Comparative genomic analysis of three *Saccharomyces cerevisiae* commercial wine strains

Rui Pedro Conceição de Carvalho

Under supervision of Nuno Gonçalo Pereira Mira and Alexandra Mendes Ferreira
IST, Lisbon, Portugal
December 2017

Abstract

The success of wine fermentation is highly dependent on the ability of the involved strains to cope with the multiple stresses that occur along the alcoholic fermentation of the must which impact not only the yield of the process as well as the sensorial properties of the product. Studies have been performed addressing the adaptive responses evolved by *S. cerevisiae* wine strains to the challenging environment of wine fermentation, however, this has been mostly done at a transcriptomic level and little is known on genomic elements driving those adaptive responses. To obtain insights into this the genome of the commercial wine *S. cerevisiae* strains QA23, VL1 and CEG was sequenced and mapped against the genome of the well annotated strains S288c (a lab strain) and EC1118 (a commercial wine yeast strain). Besides focusing the alterations occurring in the coding sequences, focus was also put on the identification occurring in non-coding regions, in particular, in gene promoter regions. The comparison with the laboratory strain focused on coding sequences elucidated on general adaptive responses evolved by wine yeasts in general, being possible to identify a number of genes diverging significantly in the three wine yeasts when compared with their S288c counter-partners, suggesting that these are genes under selective pressure in the context of wine fermentation. Consistently, the expression of some of these genes was found to influence yeast fermentation in simulated grape juice medium. The comparison with the genome of the EC1118 wine yeast revealed a closer similarity between this strain and QA23, consistent with previous studies. Evidences were also obtained suggesting that three genomic regions present in the genome of EC1118 strain encompassing several "foreign genes" acquired through horizontal gene transfer, are also present in the genome of the QA23, CEG and VL1 strains. The analysis of the non-coding genomic sequence (corresponding to all genomic regions where genes are not annotated) of the three wine strains suggest the occurrence of significant alterations in the networks that may govern genomic expression as a result of a significant number of gains/losses of DNA binding sites in gene promoters.

**Key words:** comparative genomics, *S. cerevisiae* wine strains, adaptation to wine fermentative environment, stress tolerance
Introduction

During wine fermentation, yeast cells are challenged with a wide diversity of stresses, that can come from the conditions of the fermentative process itself as well from the own metabolism of the yeast during this process. Temperature variations, ethanol inhibitory concentrations, osmotic pressure and nutrient starvation are the main stresses occurring during wine fermentation, which affect the physiological behaviour and fermentative capacity of the wine yeasts (Bauer & Pretorius, 2000).

Consequently, the success of wine fermentation is largely dependent on the capacity of the yeast strains to maintain their physiological parameters within appropriate limits to face such conditions. In this context, it’s clear that wine yeast strains have developed mechanisms not only to sense, but also to respond and adapt to the environment, a few of these being described in this monograph. Otherwise, yeasts with low capacity to adapt to this challenging environment, are more set to die and affect the whole fermentative process. Previous studies had shown that there is an inverse correlation between the inherent stress resistance of specific strains and the frequency of occurrence with “stuck” or sluggish fermentations (Attfield (1997); Henschke, (1998); Bisson, (1999)). These undesirable situations, are a major cause of significant losses in the process rentability. Logically, all that concerns the resistance of the yeasts to face these stresses is of great interest to the winemaking industry since a small improvement in the response to a specific stress, could result in significant profits to the winemaking industry.

The challenging environment of grape must during wine fermentation has led to a selective pressure of wine strains. This pressure acts not only but also at the genome level, and in order that wine yeast strains could adapt to this stressful condition, they underwent mutations that led to modifications in their genome, allowing over time these to present a better fitness regarding the wine fermentative process. In order to adapt to the challenging and stressful environment of grape must, wine yeasts have several response mechanisms at the genome level. This adaptation can occur from a small-scale nucleotide changes such as single nucleotide polymorphisms (SNPs), insertions, deletions or substitution of bases, which can change the final protein structure, interaction protein-protein or the expression of a given gene since these changes can also affect transcription factors. Gene duplications are also very described in the literature. It’s also possible to occur large-scale genome rearrangements events such as chromosome duplication, aneuploidy, translocations, copy number variation or interspecific hybridization which will have a significant impact in gene expression. The previously mentioned molecular mechanisms at the genome level have already been described as being advantageous since they contribute to the adaptation of wine strains (Pretorius.2000).

According to the results obtained by Novo et al (2009) in the complete genome sequencing of EC1118 wine strain and simultaneous comparison against the laboratory strain S288C, it was possible to identify a group of 111 genes present in S288C strain, but the same genes were not found in EC1118 strain. Simultaneously EC1118 strain was found to possess 34 unique genes that are not present in S288C strain. Those 34 genes are essential to the wine fermentation process and are encompassed in 3 large regions (A, B and C), two of which are subtelomeric (chromosome VI and XV) and the remaining region is inserted in the chromosome XIV. It was found that one of those regions is originated from very close species to Saccharomyces genus and the other 2 regions were originated from non-Saccharomyces genus/species. Zygosaccharomyces bailii was identified as being a donor for 1 of these 2 regions, namely region C, providing evidences that horizontal gene transfer has occurred between Z. bailii and S.cerevisiae. Horizontal gene transfer had never been reported in eukaryotes before. Most of the modifications mentioned in the studies focused in this introduction up to this point are related mainly with modifications at the coding genome. Analysis of changes in non-coding regions is important since it may allow for a deeper understanding of how regulatory transcription network is established in wine strains compared to other yeast strains. Gene expression occurs essentially due to the presence of flanking sequences of a given gene that are non-coding sequences that contain
binding sites that are recognized by sequence-specific nucleotide-binding proteins, transcription factors that modulate the transcription. New possible adaptation mechanisms can arise from variations at the genetic level that contribute for different expression, which can confer the ability to survive in challenging environments. It is then expected that the study of the non-coding regions, lead to a better understanding and identification of resistance determinants with respect to grape must.

**Materials and Methods**

The reads obtained after genome sequencing by SOLiD were kindly provided by Prof Alexandra Mendes-Ferreira (Universidade de Trás-os-Montes e Alto Douro). The read files (fastaq format) obtained for the three strains were imported into the software CLC Genomics Workbench. The reads were trimmed based on quality, as detailed in Table 1. The resulting trimmed sequences were then assembled using “de novo assembly” and mapped against the genome of the reference strain S288C (version R64-2-1) which was obtained from Saccharomyces Genome Database (SGD) and previously imported to the software with the respective annotation. Additionally, variance detection was performed from the mapped reads using both probabilistic and quality-based variant detection in order to obtain more accurate results. The parameters established in the different steps when using the software CLC Genomics Workbench are summarized in Table 1.

**Results**

The number of SNPs obtained in the SNP calling concerning the EC1118 strains was significantly lower in the 3 strains compared to the SNP calling for the S288C strain as highlighted in Table 2. A total of 20357, 22594 and 23011 were identified for QA23, VL1 and CEG respectively in the mapping against S288C whereas a total of 2681, 8513 and 8141 were identified in the same strain order for mapping against EC1118.

---

### Table 1 – Summary of the parameters used in the CLC Genomic workbench operations

<table>
<thead>
<tr>
<th>Parameters</th>
<th>CLC Genomics Workbench operation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Import and sequencing</strong></td>
<td>Import high-throughput sequencing data; SOLiD from life technologies; Paired reads: Yes; Discard read names: Yes; Discard quality scores: No</td>
</tr>
<tr>
<td><strong>Quality trimming</strong></td>
<td>Ambiguous trim: yes; ambiguous limit: 2; quality trim: 0.05; Remove 5’/3’ terminal nucleotides: no; discard short/long reads: no; mismatch cost: 2; insertion cost: 3; deletion cost: 3; length fraction: 0.5; similarity fraction: 0.8</td>
</tr>
<tr>
<td><strong>De novo assembly</strong></td>
<td>Word size: 20; bubble size: 50; minimum contig length: 200; perform scaffolding: yes</td>
</tr>
<tr>
<td><strong>Mapping reads to reference</strong></td>
<td>Reference: Saccharomyces cerevisiae S288C; mismatch cost: 2; insertion cost: 3; deletion cost: 3; length fraction: 0.5; similarity fraction: 0.8; non-specific match handling: map randomly</td>
</tr>
<tr>
<td><strong>Probability variant detection</strong></td>
<td>Ignore non-specific matches: yes; ignore broken pairs: yes; Minimum coverage: 10; Variant probability: 50; Required variant count: 2; Required presence in forward and reverse reads: yes; Ignore variants in non-specific matches: yes; Filter 454/Ion Homopolymer indels: No; Maximum expected variations: 2; Genetic code: 1-Standard Filters applied: Frequency: ≥50; Average quality ≥20</td>
</tr>
<tr>
<td><strong>Quality-based variant detection</strong></td>
<td>Neighbourhood radius: 5; Maximum gap and mismatch count: 2; Minimum neighbourhood quality: 15; Minimum central quality: 20; Ignore non-specific matches: yes; Ignore broken pairs: yes; Minimum coverage: 10; Variant probability: 50; Maximum expected variations: 2; Required presence in forward and reverse reads: Yes; Ignore variants in non-specific matches: yes; Filter 454/Ion homopolymer indels: No; Genetic code = 1-Standard Filters applied: Frequency ≥50; Average quality ≥20</td>
</tr>
</tbody>
</table>
One of the clear observations that comes out from these results is the significantly lower number of SNPs in all cases compared with the number of SNPs observed in the case of S288C. This could be justified by the fact that the QA23, VL1 and CEG strains are wine strains and therefore, when mapping against a strain that is also a wine strains, less differences/SNPs are verified. When a comparison was made among the three wine strains considering the genes affected by non-synonymous SNPs resulting from the mapping against the S288C genome, about 2347 genes were found to harbour non-synonymous SNPs commonly in the three wine strains and in the mapping against EC1118 strain 728 were found to harbour non-synonymous SNPs.

In a second step, the genes harbouring SNPs in the three strains were investigated in further detail. In specific, it was analysed the distribution of SNPs throughout the predicted ORFeome of the three wine strains. Therefore, the goal of this analysis undertaken for the three strains with respect to the mapping against S288C and EC1118, is to understand which genes are under a higher selective pressure. In the case of the mapping against the S288C strain, it was considered as relevant the cases in which a given gene has a total of 15 or more SNPs.

A closer inspection of the results obtained reveals that about 10 genes harbour a similar number of SNPs (≥15) in the three wine strains with respect to the mapping against S288C strain: ZRT1, ALR2, EG12, GCN1, PEP1, SLC1, IRA2, YM317W, R1F1 and YHL008C.

In the case of the mapping against EC1118, about 5 genes harbour a similar number of SNPs (≥5) in the three wine strains: ZRT1, TRE2, IRA1, YMR317W and REV1.

A closer analysis was performed on SNPs that lead to the appearance of premature STOP codons in the three wine strains since this could result in protein truncation. The occurrence of truncated proteins could result from SNPs that change the reading frame (fs) or SNPs that result in the appearance of a STOP codon. In that sense, the genes that were found to harbour a premature STOP codon in the three strains and not in the lab strain were identified and compared. For this analysis, it was only considered those proteins whose level of truncation was 50% or more. An important observation that comes out of this analysis is the fact that 17 genes were commonly associated with SNPs leading to truncated proteins when the three wine strains are compared with S288C genome (Figure 1).
number obtained when the laboratory strain S288c was used (2.57). Altogether these results suggest a closer proximity of QA23, VL1 and CEG with the EC1118 strain than with the S288c strain. It was of notice the fact that the QA23 strain showed a higher similarity to EC1118 than VL1 or CEG, as suggested by the lower number of variations observed. These results are consistent with previous phenotypic and genetic screenings that indeed confirm a closer proximity between QA23 and EC1118 (Barbosa et al. 2014, Barbosa et al. 2015b).

When this comparison is made taking the EC1118 genome as reference, only 2 genes associated with the appearance of premature STOP codon are shared by QA23, VL1 and CEG.

Figure 2 - Venn diagram relative to genes affected by SNPs non-synonymous leading to truncated proteins in EC1118 mapping (truncation ≥ 50%). The truncation percentage is indicated for each gene.

Surprisingly, none of the genes identified in this report associated with these kinds of SNPs were present /identified in the Fermentome (Walker et al. 2014). It is in fact, quite interesting to note that there are genes associated with non-synonymous SNPs that lead to the appearance of premature STOP codon and whose truncation of the respective protein is greater than 90%. By that observation, it can be inferred that the translation of these genes results in proteins that will be inactive or incomplete due to such degree of truncation. The truncation of those proteins suggests that there is selective pressure to truncate those proteins whose function could be inactivated. The elimination of this genes or truncation of the proteins associated with these genes was not already described as being an advantage in the context of wine fermentation. Therefore, it will be interesting to verify if the truncation, which is assumed that lead to protein inactivation, is advantageous for yeasts performance during the alcoholic fermentation process.

The complete genome sequencing of EC1118 wine strain revealed, for the first time, the occurrence of gene transfer events between eukaryotes (Novo et al. 2009), being specifically identified 34 genes involved in key fermentation functions coming from Zygosacharomyces bailii were present in the genome of the EC1118 wine strain, as detailed in the introduction section. Within this context, it was thought whether the genome of QA23, VL1 and CEG strains could also harbour new genes having an origin outside of S. cerevisiae and, in particular, if they harboured the “foreign” genes identified in the EC1118 strain. Indeed, we could detect with confidence reads mapping against all EC1118 foreign genes which is a good indicative that these regions, or at least a subset of them, are also present in the genome of VL1, CEG and QA23 (Table 12 and annex). In some cases, the genes encoded by the QA23, VL1 and CEG strains were found to harbour non-synonymous SNPs when compared with the EC1118 orthologues. It was of remark the observation that in the CEG strain several of these genes (4) are truncated, which could be of relevance considering that these “foreign” genes were found to play a role in improving fermentation undertaken by EC1118. In a second step we have also tried to identify whether the QA23, VL1 and CEG strains could encode “foreign” genes not present in the genome of the EC1118 strain. This was difficulted by the utilization of SoLiD for the whole-genome sequencing of the strains since the short-reads gave rise to very short contigs rendering difficult to obtain larger contigs in which the gene detection would be facilitated. In this sense, it was decided to assemble only the reads coming from QA23, VL1 and CEG that could not be mapped against the genome of the S288c strain, which would lead to small contigs having a size similar to the one observed in genes. The overall idea of this approach was to see on whether we could reconstruct eventual “foreign” new genes sequences, which would then be used for BLAST at NCBI. A schematic representation of the workflow used is shown in Figure 3.
Figure 3 - Scheme of the Workflow undertaken concerning the analysis of the unmapped reads against the S288c strain.

To identify and subsequently characterize the alterations occurring in the non-coding genome of the three commercial wine strains, the assembled reads obtained after sequencing these strains genome were compared with the publicly available genome of the S288C reference strain. The results of this comparative analysis regarding the non-coding regions are schematically represented in Figure 4 and summarized in Table 3. A total of 9066, 10684 and 10520 SNPs were identified in the non-coding genome of QA23, VL1 and CEG respectively, the vast majority of these being found in gene promoters (which were considered to be composed by the 1000 bp located upstream of each ATG start codon).

Table 3 – Overview of the number of SNPs identified in the coding and non-coding genome of the three commercial wine strains, when compared with the genome

<table>
<thead>
<tr>
<th>SNPs/Strain</th>
<th>QA23</th>
<th>VL1</th>
<th>CEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs in Non-coding regions</td>
<td>11678</td>
<td>13862</td>
<td>13767</td>
</tr>
<tr>
<td>SNPs in gene promoter</td>
<td>9066</td>
<td>10684</td>
<td>10520</td>
</tr>
<tr>
<td>Total</td>
<td>29463</td>
<td>33313</td>
<td>33570</td>
</tr>
</tbody>
</table>

Closer analysis on the results obtained with the comparative analysis of QA23, VL1 and CEG gene promoters with those present in S288c showed that around 43, 44 and 45% of all genes predicted by the genome of the three wine strains harboured at least one SNP in their promoter, when compared with their S288c counter-partners. Most gene promoters (an average of 3582 promoters) showed no SNPs and among those that did showed differences among the strains most varied by one (an average of 976 promoters) or two SNPs (an average of 595 promoters) (Figure 4).

Figure 4 - Representation of the number of SNPs found in the promoter region of genes encoded by QA23, VL1 and CEG when compared to S288C strain.

To take a hint into how the SNPs found were distributed throughout QA23, VL1 and CEG promoters, these were divided in 50bp windows and the number of mutations in each of these windows was computed. The results of this analysis are shown in Figure 5 and in the heatmaps of the figure 6. The results obtained showed no significant differences observed in the different windows of the promoter that the SNPs in the promoter regions seem to be distributed throughout the entire gene promoters, which is consistent with the idea that these are regions that evolve rapidly. A previous similar approach with respect to the Saccharomycetae family has shown that this initial 350 bp window of yeast’s gene promoters is particularly enriched in binding sites for transcription factors (Gasch et al. 2004). If this is the case for the three commercial wine strains, and taking account the specific strains of yeasts to wine fermentation environment, then such a high rate of mutation in those regions might reflect the capacity of the wine strains to rewrite its genomics expression program by extensively affecting the transcription factor-DNA interactions.
As said above it had not been determined the transcript levels of CEG/QA23/VL1 relative to the S288c strain therefore rendering difficult to link the observed phenotypes with differences in gene expression. However, a previous transcriptomic analysis has profiled the genomic expression of CEG/QA23/VL1 strains during fermentation in nitrogen -replete or – depleted medium, thereby allowing the identification of genes differently expressed in the three strains (Barbosa et al. 2015b). On the overall, it was possible to identify a set of 89, 84, and 84 genes (for QA23, VL1 and CEG, respectively) over-expressed during fermentation of nitrogen depleted grape juice medium medium and 190, 199 and 201 genes of QA23 and to VL1 and CEG repressed during the same conditions. It is important to stress that this is only a snapshot of the overall transcriptomic landscape of the three commercial wine strains and many other differences in genomic expression could be obtained; however, it would be impossible to perform such assessment for a great variety of experimental conditions. To establish an eventual association between alterations in gene promoters and changes in transcript levels, we have identified those promoters that were only found to be changed specifically in one of the strains) and crossed that list with the gene expression data. The result of that crossing is shown in figure 7.

Figure 5 - Distribution of the SNPs throughout QA23, VL1 and CEG promoters. Each promoter region was divided in 50bp segments and the number of SNPs (assessed upon comparing each promoter with its corresponding S288C counter-partner) was computed.

Figure 6 – Heatmap representation of the distribution of SNPs found in 5 promoters that changed between S288C and the three strains. In this case, the first two gene promoters (PAU10 and YER188W) show how the distribution of SNPs per base window is similar for these two gene promoters whereas the last 3 gene promoters (YER187W, YOL159C-A and AIF1) reflect how the distribution of SNPs per base window differs among QA23, VL1 and CEG. Legend: Grey - number of mutations in the base pair window equal to 0; Green - number of mutations in the base pair window greater than 0 and lesser than 5; Yellow – Number of mutations in the base pair window greater than 5 but lesser than 7; Red – number of mutations in the base pair window greater than 7.

Figure 7 – Correlation between the number of mutations found harbouring promoter regions in the three strains when compared to S288C genome and the expression or repression of the corresponding genes in Low-nitrogen conditions. QA23 is represented in blue points, VL1 is represented in green points and CEG strain is represented in red points. In this case, it was considered the SNPs in promoter that were specifically in each strain with the corresponding fold-change value higher (for up-regulation) than 3 and -3 (down-regulation).

The fact that one strain has SNPs that do not occur in the other 2 strains with the fact that SNPs occurring at level of genes that have altered it expression may indicate that there is, possibly, a connection between the occurrence of SNPs and the expression change in these genes. However, additional studies should be developed in order to understand whether or not
this change in expression is directly related to the occurrence of these SNPs.

As mentioned previously, gene promoters are rapidly evolving regions and as such it is difficult to understand on whether or not the functional relevance of the differences identified. However, if these SNPs fall inside regions that are known to be recognized by transcription factors the alteration could result in an improvement or a loss of binding of the regulator to that promoter; thereby resulting in an alteration of gene expression. Taking advantage on the huge amount of information that has been gathered in S. cerevisiae concerning the DNA motifs that serve as binding sites for transcription factors (TFs) and also on the regulatory associations established between TFs and their target genes, largely compiled in the YEASTRACT database (http://www.yeastract.com/consensuslist.php) ((Abdulrehman et al. 2011, Monteiro et al. 2008, Teixeira et al. 2006, Teixeira et al. 2014) it is possible to examine on whether the SNPs found in the promoter region of the wine strains affected putative TF binding sites. For this, the reconstructed gene promoters present in QA23, VL1 and CEG strains were searched for the 326 described TF binding sites described in YEASTRACT. The same analysis was performed in the promoters of the S288c strain. As it can be seen by the results shown in Figure 8, in the vast majority of the cases the number of aggregated binding sites (that is, the sum of all binding sites found in a promoter) was identical in the wine strains and in the S288c strain, probably reflecting the lower rate of change occurring in DNA binding sites when compared with the surrounding regions of the promoter (Gasch et al. 2004) Nonetheless, it was also possible to detect a high number of genes that showed in the wine strains a higher or a lower number of aggregated binding sites indicating TFs that lost of gained binding sites as the result of the changes occurring in the promoter region.

To gain further insights into this in Figure 9 and 10 it is shown a representation of those transcription factors that lost or gained binding sites in the promoters of the wine strains, when compared with the frequency that was observed in the S288c counter-partner.

To gain further insights on how the regulation network of the three strains could be impacted by these “gained” or “lost” binding sites, a closer look was taken into the putative regulatory
networks controlled by Zap1 (Figure 11) Uga3 and Gat1 transcription factors. These transcription factors were selected because they were identified as being relevant in the context of wine fermentation in the fermentome analysis (Zap1) or due to their involvement in response to alterations in nitrogen availability, a key trait that is known to impact fermentations.

Figure 11 - Zap1 regulatory network for S288C, QA23, VL1 and CEG. Genes that were gained by Zap1 in each strain are highlighted in green. Consensus sequence for Zap1 transcription factor: ACCYYNAAGGT

It now remains to be tested whether or not Zap1 transcription factor is able to influence the expression of HTZ1 (and the same applies for the uncovered potential associations between these regulators and their “new” target genes) with further experimental work.

Concluding Remarks

One of the clear observations that comes out of the genomic analysis herein presented is the massive alterations observed between the three wine strains QA23, CEG and VL1 and the laboratory S288c strain. Although some of these modifications are likely to be mere the result of evolution and may not represent meaningful events affecting the physiology of the three commercial wine strains during fermentation, it is clear that many of these modifications can in fact play an important role for the evolution and different adaptation of QA23, VL1 and CEG to wine fermentation context. For example, it was possible to demonstrate here the relevance of the expression of RIF1, IRA2, GCN1, EQT2, SPO77, MDS3, PEP1 and YOL159C genes, these being some of the genes found to diverge more in the wine strains and in the lab strain. It was also herein shown that truncation of genes PRM9, SRN2, CRS5, ZPS1, HXT8, YBR182C-A, YMR316C-A, SDL1, FYV12, YJL028W and YBR090C is apparently beneficial for the progress of the fermentation of the laboratory strain BY4741, these being genes that are naturally truncated in the wine strains. Another relevant observation was the identification in the genomes of QA23, VL1 and CEG of a set of putative “foreign” genes, a subset of them also identified in the genome of EC1118 strain as being the result of horizontal gene transfer. Because these genes were found to play an important role in wine fermentation, their acquisition by the CEG, VL1 and QA23 strains could also represent an advantage for these strains. Interestingly, several new “foreign” genes were also identified in this thesis, although further experimental work is required to confirm that these genes are present and that they do play a role in fermentative capacity of the strains. Analysis of the non-coding genomic sequences of the wine strains, in comparison with the one of the laboratory strain, also unveiled significant differences including several modifications in the set of genes that could be potentially regulated by the different transcription factors. Further experimental work is now required to confirm the relevance of these new regulatory associations between TFs and target genes, which could of great relevance by changing the architecture of the regulatory networks governing genomic expression in these wine yeasts.

References


