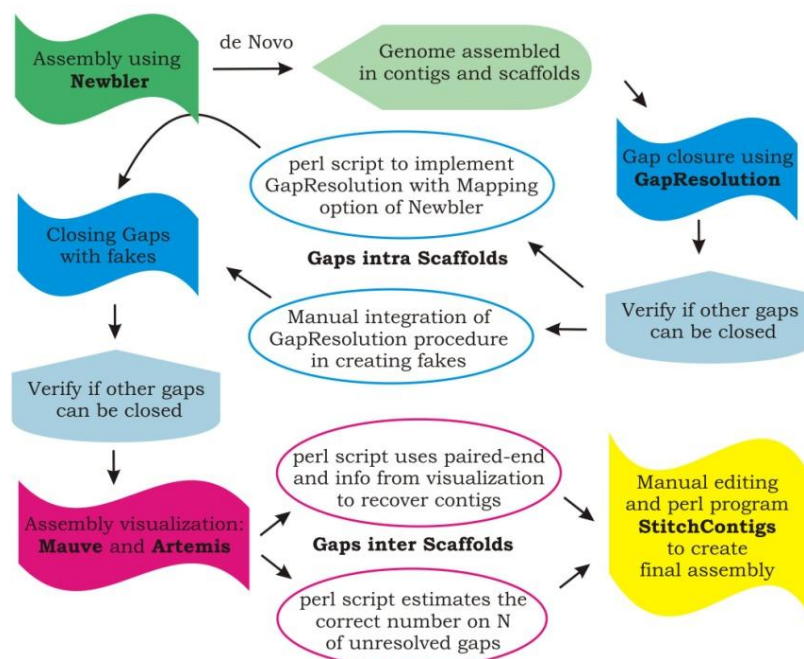


Figure 3.4 Flowchart resuming the strategy used for the finishing process. After *de novo* assembly by Newbler, gaps were progressively closed using different scripts launching locally both *de novo* and reference assemblies to create fake sequences that replace gaps.

Visualization tool and manual editing were used to identify sequences to fill the gaps and to stitch sequences. The last step of the finishing process was done with another script called StitchContigs. This script takes as input a list with the order of all the feature obtained using Mauve and Artemis software and manually verified.

It position correctly all scaffolds plus contigs that should be inserted into gaps, fakes sequences that were not replaced automatically and correct the numbers of “N” that should be put on the two sides of the contigs inserted into gaps, between scaffolds or in gaps where the original number of N estimated by Newbler was wrong. Then giving to the program a multiFasta file with all the sequences of scaffolds, fakes sequences and contigs, it uses the list to order and modify them and it gives as output a multiFasta file containing the definitive assembly of the genome.

Genomes Alignment and Visualization

To align and compare genomes we tried several programs and we eventually choose two of them: Mauve and Mugsy. Most of the programs we tried were not suitable for our needs because they align only couples of sequences at time or they can't use multiple scaffolds or chromosomes from different strains as inputs. On the contrary, Mauve and Mugsy are able to align several genomes and they accept as input multi-Fasta files. These characteristics were important in our analysis because we had to align several entire genomes composed by lots of scaffolds. Mauve and Mugsy both rely on the concept of LCBs (Locally Collinear Blocks) which represent homologous regions without rearrangements among the input genomes. Each LCB must be separated from the next by rearrangement in at least one genome. During the course of evolution, genomes undergo both local and large-scale mutational processes. Local mutations affect only a small number of base and include nucleotide substitution and insertion or deletion of nucleotides. Large-scale mutations can include gain and loss or duplication of large segments. LCBs allow the identification of conserved regions among the analyzed genomes and highlight large scale rearrangements such as gain or loss, duplication and inversion of large segments. Small indels and SNPs do not interrupt the extension of the LCBs (29)(30).

Mauve calculates LCBs using a process composed by several steps. It initially identifies multi-MUMs (Multiple Maximal Unique Matches) which are exactly matching subsequences shared by two or more genomes. They are used to infer a phylogenetic guide tree between the genomes to estimate the sequence similarity and set the weight criteria for the following steps. This value is called LCB weight and it sets the minimum number of matching nucleotides identified in a collinear region for that region to be considered true homology. A subset of multi-MUMs is then used as seeds and are extended and clustered together to create LCBs. Each LCB is required to meet weight criteria. Further analysis is performed between pair of genomes: the program search regions outside LCBs to extend them or to create new ones (29). An LCB is composed solely of regions shared by a subset of the genomes. Remaining unaligned regions are those which are duplicated or distinctive of a genome. Mauve was also chosen for two interesting tools: the “Order Contig”, and the Viewing System. Contig (or scaffold) boundaries represent potentially artificial LCB edges. Therefore, finding the contig order that minimizes the number of LCBs caused by contig edges is equivalent to finding a likely contig order. The tool “Order Contig” of Mauve use a reference genome to place the contigs of the query genome in the order that allow to reduce the number of LCBs extending their boundaries beyond contigs' edges (31)(4). This tool was used to align multi-Fasta files of the scaffolds of each sequenced genome with the reference genome of *S. cerevisiae* S288c to identify their probable order.

Mauve Viewing System is a graphic tool which displays the LCBs and rearrangements among aligned genomes in a user-friendly way. LCBs are represented by colored blocks and scaffolds/chromosome boundaries are represented by a red line. The viewer uses the first sequence to assign a reference orientation to LCBs in the remaining sequences. This tool was used to fasten the visualization of Mauve output and to have an immediate idea of the large scale differences between the aligned sequenced genomes.

Mugsy is a tool that combine several programs to optimize whole genomes multiple alignment process. The first step is an all-against-all pairwise alignment between the input genomes for identifying homology, rearrangements and duplications. A filter is then applied to identify matches likely to be orthologous and to report duplicated sequences present in only one of the genomes. Data from the pairwise alignment are used to build an alignment graph where each vertex represents an ungapped genomic segment and edges represent homology statements that passed the orthology filtering criteria. The alignment graph is then processed to identify LCBs. Then, as in Mauve, a multiple alignment for each LCB is calculated (32). Mugsy was used to align the four finished genomes with the reference genome of *S. cerevisiae* S288c and the genome of the enological strain EC1118 to get as output the alignment of all the homologous and non homologous regions of the genomes. We choose this program instead of Mauve because the output is given in a MAF format which was easier to parse for the next analysis, and because the coordinates of the start and end position of each alignment are relative to scaffolds and not to the beginning of the genome as in Mauve.

Artemis is a DNA sequence visualization tool that allows to examine the results of any analysis in the context of the sequence and its six-frame translation (33). A Fasta or GenBank format file is provided as reference to the tool and data from EMBL, GenBank and BAM format files are represented relatively to the reference sequence. This tool was utilized for several purposes. All the contigs, the shotgun and paired-end reads from each sequenced genome were mapped to the reference genome of *S. cerevisiae* S288c using BLAST. The output of BLAST was converted into a EMBL or BAM file containing all the relative positions of start and stop of each aligned feature. These files were visualized with Artemis to have an immediate representation of the position of all contigs, shotgun and paired-end sequences in relation to the reference. This procedure was used to clarify contigs order across chromosomes and to identify those positioned inside intrascaffold gaps. To validate the position we analyzed not only the mapped contigs but also shotgun and paired-end sequences. Translocation and gain or loss of segments of DNA were highlighted by the presence of lots of shotgun sequences aligned only partially in the region of the rearrangement or by paired-end sequences with the two ends mapping too far or too close each other (respect to the more probable insert size). The identification of contigs mapping inside intrascaffold gaps was also validate with a script written in Perl. This program calculates gaps dimension using the length of paired-ends sequences mapping in the two flanking contigs. Once estimated the gap size it looks for all the contigs with that size connected (i.e. one of the flanking contig, or the contig in the gap) by paired-end sequences to find the proof to the positioning of contigs. Artemis software have been used also to analyze transcriptome data. Reads obtained from SOLiD sequencing were aligned to the genome and their distribution across the chromosome have been visualized thanks to this tool to detect transcribed regions.

Gene Prediction and Annotation

Genome annotation was based on a combination of methods including transferring of S288c annotated ORFs with RATT software and *de novo* gene prediction with GeneMark software. The tool RATT (Rapid Annotation Transfer Tool), transfers annotations from a high-quality reference to a new genome on the basis of conserved synteny (34). We used this program to transfer the annotation of *S.cerevisiae* S288c downloaded from the Saccharomyces Genome Database (www.yeastgenome.org) to our sequenced genomes. RATT compares the query sequence to the reference genome to define regions sharing synteny. Then the annotation-mapping step associates each feature within a reference EMBL file with the new coordinates on the query. A feature is not transferred if it bridges a synteny break and if its coordinate boundaries match different chromosomes, different DNA strands, or if the mapped distance of its coordinates has increased by more than 20 kb. This program was used to transfer all the conserved genes between S288c and each assembled genome.

The GeneMark.hmm (35) algorithm was designed to improve the gene prediction quality in terms of finding exact gene boundaries. This program takes as input scaffold sequences and it defines a functional role of each nucleotide in the sequence specifying if they are part of non-coding region, if they reside in a gene sequence in the direct or in the complementary DNA strand. This prediction is performed using Hidden Markov Models based on probabilities calculated in the training step of the program and it generates stretches of DNA sequence with coding or non-coding statistical patterns. We have used a software version previously trained on the *S. cerevisiae* genome in order to reduce false positive identification and to obtain a high quality gene prediction. This program was used to validate RATT data and to predict genes not present in the genome of S288c, and to annotate possible gained regions of the oenological strains sequenced. Anyway the analysis on gained genes have not been performed yet because new regions should be checked by PCR before being considered to be sure that they are not assembly mistakes.

Comparison of Intergenic Regions

The aim of this project is to see if differences in promoter regions are correlated to differential gene expression. As reported in the introduction, we decided to look for differences in the entire intergenic regions upstream the transcription start sites of genes, considering genes having in their intergenic region transcription factor binding sites, tandem repeats or both. To identify differences in the intergenic regions among the four sequenced strains, and the two references S288c and EC1118, it was necessary to align the sequences in order to identify for each position of each genome the corresponding position on the other genomes. For this reason I wrote a Perl script that takes the output of the multiple alignment of the six genomes performed by Mugsy, and uses these information to create a database. Each line of the database corresponds to a position of the multiple alignment that can be present in all the genomes or only in a subset of them.

For each position the information provided by the database are: the consensus base among the genomes aligned in that position (base or “-” if that position is not present in most of the six genomes); the chromosome and the coordinate relative to the beginning of the chromosome where that aligned position is found in each genome; the differences in the base identity in that position with respect to the consensus for each genome;

the gain or loss of bases in that position for each genome. Positions included in a unique aligned block are consecutive in the database and each block is separated from the former one by a tag (see Fig. 3.5).

consensus	coordref	coordEC1118	coord1	coord2	coord3	coord4	SNPref	SNPEC1118	SNP1	SNP2	SNP3	SNP4	INDEC1118	IND1	IND2	IND3	IND4
>																	
G	chr13 NC001145*626001	chr13_1 FN393082_1M3*596080	chr13*592854	chr13*603492	chr13*2*343053	chr13*591294											
T	chr13 NC001145*626002	chr13_1 FN393082_1M3*596081	chr13*592855	chr13*603493	chr13*2*343054	chr13*591295										A	
A	chr13 NC001145*626003	chr13_1 FN393082_1M3*596082	chr13*592856	chr13*603494	chr13*2*343055	chr13*591296											
G	chr13 NC001145*626004	chr13_1 FN393082_1M3*596083	chr13*592857	chr13*603495	chr13*2*343056	chr13*591297											
G	chr13 NC001145*626005	chr13_1 FN393082_1M3*596084	chr13*592858	chr13*603496	chr13*2*343057	chr13*591298										C	
T	chr13 NC001145*626006	chr13_1 FN393082_1M3*596085	chr13*592859	chr13*603497	chr13*2*343058	chr13*591299											
C	chr13 NC001145*626007	chr13_1 FN393082_1M3*596086	chr13*592860	chr13*603498	chr13*2*343059	chr13*591300											
G	chr13 NC001145*626008	chr13_1 FN393082_1M3*596087	chr13*592861	chr13*603499	chr13*2*343060	chr13*591301											
T	chr13 NC001145*626009	chr13_1 FN393082_1M3*596088	chr13*592862	chr13*603500	chr13*2*343061	chr13*591302											
-	-	-	chr13_2*343062	-	C								ins				
C	chr13 NC001145*626010	chr13_1 FN393082_1M3*596089	chr13*592863	chr13*603501	chr13*2*343063	chr13*591303											
A	chr13 NC001145*626011	chr13_1 FN393082_1M3*596090	chr13*592864	chr13*603502	chr13*2*343064	chr13*591304											
G	chr13 NC001145*626012	-	chr13*592865	chr13*603503	chr13_2*343065	chr13*591305											
G	chr13 NC001145*626013	chr13_1 FN393082_1M3*596091	chr13*592866	chr13*603504	chr13*2*343066	chr13*591306										del	

Figure 3.5 Example of the database created using Mugsy output. For each genome the coordinates of the orthologous position are reported together with differences from the consensus sequence (SNPs, insertions or deletions).

Different scripts were written to analyze this database and extract information. The first two scripts were developed to find the corresponding coordinates of the transcription factor binding sites and of the tandem repeat sequences of the reference genome of S288c in the other strains, and to identify possible differences among the six genomes. For this step I used the transcription factor binding site annotated by Harbison (36) on the genome of S288c, while tandem repeats were identified in the genome of S288c using the program Tandem Repeat Finder (37)(11). This program takes as input DNA sequences and precisely identifies all the two or more contiguous and approximate copies of a pattern of nucleotides without the need to specify either the pattern or the pattern size. For each identified tandem repeat it calculates the percent identity, the presence of indels and other statistics.

The outputs of the two scripts highlighted all the differences present in these sequences in the six genomes and provided lots of data that were successively analyzed with statistical methods. Annotated features of S288c genome were downloaded from the Saccharomyces Genome Database (SGD) and used to identify intergenic regions regulating each gene. A Perl script was written to identify intergenic regions and state all their differences in TF binding sites, tandem repeats and other sequences. This script selects among the annotated features only coding sequences and tRNAs and uses their coordinates to define intergenic regions. Then the coordinates of the TF binding sites and tandem repeats formerly analyzed are used to allocate each of these sequences in the corresponding intergenic region. Sequences localized inside coding sequences are filtered in this step and only intergenic regions containing at least one TF binding sites or tandem repeats are kept. Regions positioned between two genes on the same strand regulate only the downstream element, indeed those between gene on different strands with start positions facing each other, regulate both the genes (see fig. 3.6).

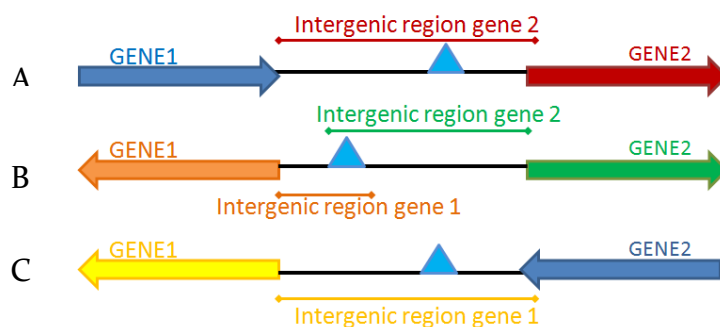


Figure 3.6 Cartoon representing the different kinds of intergenic regions: between genes on the same strain regulate only the downstream gene (A and C), between genes with transcription start sites facing each other regulate both the genes (B).

Differences among tandem repeats in intergenic regions were compared to those positioned in the entire genome. Variation in number of units and the percentage of conserved and mutated tandem repeat sequences were analyzed with respect to the number of repeated units and to the tandem repeat unit length. These data were analyzed, compared and tested using the R environment for statistical computing and graphics. The linear regression was performed to modeling the relationship between dependent and independent variables.

Differences identified among TF binding sites were analyzed using the hypergeometric distribution to see if there are sites bound by specific transcription factors mutating more than the others and to define if the increased or decreased probability of mutation is statistically significant. This distribution defines the probability to find randomly the same number of mutated sequences in a set of elements composed by the same number of the elements of the class taken from all the TF binding sites. The probability is given as a value comprised between 0 and 1 and values < 0.05 and > 0.95 are considered significant. The last step performed by the script is the identification of the differences present in the portions of the intergenic regions not comprised in TF binding sites or tandem repeat sequences. Only indels larger than 5 bps in these portions were considered because smaller deletions probably could not affect significantly the distance between regulative sequences and the transcription start site of the genes.

The script produces several outputs highlighting all the differences identified in the successive steps. The final output is the list of all the genes regulated by one of the considered intergenic regions, with a resume of all the identified differences in the six genomes. These information were used to perform pairwise comparisons between the reference S288c and the other strains.

Neighbor Joining Tree and SNPs

Phylogenetic relationship between the six strains have been computed using the tools Neighbor and Drawtree of the PHYLIP Package (38). The program Neighbor takes as input a matrix of values representing distances between strains calculated considering all the possible couples of them. This program implements the Neighbor-Joining method and the UPGMA method of clustering and computes unrooted trees by successive clustering of lineages. Distances calculated by this program are then given as input to Drawtree that draw an unrooted tree diagram.

These programs were used to create two kinds of trees. The first was produced giving to the program an input matrix calculated counting SNPs between couples of strains from whole genome alignments performed by Mauve. The second matrix was created determining the distance between different expression profiles calculated using the Pearson correlation coefficient and obtained by the comparisons of all the possible couples of strains. The first tree should highlight genetic distances between strain.

Mauve software produced a list of SNPs for each strain referred to genomic positions. Dedicated perl scripts were used to identify oenological specific SNPs and SNPs density along chromosomes, implementing R package to produce graphs.

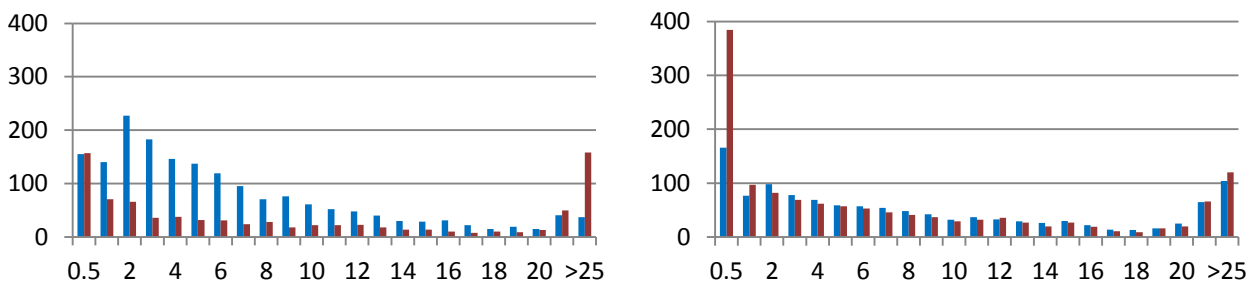
RESULTS AND DISCUSSION

Copies of the mitochondrial genome present in yeast cells are a major problem for the sequencing of the genome, because a large number of redundant sequences can significantly reduce the coverage of genomic DNA and greatly increases the mitochondrial DNA. This problem was solved using the CsCl gradient purification, which has allowed us to eliminate completely the mt DNA. As previously stated the sequenced strains are the derivative homozygous from natural heterozygous, from this point on they will be called P283, Roo8, R103 and P301.

Sequence Assemblies

The 454 sequencing results ensured a satisfying depth of coverage, calculated as sum of the shotgun and the paired-end (PE) sequences without redundancy. For PE sequences the coverage was calculated using all the sequences and also eliminating the redundancy due to the high number of PCR cycles performed during sample preparation. The total coverage was used to infer the theoretic percentage of sequenced bases of the genome using the simplified formula of the Poisson distribution $C=1-e^{-r}$ where r is the theoretic coverage. The depth of coverage obtained should ensure that more than the 99,999% of the genome should be sequenced for each strain.

The distributions of contig lengths obtained from the assembly without and with PE sequences are shown in the graphs below. The strain Roo8 is the strain with less shotgun sequences, but it is characterized by having the higher number of unique PE sequences among the other genomes. On the contrary the strain R103 has a high number of shotgun sequences but very few unique PE sequences. Comparing the two assemblies it is clear that in the strain with less shotgun sequences, the number of contigs is higher and the distribution of contig lengths is shorter. Good quality PE sequences can improve the quality of the assembly more than a higher number of shotgun sequences.



Roo8	Shogun only	shotgun & PE
Contigs	1789	872
Max length	45537	146514

R103	Shogun only	shotgun & PE
Contigs	1194	1360
Max length	81594	79799

Figure 3.7 Distribution of contig lengths obtained from the assembly of the shotgun sequences (blue) and of the shotgun plus the paired-end sequences (red) for the two strains Roo8 and R103.

All the strains have a similar contig length distribution but strain Roo8 having few shotgun and lots of PE sequences, displays the best assembly characterized by less contigs with higher lengths.

Maximum contig lengths are similar in all other strains and the number of contigs varies between 1360 and 1864. P284 and R103 display a similar contig lengths distribution. On the

other hand, the strain P301 has an higher number of short contigs (between 1 and 6 kb) and less contigs longer than 25 kb if compared with the other strains. This is probably due to the low number of shotgun sequences that were not balanced by unique PE.

The number of PE sequences obtained for each strain was very important for the quality of scaffold assemblies, in the R103 strain in fact most of the contigs could not be linked together into bigger scaffolds. It is important to note that the final quality of the assembly is not merely determined by the distribution of the scaffold size but also by the general quality of the sequence because a large number of gaps in the scaffolds could compromise the subsequent analyses such as the gene finding process.

Table 3.2 Statistics for assemblies determined using Newbler software.

	P283	R008	R103	P301
Scaffolds	147	67	514	215
Bp	11893389	11803514	11783464	12213932
Max length	898747	1127943	336729	944015
% sequenced genome	99.11	98.36	98.20	101.78
N50 scaffold	359400	662564	96711	529861
Contigs into scaffolds	928	597	819	1201
% contigs into scaffolds	60.03	68.46	60.22	64.43
contigs:scaffolds ratio	6.31	8.91	1.59	5.59

The number of sequenced bases predicted by the Poisson distribution was overestimated. This formula, indeed, is optimal for Sanger sequencing where all the considered sequences are unique, but not for next generation sequencing chemistries where lots of sequences are clonal because of the presence of PCR cycles in their protocols. Instead of the predicted theoretic 99,999% of sequenced bases, the sequenced bases are approximately the 95% of the genome considering the assembled contigs. The percentage of contigs assembled into scaffolds by Newbler varies from 60 to 68% in the different strains and is higher in strains with a greater number of PE. The total number of gaps (between scaffold and intrascaffold) left by Newbler in *de novo* assembly was quite different in the four sequenced strains and it obviously mirrored the shotgun and PE sequences qualities of each genome. To obtain the sequence of complete genomes by filling the remaining gaps in the assembly was very important for the following analysis on the genomes. We were interested in achieving a good assembly to have the possibility to identify large scale rearrangements and possible gained or lost regions by comparing the different strains. We were also looking at small scale differences such as SNPs and small INDELs, for this reason it was important to replace the greater number un unsequenced bases (N) in the scaffolds with suitable sequences to identify differences aligning the genomes

Gap Filling Results

To obtain the sequence of complete genomes by filling the remaining gaps in the assembly was very important for the following analysis on the genomes. We were interested in achieving a good assembly to have the possibility to identify large scale rearrangements and possible gained or lost regions by comparing the different strains.

We were also looking at small scale differences such as SNPs and small INDELs, for this reason it was important to replace the greater number un unsequenced bases (N) in the scaffolds with suitable sequences to identify differences aligning the genomes. The

procedure of filling gaps was quite challenging, especially for repeats and homopolymer, but it was important to try to solve them because one of my work aim was to identify differences in promotorial tandem repeats. The finishing process allowed to solve a great number of gaps using only bioinformatics methods. The percentages of solved intrascaffold gaps are quite similar between the different strains and it is a bit higher in the strains having more paired-end sequences. The percentages of solved interscaffold gaps (see table 3.3) are inversely correlated to the number of paired-end sequences (for the strain Roo8 the program for local reassembly was not used) because a large number of non-redundant paired-end sequences helped to generate very large scaffolds during Newbler assembly and local reassembly not improved substantially the final result. After finishing we got four high quality genomic assemblies (see table 3.3) composed by a low number of scaffolds which includes all chromosomes sequences plus some mitochondrial and 2-micron plasmid regions and some repeated regions such as telomeric and ribosomal (not all the assembled genomes have all the sequences of these regions). A rough comparison our result with the S288c genome indicates that the percentage of sequenced genome is higher than the 95% for all the strains, and the percentage of undefined bases left in the assembly is very low (less than the 2%). The results obtained are quite similar to those reported for other high quality yeast genome assemblies like EC1118 (4)(39) that was sequenced with the Sanger method or a combination of Sanger and 454-FLX methods.

Table 3.3 Gaps solved by the different programs and statistics after the finishing process.

	P283	Roo8	R103	P301
Scaffolds	147	67	514	215
Intrascaffold gaps	779	526	545	966
Gaps between scaffolds (approx.)¹	129	49	496	197
Gaps containing contigs	74	35	55	162
Intrascaffold Gaps				
GapResolution implemented	216	239	145	148
Local Reassembly	106	84	80	346
Solved intrascaffold gaps (%)	322 (41%)	323 (61%)	225 (41%)	494 (51%)
Interscaffold Gaps				
Solved by local reassembly (%)	39 (30%)	-	386 (78%)	12 (16%)
After Finishing				
Scaffolds	34	32	73	41
Genome size	11409448	11600348	11484928	11485677
% sequenced genome ¹	95.08%	96.67%	95.71%	95.71%
Number of "N" undefined bases	165915	173320	136495	212968
% of undefined bases	1.45%	1.49%	1.19%	1.85%

¹ - the percentage of sequenced genome was calculated with respect to the S288c haploid genome (12 Mb).

Once the finishing process was completed, good assemblies of the four genomes were obtained and it was possible to compare them each other, with S288c and with all other sequenced strains.

SNPs Distribution and Phylogenesis

The number of SNPs can be considered a measure of strain relatedness. From this measure we obtained the tree reported below. Oenological strains are strictly related on the basis of the number of SNPs identified; strains derived from other technological environments (beer, laboratory, sake, pathogens) are more distantly related to oenological strains. For SNPs analysis we have selected 18 *S. cerevisiae* strains among those with the best assembly quality in order to simplify the alignment process. Our aim was to classify our strains in comparison with other yeasts having different geographical location, ecology or associated with different fermentation technologies but we were not interested in a global population structure analysis since this is already been done (15,16,40). Strains selected comprise 11 wine strains (4 of these are the strains sequenced in this work) having different origin (commercial and wild type –ecotypical- isolates) (EC1118; P283; Roo8; R103; P301; AWR11796; RM11; QA23; VL3; VIN13; AWR11631), two strains involved in beer fermentation (FosterO; FosterB), one used in Sake production (Kyokai7), one used for bioethanol production, a clinical isolate (YJM789) and two laboratory strains (S288c and Σ 1278b). Polymorphisms were identified after genome alignment using MAUVE software for a total of 368408 SNPs. Pairwise SNPs difference in alignments were determined using dedicated PERL script and were used to determine a neighbour-joining tree using Phylip package. Heterozygous positions in the genome of diploid and tetraploid strains (27) were also taken into consideration as SNPs differences. It is clear from the phylogenetic tree that ecotypical strains clustered in the same lineage with all other wine strains (Fig 3.8) independently from their geographic origin, in fact in the same group we found strains isolated from Champagne fermentations (EC1118), AWR11631 (descended from N96 that is similar to EC1118), RM11 isolated from a California vineyard and QA23 that was selected in Portugal.

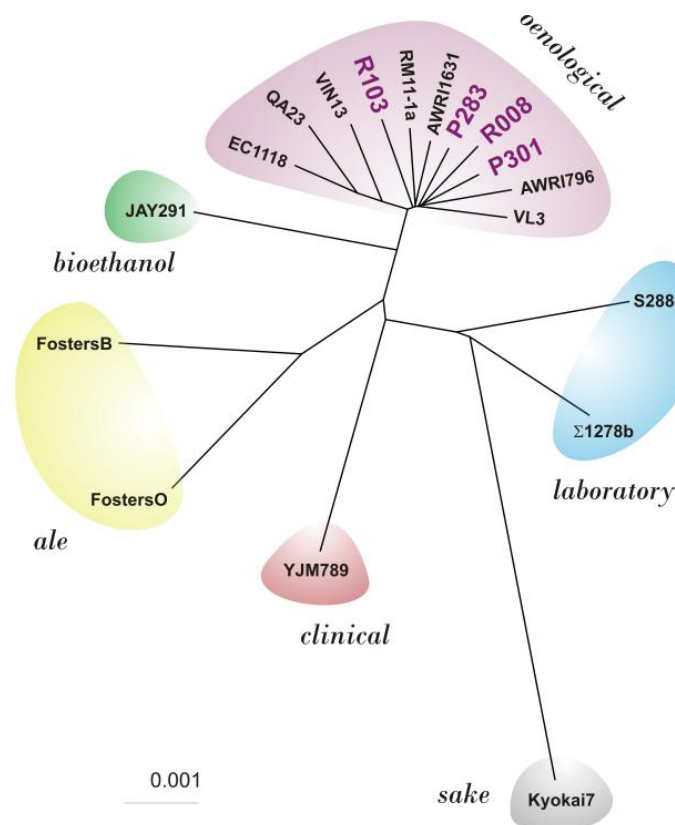


Figure 3.8 Neighbour-joining tree of 18 high quality assembly *S. cerevisiae* strains.

Since it is known that SNPs distribution in the *S. cerevisiae* genome is quite complex due to human traffic and subsequent recombination between strains of different geographic origin (15,41) we have analyzed this feature within a 10 kb block (window) along the 16 chromosomes using 1 kbp step. Analysis of the number of the 10 kb blocks having a number of SNPs ranging from 0 to 100 reveals three main distributions:

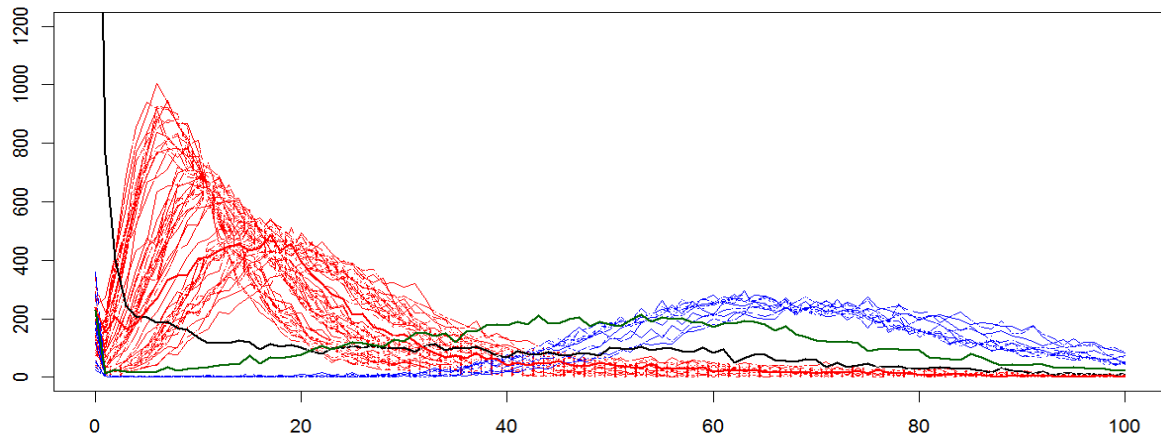


Figure 3.9 Each line represents a comparison between two strains, here only a selection of comparisons is shown. Each line reports the number of 10 kb regions (y) containing a given number of SNPs (x axe). Black line reports the comparison between S288c and Σ 1278b, red ones the comparison between oenological strains, green one the comparison between FosterO and FosterB and blue ones the comparison between sake strain Kyokay7 and oenological strains.

Strains derived from recent crosses, like S288c and Σ 1278b (black line), should show a non-random distribution of a high SNPs percentage. As previously evidenced (25) SNPs identified in comparison between these two strains are clustered together and, in fact, approximately 45% of the Σ 1278b genome have less than 1 SNP every 1000 bases.

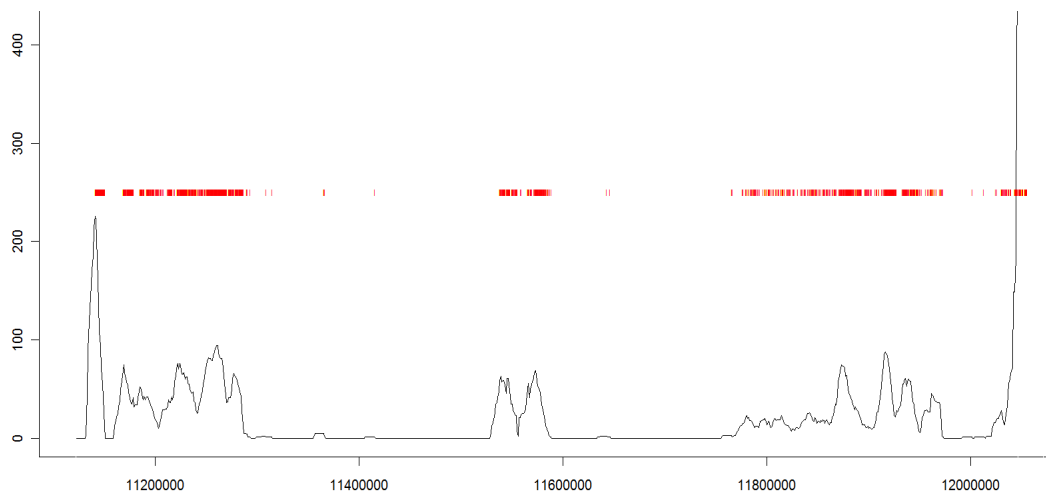


Figure 3.10 SNPs distribution along the genome determined using a 10 kb sliding window. Here we report the comparison between S288c and Σ 1278b performed on chromosome 16, red lines indicate SNPs positions along the chromosome.

A second interesting situation was found comparing commercial wine strains, in fact a small fraction of their genome has a very small SNPs frequency, while a large part has 5-40 SNPs in a 10 kb window. The black line in Fig. 3.11 represents the comparison between EC1118 and QA23 strains.

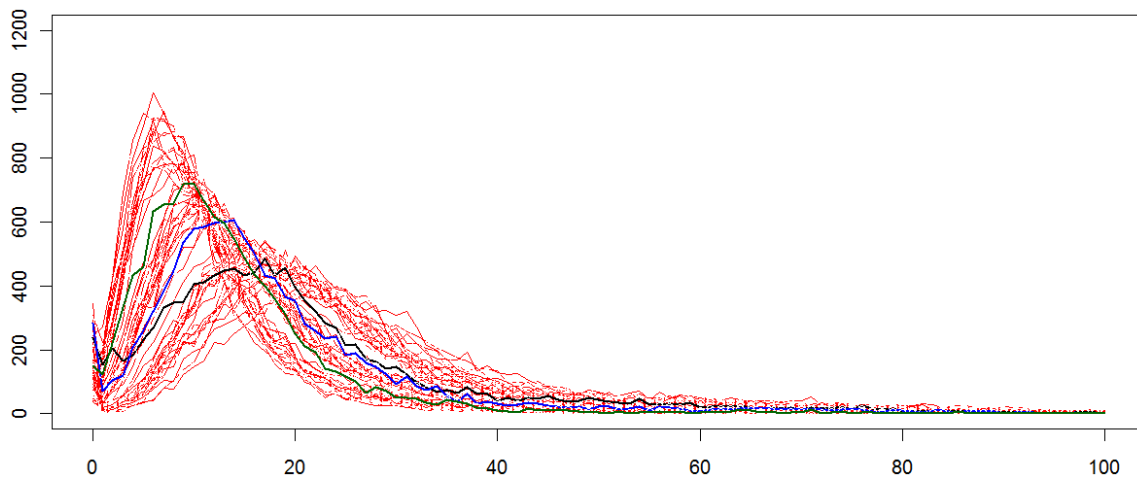


Figure 3.11 Each line represents a comparison between two strains, here only wine strains are reported. Each line reports the number of 10 kb regions (y axes) containing a given number of SNPs (x axes). Lines coloured in green, blue and black report the comparison between Roo8 vs. AWRI796, VL3 vs. AWRI1631 and EC1118 vs. QA23

A closer inspection revealed that chromosomes VIII and XVI of QA23 strain are very similar to those of EC1118 and together constitute the first “peak” in Fig. 3.11. The “mixed architecture” of the genome is also evident from the discrete genomic regions in chromosomes IV and XI that have approximately 50 SNPs per 10 kbp.

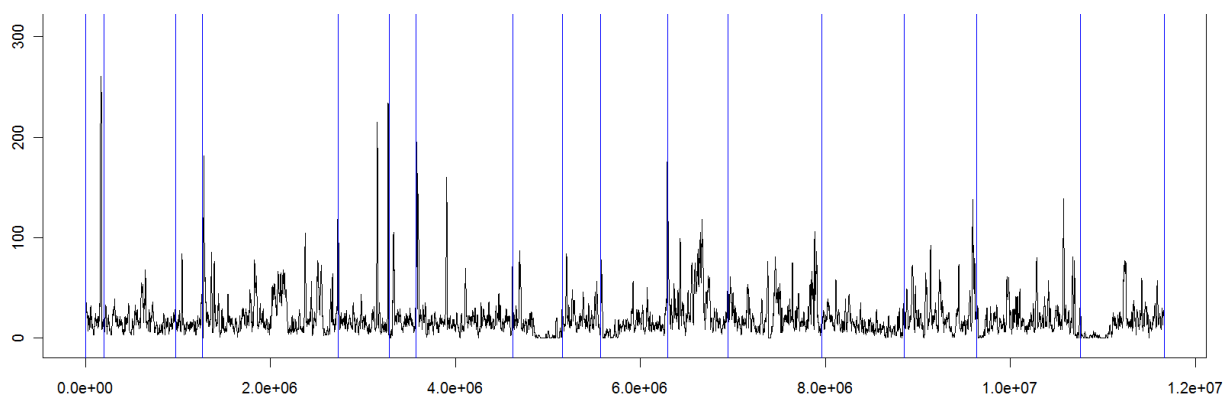


Figure 3.12 Whole genome comparison between EC1118 and QA23 strains.

Similar but less evident result was obtained for the comparison between VL3 and AWRI1631 strains (blue line in Fig. 3.11) and comparing some ecotypical and commercial strains (for example EC1118 and Roo8 – data not shown). On the contrary the blue lines in Fig. 3.9 report the comparisons between Kyokay7 (a strain used in Sake fermentation) and oenological strains. This distribution indicates that oenological and sake strains are very distantly related and do not have closely related genomic regions.

We have also analyzed SNPs identified in oenological strains and compared all bases present in oenological strains compared to all other strains. We found 315 positions that are conserved in all oenological strains but diverged in at least one of the other strains. Despite these position could be conserved because large part of wine yeasts are members of a single well-defined subpopulation and probably derive from a single (or a very small number) of domestication events (16,40) we cannot exclude that these are connected to the function of some genes in oenological environments.

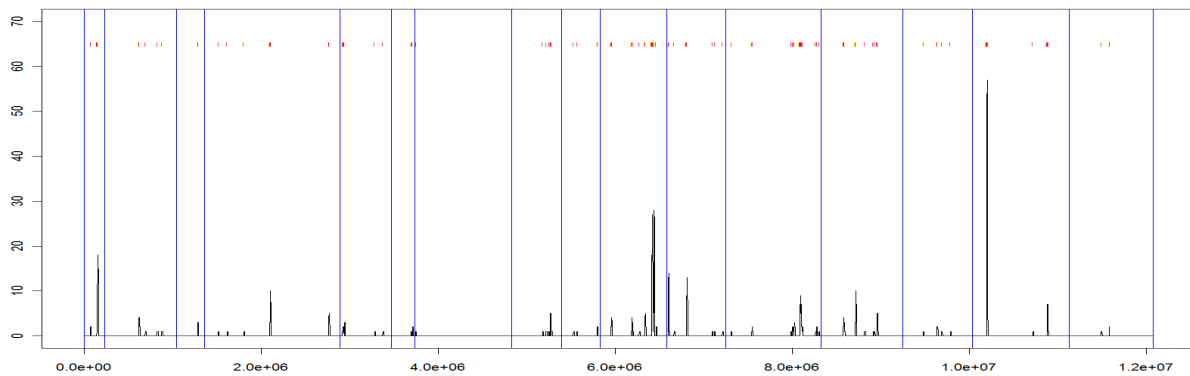


Figure 3.13 distribution of the 315 “oenological SNPs” respect to S288c genomic positions.

In order to gain a better understand of this point, we have analyzed these data using SNPeff software (<http://snpeff.sourceforge.net>) in order to classify SNPs respect to their effect on protein-coding genes (synonymous and non-synonymous changes, changes in upstream and downstream regions). Analysis reveals 89 non-synonymous changes (localized on 58 genes) considering S288c as a reference, three STOP codons gained (three genes), 108 synonymous changes (70 genes), remaining SNPs are localized in intergenic regions. As expected from results reported in Fig. 3.13, both SNPs determining synonymous and non-synonymous changes preferentially affect genes on chromosome X.

GO analysis performed using YeastMine website on genes groups showing synonymous and non-synonymous changes did not show highly significant results. Among genes showing non-synonymous changes, 19 belong to the “response to stimulus” class (p-value 0.0026), 6 to the “cellular nitrogen compound catabolic process” (p-value 0.0036) and 6 to the “response to organic substance” (p-value 0.005).

This result is not surprising because, as previously mentioned, during grape must fermentation, yeast are exposed to an hostile environment (high concentrations of sugar, high levels of ethanol, low pH, the presence of sulfites, and limiting quantities of nitrogen, lipids and vitamins, under strong anaerobic conditions). Since we have previously stated that oenological strains are strictly related from a technological point of view but have different geographical origin (Fig. 3.8), it is at a certain extent expected that genes involved in nitrogen utilization and catabolic process or in the response to specific organic substances or external *stimuli* can reveal evidence of natural selection.

Structural Variations

The genomes of the six strains were aligned using the program Mauve and the alignment was analyzed thank to the viewer tool. From the manual inspection of the alignment we identified some translocations and gained or lost sequences typical of a specific strain or conserved in more than one genome. Among translocations, in the genome of Roo8, nearly half of the chromosome XVII seems to be translocated to the chromosome VIII if compared to the genome of S288c. Other portions with variable length from 25 to 150 kb are translocated from the chromosome XV of S288c to the chromosome XVI and from IX to XIV in the genome of the strain P284, and from IX to XII in the genome of R103. Possible gained regions could be present on the chromosomes VIII, IX and XVI in more than one strain. Specific primers have been produced for four selected regions and the presence of these rearrangements have been successfully tested by PCR.

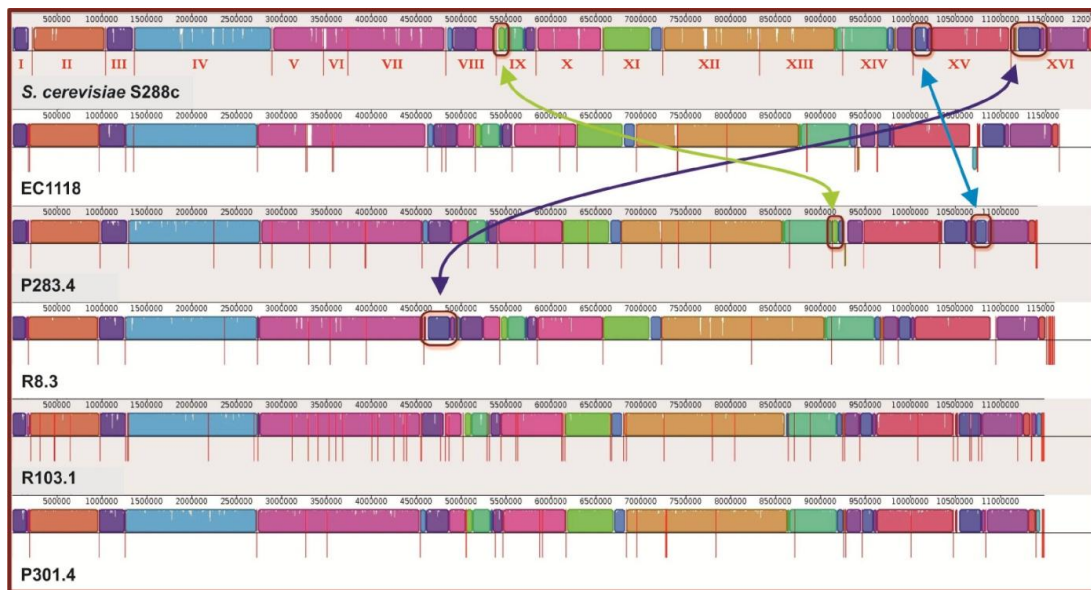


Figure 3.14 Visualization with the viewer tool of Mauve of the alignment of the six genomes.

Mauve tool allows to easily identify large rearrangements, and gained and lost regions by comparing the LCB blocks of the genomes. In the genome of the reference strain S288c which is completely sequenced, red lines represent chromosome boundaries, in the other genomes they represent scaffold boundaries. Most of the LCBs are conserved in the different genomes and are positioned in the same order. Small rearrangements are present in at least one genome between LCB boundaries. Large rearrangements are easy to identify. A schematic report of PCR that confirmed four translocation is listed in the following table:

Table 3.4 Schematic representation of PCR used to verify the four major traslocations

	primers	C	C-	primers	C+	C-	primers	C	C-	primers	C	C-
		+						+			+	
S288c	A-B chr III	A B	B C	A e B chr VII	AB	AC	A-B chrXVI	A B	AC	A-B chrXV	A B	A C
EC1118		-	-		AC			A B			A B	
P283		A B	A-C chr7_2		AC	AB/D C			AB/A C	A-C chr16_2	A C	A B
Roo8		A B	D-C chr7_2		AB/D C		A-C chr8	A C			A B	
R103		A B			AC	AB		A C			A B	
P301	B-C chr11	B C	A B		AB/D C			A B			A B	

All primers and corresponding sequences are reported in the Appendix II. A complete list of all translocation found can be found in the following table:

Table 3.5 List of all translocation found between chromosomes in the four strains

Traslocation	From	To	size (bp)	P283	Roo8	R103	P301
TR3-11	3	11	11000			x	x
TR6-10	6	10	18000	x	x	x	x
TR8-16	8	16	10000	smaller	x	chr8_1	x
TR16-8a	16	8	100kb		x	chr8or16	

TR16-8b	16	8	big	x
TR9-14	9	14	70000	x
TR9-13	9	13	25000	x
TR11-8	11	8	10000	x
TR15-16	15	16	150000	x
TR16-9	16	9	10000	x

Effect of this structural variation on gene expression will be discuss in the next chapter. A variety of regions absent in the reference genome and some major deletions have been found.

Some of these are common to other sequenced strains, some other are specific of our ones. In the following table all regions are listed and the correspondence for each strain is reported.

Table 3.6 List of all specific regions absent in the reference S288c strain and major deletion

Region	chr	length (Kb)	EC1118	P283	Roo8	R103	P301
A	6	38	EC1118_1F14	-	-	-	-
B	14	17	EC1118_1N26	-	-	-	-
C	15/9	65	EC1118_1O4	-	-	9_1	possible_chr9_tel
Roo8Oo1	15	39	-	-	15	-	-
P283Ao1	1	16	EC1118_1A28	1			
R103Io1	9	11	-	-	-	9_1	-
Roo8Ao1	1/8	10	-	1	chr8_tel	chr10r8	chr8or1
Roo8Ao1		10				8_8	chr8_tel
R103Po1	16	10	-	-	-	possible_chr16_1	16_1
Roo8Xo1	?	9	-	-	unknown	-	-
Roo8Jo1	10	7	-	10_1	10	-	10_1
R103Oo1	15	6	-	-	-	15_1	-
Roo8Go1	7	5	-	-	7_2	-	7_1
P301No1	10/2/14	5	EC1118_1J19	2_tel	14_tel	10_5	14_2
Roo8Po1	16	5	-	-	16	-	-
Roo8Bo1	2/9	5	-	-	chr20r9_tel	-	-
P283Go1	7	3	EC1118_1G1	7_2	7_2	7_8	possible_chr7_tel
Roo8Oo1	15	3	-	-	15	15_1	-
P301Xo1	?	4	-	-	-	-	scaffold00213_tel
Del 1	7	10	-	7	-	7	-
Del 2	14	5	-	-	14	-	-
Del 3	15	15	-	15	-	-	15
Del 4	1	15	-	1	1	1	1
Del 5	7	12	-	7	7	7	-

Genomes Annotation

The four sequenced genomes were annotated using the program RATT (34) which transfers annotations from a reference (in this case we used the annotation of S288c) to a new genome on the basis of conserved synteny. The program transfers all the annotated features, including coding sequences, tRNAs, ncRNAs and other sequences such as repeats, LTR and rRNAs. One of the main aims of the project is to identify orthologous genes that are differentially expressed between different strains, to do this it is important to have the greater number of annotated genes in each genome. From the annotated 6607 ORFs of S288c genome, RATT transferred from 5580 to 5722 features in the different strains. This is a good result considering that more than 800 ORFs of S288c are dubious and that we sequenced approximately the 96% of the total genome of each strain.

Table 3.7 List of all annotated features to the genomes of the four sequenced strains.

	P283	Roo8	R103	P301	EC1118	S288C
Protein coding genes transferred with RATT	6350	6370	6384	6384	6524	6711
Stain-specific protein coding genes	4	13	13	19	19	-
LTR	179	205	209	196	231	382
tRNA	246	257	257	249	271	299
rRNA	-	-	9	-	21	27
ncRNA	1687	1698	1712	1726	1701	1740
Protein coding genes of S228c missing	343	334	323	315	245	-
Total annotated features	8809	8877	8907	8889	9012	9159

The total number of transferred annotations of each strains varies from 6462 to 6546 and it is proportional to the number of sequenced bases of the genome and in inverse proportion to the number of undefined bases.

All 33 genes newly found in our strain and absent in S288c reference strain are reported in Appendix II. Most of them has been annotated while some remain with unknown function. Some very interesting genes have been found, such as a Putative fructose symporter (similarity with *Z. rouxii*), a medium chain alcohol dehydrogenase (similarity with *S. cerevisiae* RM11-1a and AWR11631), a fungal specific transcription factor domain, c6 zinc (similarity with *P. marneffeii*) and a putative glucose transporter of the major facilitator superfamily (low similarity with *C. dubliniensis*). In the next chapter their expression level will be discuss.

Finally the number of LTR and Ty elements have been compared to that of EC1118 and S288c strains. As reported previously (27, 15) wild type and oenological strains present a lower number of this elements when compared to the laboratory strain. Main results are reported in the following table while their influence on gene expression will be discuss in the next Chapter.

Table 3.8 Total number of Ty elements and LTR for each category

LTR							
Strain	total	delta	sigma	tau	omega	other	unique
S288c	368	287	41	34	6	0	163
EC1118	202	145	24	14	4	15	22
P283	175	124	19	15	2	14	17
Roo8	193	136	16	23	4	14	5

R103	187	127	27	16	3	14	18
P301	171	121	21	14	3	12	25
Ty elements							
Strain	total	1 or 2	3	4	5	other	unique
S288c	49	41	2	3	0	3	38
EC1118	6	6	0	0	0	0	2
P283	9	9	0	0	0	0	3
Roo8	9	9	0	0	0	0	3
R103	5	4	1	0	0	0	3
P301	5	5	0	0	0	0	0

Transcription Factor Binding Sites

For this analysis we started from transcription factor binding sites annotated by Harbison *et al.* (36) in S288c. Among the 3337 sites 88.2% are present in all considered strains and we found that 6% of these are mutated in at least one strain. These mutations are comprehensive of 98 SNPs, 26 INDELs and 39 are both SNP and INDEL. It has been calculated if the frequency of mutation of each kind of TFBS is higher (red) or lower (green) than the expected with an hypergeometric distribution. Using GO categories we have seen that over-mutated TFBS regulate genes involved in metabolism regulation and response to environmental stresses. Classes of genes regulated by these transcription factors regard for example sterol transport, fatty acid metabolic process in response to cold, salt tolerance, amine transporters, alkaline pH response, drug resistance and growth in response to glucose limitation.

Three classes of TF binding sites results less mutated than the expected and all of them are involved in regulation of the basal functioning of the cell, such for example those controlling the regulation of cell cycle progression from G₁ to S phase and regulation of the transcription by RNA polymerase I and RNA polymerase II and amino acid biosynthesis.

Table 3.9 TF with binding sites more (red) or less (green) mutated than the expected in the six genomes.

P-values were calculated using hypergeometric distribution.

TF	mutated	total	p-value	Regulated Functions
SPT23	8	114	1.6E ⁻⁷	fatty acid metabolic process, response to cold
SUT1	8	138	7.9E ⁻⁶	sterol transport
CIN5	12	408	1.7E ⁻⁴	mediates pleiotropic drug resistance and salt tolerance
NRG1	10	306	2.0E ⁻⁴	mediates glucose repression and negatively regulates filamentous growth and alkaline pH response
DIG1	12	456	5.1E ⁻⁴	negative regulation of invasive growth in response to glucose limitation
SNT2	3	78	6.9E ⁻³	computational analysis suggests a role in regulation of expression of genes encoding amine transporters
MBP1	1	690	0,99	regulation of cell cycle progression from G ₁ to S phase
REB1	2	1014	0,99	DNA binding protein which binds to genes transcribed by both RNA polymerase I and RNA polymerase II
GCN4	0	720	0,99	transcriptional activator of amino acid biosynthetic genes in response to amino acid starvation

For each TFBS it has been identified the corresponding gene putatively regulated. A region of 500bp upstream and 100bp downstream the transcription start site has been considered for all protein coding genes and ncRNAs to verify the presence of a TFBSs.

It has been selected 4106 couples of gene and regulative TFBS formed by 1967 features and 1423 TFBSs eventually redundant. Differences in the expression of these features have been analyzed to understand the influence of mutated TFBSs on their expression.

Tandem Repeats

For each TR it has been identified the corresponding gene putatively regulated. A region of 300bp upstream and 100bp downstream the transcription start site has been considered for all protein coding genes and ncRNA to verify the presence of a TR. It has been selected 374 feature TR-regulated considering both 318 unique features and 287 unique TR eventually redundant.

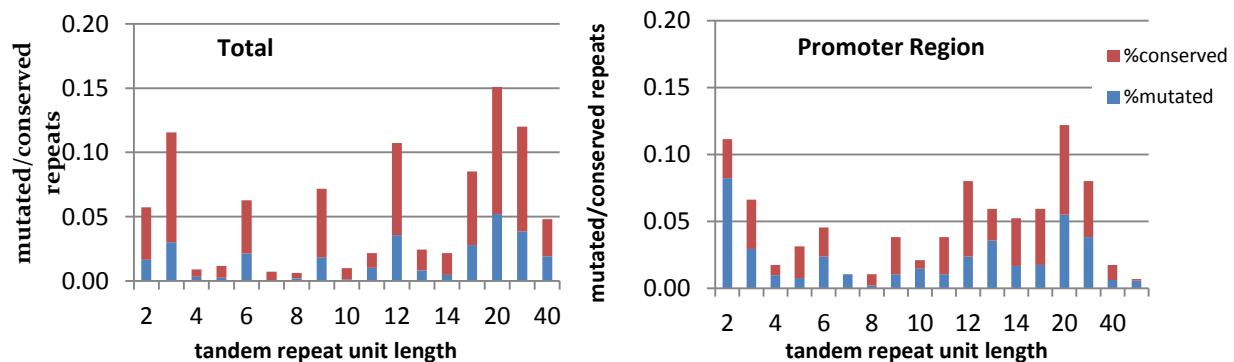


Figure 3.15 Percentage of differentially expressed genes with an expression variation between strains higher than 4 and 8 times than the reference S288c calculated for genes with intergenic region without tandem repeats and for those with tandem repeats with different level of differences in repeat length

The percentage of conserved and mutated tandem repeats was calculate for the two set of data also with respect to the specific unit length (see figure). As it concern the total set of tandem repeats, all the different classes of unit length seem to have the same ratio between mutated and conserved sequences. Repeats composed by di- and tri-nucleotides are significantly more mutated then those constituted by longer units. Short repeated units are possibly more prone to mutate thanks to slippage.

Gene Ontology of putative regulated genes

Mutated TR putative regulation of 133 genes enriched in transport, expression and biosintesis classes. Non mutated TR putative regulation of 94 genes enriched in lipids, membrane molecules, transport classes.

REFERENCES

- (1) Mortimer RK, Johnston JR. Genealogy of principal strains of the yeast genetic stock center. *Genetics* 1986 May;113(1):35-43.
- (2) Gu Z, David L, Petrov D, Jones T, Davis RW, Steinmetz LM. Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 2005 Jan 25;102(4):1092-1097.
- (3) Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, et al. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci U S A* 2007 Jul 31;104(31):12825-12830.
- (4) Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, et al. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci U S A* 2009 Sep 22;106(38):16333-16338.
- (5) Mortimer RK. Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res* 2000 Apr;10(4):403-409.
- (6) Bakalinsky AT, Snow R. Conversion of Wine Strains of *Saccharomyces cerevisiae* to Heterothallism. *Appl Environ Microbiol* 1990 Apr;56(4):849-857.
- (7) Bradbury JE, Richards KD, Niederer HA, Lee SA, Rod Dunbar P, Gardner RC. A homozygous diploid subset of commercial wine yeast strains. *Antonie Van Leeuwenhoek* 2006 Jan;89(1):27-37.
- (8) Ibeas JI, Jimenez J. Mitochondrial DNA loss caused by ethanol in *Saccharomyces flor* yeasts. *Appl Environ Microbiol* 1997 Jan;63(1):7-12.
- (9) Rachidi N, Barre P, Blondin B. Multiple Ty-mediated chromosomal translocations lead to karyotype changes in a wine strain of *Saccharomyces cerevisiae*. *Mol Gen Genet* 1999 Jun;261(4-5):841-850.
- (10) Carro D, Bartra E, Pina B. Karyotype rearrangements in a wine yeast strain by rad52-dependent and rad52-independent mechanisms. *Appl Environ Microbiol* 2003 Apr;69(4):2161-2165.
- (11) Dunn B, Levine RP, Sherlock G. Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC Genomics* 2005 Apr 16;6:53.
- (12) Louis EJ. The chromosome ends of *Saccharomyces cerevisiae*. *Yeast* 1995 Dec;11(16):1553-1573.
- (13) Longo E, Vezinhet F. Chromosomal rearrangements during vegetative growth of a wild strain of *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 1993 Jan;59(1):322-326.

- (14) Perez-Ortin JE, Querol A, Puig S, Barrio E. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res* 2002 Oct;12(10):1533-1539.
- (15) Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. *Nature* 2009 Mar 19;458(7236):337-341.
- (16) Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 2009 Mar 19;458(7236):342-345.
- (17) Roach JC, Boysen C, Wang K, Hood L. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* 1995 Mar 20;26(2):345-353.
- (18) Romano P, Fiore C, Paraggio M, Caruso M, Capece A. Function of yeast species and strains in wine flavour. *Int J Food Microbiol* 2003 Sep 1;86(1-2):169-180.
- (19) Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010 Jun;95(6):315-327.
- (20) Winde JHd. Functional genetics of industrial yeasts. Berlin: Springer; 2003.
- (21) Birney E, Durbin R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* 2000 Apr;10(4):547-548.
- (22) Barnett JA. A quick procedure for anaerobic fermentation tests in the identification of yeasts. *Arch Mikrobiol* 1972;84(3):266-269.
- (23) Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual. 2nd ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 1989.
- (24) Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J. The next-generation sequencing technology and application. *Protein Cell* 2010 Jun;1(6):520-536.
- (25) Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, et al. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* 2003 Jan;163(1):79-89.
- (26) Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005 Sep 15;437(7057):376-380.
- (27) Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, et al. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. *PLoS Genet* 2011 Feb 3;7(2):e1001287.
- (28) Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 2004 Jun;14(6):1147-1159.

- (29) Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010 Jun 25;5(6):e11147.
- (30) Pretorius IS. Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. *Yeast* 2000 Jun 15;16(8):675-729.
- (31) Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 2009 Aug 15;25(16):2071-2073.
- (32) Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 2011 Feb 1;27(3):334-342.
- (33) Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. *Bioinformatics* 2000 Oct;16(10):944-945.
- (34) Otto TD, Dillon GP, Degraeve WS, Berriman M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* 2011 May;39(9):e57.
- (35) Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 1998 Feb 15;26(4):1107-1115.
- (36) Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004 Sep 2;431(7004):99-104.
- (37) Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999 Jan 15;27(2):573-580.
- (38) Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981;17(6):368-376.
- (39) Walker GM. Yeast physiology and biotechnology. Chichester, West Sussex: Wiley; 1998.
- (40) Legras JL, Merdinoglu D, Cornuet JM, Karst F. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol Ecol* 2007 May;16(10):2091-2102.
- (41) Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 2009 Mar 19;458(7236):342-345.

4. TRANSCRIPTIONAL PROFILES

INTRODUCTION

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods.

RNA Sequencing

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease. The key aims of transcriptomics are: to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5' and 3' ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions (1).

The SOLiD™ 3 platform, developed by Applied Biosystems, allows an enormous throughput (more than 20 Gb) but it produces short sequences (400 million sequences 50 bp long). The high number of sequences produced and the possibility to align them on the reference genome using specific algorithms (2) allows both the identification of the absolute expression level of the transcripts and the determination of their structure (3). Concerning oenological yeasts, only few published researches use the novel genomic approach (nobody cDNA sequencing for transcriptome analysis).

This method allows the identification of the 3' and 5'-ends of the transcripts, the study of intron/exon boundaries and analysis of genes that are difficult to identify using bioinformatics (such as for example small RNAs). These sequencing strategies are imposing a new standard in gene expression projects, in fact the dynamic range is higher than in microarray experiments allowing the analysis of genes expressed at very different levels. Moreover, gene expression is no more limited by oligos that are restricted to specific genomic regions such as in microarray experiments but is unbiased and directed to all the transcripts at a single base resolution. Recently developed genomic techniques allowed to carry out the precise mapping of both Mendelian and quantitative traits (QTL). In these projects the conventional breeding of haploid parental strains and phenotypical analysis of segregants are coupled with genomes sequencing to correlate the presence of DNA polymorphic sequences (SNPs) to phenotypic characters. All these methods can map the traits with a resolution ranging from 6 to 64 kb but the bulk segregant analysis seems faster and more cost-effective (4). This peculiar use of the modern sequencing methods is particularly effective when complex phenotypic traits.

Transcription Factors

Being short and having sequences that are directly recognized by proteins, these sites represent sequences that if mutated should induce a significant alteration in gene expression. The vast majority of the transcription factor (TF) binding sites lie between 100 and 500 bp in promoter regions upstream of protein-coding sequences. Detailed comparisons between different yeast species (5) (6) showed that changes in TF binding sites regulating stress related genes, are associated with higher expression variance than the other genes, anyway the difference was not significantly high for none of them. They verify if mutations affected the interaction between TFs and their binding sites and they found that the binding to mutated sites was lower. Anyway gene expression in most cases remained conserved suggesting that compensatory mechanisms rapidly evolve maintaining a stable expression patterns. Moreover, they highlighted that genes with unexplained differential up-regulation are often characterized by differences in the regions flanking TF binding sites suggesting that surrounding regions can be important for the binding of the transcription factor and possibly to modulate chromatin structure. These results suggest that to study if differences in promoter regions affect gene expression, the analysis of TF binding sites is not sufficient. Regions flanking these sites seem as much important to regulate the expression and there are other elements in intergenic regions that can have regulatory roles. Furthermore 25% of all gene promoters contain tandem repeat sequences. The comparison of 33 promoters containing tandem repeat (TR) sequences in seven different yeast strains displayed that 25 TRs differed in repeat units in at least one strains (7)(8). These sequences have higher mutation frequencies than the other genomic regions because during DNA replication they are more prone to slippage. Repeat variability was compared to expression variance and there are evidences showing that genes driven by these repeat-containing promoters show higher rates of transcriptional divergence (7). Variations in repeat length resulting in changes in expression have been reported to be correlate to changes in local nucleosome positioning that can affect the accessibility to the other promoter sequences. To unravel this complex problem, it should be necessary to analyze the complete intergenic region localized upstream the transcription start site of each gene in order to elucidate the role of these elements in regulating gene expression. This should allow to identify SNPs and insertions or deletions in TF binding sites, variation in tandem repeat lengths, and variation in the length of the DNA regions between these element and the transcription start sites that could affect the positioning of the nucleosomes or transcription factor proximity to the transcription start site. We expect that this analysis could provide a first glimpse on how differences in these elements can modify gene expression.

MATERIALS AND METHODS. MOLECULAR BIOLOGY

Total RNA extraction

The total RNA has been extracted from each sample using the RiboPure™-Yeast kit (Ambion) that combines cell disruption, phenol extraction and RNA purification. Extraction have been performed as explained by the protocol of the kit, starting from samples containing approximately 3×10^8 cells. Cells were resuspended into lysis buffer, 10% SDS and phenol-chlorophorm and were disrupted thanks to the mechanical action of the Zirconia Beads added to the sample and to the vortexing step. The aqueous phase containing the RNA was separated by centrifugation and collected. RNA was then purified using the filter cartridges provided with the kit. The quality and the quantity of the purified total RNA samples were measured using the Agilent 2100 bioanalyzer, the Nanodrop and running samples on denaturing gels. 4µg of each replica for each strain were pooled together and freeze-dried. The three replicates for each strain should ensure the minimization of random fluctuation in gene expression due to external conditions.

When purification was carried out to obtain RNA for library construction and sequencing, acidic phenol was used (phenol solution saturated with 0.1 M citrate buffer, pH 4.3 for molecular biology, Sigma) while in all the other cases standard basic phenol was used (phenol solution equilibrated with 10 mM Tris HCl, pH 8.0, 1 mM EDTA, for molecular biology, Sigma).

rRNA Subtraction

The total RNA extracted from cells includes the complete collection of all transcribed elements of the genome, comprising mRNAs, rRNAs, and regulatory RNA molecules such as microRNAs and short interfering RNAs, snRNAs, and other RNA transcripts of yet unknown function. Large rRNAs constitutes 90-95% RNA species in total RNA so to sequence the transcriptome it is important to eliminate as much as possible rRNA molecules because being so numerous most of the reads produced would be sequences of these molecules. mRNA enrichment using polyA-selection methods is the most common approach used to eliminate rRNA and collect mRNA molecules, but this technique do not enrich the complete transcriptome because most of the regulatory RNA molecules do not have the polyA sequence so they can't be present in the samples. To get the complete set of transcribed RNA molecule, we chose a different approach. The RiboMinus™ Transcriptome Isolation Kit (Invitrogen) was used to selectively remove large rRNAs (18S and 26S in yeast) from total RNA. More than the 98% of rRNA molecules should be removed using this approach, and all the other kinds of RNA should remain in the enriched fraction. Large rRNAs depletion have been performed as suggested by the RiboMinus™ Transcriptome Isolation Kit protocol. RiboMinus™ Probes labeled with a biotin tag plus the hybridization buffer were added to the samples of purified RNA. The probes selectively bind rRNA molecules in solution. Then streptavidin coated magnetic beads are added to bind the biotin tags of the probes molecules. Using a magnet is then possible to separate the beads and everything bound to them and collect only the aqueous solution containing the total RNA without the contaminating large rRNA molecules. RNA samples are then purified and concentrated using silica-based membrane columns (RiboMinus Concentration Module from Invitrogen).

mRNA deCAPping

Each sample then underwent Tobacco Acid Pyrophosphatase (TAP) and DNase treatment. These steps allowed to eliminate the 5'-CAP in the molecule of RNA to be sure to sequence also the 5'-ends of the transcripts. In fact the CAP can interfere with the ligation of the adapters used for the library preparation. Treatment with the DNase was instead used to remove contaminating DNA, and to subsequently remove the DNase and divalent cations from the samples. The treatment with the TAP enzyme (Wako Chemicals) have been performed resuspending the samples of purified total RNA into 12µl of water and adding the 5X TAP buffer plus 10 U of TAP enzyme suggested by the protocol, and incubating the reaction at 37°C for 40 minutes. The contaminant DNA was removed using the DNA-free™ kit from Applied Biosystems. The decontamination was performed as suggested in the protocol, adding the 10X DNase buffer and the DNase enzyme to each sample. Tests showed that the TAP buffer added in the former step didn't interfere with this reaction. After incubation at 37°C for 20 minutes, the inactivation reagent was added to the samples to inactivate the enzyme and then by centrifugation the aqueous phase containing the RNA was separated and collected. Pellet was resuspended in 11µl. 1µl was diluted 1:2 and used for Nanodrop quantification.

DecAPPED mRNA was checked for integrity by running the diluted sample in the Agilent Bioanalyzer chip.

Readings of RNA were done using Agilent bioanalyzer 2100 by the Microcribi and BMR genomics services using either nano or pico chips according to the concentration of the samples submitted. See the following table for details about concentration range:

Table 4.1 Agilent bioanalyzer chip in relation to sample concentration and type

Chip format	Total RNA	mRNA
NANO	Range 50-500 ng/µl	Range 25-250 ng/µl
PICO and 6000pico	Range 200-5000 pg/µl	Range 500-5000pg/µl

Samples were quantified using Nanodrop and diluted in mQ nuclease free water for submission.

SOLiD Libraries preparation

The RNA obtained was used to prepare the libraries using the SOLiD Whole Transcriptome Analysis Kit protocol. RNA was initially chemically fragmented adding the RNaseIII enzyme plus the provided 10X buffer and incubating the reactions at 37°C for 10 minutes. Fragmented RNA was then purified and concentrated using silica-based membrane columns (RiboMinus Concentration Module from Invitrogen). Yield and size distribution of the fragmented RNA was assessed using the Qubit Fluorometer (Invitrogen) and the Agilent 2100 Bioanalyzer. The optimal fragment sizes range is from 35 to 500 nucleotides, and the average size should be 100–200 nt.

Reverse transcription of the RNA to cDNA require the ligation of specific adapters to the RNA molecules. This step was performed adding to the fragmented RNA the Adaptor Mix, the provided buffers and the Ligation Enzyme and incubating the reaction overnight at 16°C. Then reverse transcription was performed adding dNTPs, the reverse transcriptase and its buffer and incubating at 42°C for 30 minutes.

The cDNA was then purified using MinElute PCR Purification columns (Qiagen). cDNA samples were run on pre-casted polyacrylamide gels to separate cDNA molecules with respect to the size. Regions of the gel containing 100–200 nt cDNA molecules were excised and saved. The cDNA from gel slices was amplified by PCR using specific primers binding the adapters. Couples of primers with different barcode sequences in one of the primer have been used for the different samples. The barcode, once sequenced, allows to assign the reads to the correct sample. The DNA obtained was then purified and its yield and size distribution was assessed again using the Agilent 2100 Bioanalyzer, NanoDrop and Qubit Fluorometer. It was important to know the concentration of each sample because they were then pooled together and the same amount of DNA should be taken from each sample to balance them and to obtain a similar number of reads for each condition and strain under analysis. Once having pooled together the right quantity of each sample, the obtained solution underwent the emulsion PCR step.

Emulsion PCR and beads enrichment

Emulsion PCR is a crucial step that allows to create beads covered by several DNA copies obtained through the amplification of the same single DNA molecule. It is important that each bead contains single strand copies obtained only from one DNA molecule and that all the obtained beads have DNA bound to them, for this reason it is important to balance accurately the number of beads and DNA molecules in the emulsion PCR.

The aqueous phase is prepared adding to the sample of pooled DNA all the elements provided and required to accomplish the PCR. Two kinds of primers are used, they specifically bind the DNA sequences of the primers used in the amplification step. Primer P₂ is present only in the solution prepared for the PCR, primers P₁ are provided in the solution but they are also bound to the magnetic beads. The magnetic beads covered by P₁, are added to the aqueous solution and then this solution is dispensed into the oil phase, and the mixture is emulsified by the ULTRATURRAX device. This instrument mixes the two phases to create small droplets of water separated by the oil. Each drop represents a micro reactor and the system is calibrated to obtain droplets containing a DNA molecule a bead and the PCR reagents. The emulsion is then dispensed in 96 well plates and amplification performed in a thermalcycler. At the end of the PCR beads are recovered and enriched. Beads enrichment allows recover only those beads which present correctly amplified DNA on themselves and discard nude and poorly DNA containing beads. This procedure uses polystyrene beads covered by single-stranded P₂ adaptors to capture template beads covered by molecules of DNA. Only the beads collected from this step can be used for sequencing. The last step before sequencing run is the modification of 3'-ends. In order to prepare the P₂-enriched beads for deposition and binding to the surface of the sequencing device, a dUTP is added to the 3'-end of the P₂ templates using a terminal transferase reaction.

Sequencing with the SOLiD system

Once 3'-ends modification is accomplished beads are ready for sequencing run. Each bead is covered by several copies of the same molecule of DNA having the structure shown in figure 2.5. The extremity having the sequence of the P1 primer is bound to the bead, the other end has the sequence of the P2 primer and is used for the binding to the surface of the sequencing device. The central part of the molecule contains the target DNA sequence, an internal adaptor and the barcode.

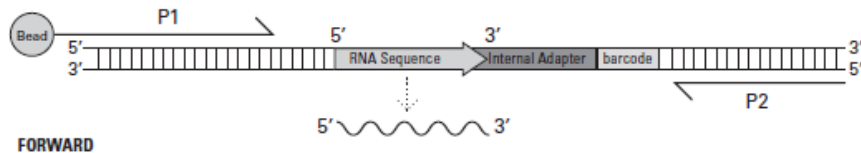


Figure 4.1 Structure of the molecules of DNA bound to the beads. The target sequence is flanked by the adapter P1 that during the sequencing is bound by the primer to start each round of ligations. On the other end of the molecule there is the barcode which is sequenced to know to which sample the sequence belong. Barcode is sequenced using the same mechanism used for the target, but ligation cycles start using primers binding the adapter P2.

An important step useful to verify the quality of the library before the sequencing run is the WFA (Work Flow Analysis). It is a quality control which is similar to the sequencing run but it uses only a small fraction of the sample to evaluate beads quality and polyclonal degree. For example, during this step the P2:P1 ratio is calculated to predict the number of optimal constructs (if the P2 adaptor is not present the DNA molecule bound to the bead it is not integer), and depending on the data from this run it is possible to predict how many beads we are going to deposit. After this procedure, the sequencing run is performed. SOLiD system is based on the sequencing-by-ligation technology (9). A primer is hybridized to the adapter sequence within the library template. Then a set of oligonucleotide octamers each labeled with a specific fluorophore among 4 colours, are added. In these octamers, the first and second bases are characterized by one of four fluorescent labels at the end of the octamer. Only the octamers complementary to the sequence of the DNA can bind the DNA molecule and only the octamers binding with the first two bases the two positions after the primer can be ligated to the primer molecule. At this point the fluorescence from the label is detected and bases 1 and 2 in the sequence are thus determined. The ligated octamer oligonucleotides are cleaved off after the fifth base, removing the fluorescent label, then hybridization and ligation cycles are repeated. Progressive rounds of octamer ligation enable sequencing of every five bases. Then the extension product is removed and the other round of ligation cycles are performed, starting from a different position in the DNA template. After five rounds the sequence is completely determinate (10). Reads obtained from the sequencing run are encoded in “Colour Space”, each base position is described by two colours and, knowing the identity of the first position (inside the adapter sequence) and using particular rules, it is possible to convert colours into base calls. For some applications sequences are used with the “colour space” coding because this facilitates reads alignment and the identification of true differences (SNPs) and sequencing errors. The SOLiD™ 3 System should generate approximately 300×10^6 reads (30-50 Gbp) per run with reads that are 50 bases long (10). With the current version of the sequencing system it is not possible to produce longer sequences because for every cycle the background noise increases and the quality of the fluorophore detection and of the sequence decrease.

MATERIALS AND METHODS. BIOINFORMATICS

Reads Alignment and Differential Expression

Standard DNA alignment programs are inadequate to manage the data produced by new generation sequencers. To answer this problem, PASS software have been developed with the objective of improving execution time and sensitivity when compared with other available programs (2). PASS performs fast gapped and ungapped alignments of short DNA sequences onto a reference DNA, typically a genomic sequence. It is designed to handle a huge amount of reads such as those generated by Solexa, SOLiD or 454 technologies. The algorithm is based on a data structure that holds in RAM the index of the genomic positions of 'seed' words (typically 11 and 12 bases) as well as an index of the precomputed scores of short words (typically seven and eight bases) aligned against each other. After building the genomic index, the program scans every query sequence performing three steps: (1) it finds matching seed words in the genome; (2) for every match checks the precomputed alignment of the short flanking regions; (3) if passes step 2, then it performs an exact dynamic alignment of a narrow region around the match. The performance of the program is very striking both for sensitivity and speed. For instance, gap alignment is achieved hundreds of times faster than BLAST and several times faster than SOAP (11), especially when gaps are allowed. Furthermore, PASS has a higher sensitivity when compared with the other available programs. This software has been used for all reads alignment performed in this work. Outputs of PASS were used as input for a script specifically developed to calculate for each gene the mean coverage of RNASeq reads mapping on them. This value is a direct measure of the level of expression of each gene and was used as input for the DegSeq (12)(13).

Up to now, there are few handy programs for comparing RNA-seq data and identifying differentially expressed genes from the data, although some recent publications have described their methods for this task (14-16). For our data analyses we preferred DEGseq, a free R package software. Two novel methods along with three existing methods have been integrated into DEGseq to identify differentially expressed genes (12). The input of DEGseq is uniquely mapped reads from RNA-seq data with a gene annotation of the corresponding genome, or gene (or transcript isoform) expression values provided by other programs like RPKM (17). The output of DEGseq includes a text file and an XHTML summary page. The text file contains the expression values for the samples, a P-value and two kinds of Q-values for each gene to denote its expression difference between libraries.

Expression data for each orthologous gene identified using RATT in every genome were matched with the tags representing the differences in the intergenic regions assigned to each gene in the genomic analysis, and these data were used to study how these differences influence the transcription. Classes containing genes with conserved intergenic regions have been chosen as reference and the other classes with mutations were compared to them. Couples of distributions were compared using statistical tests of the R environment. We used the F Test to compare variances of two distribution from normal populations. The null hypothesis is that the ratio of the variances is equal to zero. However these sets of data don't have normal distributions, so results from this test don't represent completely the model. For this reason we used the Kolmogorov Smirnov Test which define if two sets of data derive from the same distribution.

Hierarchical Clustering using TMEV

TIGR MultiExperiment Viewer (TMEV), one member of the suite of microarray data analysis programs is an application that allows the visualization of gene expression data (RNA-seq or microarrays) and the identification of genes and expression patterns of interest (18).

TMEV is composed by several modules, useful to perform different types of analysis in the same work session. Each program implemented in TMEV has a dialog window where the user can insert the parameters of interest.

MEV can interpret different file formats, including the MultiExperiment Viewer format (.mev), the TIGR ArrayViewer format (.tav), the TDMS file format (Tab Delimited, Multiple Sample format), the Affymetrix file format, and GenePix fileformat (.gpr). In my analysis the input file, a TDMS file, contains a matrix of $\log_2$ ratio expression values for each gene (rows) in each strain or condition examined (columns). $\log_2$ ratio expression values were calculated considering absolute expression values (number of uniquely mapped reads in the coding region of each gene identified) respect to the average value of each gene in all strains and conditions considered in gene expression experiments.

$$\log_2 (N_i/N_{iav})$$

“ N_i ” is the number of reads for the gene “i” in one strain and in one of the two conditions analyzed, while “ N_{iav} ” is the average number of reads of the gene “i” calculated considering all strains (in which the genes is present) and conditions. To perform an unsupervised cluster analysis I used the HCL (Hierarchical Clustering) module of TMEV, an agglomerative algorithm that arranges genes and strains according to similarity in the gene expression pattern. The object of a hierarchical clustering is to compute a dendrogram that assembles all elements into a single tree. For any set of “n” genes, an upper-diagonal similarity matrix is computed, which contains similarity scores for all pairs of genes. The matrix is scanned to identify the highest value (representing the most similar pair of genes). A node is created joining these two genes, and a gene expression profile is computed for the node by averaging observation for the joined elements. The similarity matrix is updated with this new node replacing the two joined elements, and the process is repeated “n-1” times until only a single element remains.

Agglomerative algorithms begin with each element as a separate cluster and merge them into larger clusters. An important step in any clustering process is to select the method to measure the distance between two clusters, which will determine how the similarity of two elements is calculated. This will influence the clustering, as some elements may be close to one another according to one distance and further away according to another. TMEV allows to calculate the distance with different approaches, in this study I chose the Euclidean distance method. Another parameter to set is the “Linkage Method” that indicates the approach used for determining cluster-to-cluster distances, when constructing the hierarchical tree. I used the “average linkage” method as a measure of cluster-to-cluster distance. The cluster analysis visualization of TMEV consists of colored rectangles, representing genes expression values. Each column represents all the genes from a single experiment, and each row represents the expression of a gene across all experiments.

The default color scheme used to represent expression level is red/green (red for overexpression, green for underexpression); black rectangles are not-differentially expressed genes and green those that do not have assigned value (NA). In the upper and left part of the graph is reported the dendrogram structure that represents the correlation between genes (or experiments).

Gene Ontology

Genes significantly differentially expressed in oenological strains with respect to the reference S288c have been selected and Gene Ontology categories significantly enriched in these genes were identified using the YeastMine tool (<http://yeastmine.yeastgenome.org/yeastmine/begin.do>). This program takes as input the two lists of genes: the total set and those with a characteristic of interest, in this case the differential gene expression, and it use the Gene Ontology database to identify biological processes, molecular functions and cellular components typical of the genes on the lists provided. This program automatically classify all the input genes in biological categories simplifying the subsequent biological data interpretation. Genes belonging to categories which are over-represented, are identified thanks to statistical test performed by the program. Output files with statistics on each gene and on the identify classes are produced (19)(20).

RESULTS AND DISCUSSION

RNA-seq Results

RNA was extracted from each sample. Approximately 95% the total RNA is constituted by large rRNA molecules. It is important to eliminate them before the sequencing because being so most of the reads produced would be sequences of these abundant transcripts. rRNAs were subtracted from the samples using a specific kit that should remove 98% of the rRNA molecules as described in the par. 2.7. This means that after the subtraction of the rRNA, at least half of the molecules of the sample will be rRNA. After rRNA subtraction the quality and quantity of the samples were measured. Figure 3.8-A and 3.8-B shows the RNA profiles of two samples calculated by the bioanalyzer (Agilent). Molecules of RNA have lengths varying from 50 to some thousands of nucleotides. Length distribution shows that the RNA is integer because most of the molecules are longer than 500 nucleotides. The two higher peaks correspond to molecules representing residual rRNAs 18S and 26S and these profiles show that after subtraction the rRNA contamination is still high, so their presence will be probably mirrored by the RNAseq results.

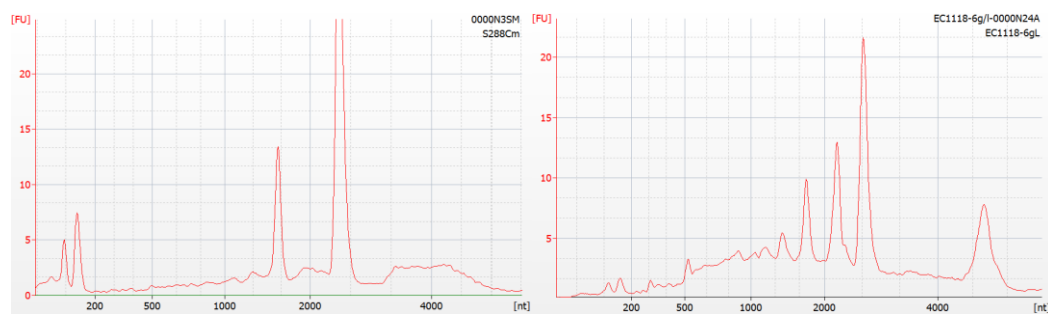


Figure 4.2 RNA profiles of two samples calculated by the bioanalyzer (Agilent). Length distribution shows that the RNA is integer and that contaminating rRNAs are still present after subtraction. In the sample of figure B the subtraction was more efficient than the sample of figure A, in fact peaks are lower and the amount of total RNA molecules is greater.

RNA-seq was performed using the SOLiD sequencer of the CRIBI Biotechnology Centre in collaboration with Prof. Giorgio Valle. Approximately 657 millions of beads were deposited on the surface of the sequencing device but only 633 millions of them were efficiently detected by the camera used to acquire the light emitted by the fluorophores during each ligation cycle. Only 585 millions of reads passed the quality controls and were effectively reported in the Fasta output file. For each analyzed strain and condition we obtain an average amount of 49 millions of reads. Some samples, such as EC1118 45g/l and R8.3 45g/l, display a quite lower or higher total number of reads (see table 4.2). This depends probably on errors in the quantification of the different samples before pooling them together. Anyway the order of magnitude is the same for each sample, so these differences doesn't cause problems during the comparison of the RNASeq data if the correct normalization is used.

Reads were aligned to the corresponding genomes using the software PASS (2)(21). PASS filters further on the reads and keeps only the high quality ones. It then uses these reads to perform the alignment. Not all the reads are successfully aligned by PASS, and this step represent a further filters on the number of reads. Among the aligned reads, a fraction of them are uniquely aligned, others align in more than one position.

Reads uniquely aligned are the target of the analysis because they are those that can be used to calculate the expression profile of each sequence. Reads mapping in more than one position represent those mapping in repetitive regions of the genome such as those coding for rRNAs and other repetitive elements. Eliminating as much rRNA as possible from the sample of total RNA is important to avoid to obtain lots of reads mapping on repeated regions at the expense of the uniquely mapped reads that are the more useful to create expression profiles. Table 4.2 clearly demonstrate that due to the high number of filters imposed during the different steps of reads detection and alignment, it is important to start from a high number of beads to be sure to get lots of uniquely mapped reads. In fact only approximately 30% of the beads produce uniquely mapped reads.

Table 4.2 Statistics from the SOLiD sequencing run and the alignment of the obtained reads to the corresponding genomes performed by PASS. The table shows how for each subsequent step of the analysis reads are filtered to get only uniquely mapped reads that can be used to calculate the expression profile of the genomes.

Strain	genome size (Kbp)	Condition	fastq reads	% filtered by PASS	reads to align	reads aligned by PASS	% aligned	unique	% unique	% unique among
P283.4	11.41	6g/l	44269000	28.5	31640757	20146259	63.7	17546620	55.5	87.1
		45g/l	57237000	28.4	40988114	21409095	52.2	12075366	29.5	56.4
R8.3	11.60	6g/l	41590000	29.5	29323095	19179784	65.4	18116495	61.8	94.5
		45g/l	64381000	24.5	48622042	18553213	38.2	18285815	37.6	98.6
R103.1	11.48	6g/l	46473000	13.4	40268187	25375647	63.0	24971785	62.0	98.4
		45g/l	46648000	30.7	32317727	15458446	47.8	15306793	47.4	99.0
P301.4	11.49	6g/l	54233000	29.6	38177940	21413852	56.1	20572267	53.9	96.3
		45g/l	47594000	30.8	32950859	13645737	41.4	13456354	40.8	98.6
EC118	11.66	6g/l	48443000	28.1	34827241	28751608	82.6	15599019	44.8	54.3
		45g/l	36288000	25.2	27126775	14605312	53.8	9161217	33.8	62.7
S288c	12.16	6g/l	53663000	27.7	38809553	34937173	90.0	11070439	28.5	31.7
		45g/l	44888000	24.9	33712429	28418143	84.3	14075549	41.8	49.5

Gene Expression Level Results

Expression data analysis, considering all strains together, shows that 34-44% of genes and ncRNA have a mean coverage major than 10 (medium-high transcriptional level) at 45 g/l, it is a 4% higher compared to 30-40% of features at 6g/l. It is interesting to notice that SUT and SAUT are generally expressed at a lower level when compared to the global gene expression. In fact the same analysis performed excluding ncRNA highlights that 41-53% of genes have medium-high coverage at 45 g/l, that is a 5% higher compared to 36-48% at 6 g/l. This trend was found even more in genes specific of oenological strains, that present a percentage of genes medium-highly expressed increased of 10% at 45 g/l compared to expression at 6 g/l.

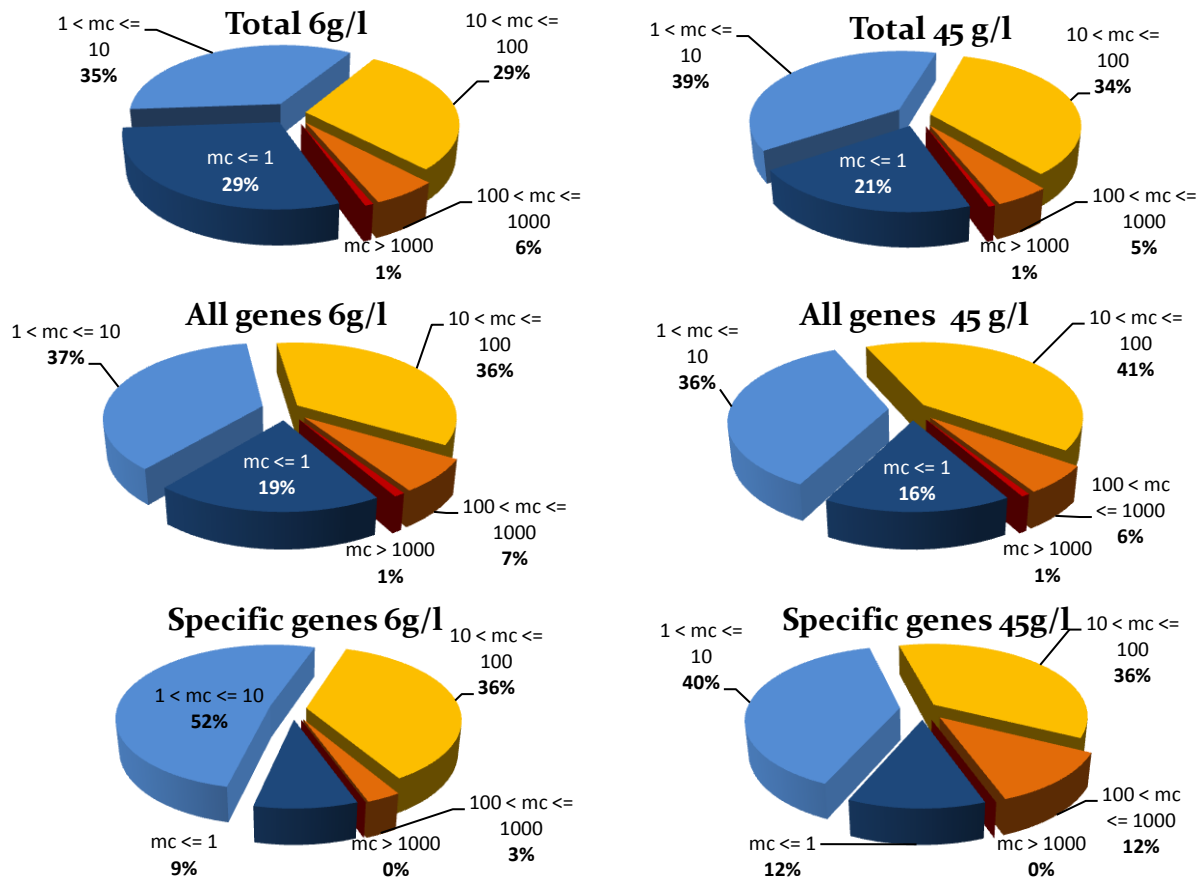


Figure 4.2 Graphs representing the mean transcription levels of total features of all strain respectively at the first time point of fermentation (6g/l) and at the second point (45g/l); same graphs are reported specific for total protein coding genes (ncRNA excluded) and for those genes found in our strains and absent in S288c

Considering the strain-specific transcriptional levels it is also possible to distinguish between strains with a lower number and a higher number of expressed genes in the two condition of fermentation.

At the beginning of fermentation process strains are in mid-exponential phase of growth curve and are exposed to high sugar concentration. It is interesting to notice that at this point a “non-fermentative” strain like S288c shows a higher percentage of medium-highly expressed genes respect to all the other “fermentative” strains. It is possible to hypothesize that this higher number of high expressed genes could be an attempt to balance the hostile environment in which other strain are already adapted. Strain P283 could be more adapted to fermentation condition due to this over-expression.

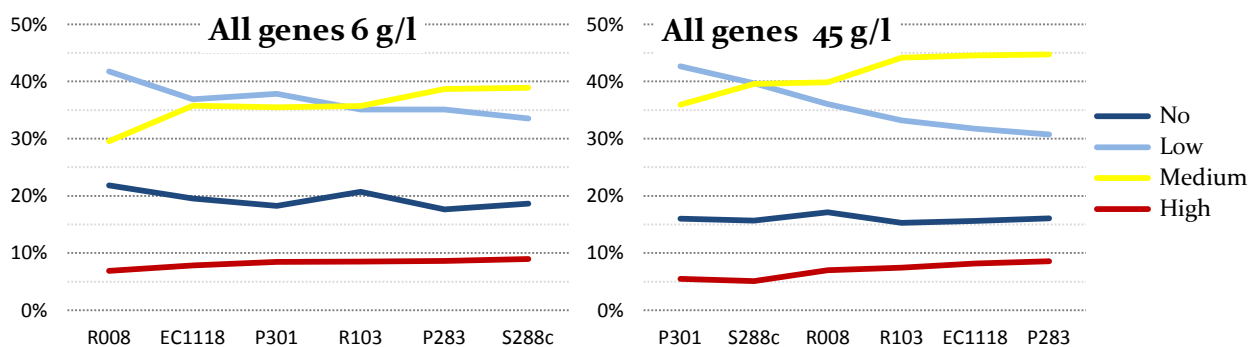


Figure 4.3 Graphs representing the strain specific transcription levels of protein coding genes respectively at the first time point of fermentation clustered according to their coverage levels in No (mean gene coverage ≤ 1), Low (1 < m.g.c. ≤ 10), Medium (10 < m.g.c. ≤ 100) and High (100 < m.g.c. ≤ 1000).

Looking at this trends we grouped strains with low (Roo8), medium (EC1118, P301, R103) and high number (P283, S288c) of expressed genes at 6 g/l. On the contrary in the second phase of fermentation curve strains are in early stationary phase, glucose concentration is reduced but ethanol level raised to 9% v/v. In this case strain P283, EC1118 and R103 show high percentage of medium-highly expressed genes, Roo8 and S288c medium number and P301 has the lower.

Trends in transcription level of ncRNA considering all strains are reported below.

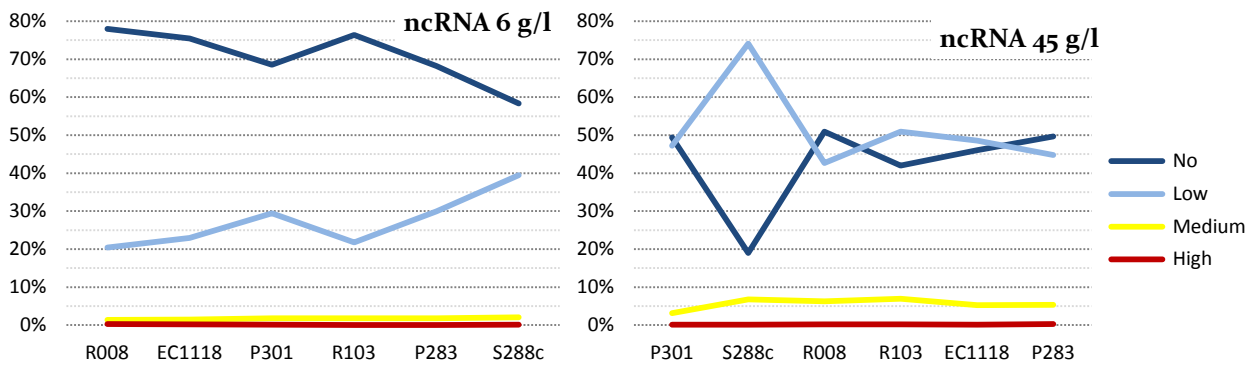


Figure 4.4 Graphs representing the strain specific transcription levels of ncRNA respectively at the first time point of fermentation clustered according to their coverage levels in No (mean gene coverage ≤ 1), Low (1 < m.g.c. ≤ 10), Medium (10 < m.g.c. ≤ 100) and High (100 < m.g.c. ≤ 1000).

Specific protein coding genes absent in S288c

We have found that genes identified in the four oenological yeast genomes and absent in S288c are differentially expressed in the two different fermentation points. Large part of these genes are more expressed at the beginning of the stationary phase respect to 6 g/l. 10 of the genes identified in EC1118, out of 28 total, are up-regulated more than 4 times at 45 g/l.

The four genes named from P283 strain are differentially expressed in other strains more than in P283, for example P283_G2_2311 is more expressed at 45 g/l in P301 strain, P283_G2_2316 is more expressed at 45 g/l in EC1118 and in R103 strains and P283_I1_0711 (coding the killer toxin) is more expressed at 6 g/l in R103.

Genes named after Roo8 strain are expressed more than 4 times at 6 g/l, both hypothetical proteins, and two (medium chain alcohol dehydrogenase and putative allantoate permease) at 45 g/l. Allantoate/ureidosuccinate permease in *S. cerevisiae* is coded by Dal5p that has been shown previously to play a role in the utilization of certain dipeptides as a nitrogen source (22). Uptake assays indicated that either Ptr2p (a di/tri-peptide transporter with very broad substrate specificity) or Dal5p was predominantly used for dipeptide transport in the common laboratory strains S288c and W303, respectively. These two dipeptide transport systems have complementary activities under different regulatory controls in common laboratory yeast strains suggesting that dipeptide transport pathways evolved to respond to different environmental conditions (22). For example DAL5 expression was down-regulated in the presence of leucine and the absence of CUP9, whereas PTR2 was up-regulated. DAL5 mRNA levels dropped precipitously when a repressive nitrogen source was provided. These control characteristics of DAL5 expression indicates that it is subjected to the nitrogen catabolite repression and this is also true for and Roo8_O1_4131 gene in Roo8 strain (23).

This effect is due to the ability of *S. cerevisiae* to use different nitrogen sources for growth but not all nitrogen sources support growth equally well. *S. cerevisiae* selects nitrogen sources that enable the best growth by a mechanism called Nitrogen Catabolite Repression (NCR) (24,25). Good nitrogen sources such as glutamine, asparagine or ammonium decrease the level of enzymes required for utilisation of poorer nitrogen sources (26).

This indicates that certain *S. cerevisiae* strains have more than two different di/tri-peptide transporters, this can help to better respond to different nitrogen sources. Finally, three genes identified in R103, out of a total of 10, are more expressed at 45 g/l but all these genes code for hypothetical proteins. All these data together indicate that these “unique genes” are regulated in response to changing environmental conditions, this is not trivial, since it was demonstrated that some of these genes are laterally transferred from other eukaryotes (27). Previous reports (27) hypothesize that these genes were involved in adaptation to oenological conditions, here we add strong evidence that supports these previous findings and hypothesize the control level for Roo8_O1_4131 gene.

Influence Of Structural Variations On The Expression Of Flanking Genes

Genes flanking structural variations (translocations, inversions and inserted or deleted regions) have been manually collected analyzing Mauve alignment of the six genomes. Structural variations in regions flanking the genes can affect gene expression due to differences in regulative elements. 18 major critical points with 41 total variations in the six genomes have been analyzed. 7 variations are specific of a single strain, the others are macroscopically common between two or more genomes but can have minor internal differences (SNPs or small indels). Among these 41 variations, there are 17 translocations, 2 inversions, 3 insertions and 3 deletions, the other 16 are highly different regions characterized by several small structural variations collectively grouped in a single region. In general we noticed a higher number of differences in the chromosome 16 due to translocations 15-16 and 8-16 also verified by PCR and sequencing (only 8-16). Strain R103 in particular shows a higher number of variations compared to the other strains. 56 genes flanking the variations have been collected and differences in gene expression respect to S288c have been analyzed to understand the influence of structural variations on the expression.

Results show a correlation between structural variations and gene expression in 10 features but it seems that this correlation is condition specific. In fact 7 of these genes (YGR287C, YHR114W, YPL093W, YAR033W, YAR035W, YOL065C, YAL008W) seems to be influenced in their expression at 6 g/l and 3 genes (YAL003W, YOL086C, YPL092W) at 45g/l in the six strains. This can be due to the different effect on genes regulated by a pattern of condition-specific factors not altered by the variation.

GO Classes Enriched in Oenological strains

GO terms enriched for genes differentially expressed between oenological strains and S288c are reported together with the p-value calculated using the Hypergeometric distribution. Holm-Bonferroni multiple test corrections have been also performed to take into account the number of tests being carried out and to correct the p-values accordingly. Pathways enriched for genes differentially expressed have also been considered from Pathway Tools.

Table 4.3 Table reporting GO categories and Patway Tools enriched in oenological strains

	6 g/l				45 g/l			
	GO Terms and Description	p-Value	Genes	Total	GO Terms and Description	p-Value	Genes	Total
up in S288c	GO:0032197 - transposition, RNA-mediated	2.97E-48	54	255	GO:0032197 - transposition, RNA-mediated	1.28E-60	56	183
	GO:0022415 - viral reproductive process	2.84E-27	26	255	GO:0022415 - viral reproductive process	5.20E-39	30	183
	GO:0000003 - reproduction	7.35E-19	65	255	GO:0000003 - reproduction	5.29E-14	47	183
	GO:0022607 - cellular component assembly	2.17E-06	48	255	GO:0006259 - DNA metabolic process	4.03E-08	22	183
	GO:0006259 - DNA metabolic process	6.59E-06	42	255	GO:0006414 - translational elongation	7.86E-08	29	183
	GO:0030476 - ascospore wall assembly	2.31E-04	8	255	GO:0019538 - protein metabolic process	4.55E-03	38	183
	GO:0034293 - sexual sporulation	1.14E-03	12	255	GO:0010033 - response to organic substance	1.02E-02	11	183
	GO:0019438 - aromatic compound biosynthetic process	6.40E-03	6	255	GO:0000128 - flocculation	1.53E-02	2	183
	GO:0006414 - translational elongation	1.70E-02	22	255	GO:0055085 - transmembrane transport	3.15E-02	16	183
	GO:0009110 - vitamin biosynthetic process	2.23E-02	6	255				
	GO:0030154 - cell differentiation	3.72E-02	13	255				
	GO:0071702 - organic substance transport	4.70E-02	12	255				
	Pathways	p-Value	Genes		Pathways	p-Value	Genes	
	Thiamine biosynthesis	5.22E-05	4		Arginine biosynthesis	1.56E-04	3	
	Tryptophan degradation via kynurenine	1.08E-02	2		Glutamate biosynthesis from ammonia	2.84E-02	1	
up in oenological strains	Glycine biosynthesis from glyoxylate	2.86E-02	1					
	Sucrose degradation	2.86E-02	1					
	GO Terms and Description	p-Value	Genes	Total	GO Terms and Description	p-Value	Genes	Total
	GO:0055114 - oxidation-reduction process	3.57E-04	16	95	GO:0055114 - oxidation-reduction process	1.04E-16	63	296
	GO:0006886 - intracellular protein transport	4.69E-02	9	95	GO:0006066 - alcohol metabolic process	1.32E-10	36	296
	GO:0008610 - lipid biosynthetic process	1.26E-02	7	95	GO:0006629 - lipid metabolic process	2.58E-07	35	296
					GO:0044281 - small molecule metabolic process	3.20E-07	83	296
	GO:0051186 - cofactor metabolic process	2.25E-02	7	95	GO:0007005 - mitochondrion organization	5.86E-07	37	296
	GO:0006612 - protein targeting to membrane	1.36E-02	4	95	GO:0006108 - malate metabolic process	1.22E-02	2	296
	GO:0000316 - sulfite transport	2.80E-02	1	95	GO:0071474 - cellular hyperosmotic response	3.91E-02	2	296
	GO:0016226 - iron-sulfur cluster assembly	3.17E-02	2	95	GO:0006811 - ion transport	4.79E-02	14	296
	Pathways	p-Value	Genes		Pathways	p-Value	Genes	
	Aerobic respiration, electron transport chain	1.99E-03	4		Superpathway of ergosterol biosynthesis	9.63E-08	12	
					Aerobic respiration, electron transport chain	1.00E-07	13	
					Arginine degradation	8.26E-03	3	

Genes Involved in Ethanol Tolerance

Ethanol is well known as an inhibitor of microorganisms growth. The toxic effects of ethanol on yeast cells involve loss of cells viability, inhibition of yeast growth and of different transport systems such as the general amino acid permease and the glucose transport system (28). Moreover, the rising of ethanol concentration during fermentation process (especially in presence of high concentration of sugar substrates) acts to reduce growth and fermentation rates and adversely affect cell viability (29). For this reason a high level of ethanol tolerance for a yeast strain is a pre-requisite for a high efficiency of fermentation.

In order to have a general overview of the different responses to increasing ethanol concentration of the strains considered in expression analysis, we have generated a list of 369 genes involved in EtOH tolerance identified in three papers and one review (30-32).

Absolute expression values (number of uniquely mapped reads in the coding region of each gene identified) were considered respect to the average value of each gene in all strains and conditions considered in gene expression experiments.

$$\log_2 (N_i/N_{iav})$$

“ N_i ” is number of reads for the gene “i” in one strain and in one of the two conditions analyzed, while “ N_{iav} ” is the average number of reads of the gene “i” calculated considering all strains (in which the genes is present) and conditions. The two fermentations points considered in RNA-seq analysis were discussed previously (cap. 2: Fermentation in Controlled Bioreactors) the first sample was taken at the beginning of the fermentation when the CO_2 produced was 6 g/l, while the second sample was taken at 45 g/l (I refer to these points using 6 and 45 g/l abbreviation). Data analysis using TMEV software (18) indicates that global expression profile of the strains at 6 and 45 g/l are very different and this is obviously due to rising of ethanol concentration (Fig. 4.5). The behaviour of strains is quite interesting: S288c (the laboratory strain) is very different from oenological strains at 45 g/l, while at 6 g/l expression profiles of S288C and R103 are quite similar.

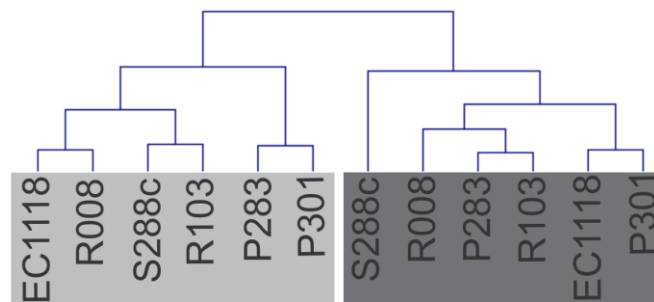


Figure 4.5 clustering of strains determined considering expression profile of “ethanol resistance” genes. Light grey box highlights the 6 g/l condition, dark grey box the 45 g/l condition.

This makes sense because S288c is very different from oenological strains at least in terms of fermentation properties and ethanol resistance and R103 strain has poor fermentation characteristics and takes more time to complete fermentation process respect to the other oenological strains considered. Ethanol stress resistance obviously recovers a very relevant role in the second part of the fermentation process (at 45 g/l) and for this reason the discrepancy of S288c expression profile at 45 g/l is particularly relevant. At 45 g/l expression profiles of oenological strains are quite similar, the more different is R008 but it remains within the “oenological cluster”. Situation is more complex for commercial strain EC1118 and for ecotypical strains but the second part of the fermentation process is more similar for strains EC1118 and P301.

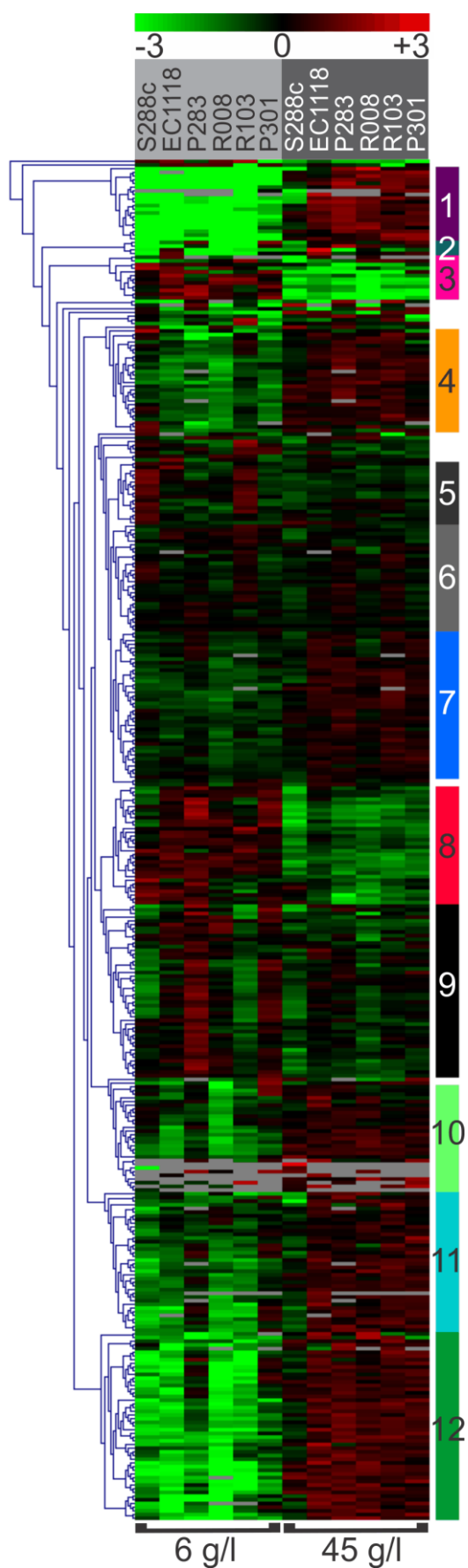


Figure 4.6 gene clusters obtained using TMEV software are highlighted using coloured boxes in the right part of the figure. Analysis was performed using euclidean distance calculation.

Hierarchical clustering of the expression values, identified 12 gene clusters that are highlighted in colours in Fig. 4.6, this put in evidence the high variability of expression of the genes involved in ethanol tolerance between strains and conditions examined. This can be easily explained because ethanol tolerance is under polygenic control as a typical quantitative trait (30). As expected, a lot of genes increase their expression values in all strains at 45 g/l (clusters 1-2-4-7-10-11-12) but there are also a large number of genes that reduce their expression (clusters 3-8). Various gene clusters display marked differences between strains, these are of particular interest (for example 5-6-9), below we discuss in detail genes of the clusters 1 and 9 that are relevant to understand differences in ethanol tolerance between strains.

Cluster 1 (20 genes): expression of these genes is markedly increased at 45 g/l. Behaviour is quite similar in all strains except in S288c at 45 g/l, in this strain expression is lower than other strains and this probably have a high relevance in determining the lower ability of S288C to face the second part of the fermentation process. The GO analysis performed using YeastMine identified 9 genes involved in oxidation-reduction process (p-value 0.0132). A more accurate analysis of these genes indicates that some of these code mitochondrial proteins relevant for growth in presence of ethanol (for example mitochondrial aldehyde dehydrogenase, YOR374W). Expression of this gene is repressed by glucose.

Other proteins are involved in glycogen production. The level of this compound in some yeast strains is important, as it is the sole source of metabolic energy for lipid synthesis and hexose transport in the first few hours of fermentation. Because of this, levels decline during the first 24 h of fermentation but then rise and peak at the end of the growth phase (immediately after the 45 g/l point), before gradually declining during the stationary phase (33).

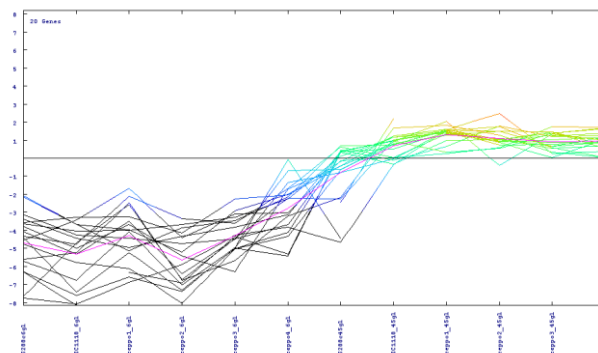


Figure 4.7 Expression profile of genes belonging to cluster 1 between strains (S288c, EC1118, P283, Roo8, R103, P301) and comparing 6g/l (left part of the figure) and 45 g/l fermentation points (right).

Cluster 9 (47 genes): this is a very important cluster because these genes have a marked difference between strains (at 6 g/l). In strains having good fermentation properties like P283 and P301 they have a high expression, in strains like EC1118 and Roo8 they have an intermediate expression and in S288c and R103 they have a low expression. GO analysis performed using YeastMine reveals that these genes represent some of the main classes previously described for their importance in ethanol stress tolerance: intracellular pH reduction (p-value 0.0013, 3 genes), protein folding (p-value 0.00037, 6 genes) and negative regulation of ribosomal protein gene transcription from RNA polymerase II promoter in response to chemical stimulus (p-value 0.009, 1 gene).

There are also genes involved in glycerol biosynthesis and induced in response to hyperosmotic and oxidative stress (DL-glycerol-3-phosphatases - YERo62C), others involved in overproduction of inositol (methylene-fatty-acyl-phospholipid synthase - YJRo73C), in vacuolar protein sorting (YPLo65W, YMLo97C), in ubiquitin-mediated proteolysis (YBR173C) and in ERGosterol biosynthesis (YERo44C). Curiously there are also two genes involved in microtubule biogenesis (YMLo94W, YELoo3W). Most of these pathways and molecular functions have been previously demonstrated to be involved in ethanol stress response (34). The markedly different expression of these genes between strains examined in this study, emphasize the relevance of ethanol stress resistance to maintain a high fermentation rate.

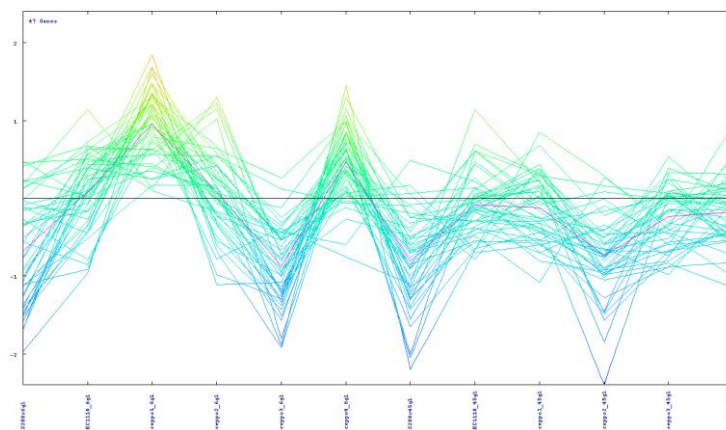


Figure 4.8 Expression profile of genes belonging to cluster 9 between strains (S288c, EC1118, P283, Roo8, R103, P301) and comparing 6g/l (left part of the figure) and 45 g/l fermentation points (right).

Transcription Factor Binding Sites

For each TFBS it has been identified the corresponding gene putatively regulated. A region of 500bp upstream and 100bp downstream the transcription start site has been considered for all protein coding genes and ncRNAs to verify the presence of a TFBSs. It has been selected 4106 couples of gene and regulative TFBS formed by 1967 features and 1423 TFBSs eventually redundant. Differences in the expression of these features have been analyzed to understand the influence of mutated TFBSs on their expression.

Results show a correlation between mutated TFBSs and gene expression in 32 couples, it seems that this correlation can be condition specific. In fact 12 of these genes (YDL049C, YMR318C, YKL180W, YML116W, YGL195W, YJR025C, YOR054C, YGL001C, YOL128C, YJR089W, YJR088C, YOL126C) involved in glyoxylate cycle, enriched in GO oxidation-reduction and alcohol metabolic process, seems to be influenced in their expression only at 6 g/l, 12 genes (YKR075C, YOR372C, YGR249W, YGL055W, YGR243W, YGL056C, YMR078C, YNL244C, YOR235W, YAL061W, YNL103W, YER028C) involved in superpathway of fatty acid biosynthesis, saturated and unsaturated and enriched in GO regulation of macromolecule biosynthetic process at 45g/l and 5 (YKR030W, YLR286C, YIL159W, YGL116W, YIL140W) enriched in GO classes of positive regulation of mitosis and cell cycle at both condition in the six strains. This can be due to the different effect on genes regulated by a pattern of condition-specific factors not altered by the variation.

It has been verified the correlation between differences in TF expression and correspondent differences in expression of genes regulated by it. For each gene it has been calculated the mean difference of expression in both condition. We found a direct correlation in expression of some TF-gene couples like those of GAL4, HAP4, INO4, TEC1 and ZAP1. CIN5, ROX1, TYE7 and SWI5 show a weak correlation. HAP2 expression is correlated with that of S288c-P301 and S288c-P283 at 6 g/l. DAL80 seems to be a negative regulator in the comparison S288c-P283 at 45g/l but has no correlation in S288c-R103 at 45 g/l. MET32 expression is direct correlated but with a strange trend in S288c-R008 comparison at 45 g/l. MOT3, YAP7 and SKO1 show a correlation only in some comparisons.

Differential expression linked to differences in TR lenght

The percentage of differentially expressed genes with an expression higher than 2 and 4 times than the reference S288c have been calculated for genes with intergenic region without tandem repeats and for those having tandem repeats with different level of differences in length. Higher the difference in TR length, higher the number of differentially expressed genes.

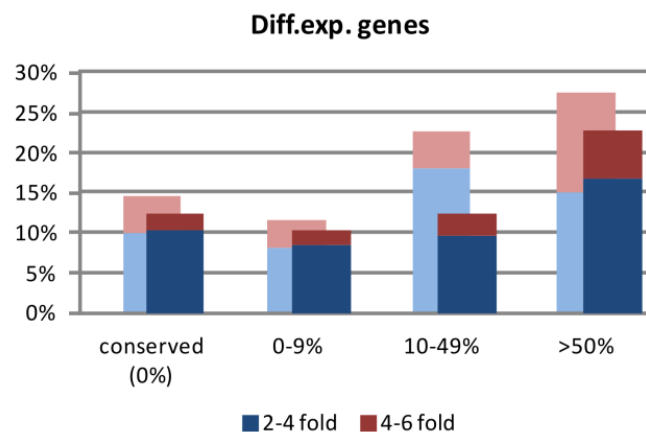


Figure 4.9 Percentage of differentially expressed genes with an expression variation between strains higher than 4 and 8 times respect to S288c calculated for genes without TR in intergenic region and for those with variable TR (0-9%, 10-49% and >50%).

Global analysis of the influence of different factors on gene expression.

Identification of the correlations between the functional information encoded in a genome (like regulatory elements) and gene expression is a central challenge in biological research. Our basic idea was to verify the influence of eight different factors (reported in the list below) on the global gene expression profile in order to obtain a clear indication of the impact of different “modifiers”.

1. Mutations on transcription factor (TFs) binding sites can change the efficiency of the binding of the TF and subsequently influence the expression of the gene downstream.
2. Changes of the length of the tandem repeat located close to the transcription start site can change the assembly efficiency of the transcription initiation complex.
- 3./4. Presence or absence of the LTRs (3a) and Ty elements (3b) in the promoter region.
5. Structural variations (deletions, insertions, translocations) located in the promoter region.

Apart from these structural changes on promoter regions, we have considered other three factors that can influence gene expression but that are not directly due to genomic differences between strains.

6- Up or down regulation of genes coding transcription factors determine a change in the transcription level of the genes that are regulated by these TFs.

7- Presence or absence of non coding transcripts (SAUT) expressed in “antisense” to specific genes coded on the “sense” strand.

8- Partially overlapped transcripts are frequent in yeast and it was hypothesized that they can have a reciprocal influence on expression levels.

Taken all together these differences can determine at least part of the transcriptional changes identified between strains using RNA-seq. Since it is impossible to perform a statistical analysis considering presence/absence of a certain character between strains because the number of biological replicates (strains) is too low, we have considered as a general indication of the influence of each character described above on expression level of a given gene the correlation coefficient (pearson) between expression level ($\log_2$ ratio calculated between each expression value and the mean expression value determined considering all the experiments and all the strains) and the presence/absence of a specific genomic difference. Genes having a correlation value between “expression” and “difference” equal or higher than 0.8 (or lower than -0.8) (threshold was chosen arbitrarily after a manual verification) are considered correlated (or anticorrelated). From this analysis we have excluded non coding transcripts (SAUT). The total number of genes differentially expressed more than 4 times and considered in this calculation is 2617 (these genes were named DiffExp). 1247 of these genes (48%) are potentially influenced by one of the eight characters described because they have at least one difference between strains (named DiffExp_Modified), while 1370 genes (52%) do not show any difference (DiffExp_NotModified). This suggests that other important factors not considered in our model have a role in determining differences of expression between strains. Analysis of all the correlation values with the eight different characters reported above (TF mutated, presence/absence of Ty elements, etc.) indicates that only 296 genes (out of 2617 = 11.3%) have a correlation value equal or higher than 0.8 (named DiffExp_Modified_Corr).

The difference between the percentage of “DiffExp_Modified_Corr” (11.3%) (*determined as described below) and (NotDiffExp_Modified_Corr) (7.3%) is not very high and we can assume that our idea is substantially correct but needs to be improved. Also the number of genes “DiffExp_Modified” (48%) is not very high respect to the number of genes “NotDiffExp_Modified” (42%) but this can be determined by the redundancy of the regulatory elements in promoter regions. Finally we have classified the eight factors considering their impact on the expression differences between strains:

Presence of SAUT in antisense of genes (7) correlates in **2.8%** of all the genes;

transcription factors differentially expressed (6) correlates in **1.97%** of the genes;

5'-3' transcripts overlap (8) correlates in **1.83%** of the genes;

long tandem repeats (3a) correlates in **0.93%** of the genes;

tandem repeats having variable length (2) correlates in **0.45%** of the genes;

transcription factor binding sites mutations (1) correlates in **0.37%** of the genes;

Ty elements (3b) correlates in **0.36%** of the genes;

structural variations (5) correlates in **0.14%** of the genes.

In this analysis we have considered both positive and negative values (correlated and anticorrelated) but, as expected, the number of correlated genes is higher than anticorrelated (data not shown). This gives us a rough indication of the global impact of different factors in influencing gene expression between strains. The quite low percentage of correlations identified (11.3%) are probably due to the presence of other factors (such as for example epigenetic effects) that have to be included in our “model” to improve our model but these preliminary results indicates also that the approach seems to be substantially correct.

(*) 4367 genes (here we do not considered SUT and SAUT) do not change their expression level between strains (NotDiffExp) and 1846 of these genes (42%) have at least one difference between strains (NotDiffExp_Modified), obviously these differences between strains do not influence gene expression. 2521 out of 4367 (58%) do not have any of the eight differences reported (NotDiffExp_NotModified). 321 of these genes (out of 4367) correlates with expression values (7.3%) (NotDiffExp_Modified_Corr).

REFERENCES

- (1) Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009 Jan;10(1):57-63.
- (2) Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, et al. PASS: a program to align short sequences. *Bioinformatics* 2009 Apr 1;25(7):967-968.
- (3) Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008 Jun 6;320(5881):1344-1349.
- (4) Brauer MJ, Christianson CM, Pai DA, Dunham MJ. Mapping novel traits by array-assisted bulk segregant analysis in *Saccharomyces cerevisiae*. *Genetics* 2006 Jul;173(3):1813-1816.
- (5) Tirosh I, Weinberger A, Bezalel D, Kaganovich M, Barkai N. On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol* 2008;4:159.
- (6) Querol A, Fernandez-Espinar MT, del Olmo M, Barrio E. Adaptive evolution of wine yeast. *Int J Food Microbiol* 2003 Sep 1;86(1-2):3-10.
- (7) Vences MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 2009 May 29;324(5931):1213-1216.
- (8) Romano P, Suzzi G. Acetoin production in *Saccharomyces cerevisiae* wine yeasts. *FEMS Microbiol Lett* 1993 Mar 15;108(1):23-26.
- (9) Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005 Sep 9;309(5741):1728-1732.
- (10) Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J. The next-generation sequencing technology and application. *Protein Cell* 2010 Jun;1(6):520-536.
- (11) Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008 Mar 1;24(5):713-714.
- (12) Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 2010 Jan 1;26(1):136-138.
- (13) Gower JC. Generalized Procrustes Analysis. *Psychometrika* 1975;40:33-51.
- (14) Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* 2009 May 12;10:221.

- (15) Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008 Sep;18(9):1509-1517.
- (16) Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009 May;6(5):377-382.
- (17) Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008 Jul;5(7):621-628.
- (18) Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, et al. TM4 microarray software suite. *Methods Enzymol* 2006;411:134-193.
- (19) Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 2003;4(4):R28.
- (20) Lopes CA, Rodriguez ME, Querol A, Bramardi S, Caballero AC. Relationship between molecular and enological features of Patagonian wine yeasts: relevance in selection protocols. *World Journal of Microbiology & Biotechnology* 2006;22:827-833.
- (21) Rodriguez ME, Lopes CA, van Broock M, Valles S, Ramon D, Caballero AC. Screening and typing of Patagonian wine yeasts for glycosidase activities. *J Appl Microbiol* 2004;96(1):84-95.
- (22) Cai H, Hauser M, Naider F, Becker JM. Differential regulation and substrate preferences in two peptide transporters of *Saccharomyces cerevisiae*. *Eukaryot Cell* 2007 Oct;6(10):1805-1813.
- (23) Rai R, Genbauffe F, Lea HZ, Cooper TG. Transcriptional regulation of the *DAL5* gene in *Saccharomyces cerevisiae*. *J Bacteriol* 1987 Aug;169(8):3521-3524.
- (24) Wiame JM, Grenson M, Arst HN, Jr. Nitrogen catabolite repression in yeasts and filamentous fungi. *Adv Microb Physiol* 1985;26:1-88.
- (25) Broach JR, Pringle JR, Jones EW. The Molecular and cellular biology of the yeast *Saccharomyces*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 1991.
- (26) ter Schure EG, van Riel NA, Verrrips CT. The role of ammonia metabolism in nitrogen catabolite repression in *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* 2000 Jan;24(1):67-83.
- (27) Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, et al. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci U S A* 2009 Sep 22;106(38):16333-16338.
- (28) Alexandre H, Plourde L, Charpentier C, Francois J. Lack of correlation between trehalose accumulation, cell viability and intracellular acidification as induced by various stresses in *Saccharomyces cerevisiae*. *Microbiology* 1998 Apr;144 (Pt 4)(Pt 4):1103-1111.

- (29) Piper PW. The heat shock and ethanol stress responses of yeast exhibit extensive similarity and functional overlap. *FEMS Microbiol Lett* 1995 Dec 15;134(2-3):121-127.
- (30) Hu XH, Wang MH, Tan T, Li JR, Yang H, Leach L, et al. Genetic dissection of ethanol tolerance in the budding yeast *Saccharomyces cerevisiae*. *Genetics* 2007 Mar;175(3):1479-1487.
- (31) Li BZ, Cheng JS, Ding MZ, Yuan YJ. Transcriptome analysis of differential responses of diploid and haploid yeast to ethanol stress. *J Biotechnol* 2010 Aug 2;148(4):194-203.
- (32) van Voorst F, Houghton-Larsen J, Jonson L, Kielland-Brandt MC, Brandt A. Genome-wide identification of genes required for growth of *Saccharomyces cerevisiae* under ethanol stress. *Yeast* 2006 Apr 15;23(5):351-359.
- (33) Quain DE, Boulton CA. Growth and metabolism of mannitol by strains of *Saccharomyces cerevisiae*. *J Gen Microbiol* 1987 Jul;133(7):1675-1684.
- (34) Ma M, Liu ZL. Mechanisms of ethanol tolerance in *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol* 2010 Jul;87(3):829-845.

5. DISCUSSION AND CONCLUSIONS

Thanks to the development of new sequencing technologies, genomic comparisons of multiple strains or organisms with different phenotypic characters is becoming common. Despite various studies have been recently performed to compare different genomes, technical barriers still constrained the comparison at a transcriptional level, and made a complex problem the association of genomic features with phenotypic characters (1)(2). The development of novel high-throughput RNA sequencing technologies (3) allows pulling down these barriers providing a new method for both mapping and quantifying transcriptomes.

Starting from the assumption that different yeast strains have different fermentation characters and produce a unique profile of volatile flavors compounds, it would be interesting to investigate phenotypic characters of ecotypical yeast strains and correlate it with their genome content. To reach this aim, four representative strains of the endemic *S. cerevisiae* population of the Veneto vineyards have been chosen using a PCA approach to discriminate phenotypic characters of interest. The selection strategy also allowed to verify the genomic structure of 20 yeast strains using classical approaches such as spore dissection and PFGE. In order to simplify the genome sequencing, derivative lines obtained from sporulation and tetrads dissection were produced and various phenotypic tests performed on parental strains and derivative homozygous to obtain a detailed oenological characterization. The genome of the four homozygous derivative lines were successfully sequenced using 454-FLX approach reaching a coverage major than 95% respect to the reference genome of S288c laboratory strain.

In this project we have firstly provided a comparison between oenological and laboratory yeast genomes and then correlated the differences identified with their gene expression. Management of the huge amount of data produced entails the use of complex custom bioinformatics pipeline. Since a lot of bioinformatics instruments for genome assembly, gene finding and annotation are available, the facility to sequence genomes at a rate previously inconceivable requires new software able to use data yet available to simplify analysis of those newly produced. With the introduction of third generation sequencing technologies (4), biologists will face even more informatics challenges, including the development of efficient methods to store, retrieve and process even larger amounts of data. Starting from pre-existing programs like Gap Resolution and Newbler we have developed a pipeline that integrates them with perl scripts written ad hoc to take advantage from the high quality genomic data of S288c strain for the finishing process. Using this pipeline we obtained four high quality assemblies of ecotypical yeast genomes with on average 2.5 scaffolds per chromosome. The good results achieved facilitated the subsequent gene finding, annotation and gene expression analysis. 95-97% of the protein-coding genes of S288c were successfully transferred using RATT software (5) to the four yeast genomes.

Identification of orthologous facilitates the subsequent identification of the genes that are specifically present in oenological strains. Variable numbers of genes were identified that are not present in the reference S288c genome but similarity search in NCBI database revealed that frequently they are present in other *S. cerevisiae* strains previously sequenced. This indicates that *S. cerevisiae* genome have been extensively sampled and the probability to identify new genomic regions in oenological strains is rapidly decreasing.

Transcriptional analysis revealed for the first time that these “oenological specific genes” are expressed on average at a level comparable to the other genes present in all strains and, more importantly, that are frequently differentially expressed comparing different points of the fermentation curve. This revealed for the first time that these genes have a role in fermentation and probably favour these strains in oenological environment.

Genome alignment performed using MAUVE software allowed identification of 368408 SNPs that are extremely useful to better investigate genetic diversity between strains and evolutionary processes acting within populations. Pairwise SNPs difference identified in genomic alignments clustered the four ecotypical strains sequenced in a group comprising all the oenological strains considered. This result confirmed previous analysis performed using next-gen sequencing, microarrays and microsatellite length polymorphisms (6-8) that tend to cluster yeast strains on the basis of their technological niche. Despite previous studies (8) indicates that part of the genetic diversity between these technological strains was associated with geographical differences, three out of four genomes analysed are similar to VL3 and AWRI796, while R103 is at certain extent “at the edge” of this group. In this study we have selected R103 as a sort of “negative control” for its reduced fermentation performances and for this reason can be considered an “atypical” oenological strain. Since the “wine group” is mainly determined by technological characteristics, we can consider that the position of R103 in the cluster is probably due to its phenotype. SNPs distribution along the genome has been evaluated using a 10 kb sliding window, results obtained confirm that *S. cerevisiae* genome is quite complex probably due to human traffic and subsequent recombination between strains of different geographic origin (6), this is particularly evident for oenological strains that frequently display large blocks of homology. For example EC1118 and QA23 have two large portions on chromosomes 8 and 16 showing a very high similarity and this is also true for Roo8 vs. AWRI796 and VL3 vs. AWRI1631 comparisons.

Analysis of SNPs that are conserved only in oenological strains identified 315 positions that can be considered a sort of “oenological signature”, 62.5% of these are localized in protein-coding regions and 28.2% determine non-synonymous changes. Despite we can not exclude that these differences between strains are determined by a common origin of wine strains, analysis of proteins having non-synonymous changes indicates that their gene ontology are related to processes relevant for the adaptation to the oenological environment such as nitrogen utilization and catabolic process or the response to specific organic substances.

The finding of the correlation between the functional information encoded in a genome, like genes and regulatory elements, and gene expression is a central challenge in biological research. Taking advantage from the list of transcriptional factor binding sites previously identified in yeast (9) we performed an analysis on the effect that mutations in promoter regions exercise on gene expression. We propose that alterations in tandem repeat length have a more important role respect to differences in transcription factor binding sites. In particular, the percentage of highly differentially expressed genes in those classes regulated by tandem repeats with highly variable repeats seems to be higher than those of classes with less mutated sequences. Difficulties in finding a strong correlation between variations in promoter regions and gene expression could be ascribed to the regulation of the transcript degradation and stability and to epigenetic effects that are not considered in our model. Moreover we found that differences of gene expression of regulatory factors have a deep effect on downstream pathways, determining expression alterations in genes, which could not be ascribed to differences in their promoter regions but to secondary effects.

Our RNA-seq analysis revealed that gene expression comparison of the orthologous genes in the oenological and laboratory strains highlights the existence of a fingerprint characterizing oenological strains. Some of these genes have been previously identified for their role in facing stressful conditions that is a typical characteristic of the oenological environment. To better investigate this point we have also analyzed the expression profile of 369 genes identified in literature and involved in ethanol tolerance (one of the more relevant element of stress during fermentation). Yeast evolution favours fermentation over respiration and this determines ethanol accumulation; this compound has significant adverse effects on cellular growth and viability and on fermentation process itself (10). Respect to other microbes that are present in natural fermentation process, yeast evolved a high ethanol tolerance, one of the key factors for this organism to dominate must fermentation. This character is also one of the most important properties of microbes to improve the efficiency and economy of ethanol production. Specially, in lignocellulosic ethanol production, increased ethanol tolerance is one of the essential traits of microbes (11).

However, different *S. cerevisiae* strains display very different ethanol tolerance and this study gave us the opportunity to investigate these differences at a transcriptomic level. Hierarchical clustering of 369 genes selected from literature for their importance in ethanol tolerance allowed a classification of the six strains in terms of transcriptional behaviour. Gene expression at 6 g/l revealed that strains having a poor fermentation properties (S288c and R103) have also similar expression, while in the second point of the fermentation curve oenological strains are clearly distinct from S288c. This indicates that in strains fermentation is influenced by ethanol tolerance and, despite at 6 g/l ethanol concentration is low, these genes already play a significant role. Since during fermentation different stress responses are scheduled, it has been postulated that yeast tends to anticipate the stress response; this suggests why differences in expression of genes involved in ethanol tolerance are evident in the first time point examined in our work.

Two gene clusters identified using TMEV software are particularly relevant to understand the different behaviour of the six strains during ethanol stress. The first one reveals that S288c is different from oenological strains at 45 g/l in terms of glycogen production. Some authors indicates that the level of this compound is important because it is an important energy source for relevant processes such as lipid synthesis and hexose transport especially in the first few hours of fermentation (12). Concentration of this compound rise and peak at the end of the growth phase (immediately after the 45 g/l point), before gradually declining during the stationary phase (12). The low level of these transcripts in S288c could reduce glycogen synthesis and this influences negatively the second part of the fermentation process. It was suggested that in certain cases lipids addition can compensate for low glycogen levels, this can be a possible test to perform in the next future. Expression of a second gene cluster reflects fermentation properties, these genes are involved in different processes related to ethanol resistance (like ERGosterol biosynthesis, vacuolar protein sorting and inositol production). In strains S288c and R103 (having bad fermentation properties) expression of these genes is particularly low at 6 g/l. This finding again suggests the importance of gene expression in the early fermentation step and indicates that probably a global reduction of various processes reduce the fermentation in some of the strains analyzed.

We can conclude that expression of genes involved in ethanol tolerance have a strong role in determining fermentation properties of the strains examined but there is not a complete overlap between these two characters. This genomic and transcriptional study pave the way to future studies that will allow to infer more specific function of features with unknown role and to identify correlations between important oenological characters such as SO₂ resistance and their genetic determinants.

REFERENCES

- (1) Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, et al. Genotype to phenotype: a complex problem. *Science* 2010 Apr 23;328(5977):469.
- (2) Souciet JL, Genolevures Consortium GDR CNRS 2354. Ten years of the Genolevures Consortium: a brief history. *C R Biol* 2011 Aug-Sep;334(8-9):580-584.
- (3) Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008 Jun 6;320(5881):1344-1349.
- (4) Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J. The next-generation sequencing technology and application. *Protein Cell* 2010 Jun;1(6):520-536.
- (5) Otto TD, Dillon GP, Degraeve WS, Berriman M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* 2011 May;39(9):e57.
- (6) Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. *Nature* 2009 Mar 19;458(7236):337-341.
- (7) Legras JL, Merdinoglu D, Cornuet JM, Karst F. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol Ecol* 2007 May;16(10):2091-2102.
- (8) Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 2009 Mar 19;458(7236):342-345.
- (9) Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004 Sep 2;431(7004):99-104.
- (10) Stanley D, Bandara A, Fraser S, Chambers PJ, Stanley GA. The ethanol stress response and ethanol tolerance of *Saccharomyces cerevisiae*. *J Appl Microbiol* 2010 Jul;109(1):13-24.
- (11) Zaldivar J, Nielsen J, Olsson L. Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. *Appl Microbiol Biotechnol* 2001 Jul;56(1-2):17-34.
- (12) Quain DE, Boulton CA. Growth and metabolism of mannitol by strains of *Saccharomyces cerevisiae*. *J Gen Microbiol* 1987 Jul;133(7):1675-1684.

ACKNOWLEDGEMENTS

This study has been funded by the University of Padua on a "Progetto di Ateneo" grant.

We would like to thank the "Provincia di Treviso" for providing the PhD fellowship and Prof. Giorgio Valle for his valuable assistance.

A tutte le persone che per scelta o per caso sono entrate nella mia vita.

*Un abbraccio infinito a chi l'ha attraversata con affetto
soffermandosi a raccogliere il meglio di me.*

A Stefano, che non ringrazierò mai abbastanza.

APPENDIX I

All the experimental procedures used in this work are reported in detail in every chapter, in order to be reproduced in any laboratory. A few of the described protocols do not refer to any of the experiments reported in the results. In those cases the results produced were of no relevance for the discussion, nevertheless the procedures are reported in the list of methods since they could be useful for future applications. In the majority of the cases indeed the reported procedures were obtained after careful testing and optimization of the experimental conditions specifically for natural yeasts.

LIST OF ABBREVIATIONS

bp	Base pairs
BSA	Bovine Serum Albumin (10 mg/ml), provided by NEB
CFU	Colony Formant Unit
dNTPs	10 mM deoxynucleotides, equimolar solution of dATP, dCTP, dGTP and dTTP
DTT	Dithiothreitol
EDTA	Ethylene-Diamine Tetraacetic Acid
EtBr	Ethidium Bromide
EtOH	ethanol
Gbp	Giga (billion) base pairs
GoTaq	Taq polymerase (5 U/μl), provided by Promega
h	hour
kbp	kilo base pairs
Mbp	Mega (million) base pairs
Microcentrifuge tube	RNase-free 1.5 ml microcentrifuge tube, provided by Eppendorf
min	minute
MOPS	3-(N-morpholino)propanesulfonic acid
O/N	over night
OD600	optical density of a sample measured at a wavelength of 600 nm
ORF	Open reading frame
P ₁ and P ₂	two different adaptors of SOLiD™ system library preparation kit
P283 – E	P283 heterozygous strain, natural isolate
P283 – O	P283 omozigote, linea derivata
P301 – E	P301 eterozigote, isolato naturale
P301 – O	P301 omozigote, linea derivata
Roo8 – E	Roo8 heterozygous strain, natural isolate
Roo8 – O	Roo8 homozygous strain, derivative line
R103 – E	R103 heterozygous strain, natural isolate
R103 – O	R103 omozigote, linea derivata
Rpm	Revolutions per minute
SDS	Sodium dodecyl sulfate
sec	seconds
TAE	Tris-Acetate-EDTA
TAP	Tobacco Acid Pyrophosphatase (10 U/μl), provided by Epicentre

TE	Tris-EDTA
TF	transcription factor
Tris	2-amino-2-hydroxymethyl-1,3-propanediol
Vortex	device used commonly to mix small vials of liquid
w/v	weight/volume

MEDIA AND SOLUTION

Standard buffers and solutions were prepared, unless otherwise stated, according to Sambrook and Russel (2001) and Frederick M. Ausubel, Roger Brent, Robert E. Kingston, David D. Moore, J.G. Seidman, John A. Smith, Kevin Struhl (eds.) Current Protocols in Molecular Biology 2003 John Wiley & Sons.

mQ filtered water nuclease free (Sigma) was routinely used to prepare all buffers and solutions. Chemicals, organic solvents and enzymes were analytical grade reagents and purchased from Sigma Aldrich Company, New England Biolabs, Promega Corporation and Invitrogen. Where necessary buffers, solutions, media and other materials were sterilized by autoclaving for at least 40 min at 121 °C (130 kPa), or in case of thermo labile reagents by filtration through 0.2µm syringe tip or bottle top filters (Nalgene). Antibiotics and IPTG were prepared using mQ water and kept as frozen stocks and stored at 20 °C until required.

Strain Preservation

Yeast strains can be stored for short periods of time at 4°C on YPD medium in petri dishes or in closed vials (slants). Although most strains remain viable at 4°C for at least one year, many strains fail to survive even for a few months. Yeast strains can be stored indefinitely in 15% (v/v) glycerol at -60°C or lower temperature. The strains are first grown on the surfaces of YPD plates; the yeast is then scraped-up with sterile applicator sticks and suspended in the glycerol solution. The caps are tightened and the vials shaken before freezing. The yeast can be revived by transferring a small portion of the frozen sample to a YPD plate.

Sanger sequencing

DNA sequencing of PCR amplicates was performed using the Sanger method by BMR genomics. Linear DNA necessary for sequencing reaction was usually 20 ng/Kbase. The reaction mixture that was sent off usually contained also 3,2 pmol of a primer. The all volume was heat dried at 65°C.

Agarose Gel Electrophoresis

This technique allows the separation of DNA fragments according to their sizes. Gels were prepared with molecular grade agarose (Sigma; final [0.5-1.5 % (w/v)]) dissolved in Tris-acetate-EDTA buffer (TAE; 40 mM Tris-acetate, 1 mM EDTA pH = 8.0) and Ethidium bromide DNA stain (final [1 µg ml]). Samples were loaded in the wells after the addition of 6x loading buffer (30 % (w/v) phicoll, 0.25 % (w/v) orange, 0.25 % (w/v) xylene cyanol; final [1x]). The gels were run in TAE-buffer at 80 to 100 V in a horizontal gel apparatus. DNA could be visualised by using a UV transilluminator and photographs were taken. To estimate

the size of unknown DNA fragments a DNA marker was loaded in one lane of the gel. We routinely used the Generuler series marker (Fermentas) or occasionally other ladders either from New England Biolabs or Promega. Specific indication about the ladder used will be always indicated in the gel pictures.

TAE buffer (50X)

Running electrophoresis buffer, gel's component, recipe for 1 L:

242 g of Tris base, 57.1 mL of acetic acid, 100 mL of 0.5 M EDTA (pH 8.0), water to 1L.

Agarose Gel 1% (50 ml)

0.5 g of agarose, 50 mL of filtered TAE 1X buffer, 2.5 µL of EtBr.

FA gel running buffer (1l)

100 ml of 10X FA gel buffer, 20 ml of 37% (12.3M) formaldehyde, 880 ml of water

Loading dye

50 mM Tris-HCl, pH 7.6, 0.25% bromphenol blue, 60% glycerol

DNA and RNA Manipulation

Digestion of mitochondrial DNA using HinfI enzyme have been performed for genetic characterizations of oenological strains (Shuller D., 2005). Total reaction volume 15 µl:

- 10 U of HinfI (Fermentas) (1µl)
- 10 µl template DNA (1, 2µg)
- 1.5 µl Buffer 10X
- 2.5µl water

Samples were incubated at 37°C for 2 h.

Acid phenol

Phenol solution saturated with 0.1 M citrate buffer, pH 4.3, provide by Sigma-Aldrich

Basic phenol

Phenol solution equilibrated with 10 mM Tris-HCl, pH 8.0, 1 mM EDTA, provided by Sigma-Aldrich

10X TA Buffer

330 mM Tris-acetate (pH 7.5), 660 mM potassium acetate, 100 mM magnesium acetate, and 5 mM DTT

10X TAP Buffer

0.5 M sodium acetate (pH 6.0), 10 mM EDTA, 1% β-mercaptoethanol, and 0.1% Triton® X-100, provided by Epicentre

Buffer SPG

10mM NaH₂PO₄ in glicerolo 50% con 25 mg/ml di enzima litico da *Rhizoctonia solani* o 2 mg/ml di lyticase

Buffer LET

500mM EDTA, 10mM Tris, pH 7.5

Buffer NDS

500mM EDTA, 500mM Tris, 1% laurylsarcosine, pH 7.5

DTT buffer

50 mM Tris-HCl (pH 8), 20 mM DTT (MW 154.2), 5 mM EDTA

Tris EDTA

50 mM Tris-HCl, 20 mM EDTA, pH 8

TE buffer

10mM Tris-HCl, 1 mM EDTA, pH 8

Sorbitol buffer

1.2 M sorbitolo (MW 182.17) in 50 ml, 20 mM K₂HPO₄ (MW 228.23) in 50 ml, pH 7.5

QBT

750 mM sodium chloride, 50 mM MOPS (morpholinepropanesulfonic acid), 15% ethanol, and 0.15% Triton X-100, pH 7

QC buffer

1 M sodium chloride, 50 mM MOPS, and 15% ethanol, and the final pH was adjusted to 7; QF was made with 1.25 M sodium chloride, 50 mM Tris, and 15% ethanol, pH 8.5

Solution I

1 M sorbitol and 100 mM EDTA, pH 7.5

Solution II

contained 50 mM Tris and 20 mM EDTA, pH 7.5

Growth Media

Cells were routinely grown in YPD medium (1% yeast extract, 2% peptone and 2% glucose) at 28 °C, with shaking. Zymolyase 20000 was purchased from Seikagaku (Seikagaku Kogyo Co., Ltd., Tokyo, Japan); glucose, sorbitol, glycerol, and all other chemicals used were purchased from Sigma Chemical Co. (St. Louis, Mo.).

YPD

Yeast extract 1%, Peptone 2% and Glucose 2%

Add water to reach the desiderated volume, sterilize by autoclaving for 20 min at 121 °C.

Add 2% of Bacto Agar (Difco) to the previous recipe for solid media.

Presporulation medium 1 (PRE₁)

1% Difco yeast extract, 1% Bacto Peptone, 1% glucose

PRE₂

1% Difco yeast extract, 1% Bacto Peptone, 1% potassium acetate

PRE₃

0.3% Difco yeast extract, 0.35% Bacto Peptone, 1% potassium acetate, 0.1% MgSO₄, 0.1% (NH₄)₂SO₄, 0.2% KH₂PO₄

PRE₅

0.8% Difco yeast extract, 0.3% Bacto Peptone, 10% glucose

PRE₆

0.8% Difco yeast extract, 0.3% Bacto Peptone, 5% potassium acetate sporulation medium 1 (SPO₁)

1% potassium acetate, 0.1% Difco yeast extract, 0.05% glucose

SPO₂

0.5% potassium acetate

SOS medium

For protoplast regeneration, 1% Difco yeast extract, 2% Bacto Peptone, 2% glucose, 10 mM Cl₂Ca, and 1.2 M sorbitol.

Biggy agar 1l

Yeast extract 1g, Glycin 10g, Glucosie 10g, Sulphite ammonium 3g, Bismuth Ammonium Citrate 5g, Agar 16g. pH 6.8, do Not heat.

Fucsina agar 1l

5 g/l Peptone (DIFCO), 3 g/l malt extrac (DIFCO), 10 g/l Glucose (PROLABO), 0,002 g/l Fucine (SIGMA), 16 g/l Agar (DIFCO).

MS300 (synthetic must) 1l

Macroelements: 200 g Glucose, 0,155 g CaCl₂*2H₂O, 0,2 g NaCl, 0,75 g KH₂PO₄, 0,25 g MgSO₄*7H₂O, 0,5 g K₂SO₄, 0,46 g (NH₄)Cl, 6 g malic acid, 6 g Acido Citrico,

Microelements: leucine 3,70 g, threonine 5,80 g, glycine 1,40 g, *Glutamine* 38,60 g, alanine 11,10 g, valine 3,40 g, *Methionine* 2,40 g, *phenyl alanine* 2,90 g, serine 6,00 g, *Histidine* 2,50 g, *Lysine* 1,30 g, *Cysteine* 1,00 g, *Proline* 46,80 g, 4 g MnSO₄*H₂O, 4 g ZnSO₄*7 H₂O, 1 g CuSO₄*5H₂O, 1 g KI, 0,4 g Co Cl₂, 1 g H₃BO₃, 1 g (NH₄)₆Mo₇O₂₄*4H₂O,

Vitamins: 20 g MYO Inositolo, 2 g Acido Nicotinico, 1,5 g Calcio Pantotenato, 0,25 g Tiamina Cloridrato, 0,25 g Piridossina Cloridrato, 0,003 g Biotina,

Aminoacids: Tirosina 1,40 g, Triptofano 13,70 g, Isoleucina 2,50 g, , Acidoaspartico 3,40g, Acidoglutammico 9,20g, Arginina 28,60g. Final pH 3.2

Synthetic Must (Delfini 1995) 1l

Macroelements: 0,1 g CaCl₂, 0,1 g NaCl, 1 g KH₂PO₄, 0,5 g MgSO₄*7H₂O, 3 g tartaric acid, 200 g Glucose, 0,2 g Hydrolyzed Casein, 2 g Malic acid

Microelements (stock 1000X), 200 mg/L NaMoO₄*2H₂O, 400 mg/L ZnSO₄*7H₂O, 500 mg/L H₃BO₃, 40 mg/L CuSO₄*5H₂O, 100 mg/L KI, 400 mg/L MnSO₄*H₂O, ,

Fe (stock 1000X) 0,4 mg FeCl₃*6H₂O

Vitamins (stock 1000X) 400 mg/L Piridossin cloridrate, 400 mg/L Tiamin cloridrate, 2 g/L Inosite, 20 mg/L Biotin, 400 mg/L calcium pantothenate, 400 mg/L nicotinic acid amide, 200 mg/L P-amino-benzoic acid, 0,3 g (NH₄)₂SO₄

primary Identifier	secondary Identifier	symbol	name	sgdAlias	length	nsy n	sto p	syn up
S000005443	YOL083W	ATG34	AuTophaGy related	ATG19-B	1239	5	10	5
S000003870	YJR109C	CPA2	Carbamyl Phosphate synthetase A		3357	1	1	1
S000001762	YKR054C	DYN1	DYNein	PAC6 DHC1 YML010W-B YML010C-B	12279		1	1
S000004469	YML009C-A				327		2	2
S000001191	YHR148W	IMP3	Interacting with Mpp10p		552		1	1
S000003842	YJR082C	EAF6	Esa1p-Associated Factor		342	2	16	16
S000001593	YKL110C	KTI12	Kluveromyces lactis Toxin Insensitive	TOT4	942		5	5
S000005042	YNL098C	RAS2	homologous to RAS proto-oncogene	CTN5 CYR3 TSL7 GLC5	969		1	1
S000004604	YMR002W	MIC17	Mitochondrial Intermembrane space Cysteine motif protein of 17 kDa		471		1	1
S000001592	YKL109W	HAP4	Heme Activator Protein		1665		3	6
S000001187	YHR144C	DCD1	dCMP Deaminase		939		1	1
S000001945	YFR049W	YMR31	Yeast Mitochondrial Ribosomal protein		372		2	2
S000000794	YEL068C				333		1	1
S000003846	YJR086W	STE18	STErile		333	1	19	19

S000003693	YJL157C	FAR1	Factor ARest	2493	2
S000001188	YHR145C			357	1
S000003839	YJR078W	BNA2	Biosynthesis of Nicotinic Acid	1362	1
S000003856	YJR095W	SFC1	Succinate-Fumarate Carrier	969	13
S000000390	YBR186W	PCH2	Pachytene CCheckpoint	1808	1
S000004662	YMR058W	FET3	FErrous Transport	1911	1
S000004453	YLR461W	PAU4	seriPAUperin family	363	1
S000004718	YMR112C	MED11	MEDIator complex	396	1
S000028603	YBR182C-A			195	1
S000003230	YGL261C	PAU11	seriPAUperin	363	1
S000006013	YPL092W	SSU1		1377	1
S000002327	YDL168W	SFA1	Sensitive to FormAldehyde	1161	1
S000003841	YJR080C	AIM24	Altered Inheritance rate of Mitochondria	1185	3
S000001381	YIL119C	RPI1	Ras-cAMP Pathway Inhibitor	1224	1
S000005815	YOR289W			756	2
S000005713	YOR187W	TUF1	tufM	1314	1

S000004467	YML008C	ERG6	ERGosterol biosynthesis	VID1 ISE1 LIS1 SED6	1152	2
S000004464	YML005W	TRM12	TRna Methyltransferase	TYW2	1389	1
S000001359	YIL097W	FYV10	Function required for Yeast Viability	GID9	1551	1
S000001929	YFR033C	QCR6	ubiQuinol-cytochrome C oxidoReductase	UCR6 COR3	444	1
S000000066	YAR008W	SEN34	Splicing ENdonuclease	tRNA splicing endonuclease subunit FUN4	828	18
S000004785	YMR173W-A				1185	1
S000001590	YKL107W				930	9
S000004991	YNL046W				519	1
S000004666	YMR062C	ARG7	ARGinine requiring	ECM40	1326	6 1
S000004356	YLR364W	GRX8	GlutaRedoXin		330	2
S000003854	YJR094C	IME1	Inducer of MEiosis		1083	5 1 8
S000003853	YJR093C	FIP1	Factor Interacting with Poly(A) polymerase		984	5 1 2 14
S000003848	YJR088C	EMC2	ER Membrane protein Complex		879	4 1 1
S000004349	YLR357W	RSC2	Remodel the Structure of Chromatin		2670	3 2

S000003844	YJR084W			CSN12	1272	3	3	11
S000003801	YJR040W	GEF1	Glycerol Ethanol, Ferric requiring	CLC	2340	3		2
S000005441	YOL081W	IRA2	Inhibitory Regulator of the RAS-cAMP pathway	CCS1 GLC4	9240	2	1	38
S000004353	YLR361C	DCR2	Dose-dependent Cell cycle Regulator		1737	2	1	2
S000003852	YJR092W	BUD4	BUD site selection		4344	2	1	1
S000003580	YJL044C	GYP6	Gtpase-activating protein of Ypt6 Protein		1377	2		
S000002553	YDR146C	SWI5	SWItching deficient		2130	2		
S000000669	YCR073C	SSK22	Suppressor of Sensor Kinase		3996	2	1	
S000000001	YAL001C	TFC3	Transcription Factor class C	TSV115 tau 138 FUN24	3573	2		4
S000005813	YOR287C	RRP36	Ribosomal RNA Processing		903	1		
S000005164	YNL220W	ADE12	ADEnine requiring	BRA9	1302	1		
S000004799	YMR187C				1296	1		4
S000004775	YMR165C	PAH1	Phosphatidic Acid phosphoHydrolase	SMIP2	2589	1		
S000004719	YMR113W	FOL3	FOLic acid synthesis		1284	1		
S000004664	YMR060C	SAM37	Sorting and Assembly Machinery	PET3027 TOM37 MAS37				
					984	1	1	7

S000004463	YML004C	GLO1	GlyOxalase	981	1	
S000004365	YLR373C	VID22	Vacuolar Import and Degradation	2706	1	1
S000004361	YLR369W	SSQ1	Stress-Seventy subfamily Q	1974	1	2
S000004354	YLR362W	STE11	STerile	2154	1	1 3
S000004350	YLR358C			564	1	1 3
S000004305	YLR313C	SPH1	SPa2 Homolog	1593	1	
S000001806	YKR098C	UBP11	UBiquitin-specific Protease	2154	1	
S000001697	YKL214C	YRA2	Yeast RNA Annealing protein	612	1	
S000001671	YKL188C	PXA2	PeroXisomal ABC-transporter	2562	1	
S000001595	YKL112W	ABF1	ARS-Binding Factor 1	2196	1	2
S000001594	YKL111C			336	1	1 5
S000001591	YKL108W	SLD2	Synthetically Lethal with Dpb11-1	1362	1	2 5
S000003849	YJR089W	BIR1	Baculoviral IAP Repeat-containing protein	2865	1	17

S000003847	YJR087W							351	1	1	18
S000003840	YJR079W							1035	1	1	2
S000003695	YJL159W	HSP150	Heat Shock Protein				ORE1 PIR2 CCW7	1242	1		
S000003692	YJL156C	SSY5	Sulfonylurea Sensitive on YPD					2100	1	1	1
S000003582	YJL046W	AIM22	Altered Inheritance rate of Mitochondria				RRG3 LIP3	1230	1		
S000001466	YIR027C	DAL1	Degradation of Allantoin					1383	1	1	
S000001211	YHR168W	MTG2	Mitochondrial Gtpase 2					1557	1	1	3
S000001210	YHR167W	THP2	THO2 - HPR1 Phenotype					786	1	1	1
S000001205	YHR162W							390	1		
S000001162	YHR120W	MSH1	MutS Homolog					2880	1		
S000001943	YFR047C	BNA6	Biosynthesis of Nicotinic Acid				QPT1	888	1	1	
S000000959	YER157W	COG3	Conserved Oligomeric Golgi complex					2406	1		
S000000791	YEL065W	SIT1	Siderophore Iron Transport				ARN3	1887	1		
S000000781	YEL055C	POL5	POLYmerase					3069	1	2	
S000002883	YDR475C	JIP4	Jumonji domain Interacting Protein				YDR474C	2631	1	2	1

S000002552	YDR145W	TAF12	TATA binding protein-Associated Factor	TAF68 TAF61 TafII61	1620	1	4	5
S000002551	YDR144C	MKC7	Multicopy suppressor of Kex2 Cold sensitivity	TafII68	1791	1		8
S000002550	YDR143C	SAN1	Sir Antagonist	YPS2	1833	1		7
S000003851	YJR091C	JSN1	Just Say No	PUF1	3276	1		7
S000005442	YOL082W	ATG19	AuTophaGy related	CVT19	1248		3	51
S000028853	YOL083C-A				141		1	35
S000003855	YJR094W-A	RPL43B	Ribosomal Protein of the Large subunit	L43B	554			21
S000005445	YOL085C				342			21
S000003843	YJR083C	ACF4	Assembly Complementing Factor		930			17
S000002129	YAR002C-A	ERP1	Emp24p/Erv25p Related Protein		660		2	16
S000001699	YKL216W	URA1	URAcil requiring		945			13
S000003845	YJR085C				318			12
S000003838	YJR077C	MIR1		PTP	936			11
S000004667	YMR063W	RIM9	Regulator of IME2		720			9

S000003857	YJR096W					849	8
S000004669	YMR065W	KAR5	KARYogamy			1515	7
S000004668	YMR064W	AEF1	ATPase Expression		FIG3 NCA1	1557	7
S000002554	YDR147W	EK11	Ethanolamine Kinase			1605	7
S000005806	YOR280C	FSH3	Family of Serine Hydrolases			801	6
S000004352	YLR360W	VPS38	Vacuolar Protein Sorting		VPL17	1320	6
S000004351	YLR359W	ADE13	ADENine requiring		BRA1 BRA8	1449	6
S000004347	YLR355C	ILV5	IsoLeucine-plus-Valine requiring			1188	6
S000001589	YKL106W	AAT1	Aspartate AminoTransferase			1356	6
S000003837	YJR076C	CDC11	Cell Division Cycle		PSL9	1248	6
S000005812	YOR286W	RDL2	RhoDanese-Like protein		AIM42 FMP31	450	1
S000005811	YOR285W	RDL1	RhoDanese-Like protein			420	5
S000005810	YOR284W	HUA2				732	5
S000005805	YOR279C	RFM1	Repression Factor of Middle sporulation element			933	5

S000003799	YJR038C						363	5
S000001213	YHR170W	NMD3	Nonsense-Mediated mRNA Decay			SRC5	1557	5
S000001212	YHR169W	DBP8	Dead Box Protein				1296	5
S000001209	YHR166C	CDC23	Cell Division Cycle			USA2 SLT21 RNA8 DNA39	1881	5
S000001208	YHR165C	PRP8	Pre-mRNA Processing			DBF3	7242	5
S000002881	YDR473C	PRP3	Pre-mRNA Processing			RNA3	1410	5
S000002549	YDR142C	PEX7	PEroXin			PAS7 PEB1	1128	5
S000005809	YOR283W						693	1 4
S000005808	YOR282W						321	4
S000005803	YOR277C						309	4
S000004358	YLR366W						306	4
S000004357	YLR365W						333	4
S000001597	YKL114C	APN1	APurinic/apyrimidinic eNdonuclease				1104	4
S000001596	YKL113C	RAD27	RADIation sensitive			RTH1 ERC11 FEN1	1149	4
S000001405	YIL143C	SSL2	Suppressor of Stem-Loop mutation			LOM3 RAD25	2532	4

S000001404	YIL142W	CCT2	Chaperonin Containing TCP-1	TCP2 BIN3	1584	4
S000001403	YIL141W				390	4
S000001402	YIL140W	AXL2	AXiaL budding pattern	BUD10 SRO4 HAF3 PAS14 CAT1 GLC2 CCR1	2472	4
S000002885	YDR477W	SNF1	Sucrose NonFermenting		1902	4
S000028533	YBR076C-A				267	4
S000007296	snR31	SNR31	Small Nucleolar RNA	SUF8	225	4
S000005807	YOR281C	PLP2	Phosducin-Like Protein		861	4 3
S000005446	YOL086C	ADH1	Alcohol DeHydrogenase	ADC1	1047	3
S000004800	YMR188C	MRPS17	Mitochondrial Ribosomal Protein, Small subunit		714	1 3
S000004665	YMR061W	RNA14	RNA synthesis		2034	3
S000004661	YMR057C				372	3
S000004660	YMR056C	AAC1	ADP/ATP Carrier	S24B YS22 rp50 S22B	930	3
S000004359	YLR367W	RPS22B	Ribosomal Protein of the Small subunit		876	3
S000007620	YLR363W-A				258	3
S000003797	YJR036C	HUL4	Hect Ubiquitin Ligase		2679	3
S000003579	YJL043W				774	3
S000003578	YJL042W	MHP1	MAP-Homologous Protein		4197	3

S000028744	YEL053W-A						348	3
S000002548	YDR141C	DOP1	homolog of A. nidulans DOPey				5097	3
S000028519	YCR075W-A						228	3
S000000718	YCR073W-A	SOL2	Suppressor Of Los1-1			YCRX13W	948	3
S000000668	YCR072C	RSA4	RiboSome Assembly				1548	3
S000000064	YAR003W	SWD1	Set1c, WD40 repeat protein			FUN16 CPS50	1281	2
S000006504	snR44	SNR44	Small Nucleolar RNA			SAF49	211	3
S000005974	YPL053C	KTR6	Kre Two Related			MNN6	1341	2
S000005971	YPL050C	MNN9	MaNNosyltransferase				1188	2
S000005969	YPL048W	CAM1	Calcium And Membrane-binding protein			CPBP TEF3	1248	2
S000005968	YPL047W	SGF11	SaGa associated Factor 11kDa				300	2
S000005068	YNL124W	NAF1	Nuclear Assembly Factor				1479	2
S000005067	YNL123W	NMA111	Nuclear Mediator of Apoptosis			YNM3	2994	2
S000004801	YMR189W	GCV2	GlyCine cleaVage			GSD2	3105	3
S000004468	YML009C	MRPL39	Mitochondrial Ribosomal Protein, Large subunit			YmL39	213	2
S000004465	YML006C	GIS4	Glg1-2 Suppressor				2325	2
S000004462	YML003W						873	2

S000004461	YML002W				2214	2
S000004443	YLR451W	LEU3	LEUCine biosynthesis		2661	2
S000004435	YLR443W	ECM7	ExtraCellular Mutant	ZRG15	1347	2
S000004433	YLR441C	RPS1A	Ribosomal Protein of the Small subunit	RP10A S1A rp10A	768	2
S000004432	YLR440C	SEC39	SECretory	DSL3	2130	2
S000004367	YLR375W	STP3	protein with similarity to Stp1p		1032	2
S000004360	YLR368W	MDM30	Mitochondrial Distribution and Morphology	DSG1	1797	2
S000004355	YLR363C	NMD4	Nonsense-Mediated mRNA Decay		657	2
S000028845	YLR361C-A				297	2
S000004346	YLR354C	TAL1	TransALdolase		1008	2
S000004318	YLR326W				723	2
S000004315	YLR323C	CWC24	Complexed With Cef1p		780	2
S000004313	YLR321C	SFH1	Snf Five Homolog		1281	2
S000004307	YLR315W	NKP2	Non-essential Kinetochore Protein		462	2
S000004303	YLR312C			QNQ1	1197	2
S000004302	YLR311C				348	2
S000004301	YLR310C	CDC25	Cell Division Cycle	CTN1 CDC25'	4770	2

S000001599	YKL116C	PRR1	Pheromone Response Regulator	1557	2
S000001598	YKL115C			393	2
S000003871	YJR110W	YMR1	Yeast Myotubularin Related	2067	2
S000003696	YJL160C			864	2
S000003694	YJL158C	CIS3	Clk1 Suppressing	684	2
S000001468	YIR029W	DAL2	Degradation of Allantoin	1032	2
S000001467	YIR028W	DAL4	Degradation of Allantoin	1908	2
S000001465	YIR026C	YVH1	Yeast vaccinia virus VH1 Homolog	1095	2
S000001463	YIR024C			651	2
S000028801	YIR023C-A			324	2
S000001214	YHR171W	ATG7	AuTophagy related	1893	2
S000001944	YFR048W	RMD8	Required for Meiotic nuclear Division	1989	2
S000001942	YFR046C	CNN1	Co-purified with NNF1p	1086	2
S000001940	YFR044C	DUG1	Deficient in Utilization of Glutathione	1446	2
S000001939	YFR043C	IRC6	Increased Recombination Centers	714	2

S000000667	YCR071C	IMG2	Integrity of Mitochondrial Genome	441	2
S000000282	YBR078W	ECM33	ExtraCellular Mutant	1620	2
S000000281	YBR077C	SLM4	Synthetic Lethal with Mss4	489	2
S000000063	YAR002W	NUP60	NUclear Pore	1620	2
S000028592	YAL037C-B			975	2
S000028732	YAL037C-A			93	2
S000000034	YAL036C	RBG1	RiBosome interacting Gtpase	1110	2
S000000032	YAL034W-A	MTW1	Mis TWelve-like	870	2
S000006732	tS(GCU)L			100	2
S000006729	tS(AGA)M			82	2
S000006020	YPL099C	AIM43	Altered Inheritance rate of Mitochondria	549	1
S000006019	YPL098C	MGR2	Mitochondrial Genome Required	342	1
S000028423	YPL096C-A	ERI1	ER-associated Ras Inhibitor	207	1
S000006014	YPL093W	NOG1	NucleOlar G-protein	1944	1
S000005966	YPL045W	VPS16	Vacuolar Protein Sorting	2397	1
S000005817	YOR291W	YPK9	Yeast PARK9	4419	1

S000005714	YOR188W	MSB1	Multicopy Suppressor of a Budding defect	3414	1
S000028715	YOR186C-A			210	1
S000005711	YOR185C	GSP2	Genetic Suppressor of Prp20-1	663	1
S000005165	YNL221C	POP1	Processing Of Precursor RNAs	2628	1
S000005162	YNL218W	MG51	Maintenance of Genome Stability	1764	1
S000005066	YNL122C			348	1
S000005063	YNL119W	NCS2	Needs Cia4 to Survive	1482	1
S000005043	YNL099C	OCA1	Oxidant-induced Cell-cycle Arrest	717	1
S000028699	YNL097C-B			123	1
S000005041	YNL097C	PHO23	PHOSphate metabolism	993	1
S000005038	YNL094W	APP1	Actin Patch Protein	1764	1
S000004994	YNL049C	SFB2	Sed Five Binding	2631	1
S000004990	YNL045W	LAP2	Leucine AminoPeptidases	2016	1
S000004784	YMR173W	DDR48	DNA Damage Responsive	1293	1
S000004783	YMR172W	HOT1	High-Osmolarity-induced Transcription	2160	1
S000004781	YMR171C	EAR1	Endosomal Adaptor of Rsp5p	1653	1

S000004780	YMR170C	ALD2	Aldehyde Dehydrogenase	1521	1
S000004777	YMR167W	MLH1	MutL Homolog	2310	1
S000004774	YMR164C	MSS11	Multicopy Suppressor of STA genes	2277	1
S000004721	YMR115W	MGR3	Mitochondrial Genome Required	1506	1
S000004717	YMR111C			1389	1
S000004716	YMR110C	HFD1	Homolog of Fatty aldehyde Dehydrogenase	1599	1
S000004670	YMR066W	SOV1	Synthesis Of Var	2697	1
S000004663	YMR059W	SEN15	Splicing ENdonuclease	387	1
S000004659	YMR055C	BUB2	Budding Uninhibited by Benzimidazole	921	1
S000004605	YMR003W	AIM34	Altered Inheritance rate of Mitochondria	597	1
S000004466	YML007W	YAP1	Yeast AP-1	1953	1
S000007621	YML007C-A			111	1
S000004449	YLR457C	NBP1	Nap1 Binding Protein	960	1
S000004437	YLR445W			649	1

S000006434	YLR438C-A	LSM3	Like SM	SMX4 USS2	270	1
S000004363	YLR371W	ROM2	Rho1 Multicopy suppressor		4071	1
S000004320	YLR328W	NMA1	Nicotinamide Mononucleotide Adenylyltransferase		1206	1
S000004316	YLR324W	PEX30	PEroXisome related		1572	2 1
S000004314	YLR322W	VPS65	Vacuolar Protein Sorting		315	1
S000004311	YLR319C	BUD6	BUD site selection	AIP3	2367	1
S000004309	YLR317W				435	1
S000004298	YLR307W	CDA1	Chitin DeAcetylase		906	1
S000004297	YLR306W	UBC12	UBiquitin-Conjugating		701	1
S000004295	YLR304C	ACO1	ACOnitase	GLU1	2337	1
S000004076	YLR086W	SMC4	Structural Maintenance of Chromosomes		4257	1 1
S000004075	YLR085C	ARP6	Actin-Related Protein		1317	1
S000004074	YLR084C	RAX2			3663	1 1
S000004073	YLR083C	EMP70		p24a TMN1	2004	1

S000003963	YLL040C	VPS13	Vacuolar Protein Sorting	VPT2 SOI1	9435	1
S000003962	YLL039C	UBI4	Ubiquitin	SCD2 UB14	1146	1
S000003961	YLL038C	ENT4	Epsin N-Terminal homology		744	1
S000003959	YLL036C	PRP19	Pre-RNA Processing	PSO4	1512	1
S000003958	YLL035W	GRC3			1899	1
S000001807	YKR099W	BAS1	BASal		2436	1
S000001764	YKR056W	TRM2	Transfer RNA Methyltransferase	NUD1 NUC2 RNC1	1920	1
S000001763	YKR055W	RHO4	Ras HOmolog		876	1
S000001755	YKR047W				306	1
S000001752	YKR044W	UIP5	Ulp1 Interacting Protein		1332	1
S000001748	YKR040C				504	1
S000001698	YKL215C	EXP1	OXoProlinase		3861	1
S000001695	YKL212W	SAC1	Suppressor of Actin	RSD1 COT2 SSU2 CAT80 SDC1	1872	1
S000003850	YJR090C	GRR1	Glucose Repression-Resistant		3456	1
S000007613	YJL156W-A				222	1

S000003691	YJL155C	FBP26	Fructose BisPhosphatase	1359	1
S000003688	YJL152W			360	1
S000003584	YJL048C	UBX6	UBiquitin regulatory X	1191	1
S000028804	YJL047C-A			135	1
S000003583	YJL047C	RTT101	Regulator of Ty1 Transposition	2529	1
S000003581	YJL045W			1905	1
S000001380	YIL118W	RHO3	Ras HOMolog	696	1
S000113587	YIL102C-A			228	1
S000001364	YIL102C			306	1
S000001363	YIL101C	XBP1	XhoI site-Binding Protein	1944	1
S000028794	YIL100C-A			339	1
S000001361	YIL099W	SGA1	Sporulation-specific GlycoAmylase	1650	1
S000001219	YHR176W	FMO1		1299	1
S000028553	YHR175W-A			150	1
S000001218	YHR175W	CTR2	Copper TRansport	570	1
S000001217	YHR174W	ENO2	ENO2	1314	1
S000001216	YHR173C			339	1

S000001206	YHR163W	SOL3	Suppressor Of Los1-1	750	1
S000001204	YHR161C	YAP1801	Yeast Assembly Polypeptide	1914	1
S000001203	YHR160C	PEX18	PEroXin	852	1
S000001166	YHR124W	NDT80	Non-DiTyrosine	1884	1
S000001165	YHR123W	EPT1		1267	1
S000001164	YHR122W		Cytosolic Iron-sulfur protein Assembly	696	1
S000001163	YHR121W	LSM12	Like SM	564	1
S000003227	YGL258W	VEL1	VELum formation	621	1
S000003225	YGL256W	ADH4	Alcohol DeHydrogenase	1149	1
S000003086	YGL118C			438	1
S000003083	YGL115W	SNF4	Sucrose NonFermenting	969	1
S000003082	YGL114W			2178	1
S000003081	YGL113W	SLD3	Synthetically Lethal with Dpb11-1	2007	1
S000001936	YFR040W	SAP155	Sit4 Associated Protein	3009	1
S000001934	YFR038W	IRC5	Increased Recombination Centers	2562	1
			ZRG5 NRC465		
			CAT3 SCI1		

S000001933	YFR037C	RSC8	Remodel the Structure of Chromatin	SWH3	1674	1
S000001931	YFR035C				345	1
S000001930	YFR034C	PHO4	PHOSphate metabolism	phoD	939	1
S000000958	YER156C				1017	1
S000000957	YER155C	BEM2	Bud EMergence	SUP9 IPL2	6504	1
S000000919	YER117W	RPL23B	Ribosomal Protein of the Large subunit	TSL1		
				L23B L17aB		
				YL32	885	1
S000000916	YER114C	BOI2	Bem1 (One) Interacting protein	BEB1	3123	1
S000000915	YER113C	TMN3	TransMembrane Nine		2121	1
S000000793	YEL067C				588	1
S000000789	YEL063C	CAN1	CANavanine resistance		1773	1
S000000785	YEL059W	HHY1	Hypersensitivity to HYgromycin B	ENV6	309	1
S000000784	YEL058W	PCM1	PhosphoaCetylgucosamine Mutase	AGM1	1674	1
S000000783	YEL057C				702	1
S000002886	YDR478W	SNM1	Suppressor of Nuclear Mitochondrial endoribonuclease		597	1

S000002330	YDL171C	GLT1	GLuTamate synthase	6438	1
S000002328	YDL169C	UGX2	Unidentified Gene X	672	1
S000002324	YDL165W	CDC36	Cell Division Cycle	576	1
S000002273	YDL115C	IWR1	Interacts With RNA polymerase II	1132	1
S000002271	YDL113C	ATG20	AuTophaGy related	1923	1
S000002163	YDL005C	MED2	MEDIator complex	1296	1
S000002159	YDL001W	RMD1	Required for Meiotic nuclear Division	1293	1
S000000419	YBR215W	HPC2	Histone Periodic Control	1962	1
S000000418	YBR214W	SDS24	homolog of S. pombe SDS23	1584	1
S000000415	YBR211C	AME1	Associated with Microtubules and Essential	975	1
S000000393	YBR189W	RPS9B	Ribosomal Protein of the Small subunit	1001	1
S000000391	YBR187W	GDT1	Gcr1 Dependent Translation factor	843	1
S000000033	YAL035W	FUN12	Function Unknown Now	3009	1
S000006451	snR60	SNR60	Small Nucleolar RNA	104	1
			ARP100		
			S13 SUP46		
			RPS13A S9B		
			rp21 YS11		
			eIF5B yIF2		

S000007300	snR36	SNR36	Small Nucleolar RNA	182	1
S000007295	snR19	SNR19	Small Nuclear RNA	568	1
S000006016	YPL095C	EEB1	Ethyl Ester Biosynthesis	1371	1
S000005970	YPL049C	DIG1	Down-regulator of Invasive Growth	1359	1
S000005816	YOR290C	SNF2	Sucrose NonFermenting	5112	1
S000005069	YNL125C	ESBP6		2022	2
S000005065	YNL121C	TOM70	Translocase of the Outer Mitochondrial membrane	1854	1
S000004992	YNL047C	SLM2	Synthetic Lethal with Mss4	1971	1
S000004451	YLR459W	GAB1	GPI and Actin Bar	1185	1
S000004442	YLR450W	HMG2	3-Hydroxy-3-MethylGlutaryl-coenzyme a reductase	3138	2
S000004434	YLR442C	SIR3	Silent Information Regulator	2937	1
S000004312	YLR320W	MMS22	Methyl MethaneSulfonate sensitivity	4365	1
S000004304	YLR312W-A	MRPL15	Mitochondrial Ribosomal Protein, Large subunit	762	1
S000004296	YLR305C	STT4	STaurosporine and Temperature sensitive	5703	1

S000001751	YKR043C	SHB17	SedoHeptulose 1,7-Bisphosphatase	816	1
S000003690	YJL154C	VPS35	Vacuolar Protein Sorting	2835	1
S000003576	YJL039C	NUP192	NUclear Pore	5052	1
S000001189	YHR146W	CRP1	Cruciform DNA-Recognizing Protein	1398	1
S000000787	YEL061C	CIN8	Chromosome INstability	3003	2
S000002270	YDL112W	TRM3	Transfer RNA Methyltransferase	4311	1
S000000417	YBR213W	MET8	METHionine requiring	825	1
S000000389	YBR185C	MBA1	Multi-copy Bypass of AFG3	837	1
S000000065	YAR007C	RFA1	Replication Factor A	1866	1
	tD(GUC)J2				1
	tK(CUU)F				1
	tK(UUU)L				1
	tL(CAA)N				1
	tP(UGG)H				1
	tR(UCU)J1				1
	tW(CCA)G1				1
	tY(GUA)J1				1
			BUF2 RPA1		
			FUN3		

secondary Identifier	Description	secondary Identifier
YOL083W	Receptor protein involved in selective autophagy during starvation; specifically involved in the transport of cargo protein alpha-mannosidase (Ams1p); Atg19p paralog	YOL083W
YJR109C	Large subunit of carbamoyl phosphate synthetase, which catalyzes a step in the synthesis of citrulline, an arginine precursor	YJR109C
YKR054C	Cytoplasmic heavy chain dynein, microtubule motor protein, required for anaphase spindle elongation; involved in spindle assembly, chromosome movement, and spindle orientation during cell division, targeted to microtubule tips by Pac1p	YKR054C
YML009C-A	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YML009C-A
YHR148W	Component of the SSU processome, which is required for pre-18S rRNA processing, essential protein that interacts with Mpp10p and mediates interactions of Imp4p and Mpp10p with U3 snoRNA	YHR148W
YJR082C	Subunit of the NuA4 acetyltransferase complex that acetylates histone H4 and NuA3 acetyltransferase complex that acetylates histone H3	YJR082C
YKL110C	Protein that plays a role, with Elongator complex, in modification of wobble nucleosides in tRNA; involved in sensitivity to G1 arrest induced by zymocin; interacts with chromatin throughout the genome; also interacts with Cdc19p	YKL110C
YNL098C	GTP-binding protein that regulates the nitrogen starvation response, sporulation, and filamentous growth; farnesylation and palmitoylation required for activity and localization to plasma membrane; homolog of mammalian Ras proto-oncogenes	YNL098C
YMR002W	Mitochondrial intermembrane space protein, required for normal oxygen consumption; contains twin cysteine-x9-cysteine motifs	YMR002W
YKL109W	Subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex, a transcriptional activator and global regulator of respiratory gene expression; provides the principal activation function of the complex	YKL109W
YHR144C	Deoxycytidine monophosphate (dCMP) deaminase required for dCTP and dTTP synthesis; expression is NOT cell cycle regulated	YHR144C
YFR049W	Mitochondrial ribosomal protein of the small subunit, has similarity to human mitochondrial ribosomal protein MRP-S36	YFR049W
YEL068C	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YEL068C
YJR086W	G protein gamma subunit, forms a dimer with Ste4p to activate the mating signaling pathway, forms a heterotrimer with Gpa1p and Ste4p to dampen signaling; C-terminus is palmitoylated and farnesylated, which are required for normal signaling	YJR086W

YJL157C	Cyclin-dependent kinase inhibitor that mediates cell cycle arrest in response to pheromone; also forms a complex with Cdc24p, Ste4p, and Ste18p that may specify the direction of polarized growth during mating; potential Cdc28p substrate	YJL157C
YHR145C	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YHR145C
YJR078W	Putative tryptophan 2,3-dioxygenase or indoleamine 2,3-dioxygenase, required for de novo biosynthesis of NAD from tryptophan via kynurenine; interacts genetically with telomere capping gene CDC13; regulated by Hst1p and Aftp	YJR078W
YJR095W	Mitochondrial succinate-fumarate transporter, transports succinate into and fumarate out of the mitochondrion; required for ethanol and acetate utilization	YJR095W
YBR186W	Nucleolar component of the pachytene checkpoint, which prevents chromosome segregation when recombination and chromosome synapsis are defective; also represses meiotic interhomolog recombination in the rDNA	YBR186W
YMR058W	Ferro-O2-oxidoreductase required for high-affinity iron uptake and involved in mediating resistance to copper ion toxicity, belongs to class of integral membrane multicopper oxidases	YMR058W
YLR461W	Member of the seripauperin multigene family encoded mainly in subtelomeric regions; active during alcoholic fermentation, regulated by anaerobiosis, negatively regulated by oxygen, repressed by heme	YLR461W
YMR112C	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential protein	YMR112C
YBR182C-A	Putative protein of unknown function; identified by gene-trapping, microarray-based expression analysis, and genome-wide homology searching	YBR182C-A
YGL261C	Putative protein of unknown function and member of the seripauperin multigene family encoded mainly in subtelomeric regions; mRNA expression appears to be regulated by SUT1 and UPC2	YGL261C
YPL092W	Plasma membrane sulfite pump involved in sulfite metabolism and required for efficient sulfite efflux; major facilitator superfamily protein	YPL092W
YDL168W	Bifunctional enzyme containing both alcohol dehydrogenase and glutathione-dependent formaldehyde dehydrogenase activities, functions in formaldehyde detoxification and formation of long chain and complex alcohols, regulated by Hog1p-Sko1p	YDL168W
YJR080C	Protein of unknown function; the authentic, non-tagged protein is detected in purified mitochondria in high-throughput studies; null mutant displays reduced respiratory growth and elevated frequency of mitochondrial genome loss	YJR080C
YIL119C	Putative transcriptional regulator; overexpression suppresses the heat shock sensitivity of wild-type RAS2 overexpression and also suppresses the cell lysis defect of an mpk1 mutation	YIL119C
YOR289W	Putative protein of unknown function; transcription induced by the unfolded protein response; green fluorescent protein (GFP)-fusion protein localizes to both the cytoplasm and the nucleus	YOR289W
YOR187W	Mitochondrial translation elongation factor Tu; comprises both GTPase and guanine nucleotide exchange factor activities, while these activities are found in separate proteins in <i>S. pombe</i> and humans	YOR187W

YML008C	Delta(24)-sterol C-methyltransferase, converts zymosterol to fecosterol in the ergosterol biosynthetic pathway by methylating position C-24; localized to both lipid particles and mitochondrial outer membrane	YML008C
YML005W	S-adenosylmethionine-dependent methyltransferase of the seven beta-strand family; required for wybutosine formation in phenylalanine-accepting tRNA	YML005W
YIL097W	Protein of unknown function, required for survival upon exposure to K1 killer toxin; involved in proteasome-dependent catabolite inactivation of FBPaase; contains CTLH domain; plays role in anti-apoptosis	YIL097W
YFR033C	Subunit 6 of the ubiquinol cytochrome-c reductase complex, which is a component of the mitochondrial inner membrane electron transport chain; highly acidic protein; required for maturation of cytochrome c1	YFR033C
YAR008W	Subunit of the tRNA splicing endonuclease, which is composed of Sen2p, Sen15p, Sen34p, and Sen54p; Sen34p contains the active site for tRNA 3' splice site cleavage and has similarity to Sen2p and to Archaeal tRNA splicing endonuclease	YAR008W
YMR173W-A	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; overlaps the verified gene DDR48/YML173W	YMR173W-A
YKL107W	Putative protein of unknown function; proposed to be a palmitoylated membrane protein	YKL107W
YNL046W	Putative protein of unknown function; expression depends on Swi5p; GFP-fusion protein localizes to the endoplasmic reticulum; deletion confers sensitivity to 4-(N-(S-glutathionylacetyl)amino) phenylarsenoxide (GSAO)	YNL046W
YMR062C	Mitochondrial ornithine acetyltransferase, catalyzes the fifth step in arginine biosynthesis; also possesses acetylglutamate synthase activity, regenerates acetylglutamate while forming ornithine	YMR062C
YLR364W	Glutaredoxin that employs a dithiol mechanism of catalysis; monomeric; activity is low and null mutation does not affect sensitivity to oxidative stress; GFP-fusion protein localizes to the cytoplasm; expression strongly induced by arsenic	YLR364W
YJR094C	Master regulator of meiosis that is active only during meiotic events, activates transcription of early meiotic genes through interaction with Ume6p, degraded by the 26S proteasome following phosphorylation by Ime2p	YJR094C
YJR093C	Subunit of cleavage polyadenylation factor (CPF), interacts directly with poly(A) polymerase (Pap1p) to regulate its activity; bridging factor that links Pap1p and the CPF complex via Yth1p	YJR093C
YJR088C	Member of a transmembrane complex required for efficient folding of proteins in the ER; null mutant displays induction of the unfolded protein response	YJR088C
YLR357W	Component of the RSC chromatin remodeling complex; required for expression of mid-late sporulation-specific genes; involved in telomere maintenance	YLR357W

YJR084W	Protein that forms a complex with Thp3p; may have a role in transcription elongation and/or mRNA splicing; identified as a COP9 signalosome component but phenotype and interactions suggest it may not be involved with the signalosome	YJR084W
YJR040W	Voltage-gated chloride channel localized to the golgi, the endosomal system, and plasma membrane, and involved in cation homeostasis; highly homologous to vertebrate voltage-gated chloride channels	YJR040W
YOL081W	GTPase-activating protein that negatively regulates RAS by converting it from the GTP- to the GDP-bound inactive form, required for reducing cAMP levels under nutrient limiting conditions, has similarity to Ira1p and human neurofibromin	YOL081W
YLR361C	Phosphoesterase involved in downregulation of the unfolded protein response, at least in part via dephosphorylation of Ire1p; dosage-dependent positive regulator of the G1/S phase transition through control of the timing of START	YLR361C
YJR092W	Involved in bud-site selection and required for the axial budding pattern; localizes with septins to bud neck in mitosis and may constitute an axial landmark for next round of budding; required for the formation of a double septin ring, and generally for the organization of septin structures; potential Cdc28p substrate	YJR092W
YJL044C	GTPase-activating protein (GAP) for the yeast Rab family member, Ypt6p; involved in vesicle mediated protein transport	YJL044C
YDR146C	Transcription factor that activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase; localization to the nucleus occurs during G1 and appears to be regulated by phosphorylation by Cdc28p kinase	YDR146C
YCR073C	MAP kinase kinase of the HOG1 mitogen-activated signaling pathway; functionally redundant with, and homologous to, Ssk2p; interacts with and is activated by Ssk1p; phosphorylates Pbs2p	YCR073C
YAL001C	Largest of six subunits of the RNA polymerase III transcription initiation factor complex (TFIIIC); part of the TauB domain of TFIIIC that binds DNA at the BoxB promoter sites of tRNA and similar genes; cooperates with Tfc6p in DNA binding	YAL001C
YOR287C	Component of 90S preribosomes; involved in early cleavages of the 35S pre-rRNA and in production of the 40S ribosomal subunit	YOR287C
YNL220W	Adenylosuccinate synthase, catalyzes the first step in synthesis of adenosine monophosphate from inosine 5'monophosphate during purine nucleotide biosynthesis; exhibits binding to single-stranded autonomously replicating (ARS) core sequence	YNL220W
YMR187C	Putative protein of unknown function; YMR187C is not an essential gene	YMR187C
YMR165C	Mg²⁺-dependent phosphatidate (PA) phosphatase, catalyzes the dephosphorylation of PA to yield diacylglycerol, responsible for de novo lipid synthesis and formation of lipid droplets; homologous to mammalian lipin 1	YMR165C
YMR113W	Dihydrofolate synthetase, involved in folic acid biosynthesis; catalyzes the conversion of dihydropteroate to dihydrofolate in folate coenzyme biosynthesis	YMR113W
YMR060C	Component of the Sorting and Assembly Machinery (SAM or TOB complex) of the mitochondrial outer membrane, which binds precursors of beta-barrel proteins and facilitates their outer membrane insertion; contributes to SAM complex stability	YMR060C

YML004C	Monomeric glyoxalase I, catalyzes the detoxification of methylglyoxal (a by-product of glycolysis) via condensation with glutathione to produce S-D-lactoylglutathione; expression regulated by methylglyoxal levels and osmotic stress	YML004C
YLR373C	Glycosylated integral membrane protein localized to the plasma membrane; plays a role in fructose-1,6-bisphosphatase (FBPase) degradation; involved in FBPase transport from the cytosol to Vid (vacuole import and degradation) vesicles	YLR373C
YLR369W	Mitochondrial hsp70-type molecular chaperone, required for assembly of iron/sulfur clusters into proteins at a step after cluster synthesis, and for maturation of Yfh1p, which is a homolog of human frataxin implicated in Friedreich's ataxia	YLR369W
YLR362W	Signal transducing MEK kinase involved in pheromone response and pseudohyphal/invasive growth pathways where it phosphorylates Ste7p, and the high osmolarity response pathway, via phosphorylation of Pbs2p; regulated by Ste20p and Ste50p	YLR362W
YLR358C	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified ORF RSC2/YLR357W	YLR358C
YLR313C	Protein involved in shmoo formation and bipolar bud site selection; homologous to Spa2p, localizes to sites of polarized growth in a cell cycle dependent- and Spa2p-dependent manner, interacts with MAPKs Mkk1p, Mkk2p, and Ste7p	YLR313C
YKR098C	Ubiquitin-specific protease that cleaves ubiquitin from ubiquitinated proteins	YKR098C
YKL214C	Member of the REF (RNA and export factor binding proteins) family; when overexpressed, can substitute for the function of Yra1p in export of poly(A)+ mRNA from the nucleus	YKL214C
YKL188C	Subunit of a heterodimeric peroxisomal ATP-binding cassette transporter complex (Pxa1p-Pxa2p), required for import of long-chain fatty acids into peroxisomes; similarity to human adrenoleukodystrophy transporter and ALD-related proteins	YKL188C
YKL112W	DNA binding protein with possible chromatin-reorganizing activity involved in transcriptional activation, gene silencing, and DNA replication and repair	YKL112W
YKL111C	Dubious open reading frame, unlikely to encode a protein; not conserved in closely related <i>Saccharomyces</i> species; partially overlaps the verified essential gene ABF1	YKL111C
YKL108W	Single-stranded DNA origin-binding and annealing protein; required for the initiation of DNA replication; phosphorylated in S phase by cyclin-dependent kinases (Cdks), promoting origin binding, DNA replication and Dpb11p complex formation; component of the preloading complex; unphosphorylated or CDK-phosphorylated Sld2p binds to the MCM2-7 complex; required for the S phase checkpoint	YKL108W
YJR089W	Subunit of chromosomal passenger complex (CPC; Ipl1p-Sli15p-Bir1p-Nbl1p), which regulates chromosome segregation; required for chromosome bi-orientation and for spindle assembly checkpoint activation upon reduced sister kinetochore tension	YJR089W

YJR087W	Dubious open reading frame, unlikely to encode a protein; not conserved in closely related <i>Saccharomyces</i> species; partially overlaps the verified genes STE18 and ECM2	YJR087W
YJR079W	Putative protein of unknown function; mutation results in impaired mitochondrial respiration	YJR079W
YJL159W	O-mannosylated heat shock protein that is secreted and covalently attached to the cell wall via beta-1,3-glucan and disulfide bridges; required for cell wall stability; induced by heat shock, oxidative stress, and nitrogen limitation	YJL159W
YJL156C	Serine protease of SPS plasma membrane amino acid sensor system (Ssy1p-Ptr3p-Ssy5p); contains an inhibitory domain that dissociates in response to extracellular amino acids, freeing a catalytic domain to activate transcription factor Stp1p	YJL156C
YJL046W	Putative lipoate-protein ligase, required along with Lip2 and Lip5 for lipoylation of Lat1p and Kgd2p; similar to <i>E. coli</i> LplA; null mutant displays reduced frequency of mitochondrial genome loss	YJL046W
YIR027C	Allantoinase, converts allantoin to allantoinate in the first step of allantoin degradation; expression sensitive to nitrogen catabolite repression	YIR027C
YHR168W	Putative GTPase, member of the Obg family; peripheral protein of the mitochondrial inner membrane that associates with the large ribosomal subunit; required for mitochondrial translation, possibly via a role in ribosome assembly	YHR168W
YHR167W	Subunit of the THO complex, which connects transcription elongation and mitotic recombination, and of the TREX complex, which is recruited to activated genes and couples transcription to mRNA export; involved in telomere maintenance	YHR167W
YHR162W	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the mitochondrion	YHR162W
YHR120W	DNA-binding protein of the mitochondria involved in repair of mitochondrial DNA, has ATPase activity and binds to DNA mismatches; has homology to <i>E. coli</i> Muts; transcription is induced during meiosis	YHR120W
YFR047C	Quinolinate phosphoribosyl transferase, required for the de novo biosynthesis of NAD from tryptophan via kynurenine; expression regulated by Hst1p	YFR047C
YER157W	Essential component of the conserved oligomeric Golgi complex (Cog1p through Cog8p), a cytosolic tethering complex that functions in protein trafficking to mediate fusion of transport vesicles to Golgi compartments	YER157W
YEL065W	Ferrioxamine B transporter, member of the ARN family of transporters that specifically recognize siderophore-iron chelates; transcription is induced during iron deprivation and diauxic shift; potentially phosphorylated by Cdc28p	YEL065W
YEL055C	DNA Polymerase phi; has sequence similarity to the human MybBP1A and weak sequence similarity to B-type DNA polymerases, not required for chromosomal DNA replication; required for the synthesis of rRNA	YEL055C
YDR475C	Protein of unknown function; previously annotated as two separate ORFs, YDR474C and YDR475C, which were merged as a result of corrections to the systematic reference sequence	YDR475C

YDR145W	Subunit (61/68 kDa) of TFIID and SAGA complexes, involved in RNA polymerase II transcription initiation and in chromatin modification, similar to histone H2A	YDR145W
YDR144C	GPI-anchored aspartyl protease, member of the yapsin family of proteases involved in cell wall growth and maintenance; shares functions with Yap3p and Kex2p	YDR144C
YDR143C	Ubiquitin-protein ligase; involved in the proteasome-dependent degradation of aberrant nuclear proteins; targets substrates with regions of exposed hydrophobicity containing 5 or more contiguous hydrophobic residues; contains intrinsically disordered regions that contribute to substrate recognition	YDR143C
YJR091C	Member of the Puf family of RNA-binding proteins, interacts with mRNAs encoding membrane-associated proteins; involved in localizing the Arp2/3 complex to mitochondria; overexpression causes increased sensitivity to benomyl	YJR091C
YOL082W	Receptor protein specific for the cytoplasm-to-vacuole targeting (Cvt) pathway; delivers cargo proteins aminopeptidase I (Lap4p) and alpha-mannosidase (Ams1p) to the phagophore assembly site for packaging into Cvt vesicles	YOL082W
YOL083C-A	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; identified by expression profiling and mass spectrometry	YOL083C-A
YJR094W-A	Protein component of the large (60S) ribosomal subunit, identical to Rpl43Ap and has similarity to rat L37a ribosomal protein	YJR094W-A
YOL085C	Dubious open reading frame unlikely to encode a protein, based on experimental and comparative sequence data; partially overlaps the dubious gene YOL085W-A	YOL085C
YJR083C	Protein of unknown function, computational analysis of large-scale protein-protein interaction data suggests a possible role in actin cytoskeleton organization; potential Cdc28p substrate	YJR083C
YAR002C-A	Protein that forms a heterotrimeric complex with Erp2p, Emp24p, and Erv25p; member, along with Emp24p and Erv25p, of the p24 family involved in ER to Golgi transport and localized to COPII-coated vesicles	YAR002C-A
YKL216W	Dihydroorotate dehydrogenase, catalyzes the fourth enzymatic step in the de novo biosynthesis of pyrimidines, converting dihydroorotic acid into orotic acid	YKL216W
YJR085C	Putative protein of unknown function; GFP-fusion protein is induced in response to the DNA-damaging agent MMS; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies	YJR085C
YJR077C	Mitochondrial phosphate carrier, imports inorganic phosphate into mitochondria; functionally redundant with Pic2p but more abundant than Pic2p under normal conditions; phosphorylated	YJR077C
YMR063W	Protein of unknown function, involved in the proteolytic activation of Rim101p in response to alkaline pH; has similarity to A. nidulans Pali; putative membrane protein	YMR063W

YJR096W	Putative xylose and arabinose reductase; member of the aldo-keto reductase (AKR) family; GFP-fusion protein is induced in response to the DNA-damaging agent MMS	YJR096W
YMR065W	Protein required for nuclear membrane fusion during karyogamy, localizes to the membrane with a soluble portion in the endoplasmic reticulum lumen, may form a complex with Jem1p and Kar2p; expression of the gene is regulated by pheromone	YMR065W
YMR064W	Protein required for expression of the mitochondrial OLI1 gene encoding subunit 9 of F1-F0 ATP synthase	YMR064W
YDR147W	Ethanolamine kinase, primarily responsible for phosphatidylethanolamine synthesis via the CDP-ethanolamine pathway; exhibits some choline kinase activity, thus contributing to phosphatidylcholine synthesis via the CDP-choline pathway	YDR147W
YOR280C	Putative serine hydrolase; likely target of Cyc8p-Tup1p-Rfx1p transcriptional regulation; sequence is similar to <i>S. cerevisiae</i> Fsh1p and Fsh2p and the human candidate tumor suppressor OVCA2	YOR280C
YLR360W	Part of a Vps34p phosphatidylinositol 3-kinase complex that functions in carboxypeptidase Y (CPY) sorting; binds Vps30p and Vps34p to promote production of phosphatidylinositol 3-phosphate (PtdIns3P) which stimulates kinase activity	YLR360W
YLR359W	Adenylosuccinate lyase, catalyzes two steps in the 'de novo' purine nucleotide biosynthetic pathway; expression is repressed by adenine and activated by Bas1p and Pho2p; mutations in human ortholog ADSL cause adenylosuccinase deficiency	YLR359W
YLR355C	Bifunctional acetoxyhydroxyacid reductoisomerase and mtDNA binding protein; involved in branched-chain amino acid biosynthesis and maintenance of wild-type mitochondrial DNA; found in mitochondrial nucleoids	YLR355C
YKL106W	Mitochondrial aspartate aminotransferase, catalyzes the conversion of oxaloacetate to aspartate and asparagine biosynthesis	YKL106W
YJR076C	Component of the septin ring that is required for cytokinesis; septins are GTP-binding proteins that assemble into rod-like hetero-oligomers that can associate with other rods to form filaments; septin rings at the mother-bud neck act as scaffolds for recruiting cell division factors and as barriers to prevent diffusion of specific proteins between mother and daughter cells	YJR076C
YOR286W	Protein with rhodanese activity; contains a rhodanese-like domain similar to Rdl1p, Uba4p, Tum1p, and Ych1p; overexpression causes a cell cycle delay; null mutant displays elevated frequency of mitochondrial genome loss	YOR286W
YOR285W	Protein of unknown function containing a rhodanese-like domain; localized to the mitochondrial outer membrane	YOR285W
YOR284W	Cytoplasmic protein of unknown function; computational analysis of large-scale protein-protein interaction data suggests a possible role in actin patch assembly	YOR284W
YOR279C	DNA-binding protein required for vegetative repression of middle sporulation genes; specificity factor that directs the Hst1p histone deacetylase to some of the promoters regulated by Sum1p; involved in telomere maintenance	YOR279C

YJR038C	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YJR038C
YHR170W	Protein involved in nuclear export of the large ribosomal subunit; acts as a Crm1p-dependent adapter protein for export of nascent ribosomal subunits through the nuclear pore complex	YHR170W
YHR169W	ATPase, putative RNA helicase of the DEAD-box family; component of 90S preribosome complex involved in production of 18S rRNA and assembly of 40S small ribosomal subunit; ATPase activity stimulated by association with Esf2p	YHR169W
YHR166C	Subunit of the Anaphase-Promoting Complex/Cyclosome (APC/C), which is a ubiquitin-protein ligase required for degradation of anaphase inhibitors, including mitotic cyclins, during the metaphase/anaphase transition	YHR166C
YHR165C	Component of the U4/U6-U5 snRNP complex, involved in the second catalytic step of splicing; mutations of human Prp8 cause retinitis pigmentosa	YHR165C
YDR473C	Splicing factor, component of the U4/U6-U5 snRNP complex	YDR473C
YDR142C	Peroxisomal signal receptor for the N-terminal nonapeptide signal (PTS2) of peroxisomal matrix proteins; WD repeat protein; defects in human homolog cause lethal rhizomelic chondrodysplasia punctata (RCDP)	YDR142C
YOR283W	Phosphatase with a broad substrate specificity and some similarity to GPM1/YKL152C, a phosphoglycerate mutase; YOR283W is not an essential gene	YOR283W
YOR282W	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps essential, verified gene PLP2/YOR281C	YOR282W
YOR277C	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; almost completely overlaps the verified gene CAF20	YOR277C
YLR366W	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the dubious ORF YLR364C-A	YLR366W
YLR365W	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps dubious gene YLR364C-A; YLR365W is not an essential gene	YLR365W
YKL114C	Major apurinic/apyrimidinic endonuclease, 3'-repair diesterase involved in repair of DNA damage by oxidation and alkylating agents; also functions as a 3'-5' exonuclease to repair 7,8-dihydro-8-oxodeoxyguanosine	YKL114C
YKL113C	5' to 3' exonuclease, 5' flap endonuclease, required for Okazaki fragment processing and maturation as well as for long-patch base-excision repair; member of the S. pombe RAD2/FEN1 family	YKL113C
YIL143C	Component of RNA polymerase transcription factor TFIIF holoenzyme; has DNA-dependent ATPase/helicase activity and is required, with Rad3p, for unwinding promoter DNA; interacts functionally with TFIIB and has roles in transcription start site selection and in gene looping to juxtapose initiation and termination regions; involved in DNA repair; homolog of human ERCC3	YIL143C

YIL142W	Subunit beta of the cytosolic chaperonin Cct ring complex, related to Tcp1p, required for the assembly of actin and tubulins in vivo	YIL142W
YIL141W	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YIL141W
YIL140W	Integral plasma membrane protein required for axial budding in haploid cells, localizes to the incipient bud site and bud neck; glycosylated by Pmt4p; potential Cdc28p substrate	YIL140W
YDR477W	AMP-activated serine/threonine protein kinase found in a complex containing Snf4p and members of the Sip1p/Sip2p/Gal83p family; required for transcription of glucose-repressed genes, thermotolerance, sporulation, and peroxisome biogenesis	YDR477W
YBR076C-A	Dubious open reading frame unlikely to encode a protein; partially overlaps verified gene ECM8; identified by fungal homology and RT-PCR	YBR076C-A
snR31	Proline tRNA (tRNA-Pro), predicted by tRNAscan-SE analysis; target of K. lactis zymocin; can mutate to suppress +1 frameshift mutations in proline codons	snR31
YOR281C	Essential protein that interacts with the CCT (chaperonin containing TCP-1) complex to stimulate actin folding; has similarity to phosducins; null mutant lethality is complemented by mouse phosducin-like protein MgcPhLP	YOR281C
YOL086C	Alcohol dehydrogenase, fermentative isozyme active as homo- or heterotetramers; required for the reduction of acetaldehyde to ethanol, the last step in the glycolytic pathway	YOL086C
YMR188C	Mitochondrial ribosomal protein of the small subunit	YMR188C
YMR061W	Component of the cleavage and polyadenylation factor I (CF I); CF 1, composed of the CF 1A complex (Rna14p, Rna15p, Clp1p, Pcf11p) and Hrp1, is involved in cleavage and polyadenylation of mRNA 3' ends; bridges interaction between Rna15p and Hrp1p in the CF I complex; mutant displays reduced transcription elongation in the G-less-based run-on (GLRO) assay; required for gene looping	YMR061W
YMR057C	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps verified ORF AAC1	YMR057C
YMR056C	Mitochondrial inner membrane ADP/ATP translocator, exchanges cytosolic ADP for mitochondrially synthesized ATP; phosphorylated; Aac1p is a minor isoform while Pet9p is the major ADP/ATP translocator	YMR056C
YLR367W	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps22Ap and has similarity to E. coli S8 and rat S15a ribosomal proteins	YLR367W
YLR363W-A	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the nucleus	YLR363W-A
YJR036C	Protein with similarity to hect domain E3 ubiquitin-protein ligases, not essential for viability	YJR036C
YJL043W	Putative protein of unknown function; YJL043W is a non-essential gene	YJL043W
YJL042W	Microtubule-associated protein involved in assembly and stabilization of microtubules; overproduction results in cell cycle arrest at G2 phase; similar to Drosophila protein MAP and to mammalian MAP4 proteins	YJL042W

YEL053W-A	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified gene YEL054C	YEL053W-A
YDR141C	Golgi-localized, leucine-zipper domain containing protein; involved in endosome to Golgi transport, organization of the ER, establishing cell polarity, and morphogenesis; detected in highly purified mitochondria in high-throughput studies	YDR141C
YCR075W-A	Putative protein of unknown function; identified by homology to <i>Ashbya gossypii</i>	YCR075W-A
YCR073W-A	Protein with a possible role in tRNA export; shows similarity to 6-phosphogluconolactonase non-catalytic domains but does not exhibit this enzymatic activity; homologous to Sol1p, Sol3p, and Sol4p	YCR073W-A
YCR072C	WD-repeat protein involved in ribosome biogenesis; may interact with ribosomes; required for maturation and efficient intra-nuclear transport or pre-60S ribosomal subunits, localizes to the nucleolus	YCR072C
YAR003W	Subunit of the COMPASS (Set1C) complex, which methylates histone H3 on lysine 4 and is required in transcriptional silencing near telomeres; WD40 beta propeller superfamily member with similarity to mammalian Rbbp7	YAR003W
snR44	Serine tRNA (tRNA-Ser), predicted by tRNAscan-SE analysis	snR44
YPL053C	Probable mannosylphosphate transferase involved in the synthesis of core oligosaccharides in protein glycosylation pathway; member of the KRE2/MNT1 mannosyltransferase family	YPL053C
YPL050C	Subunit of Golgi mannosyltransferase complex also containing Anp1p, Mnn10p, Mnn11p, and Hoc1p that mediates elongation of the polysaccharide mannan backbone; forms a separate complex with Van1p that is also involved in backbone elongation	YPL050C
YPL048W	Nuclear protein required for transcription of MXR1; binds the MXR1 promoter in the presence of other nuclear factors; binds calcium and phospholipids; has similarity to translational cofactor EF-1 gamma	YPL048W
YPL047W	Integral subunit of SAGA histone acetyltransferase complex, regulates transcription of a subset of SAGA-regulated genes, required for the Ubp8p association with SAGA and for H2B deubiquitylation	YPL047W
YNL124W	RNA-binding protein required for the assembly of box H/ACA snoRNPs and thus for pre-rRNA processing, forms a complex with Shq1p and interacts with H/ACA snoRNP components Nhp2p and Cbf5p; similar to Gar1p	YNL124W
YNL123W	Serine protease and general molecular chaperone; involved in response to heat stress and promotion of apoptosis; may contribute to lipid homeostasis; sequence similarity to the mammalian Omi/HtrA2 family of serine proteases	YNL123W
YMR189W	P subunit of the mitochondrial glycine decarboxylase complex, required for the catabolism of glycine to 5,10-methylene-THF; expression is regulated by levels of 5,10-methylene-THF in the cytoplasm	YMR189W
YML009C	Mitochondrial ribosomal protein of the large subunit	YML009C
YML006C	CAAX box containing protein of unknown function, proposed to be involved in the RAS/cAMP signaling pathway	YML006C
YML003W	Putative protein of unknown function	YML003W

YML002W	Putative protein of unknown function; expression induced by heat and by calcium shortage	YML002W
YLR451W	Zinc-knuckle transcription factor, repressor and activator; regulates genes involved in branched chain amino acid biosynthesis and ammonia assimilation; acts as a repressor in leucine-replete conditions and as an activator in the presence of alpha-isopropylmalate, an intermediate in leucine biosynthesis that accumulates during leucine starvation	YLR451W
YLR443W	Non-essential putative integral membrane protein with a role in calcium uptake; mutant has cell wall defects and Ca ⁺ uptake deficiencies; transcription is induced under conditions of zinc deficiency	YLR443W
YLR441C	Ribosomal protein 10 (rp10) of the small (40S) subunit; nearly identical to Rps1Bp and has similarity to rat S3a ribosomal protein	YLR441C
YLR440C	Component of the Dsl1p tethering complex that interacts with ER SNAREs Sec20p and Use1p; proposed to be involved in protein secretion; localizes to the ER and nuclear envelope	YLR440C
YLR375W	Zinc-finger protein of unknown function, possibly involved in pre-tRNA splicing and in uptake of branched-chain amino acids	YLR375W
YLR368W	F-box component of an SCF ubiquitin protein ligase complex; associates with and is required for Fzo1p ubiquitination and for mitochondria fusion; stimulates nuclear export of specific mRNAs; promotes ubiquitin-mediated degradation of Gal4p in some strains	YLR368W
YLR363C	Protein interacting with Nam7p, may be involved in the nonsense-mediated mRNA decay pathway	YLR363C
YLR361C-A	Putative protein of unknown function	YLR361C-A
YLR354C	Transaldolase, enzyme in the non-oxidative pentose phosphate pathway; converts sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate to erythrose 4-phosphate and fructose 6-phosphate	YLR354C
YLR326W	Putative protein of unknown function, predicted to be palmitoylated	YLR326W
YLR323C	Essential protein, component of a complex containing Cef1p; has similarity to <i>S. pombe</i> Cwf24p	YLR323C
YLR321C	Component of the RSC chromatin remodeling complex; essential gene required for cell cycle progression and maintenance of proper ploidy; phosphorylated in the G1 phase of the cell cycle; Snf5p paralog	YLR321C
YLR315W	Non-essential kinetochore protein, subunit of the Ctf19p-Mcm21p-Okp1p-Mcm22p-Mcm16p-	YLR315W
YLR312C	Ctf3p-Chl4p-Mcm19p-Nkp1p-Nkp2p-Ame1p-Mtw1p) Putative protein of unknown function	YLR312C
YLR311C	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data	YLR311C
YLR310C	Membrane bound guanine nucleotide exchange factor (GEF or GDP-release factor); indirectly regulates adenylate cyclase through activation of Ras1p and Ras2p by stimulating the exchange of GDP for GTP; required for progression through G1	YLR310C

YKL116C	Serine/threonine protein kinase that inhibits pheromone induced signalling downstream of MAPK, possibly at the level of the Ste12p transcription factor	YKL116C
YKL115C	Dubious open reading frame, unlikely to encode a protein; partially overlaps the verified gene PRR1	YKL115C
YJR110W	Phosphatidylinositol 3-phosphate (PI3P) phosphatase; involved in various protein sorting pathways, including CVT targeting and endosome to vacuole transport; has similarity to the conserved myotubularin dual specificity phosphatase family	YJR110W
YJL160C	Putative protein of unknown function; member of the PIR (proteins with internal repeats) family of cell wall proteins; non-essential gene that is required for sporulation; mRNA is weakly cell cycle regulated, peaking in mitosis	YJL160C
YJL158C	Mannose-containing glycoprotein constituent of the cell wall; member of the PIR (proteins with internal repeats) family	YJL158C
YIR029W	Allantoicase, converts allantoate to urea and ureidoglycolate in the second step of allantoin degradation; expression sensitive to nitrogen catabolite repression and induced by allophanate, an intermediate in allantoin degradation	YIR029W
YIR028W	Allantoin permease; expression sensitive to nitrogen catabolite repression and induced by allophanate, an intermediate in allantoin degradation	YIR028W
YIR026C	Protein phosphatase involved in vegetative growth at low temperatures, sporulation, and glycogen accumulation; mutants are defective in 60S ribosome assembly; member of the dual-specificity family of protein phosphatases	YIR026C
YIR024C	Protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies; interacts with Arh1p, a mitochondrial oxidoreductase; deletion mutant has a respiratory growth defect	YIR024C
YIR023C-A	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YIR023C-A
YHR171W	Autophagy-related protein and dual specificity member of the E1 family of ubiquitin-activating enzymes; mediates the conjugation of Atg12p with Atg5p and Atg8p with phosphatidylethanolamine, required steps in autophagosome formation	YHR171W
YFR048W	Cytosolic protein required for sporulation	YFR048W
YFR046C	Kinetochore protein of unknown function; associated with the essential kinetochore proteins Nnf1p and Spc24p; phosphorylated by both Clb5-Cdk1 and, to a lesser extent, Clb2-Cdk1.	YFR046C
YFR044C	Cys-Gly metallo-di-peptidase; forms a complex with Dug2p and Dug3p to degrade glutathione (GSH) and other peptides containing a gamma-glu-X bond in an alternative pathway to GSH degradation by gamma-glutamyl transpeptidase (Ecm38p)	YFR044C
YFR043C	Putative protein of unknown function; null mutant displays increased levels of spontaneous Rad52p foci	YFR043C

YCR071C	Mitochondrial ribosomal protein of the large subunit	YCR071C
YBR078W	GPI-anchored protein of unknown function, has a possible role in apical bud growth; GPI-anchoring on the plasma membrane crucial to function; phosphorylated in mitochondria; similar to Sps2p and Pst1p	YBR078W
YBR077C	Component of the EGO complex, which is involved in the regulation of microautophagy, and of the GSE complex, which is required for proper sorting of amino acid permease Gap1p; gene exhibits synthetic genetic interaction with MSS4	YBR077C
YAR002W	Subunit of the nuclear pore complex (NPC), functions to anchor Nup2p to the NPC in a process controlled by the nucleoplasmic concentration of Gsp1p-GTP; involved in nuclear export and cytoplasmic localization of specific mRNAs such as ASH1	YAR002W
YAL037C-B	Dubious open reading frame unlikely to encode a protein; identified by gene-trapping, microarray-based expression analysis, and genome-wide homology searching	YAL037C-B
YAL037C-A	Putative protein of unknown function	YAL037C-A
YAL036C	Member of the DRG family of GTP-binding proteins; associates with translating ribosomes; interacts with Tma46p, Ygr250cp, Gir2p and Yap1p via two-hybrid	YAL036C
YAL034W-A	Essential component of the MIND kinetochore complex (Mtw1p Including Nnf1p-Nsl1p-Dsn1p) which joins kinetochore subunits contacting DNA to those contacting microtubules; critical to kinetochore assembly	YAL034W-A
tS(GCU)L	Tyrosine tRNA (tRNA-Tyr), predicted by tRNAscan-SE analysis; can mutate to suppress ochre nonsense mutations	tS(GCU)L
tS(AGA)M	tRNA of undetermined specificity, predicted by tRNAscan-SE analysis; very similar to serine tRNAs	tS(AGA)M
YPL099C	Protein of unknown function; the authentic, non-tagged protein is detected in purified mitochondria in high-throughput studies; null mutant displays elevated frequency of mitochondrial genome loss	YPL099C
YPL098C	Protein required for growth of cells lacking the mitochondrial genome	YPL098C
YPL096C-A	Endoplasmic reticulum membrane protein that binds to and inhibits GTP-bound Ras2p at the ER; component of the GPI-GnT complex which catalyzes the first step in GPI-anchor biosynthesis; probable homolog of mammalian PIG-Y protein	YPL096C-A
YPL093W	Putative GTPase that associates with free 60S ribosomal subunits in the nucleolus and is required for 60S ribosomal subunit biogenesis; constituent of 66S pre-ribosomal particles; member of the ODN family of nucleolar G-proteins	YPL093W
YPL045W	Subunit of the vacuole fusion and protein sorting HOPS complex and the CORVET tethering complex; part of the Class C Vps complex essential for membrane docking and fusion at Golgi-to-endosome and endosome-to-vacuole protein transport stages	YPL045W
YOR291W	Vacuolar protein with a possible role in sequestering heavy metals; has similarity to the type V P-type ATPase Spf1p; homolog of human ATP13A2 (PARK9), mutations in which are associated with Parkinson disease and Kufor-Rakeb syndrome	YOR291W

YOR188W	Protein involved in positive regulation of both 1,3-beta-glucan synthesis and the Pkc1p-MAPK pathway, potential Cdc28p substrate; multicopy suppressor of temperature-sensitive mutations in CDC24 and CDC42, and of mutations in BEM4	YOR188W
YOR186C-A	Identified by gene-trapping, microarray-based expression analysis, and genome-wide homology searching	YOR186C-A
YOR185C	GTP binding protein (mammalian Ranp homolog) involved in the maintenance of nuclear organization, RNA processing and transport; interacts with Kap121p, Kap123p and Pdr6p (karyophilin betas); Gsp1p homolog that is not required for viability	YOR185C
YNL221C	Subunit of both RNase MRP and nuclear RNase P; RNase MRP cleaves pre-rRNA, while nuclear RNase P cleaves tRNA precursors to generate mature 5' ends and facilitates turnover of nuclear RNAs; binds to the RPR1 RNA subunit in RNase P	YNL221C
YNL218W	Protein with DNA-dependent ATPase and ssDNA annealing activities involved in maintenance of genome; interacts functionally with DNA polymerase delta; homolog of human Werner helicase interacting protein (WHIP)	YNL218W
YNL122C	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to mitochondria; YNL122C is not an essential gene	YNL122C
YNL119W	Protein required for thiolation of the uridine at the wobble position of Lys(UUU) and Glu(UUC) tRNAs; has a role in urmylation and in invasive and pseudohyphal growth; inhibits replication of Brome mosaic virus in <i>S. cerevisiae</i>	YNL119W
YNL099C	Putative protein tyrosine phosphatase, required for cell cycle arrest in response to oxidative damage of DNA	YNL099C
YNL097C-B	Putative protein of unknown function	YNL097C-B
YNL097C	Probable component of the Rpd3 histone deacetylase complex, involved in transcriptional regulation of PHO5; affects termination of snoRNAs and cryptic unstable transcripts (CUTs); C-terminus has similarity to human candidate tumor suppressor p33(ING1) and its isoform ING3	YNL097C
YNL094W	Protein of unknown function, interacts with Rvs161p and Rvs167p; computational analysis of protein-protein interactions in large-scale studies suggests a possible role in actin filament organization	YNL094W
YNL049C	Component of the Sec23p-Sfb2p heterodimer of the COPII vesicle coat, required for cargo selection during vesicle formation in ER to Golgi transport; homologous to Sec24p and Sfb3p	YNL049C
YNL045W	Leucyl aminopeptidase yscIV (leukotriene A4 hydrolase) with epoxide hydrolase activity, metalloenzyme containing one zinc atom; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm and nucleus	YNL045W
YMR173W	DNA damage-responsive protein, expression is increased in response to heat-shock stress or treatments that produce DNA lesions; contains multiple repeats of the amino acid sequence NNNDSYGS	YMR173W
YMR172W	Transcription factor required for the transient induction of glycerol biosynthetic genes GPD1 and GPP2 in response to high osmolarity; targets Hog1p to osmoresponsive promoters; has similarity to Msn1p and Gcr1p	YMR172W
YMR171C	Specificity factor required for Rsp5p-dependent ubiquitination and sorting of specific cargo proteins at the multivesicular body; mRNA is targeted to the bud via the mRNA transport system involving She2p	YMR171C

YMR170C	Cytoplasmic aldehyde dehydrogenase, involved in ethanol oxidation and beta-alanine biosynthesis; uses NAD ⁺ as the preferred coenzyme; expression is stress induced and glucose repressed; very similar to Ald3p	YMR170C
YMR167W	Protein required for mismatch repair in mitosis and meiosis as well as crossing over during meiosis; forms a complex with Pms1p and Msh2p-Msh3p during mismatch repair; human homolog is associated with hereditary non-polyposis colon cancer	YMR167W
YMR164C	Transcription factor involved in regulation of invasive growth and starch degradation; controls the activation of MUC1 and STA2 in response to nutritional signals	YMR164C
YMR115W	Subunit of the mitochondrial (mt) i-AAA protease supercomplex, which degrades misfolded mitochondrial proteins; forms a subcomplex with Mgr1p that binds to substrates to facilitate proteolysis; required for growth of cells lacking mtDNA	YMR115W
YMR111C	Protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the nucleus; YMR111C is not an essential gene	YMR111C
YMR110C	Putative fatty aldehyde dehydrogenase, located in the mitochondrial outer membrane and also in lipid particles; has similarity to human fatty aldehyde dehydrogenase (FALDH) which is implicated in Sjogren-Larsson syndrome	YMR110C
YMR066W	Mitochondrial protein of unknown function	YMR066W
YMR059W	Subunit of the tRNA splicing endonuclease, which is composed of Sen2p, Sen15p, Sen34p, and Sen54p	YMR059W
YMR055C	Mitotic exit network regulator, forms GTPase-activating Bfa1p-Bub2p complex that binds Tem1p and spindle pole bodies, blocks cell cycle progression before anaphase in response to spindle and kinetochore damage	YMR055C
YMR003W	Protein of unknown function; GFP-fusion protein localizes to the mitochondria; null mutant is viable and displays reduced frequency of mitochondrial genome loss	YMR003W
YML007W	Basic leucine zipper (bZIP) transcription factor required for oxidative stress tolerance; activated by H ₂ O ₂ through the multistep formation of disulfide bonds and transit from the cytoplasm to the nucleus; mediates resistance to cadmium	YML007W
YML007C-A	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to mitochondria	YML007C-A
YLR457C	Spindle pole body (SPB) component, required for the insertion of the duplication plaque into the nuclear membrane during SPB duplication; essential for bipolar spindle formation; component of the Mps2p-Bbp1p complex	YLR457C
YLR445W	Putative protein of unknown function; transcription is regulated by Ume6p and induced in response to alpha factor	YLR445W

YLR438C-A	Lsm (Like Sm) protein; part of heteroheptameric complexes (Lsm2p-7p and either Lsm1p or 8p): cytoplasmic Lsm1p complex involved in mRNA decay; nuclear Lsm8p complex part of U6 snRNP and possibly involved in processing tRNA, snoRNA, and rRNA	YLR438C-A
YLR371W	GDP/GTP exchange factor (GEF) for Rho1p and Rho2p; mutations are synthetically lethal with mutations in rom1, which also encodes a GEF	YLR371W
YLR328W	Nicotinic acid mononucleotide adenyllyltransferase, involved in pathways of NAD biosynthesis, including the de novo, NAD(+) salvage, and nicotinamide riboside salvage pathways	YLR328W
YLR324W	Peroxisomal integral membrane protein, involved in negative regulation of peroxisome number; partially functionally redundant with Pex31p; genetic interactions suggest action at a step downstream of steps mediated by Pex28p and Pex29p	YLR324W
YLR322W	Dubious open reading frame, unlikely to encode a protein; not conserved in closely related <i>Saccharomyces</i> species; 75% of ORF overlaps the verified gene SFH1; deletion causes a vacuolar protein sorting defect and blocks anaerobic growth	YLR322W
YLR319C	Actin- and formin-interacting protein; stimulates actin cable nucleation by recruiting actin monomers to Bni1p; involved in polarized cell growth; isolated as bipolar budding mutant; potential Cdc28p substrate	YLR319C
YLR317W	Dubious open reading frame; may be part of a bicistronic transcript with NKP2/YLR315W; overlaps the verified ORF TAD3/YLR316C	YLR317W
YLR307W	Chitin deacetylase, together with Cda2p involved in the biosynthesis ascospore wall component, chitosan; required for proper rigidity of the ascospore wall	YLR307W
YLR306W	Enzyme that mediates the conjugation of Rub1p, a ubiquitin-like protein, to other proteins; related to E2 ubiquitin-conjugating enzymes	YLR306W
YLR304C	Aconitase, required for the tricarboxylic acid (TCA) cycle and also independently required for mitochondrial genome maintenance; phosphorylated; component of the mitochondrial nucleoid; mutation leads to glutamate auxotrophy	YLR304C
YLR086W	Subunit of the condensin complex; reorganizes chromosomes during cell division; forms a complex with Smc2p that has ATP-hydrolyzing and DNA-binding activity; required for tRNA gene clustering at the nucleolus; potential Cdc28p substrate	YLR086W
YLR085C	Actin-related protein that binds nucleosomes; a component of the SWR1 complex, which exchanges histone variant H2AZ (Htz1p) for chromatin-bound histone H2A	YLR085C
YLR084C	N-glycosylated protein involved in the maintenance of bud site selection during bipolar budding; localization requires Rax1p; RAX2 mRNA stability is regulated by Mpt5p	YLR084C
YLR083C	Protein with a role in cellular adhesion, filamentous growth, and endosome-to-vacuole sorting; similar to Tmn2p and Tmn3p; member of Transmembrane Nine family of proteins with 9 transmembrane segments	YLR083C

YLL040C	Protein of unknown function; heterooligomeric or homooligomeric complex; peripherally associated with membranes; homologous to human COH1; involved in sporulation, vacuolar protein sorting and protein-Golgi retention	YLL040C
YLL039C	Ubiquitin, becomes conjugated to proteins, marking them for selective degradation via the ubiquitin-26S proteasome system; essential for the cellular stress response; encoded as a polyubiquitin precursor comprised of 5 head-to-tail repeats	YLL039C
YLL038C	Protein of unknown function, contains an N-terminal epsin-like domain; proposed to be involved in the trafficking of Arn1p in the absence of ferrichrome	YLL038C
YLL036C	Splicing factor associated with the spliceosome; contains a U-box, a motif found in a class of ubiquitin ligases, and a WD40 domain	YLL036C
YLL035W	Polynucleotide kinase present on rDNA that is required for efficient transcription termination by RNA polymerase I; required for cell growth; mRNA is cell-cycle regulated	YLL035W
YKR099W	Myb-related transcription factor involved in regulating basal and induced expression of genes of the purine and histidine biosynthesis pathways; also involved in regulation of meiotic recombination at specific genes	YKR099W
YKR056W	tRNA methyltransferase, 5-methylates the uridine residue at position 54 of tRNAs and may also have a role in tRNA stabilization or maturation; endo-exonuclease with a role in DNA repair	YKR056W
YKR055W	Non-essential small GTPase of the Rho/Rac subfamily of Ras-like proteins, likely to be involved in the establishment of cell polarity	YKR055W
YKR047W	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the verified gene NAP1	YKR047W
YKR044W	Protein of unknown function that interacts with Ulp1p, a Ubl (ubiquitin-like protein)-specific protease for Smt3p protein conjugates	YKR044W
YKR040C	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data; partially overlaps the uncharacterized ORF YKR041W	YKR040C
YKL215C	5-oxoprolinase; enzyme is ATP-dependent and functions as a dimer; similar to mouse Oplah gene; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm	YKL215C
YKL212W	Phosphatidylinositol phosphate (PtdInsP) phosphatase involved in hydrolysis of PtdIns[4]P; transmembrane protein localizes to ER and Golgi; involved in protein trafficking and processing, secretion, and cell wall maintenance	YKL212W
YJR090C	F-box protein component of the SCF ubiquitin-ligase complex; involved in carbon catabolite repression, glucose-dependent divalent cation transport, high-affinity glucose transport, morphogenesis, and sulfite detoxification	YJR090C
YJL156W-A	Dubious open reading frame unlikely to encode a protein, based on available experimental and comparative sequence data	YJL156W-A

YJL155C	Fructose-2,6-bisphosphatase, required for glucose metabolism	YJL155C
YJL152W	Dubious ORF unlikely to encode a functional protein, based on available experimental and comparative sequence data	YJL152W
YJL048C	UBX (ubiquitin regulatory X) domain-containing protein that interacts with Cdc48p, transcription is repressed when cells are grown in media containing inositol and choline	YJL048C
YJL047C-A	Putative protein of unknown function	YJL047C-A
YJL047C	Cullin subunit of a Roc1p-dependent E3 ubiquitin ligase complex with a role in anaphase progression; implicated in Mms22-dependent DNA repair; involved with Mms1p in nonfunctional rRNA decay; modified by the ubiquitin-like protein, Rub1p	YJL047C
YJL045W	Minor succinate dehydrogenase isozyme; homologous to Sdh1p, the major isozyme responsible for the oxidation of succinate and transfer of electrons to ubiquinone; induced during the diauxic shift in a Cat8p-dependent manner	YJL045W
YIL118W	Non-essential small GTPase of the Rho/Rac subfamily of Ras-like proteins involved in the establishment of cell polarity; GTPase activity positively regulated by the GTPase activating protein (GAP) Rgd1p	YIL118W
YIL102C-A	Putative protein of unknown function, identified based on comparisons of the genome sequences of six <i>Saccharomyces</i> species	YIL102C-A
YIL102C	Putative protein of unknown function	YIL102C
YIL101C	Transcriptional repressor that binds to promoter sequences of the cyclin genes, CYS3, and SMF2; expression is induced by stress or starvation during mitosis, and late in meiosis; member of the Swi4p/Mbp1p family; potential Cdc28p substrate	YIL101C
YIL100C-A	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YIL100C-A
YIL099W	Intracellular sporulation-specific glucoamylase involved in glycogen degradation; induced during starvation of a/a diploids late in sporulation, but dispensable for sporulation	YIL099W
YHR176W	Flavin-containing monooxygenase, localized to the cytoplasmic face of the ER membrane; catalyzes oxidation of biological thiols to maintain the ER redox buffer ratio for correct folding of disulfide-bonded proteins	YHR176W
YHR175W-A	Putative protein of unknown function; identified by fungal homology and RT-PCR	YHR175W-A
YHR175W	Putative low-affinity copper transporter of the vacuolar membrane; mutation confers resistance to toxic copper concentrations, while overexpression confers resistance to copper starvation	YHR175W
YHR174W	Enolase II, a phosphopyruvate hydratase that catalyzes the conversion of 2-phosphoglycerate to phosphoenolpyruvate during glycolysis and the reverse reaction during gluconeogenesis; expression is induced in response to glucose	YHR174W
YHR173C	Dubious ORF unlikely to encode a functional protein, based on available experimental and comparative sequence data	YHR173C

YHR163W	6-phosphogluconolactonase, catalyzes the second step of the pentose phosphate pathway; weak multicopy suppressor of <i>Ios1-1</i> mutation; homologous to <i>Sol2p</i> and <i>Sol1p</i>	YHR163W
YHR161C	Protein involved in clathrin cage assembly; binds <i>Pan1p</i> and <i>clathrin</i> ; homologous to <i>Yap1802p</i> , member of the <i>AP180</i> protein family	YHR161C
YHR160C	Peroxin required for targeting of peroxisomal matrix proteins containing <i>PTS2</i> ; interacts with <i>Pex7p</i> ; partially redundant with <i>Pex21p</i>	YHR160C
YHR124W	Meiosis-specific transcription factor required for exit from pachytene and for full meiotic recombination; activates middle sporulation genes; competes with <i>Sum1p</i> for binding to promoters containing middle sporulation elements (<i>MSE</i>)	YHR124W
YHR123W	<i>sn-1,2-diacylglycerol ethanolamine-</i> and <i>cholinephosphotransferase</i> ; not essential for viability	YHR123W
YHR122W	Protein of unknown function required for establishment of sister chromatid cohesion; synthetically lethal with <i>RFC5</i> , an <i>RF-C</i> subunit that links replication to cohesion establishment; <i>YHR122W</i> is an essential gene	YHR122W
YHR121W	Protein of unknown function that may function in RNA processing; interacts with <i>Pbp1p</i> and <i>Pbp4p</i> and associates with ribosomes; contains an RNA-binding <i>LSM</i> domain and an <i>AD</i> domain; <i>GFP-fusion</i> protein is induced by the DNA-damaging agent <i>MMS</i>	YHR121W
YGL258W	Protein of unknown function; highly induced in zinc-depleted conditions and has increased expression in <i>NAP1</i> deletion mutants	YGL258W
YGL256W	Alcohol dehydrogenase isoenzyme type IV, dimeric enzyme demonstrated to be zinc-dependent despite sequence similarity to iron-activated alcohol dehydrogenases; transcription is induced in response to zinc deficiency	YGL256W
YGL118C	Dubious open reading frame unlikely to encode a functional protein, based on available experimental and comparative sequence data	YGL118C
YGL115W	Activating gamma subunit of the AMP-activated <i>Snf1p</i> kinase complex (contains <i>Snf1p</i> and a <i>Sip1p/Sip2p/Gal83p</i> family member); activates glucose-repressed genes, represses glucose-induced genes; role in sporulation, and peroxisome biogenesis	YGL115W
YGL114W	Putative protein of unknown function; predicted member of the oligopeptide transporter (<i>OPT</i>) family of membrane transporters	YGL114W
YGL113W	Protein involved in the initiation of DNA replication, required for proper assembly of replication proteins at the origins of replication; interacts with <i>Cdc45p</i>	YGL113W
YFR040W	Protein that forms a complex with the <i>Sit4p</i> protein phosphatase and is required for its function; member of a family of similar proteins including <i>Sap4p</i> , <i>Sap185p</i> , and <i>Sap190p</i>	YFR040W
YFR038W	Putative ATPase containing the <i>DEAD/H</i> helicase-related sequence motif; null mutant displays increased levels of spontaneous <i>Rad52p</i> foci	YFR038W

YFR037C	Component of the RSC chromatin remodeling complex; essential for viability and mitotic growth; homolog of SWI/SNF subunit Swi3p, but unlike Swi3p, does not activate transcription of reporters	YFR037C
YFR035C		YFR035C
YFR034C	Putative protein of unknown function, deletion mutant exhibits synthetic phenotype with alpha-synuclein Basic helix-loop-helix (bHLH) transcription factor of the myc-family; activates transcription cooperatively with Pho2p in response to phosphate limitation; binding to 'CACGTG' motif is regulated by chromatin restriction, competitive binding of Cbf1p to the same DNA binding motif and cooperation with Pho2p,; function is regulated by phosphorylation at multiple sites and by phosphate availability	YFR034C
YER156C	Putative protein of unknown function; interacts with Hsp82p and copurifies with Ipl1p; expression is copper responsive and downregulated in strains deleted for MAC1, a copper-responsive transcription factor; similarity to mammalian MYG1	YER156C
YER155C	Rho GTPase activating protein (RhoGAP) involved in the control of cytoskeleton organization and cellular morphogenesis; required for bud emergence	YER155C
YER117W	Protein component of the large (60S) ribosomal subunit, identical to Rpl23Ap and has similarity to E. coli L14 and rat L23 ribosomal proteins	YER117W
YER114C	Protein implicated in polar growth, functionally redundant with Boi1p; interacts with bud-emergence protein Bem1p; contains an SH3 (src homology 3) domain and a PH (pleckstrin homology) domain	YER114C
YER113C	Protein with a role in cellular adhesion and filamentous growth; similar to Emp70p and Tmn2p; member of Transmembrane Nine family with 9 transmembrane segments; localizes to Golgi; induced by 8-methoxypsoralen plus UVA irradiation	YER113C
YEL067C	Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies	YEL067C
YEL063C	Plasma membrane arginine permease, requires phosphatidyl ethanolamine (PE) for localization, exclusively associated with lipid rafts; mutation confers canavanine resistance	YEL063C
YEL059W	Dubious open reading frame unlikely to encode a functional protein; mutant is hypersensitive to hygromycin B indicative of defects in vacuolar trafficking	YEL059W
YEL058W	Essential N-acetylglucosamine-phosphate mutase; converts GlcNAc-6-P to GlcNAc-1-P, which is a precursor for the biosynthesis of chitin and for the formation of N-glycosylated mannoproteins and glycosylphosphatidylinositol anchors	YEL058W
YEL057C	Protein of unknown function involved in telomere maintenance; target of UME6 regulation	YEL057C
YDR478W	Subunit of RNase MRP, which cleaves pre-rRNA and has a role in cell cycle-regulated degradation of daughter cell-specific mRNAs; binds to the NME1 RNA subunit of RNase MRP	YDR478W

YDL171C	NAD(+)-dependent glutamate synthase (GOGAT), synthesizes glutamate from glutamine and alpha-ketoglutarate; with Gln1p, forms the secondary pathway for glutamate biosynthesis from ammonia; expression regulated by nitrogen source	YDL171C
YDL169C	Protein of unknown function, transcript accumulates in response to any combination of stress conditions	YDL169C
YDL165W	Component of the CCR4-NOT complex, which has multiple roles in regulating mRNA levels including regulation of transcription and destabilizing mRNAs by deadenylation; basal transcription factor	YDL165W
YDL115C	RNA polymerase II transport factor, conserved from yeast to humans; involved in both basal and regulated transcription from RNA polymerase II (RNAP II) promoters, but not itself a transcription factor; interacts with most of the RNAP II subunits; nucleocytoplasmic shuttling protein; deletion causes hypersensitivity to K1 killer toxin	YDL115C
YDL113C	Sorting nexin family member required for the cytoplasm-to-vacuole targeting (Cvt) pathway and for endosomal sorting; has a Phox homology domain that binds phosphatidylinositol-3-phosphate; interacts with Snx4p; potential Cdc28p substrate	YDL113C
YDL005C	Subunit of the RNA polymerase II mediator complex; associates with core polymerase subunits to form the RNA polymerase II holoenzyme; essential for transcriptional regulation	YDL005C
YDL001W	Cytoplasmic protein required for sporulation	YDL001W
YBR215W	Subunit of the HIR complex, a nucleosome assembly complex involved in regulation of histone gene transcription; mutants display synthetic defects with subunits of FACT, a complex that allows passage of RNA Pol II through nucleosomes	YBR215W
YBR214W	One of two <i>S. cerevisiae</i> homologs (Sds23p and Sds24p) of the <i>S. pombe</i> Sds23 protein, which is implicated in APC/cyclosome regulation; involved in cell separation during budding; may play an indirect role in fluid-phase endocytosis	YBR214W
YBR211C	Essential kinetochore protein associated with microtubules and spindle pole bodies; component of the kinetochore sub-complex COMA (Ctf19p, Okp1p, Mcm21p, Ame1p); involved in spindle checkpoint maintenance	YBR211C
YBR189W	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps9Ap and has similarity to <i>E. coli</i> S4 and rat S9 ribosomal proteins	YBR189W
YBR187W	Putative protein of unknown function; expression is reduced in a <i>gcr1</i> null mutant; GFP-fusion protein localizes to the vacuole; expression pattern and physical interactions suggest a possible role in ribosome biogenesis	YBR187W
YAL035W	GTPase, required for general translation initiation by promoting Met-tRNAiMet binding to ribosomes and ribosomal subunit joining; homolog of bacterial IF2	YAL035W
snR60	Tryptophan tRNA (tRNA-Trp), predicted by tRNAscan-SE analysis	snR60

snR36	Arginine tRNA (tRNA-Arg), predicted by tRNAscan-SE analysis; one of 11 nuclear tRNA genes containing the tDNA-anticodon UCU (converted to mcm5-UCU in the mature tRNA), decodes AGA codons into arginine, one of 19 nuclear tRNAs for arginine	snR36
snR19	Leucine tRNA (tRNA-Leu), predicted by tRNAscan-SE analysis	snR19
YPL095C	Acyl-coenzymeA:ethanol O-acyltransferase responsible for the major part of medium-chain fatty acid ethyl ester biosynthesis during fermentation; possesses short-chain esterase activity; may be involved in lipid metabolism and detoxification	YPL095C
YPL049C	MAP kinase-responsive inhibitor of the Ste12p transcription factor, involved in the regulation of mating-specific genes and the invasive growth pathway; related regulators Dig1p and Dig2p bind to Ste12p	YPL049C
YOR290C	Catalytic subunit of the SWI/SNF chromatin remodeling complex involved in transcriptional regulation; contains DNA-stimulated ATPase activity; functions interdependently in transcriptional activation with Snf5p and Snf6p	YOR290C
YNL125C	Protein with similarity to monocarboxylate permeases, appears not to be involved in transport of monocarboxylates such as lactate, pyruvate or acetate across the plasma membrane	YNL125C
YNL121C	Component of the TOM (translocase of outer membrane) complex responsible for recognition and initial import steps for all mitochondrially directed proteins; acts as a receptor for incoming precursor proteins	YNL121C
YNL047C	Phosphoinositide PI4,5P(2) binding protein, forms a complex with Slm1p; acts downstream of Mss4p in a pathway regulating actin cytoskeleton organization in response to stress; phosphorylated by the TORC2 complex	YNL047C
YLR459W	GPI transamidase subunit, involved in attachment of glycosylphosphatidylinositol (GPI) anchors to proteins; may have a role in recognition of the attachment signal or of the lipid portion of GPI	YLR459W
YLR450W	One of two isozymes of HMG-CoA reductase that convert HMG-CoA to mevalonate, a rate-limiting step in sterol biosynthesis; overproduction induces assembly of peripheral ER membrane arrays and short nuclear-associated membrane stacks	YLR450W
YLR442C	Silencing protein that interacts with Sir2p and Sir4p, and histone H3 and H4 tails, to establish a transcriptionally silent chromatin state; required for spreading of silenced chromatin; recruited to chromatin through interaction with Rap1p	YLR442C
YLR320W	Protein that acts with Mms1p in a repair pathway that may be involved in resolving replication intermediates or preventing the damage caused by blocked replication forks; required for accurate meiotic chromosome segregation	YLR320W
YLR312W-A	Mitochondrial ribosomal protein of the large subunit	YLR312W-A
YLR305C	Phosphatidylinositol-4-kinase that functions in the Pkc1p protein kinase pathway; required for normal vacuole morphology, cell wall integrity, and actin cytoskeleton organization	YLR305C

YKR043C	Sedoheptulose biphosphatase involved in riboneogenesis; dephosphorylates sedoheptulose 1,7-bisphosphate, which is converted via the nonoxidative pentose phosphate pathway to ribose-5-phosphate; facilitates the conversion of glycolytic intermediates to pentose phosphate units; also has fructose 1,6-bisphosphatase activity but this is probably not biologically relevant, since deletion does not affect FBP levels; GFP-fusion protein localizes to the cytoplasm and nucleus	YKR043C
YJL154C	Endosomal subunit of membrane-associated retromer complex required for retrograde transport; receptor that recognizes retrieval signals on cargo proteins, forms subcomplex with Vps26p and Vps29p that selects cargo proteins for retrieval	YJL154C
YJL039C	Essential structural subunit of the nuclear pore complex (NPC), localizes to the nuclear periphery of nuclear pores, homologous to human p205	YJL039C
YHR146W	Protein that binds to cruciform DNA structures	YHR146W
YEL061C	Kinesin motor protein involved in mitotic spindle assembly and chromosome segregation	YEL061C
YDL112W	2'-O-ribose methyltransferase, catalyzes the ribose methylation of the guanosine nucleotide at position 18 of tRNAs	YDL112W
YBR213W	Bifunctional dehydrogenase and ferrochelatase, involved in the biosynthesis of siroheme, a prosthetic group used by sulfite reductase; required for sulfate assimilation and methionine biosynthesis	YBR213W
YBR185C	Membrane-associated mitochondrial ribosome receptor; forms a complex with Mdm38p that may facilitate recruitment of mRNA-specific translational activators to ribosomes; possible role in protein export from the matrix to inner membrane	YBR185C
YAR007C	Subunit of heterotrimeric Replication Protein A (RPA), which is a highly conserved single-stranded DNA binding protein involved in DNA replication, repair, and recombination	YAR007C
tD(GUC)J2		tD(GUC)J2
tK(CUU)F		tK(CUU)F
tK(UUU)L		tK(UUU)L
tL(CAA)N		tL(CAA)N
tP(UGG)H		tP(UGG)H
tR(UCU)J1		tR(UCU)J1
tW(CCA)G1		tW(CCA)G1
tY(GUA)J1		tY(GUA)J1

Exp 6g/l

Exp 45g/l

up_6g/l
up_6g/l
up_6g/l

down_6g/l_except_strain3
down_6g/l_except_strain3
down_6g/l_except_strain3

down_6g/l_except_strain3

down_6g/l
down_6g/l
down_6g/l
down_6g/l
down_6g/l
down_6g/l

best_up_6g/l

best_up_45g/l

best_up_6g/l

best_up_45g/l

best_up_6g/l

best_up_45g/l

best_up_6g/l

best_up_6g/l

best_up_6g/l

best_up_6g/l_except_strain4

best_up_6g/l_except_strain3

best_up_45g/l_except_strain3

best_down_6g/l_except_strain3

best_down_6g/l_except_strain3

best_down_6g/l

best_down_6g/l

best_down_45g/l

best_down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l

down_45g/l_except_strain4

down_45g/l_except_strain2

up_45g/l

up_45g/l

up_45g/l

poco espresso (mean 5 reads)

non signif. (mean 72 reads)

significativo ma diff exp sotto a 1

non signif. (mean 479 reads)

significativo ma diff exp sotto a 1

significativo e diff circa 1 a 6g/l tranne ceppo3, non signif a 45 g/l

significativo ma diff exp sotto a 1

non signif. (mean 460 reads)

poco diff exp a 6g/l

significativo e diff circa 1 a 6g/l tranne ceppo1e3, non signif a 45 g/l

tendenzialmente sovraespresso ma non sempre significativo

significativo e diff circa 1 a 6g/l tranne ceppo3, non signif a 45 g/l

significativo e diff oltre 1 a 6g/l tranne ceppo3, non signif a 45 g/l

significativo e diff oltre 1 a 45g/l, non signif a 6 g/l

significativo e diff oltre 1 a 45g/l, inferiore a 1 a 6 g/l

significativo e diff circa 1 a 6g/l tranne ceppo2, non signif a 45 g/l

significativo e diff circa 1 a 6g/l, non signif a 45 g/l

poco espresso (mean 1 read)

significativo e diff circa 1 a 45g/l tranne ceppo1, poco signif a 6 g/l

significativo e diff circa 1 a 45g/l, poco signif a 6 g/l

n°	freq	predicted ORFs	EC1118	P283	R008	R103	P301	AWRI796	AWRI1631	RM11	QA23	VL3	VIN13	FostersO
3	13	P283_G2_2316	<u>EC1118_1G1_6284g</u>	P283_G2_2316	R008_G2_2346	R103_G8_0531	P301_G2_0006	x		x	x	x	x	
2	12	P283_G2_2311	<u>EC1118_1G1_6283g</u>	P283_G2_2311	R008_G2_2341		P301_G2_0001	x		x	x		x	
4	9	P283_I1_0711	EC1118_1I12_1684g	P283_I1_0711		R103_M2_2196	P301_I1_0967		x		x	x	x	
1	8	P283_J1_0001	EC1118_1O30_0012g	P283_J1_0001	R008_J1_0001			x	x	x	x			
12	8	R008_G2_2336			R008_G2_2336		P301_G1_3696		x	x	x	x	x	
22	8	R103_X2_0001	<u>EC1118_1M36_0046g</u>			R103_X2_0001	P301_P1_0001		x		x		x	
20	7	R103_P2_0001	<u>EC1118_1P2_0178g</u>			R103_P2_0001					x		x	
23	7	R103_X2_0006	EC1118_1M36_0045g			R103_X2_0006	P301_P1_0006	x	x			x	x	
24	7	R103_X2_0011	EC1118_1M36_0045g			R103_X2_0011	P301_P1_0006	x	x		x	x	x	
25	7	P301_P1_0011	EC1118_1M36_0034g			R103_X2_0011	P301_P1_0011	x	x			x	x	
10	6	R008_O1_4151	EC1118_1O4_6568g		R008_O1_4151		P301_O3_0061	xx		x				
19	6	R103_J1_0001				R103_J1_0001		x		x		x		
15	5	R103_I1_0011	EC1118_1O4_6480g			R103_I1_0011	P301_O3_0091		x					
21	5	R103_I1_0001	EC1118_1O4_6502g, EC1118_1O4_6513g			R103_I1_0001		x				xx		
27	5	P301_O3_0001	EC1118_1O4_6667g				P301_O3_0001		x		x			
32	5	P301_O3_0056	<u>EC1118_1O4_6569g</u>				P301_O3_0056		x		x			
5	4	R008_O1_4111			R008_O1_4111			x						
6	4	R008_O1_4116			R008_O1_4116				x			x		
11	4	R008_O1_4156			R008_O1_4156			x		x				
17	4	R103_I1_0006	EC1118_1O4_6491g			R103_I1_0006	P301_O3_0086							
28	4	P301_O3_0006	EC1118_1O4_6656g				P301_O3_0006				x			
31	4	P301_O3_0031	EC1118_1O4_6612g				P301_O3_0031				x			
13	3	R008_I1_0016			R008_I1_0016									
14	3	R008_O1_4106			R008_O1_4106			x						
29	3	P301_O3_0021	EC1118_1O4_6634g				P301_O3_0021							
30	3	P301_O3_0026	EC1118_1O4_6623g				P301_O3_0026							
7	2	R008_O1_4121			R008_O1_4121									
8	2	R008_O1_4131			R008_O1_4131									
9	2	R008_O1_4136			R008_O1_4136									
16	2	R103_I1_0021				R103_I1_0021								

n°	freq	predicted ORFs	EC1118	P283	R008	R103	P301	AWRI796	AWRI1631	RM11	QA23	VL3	VIN13	FostersO
18	2	R103_I1_0016				R103_I1_0016								
26	2	P301_O3_0036					P301_O3_0036							
33	2	P301_J1_0002					P301_J1_0002							

n°	FostersB	JAY291	S278b	YJM789	Length (aa)	Species	Potential Function
3	x	x			105	<i>Saccharomyces cerevisiae</i>	MAL-activator protein (low similarity with <i>S. cerevisiae</i> YJM789)
2	x	x		xx	114	-	Hypothetical protein
4	x			x	297	<i>Saccharomyces cerevisiae</i>	Killer toxin (low similarity with <i>S. cerevisiae</i> YJM789)
1		x	xx		158	<i>Saccharomyces paradoxus</i>	Azetidine-2-carboxylic acid acetyltransferase (low similarity with <i>S. cerevisiae</i> RM111)
12	x				111	-	Hypothetical protein
22	x				222	-	Hypothetical protein
20	x	x		x	247	-	Hypothetical protein
23					104	-	Hypothetical protein
24					106	-	Hypothetical protein
25					106	-	Hypothetical protein
10					276	<i>Saccharomyces cerevisiae</i>	GPR1/FUN34/yaah family. Putative acetate transporter (low similarity with <i>S. cerevisiae</i> AWRI796)
19		x	x		155	-	Hypothetical protein
15					583	<i>Penicillium marneffeii</i>	Fungal specific transcription factor domain, c6 zinc finger domain (low similarity with <i>S. cerevisiae</i> Lalin QA23)
21					331	<i>Pichia angusta</i>	Putative amino acid transporter (low similarity with <i>S. cerevisiae</i> VL3)
27					209	-	Hypothetical protein
32					107	-	Hypothetical protein
5		x			267	<i>Saccharomyces cerevisiae</i>	Alpha-galactosidase, melibiase (low similarity with <i>S. cerevisiae</i> AWRI796)
6					360	<i>Saccharomyces cerevisiae</i>	Medium chain alcohol dehydrogenase (low similarity with <i>S. cerevisiae</i> RM111-1a and AWRI1631)
11					584	<i>Schizosaccharomyces pombe</i>	Putative aspartyl/glutamyl-tRNA amidotransferase subunit A (low similarity with <i>S. cerevisiae</i> AWRI796)
17					593	<i>Saccharomyces cerevisiae</i>	Putative permease, member of the allantate transporter subfamily (low similarity with <i>S. cerevisiae</i> S288c)
28					1394	<i>Saccharomyces cerevisiae</i>	Helicase encoded by the Y' element of subtelomeric regions (low similarity with <i>S. cerevisiae</i> AWRI1631)
31					261	<i>Spathaspora passalidarum</i>	NADP-mannitol dehydrogenase domain, carbonyl reductase (low similarity with <i>L. thermotolerans</i>)
13	x				835	-	Hypothetical protein
14					608	-	Hypothetical protein
29					567	<i>Zygosaccharomyces rouxii</i>	Putative fructose symporter (low similarity with <i>L. thermotolerans</i>)
30					359	<i>Saccharomyces cerevisiae</i>	Sorbitol dehydrogenase (low similarity with <i>S. cerevisiae</i> S288c)
7					828	<i>Saccharomyces cerevisiae</i>	Transcriptional regulator (low similarity with <i>S. cerevisiae</i> YJM789)
8					506	<i>Penicillium marneffeii</i>	Putative allantate permease (low similarity with <i>S. cerevisiae</i> AWRI796)
9					305	<i>Penicillium marneffeii</i>	Putative branched-chain amino acid aminotransferase (low similarity with <i>S. cerevisiae</i> AWRI796)
16					556	<i>Zygosaccharomyces rouxii</i>	Plasma membrane multidrug transporter of the major facilitator superfamily (low similarity with <i>S. cerevisiae</i> S288c)

n°	FostersB	JAY291	S278b	YJM789	Length (aa)	Species	Potential Function
18					527	<i>Candida dubliniensis</i>	Putative glucose transporter of the major facilitator superfamily (low similarity with <i>L. thermotolerans</i>)
26					343	<i>Saccharomyces cerevisiae</i>	Similar to full-length MRP-type transporter 1 (low similarity with <i>S. cerevisiae</i> Lalvin QA23)
33					268	<i>Saccharomyces cerevisiae</i>	Haze Protective Factor (low similarity with <i>S. cerevisiae</i> Vin13)

PCR1_chr3_A_F	cccactatcgacctttcttat
PCR1_chr11_B_F	ccaaacgtatcaaacttcagca
PCR1_chr11_C_R	tagcgtcctggctccactaa
PCR2_chr7_A_F	gcttggcgaatctctgaatc
PCR2_chr7_B_R	cgtttggttagacgcctgtt
PCR2_chr7_C_R	acaccacttgcgatcaaca
PCR2_chr7_D_F	ggaaacactcgcttttggt
PCR3_chr16_A_F	agaaccgtgctgctcgttaag
PCR3_chr16_B_R	gcaagcgatagcaaacaatga
PCR3_chr8_C_R	catggcagctagaacctca
PCR4_chr15_A_F	gccgtataccgttgctcatt
PCR4_chr15_B_R	caaggtttaccctgcgctaa
PCR4_chr16_C_R	accagcggaatgatatccag