BMC Genomics

Research article

Open Access

A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization

Emmanuel Talla*[1], Fredj Tekaia[1], Laurent Brino[2] and Bernard Dujon[1]

Address: [1]Institut Pasteur, Unité de Génétique Moléculaire des Levures (URA 2171 CNRS, UFR 927 Université PM Curie), 25 rue du Docteur Roux, F-75724 Paris cedex 15, France and [2]Eurogentec s.a., Parc Scientifique du Sart Tilman, B-4102 Seraing, Belgium

Email: Emmanuel Talla* - etalla@pasteur.fr; Fredj Tekaia - tekaia@pasteur.fr; Laurent Brino - l.brino@eurogentec.com; Bernard Dujon - bdujon@pasteur.fr

* Corresponding author

Published: 22 September 2003

BMC Genomics 2003, **4**:38

Received: 09 July 2003

Accepted: 22 September 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/38

Abstract

Background: Numerous DNA microarray hybridization experiments have been performed in yeast over the last years using either synthetic oligonucleotides or PCR-amplified coding sequences as probes. The design and quality of the microarray probes are of critical importance for hybridization experiments as well as subsequent analysis of the data.

Results: We present here a novel design of *Saccharomyces cerevisiae* microarrays based on a refined annotation of the genome and with the aim of reducing cross-hybridization between related sequences. An effort was made to design probes of similar lengths, preferably located in the 3'-end of reading frames. The sequence of each gene was compared against the entire yeast genome and optimal sub-segments giving no predicted cross-hybridization were selected. A total of 5660 novel probes (more than 97% of the yeast genes) were designed. For the remaining 143 genes, cross-hybridization was unavoidable. Using a set of 18 deletant strains, we have experimentally validated our cross-hybridization procedure. Sensitivity, reproducibility and dynamic range of these new microarrays have been measured. Based on this experience, we have written a novel program to design long oligonucleotides for microarray hybridizations of complete genome sequences.

Conclusions: A validated procedure to predict cross-hybridization in microarray probe design was defined in this work. Subsequently, a novel *Saccharomyces cerevisiae* microarray (which minimizes cross-hybridization) was designed and constructed. Arrays are available at Eurogentec S. A. Finally, we propose a novel design program, *OliD*, which allows automatic oligonucleotide design for microarrays. The *OliD* program is available from authors.

Background

DNA array-based technologies enable the determination of thousands gene expression patterns in a single experiment [1]. Such technologies involve either (*i*) the *in situ* synthesis of oligonucleotides using photochemical techniques or ink-jet oligonucleotide synthesizer [2,3], or (*ii*) the "spotting" of presynthesized DNA molecules or oligomers on glass slides or filter membranes [4,5]. In *Saccharomyces cerevisiae*, a large number of genome-wide expression analysis have already been performed [6,7]. Two DNA chip formats are currently in wide use in *S. cerevisiae*: the PCR product microarrays made from entire coding sequences (CDS) or synthetic oligonucleotide microarrays made from either short (20–25 mers) or long

(60–70 mers) oligonucleotides. In all cases, microarrays contain probes for all computer predicted-CDS from the original yeast sequence.

Seven years after the complete sequencing of the *S. cerevisiae* genome [8], several studies have refined the original annotation of the yeast genome and identified new genes [9]. Those studies involve computational methods [10,11], comparative methods [12–14], transposon tagging [15] and expression profiling combined with mass spectroscopy of proteins [16].

Compared to the numerous improvements in microarray data analysis and clustering methods [17–19], the problem of cross-hybridization between genes of a given organism has rarely been addressed so far. An automated approach (ProbeWiz) to design PCR products with minimal homology to other expressed sequences from a given organism was reported [20]. In this algorithm, the PCR primer sets are evaluated according to the sum weight of three penalty parameters (paralogy, primer quality and 3' proximity penalties) out of which the chances of cross-hybridization must be deduced by the user. Moreover, softwares for genome-scale PCR primer or oligonucleotide design (PrimerArray, OligoArray and OligoPicker) were recently published [21–23]. These programs compute gene-specific and secondary structure-free oligonucleotides for microarray construction.

Here, we describe a novel design of microarray probes based on a refined definition of *S. cerevisiae* genes and the analysis of their potential cross-hybridization. The dynamic range, sensitivity and reproducibility of results using this novel design was assessed and the actual cross-hybridization was experimentally examined. Following the experimental validation of our cross-hybridization procedure, we have written a novel program, *OliD*, which allows the automatic design of long oligonucleotides for microarray hybridization experiments.

Results

Establishing the list of protein coding sequences in Saccharomyces cerevisiae useful for microarray analysis

Most microarrays commonly used until now were designed according to the initial yeast genome annotation which included about 6200 protein coding genes [8,24]. The *S. cerevisiae* coding sequences (CDS) were originally predicted using, as primary criteria, their size (> = 100 codons from ATG to stop) and ignoring all reading frames entirely included within longer ones [25]. Additional criteria such as the codon adaptation index, CAI [26] were sometimes used, in particular to identify short CDS without obvious similarity or to estimate the likelihood of partially overlapping CDS. Shorter CDS were considered only when previously described or when their predicted product showed similarity in public databases. In addition, original annotations were not carried out uniformly for all chromosomes (e.g. chromosome I annotations [27] differed from chromosome VIII [28]).

Before starting the design of the novel microarray probes, we reviewed the entire yeast genome sequence and established a novel list of *S. cerevisiae* CDS. A first working list (6252 CDS) was composed of all predicted CDS exceeding 99 codons which are not entirely included within longer CDS and do not overlap other defined elements such as tRNA genes, LTRs, ... plus novel short CDS identified by comparative genomics in the Genolevures project [13]. This list includes 993 partially overlapping CDS (83% of them are in antiparallel orientation excluding frameshift sequencing errors), most of which resulting from over-prediction and/or mirror effects as judged by the fact that 58% of them do not show similarity with other hemiascomycete yeasts. In 449 cases (see additional data file 1), a clear-cut distinction between the two partially overlapping CDS was obtained by comparing their translation product against our Genolevures data set. In such cases, one CDS (conserved in the list) showed similarity to sequences of several other yeast species whereas its partner (eliminated from the list) remained entirely devoid of homolog. A final list of 5803 CDS is used for the present microarray design. This list includes 93 partially overlapping CDS that are real genes or could not be resolved based on similarity results.

Design of microarray probes

Three major considerations lead us to the present probe design: the attempt to have uniform probe lengths; the desire to locate probes preferentially near the 3' end of the CDS; and the desire to eliminate cross-hybridization. Note that we refer to the spots on the glass slide as *probe* and to the sample hybridized to the glass-bound array as *target*.

Prediction of potential cross-hybridizing regions within our selected CDS

In classical hybridization experiments, cross-hybridizing sequences can be distinguished from identical sequences by increasing the stringency of hybridization conditions. In the present case, however, it is not possible to apply the most stringent conditions for each of the 5803 genes at the same time. We have examined sequence similarity between each of the selected CDS and the entire yeast genome sequence in order to determine the lengths of possible sequence alignments, the percentage of identity between them, and therefore the best probe not susceptible of cross-hybridization.

Cross-hybridizations of a CDS with other yeast sequences may results from (1) a single continuous alignment

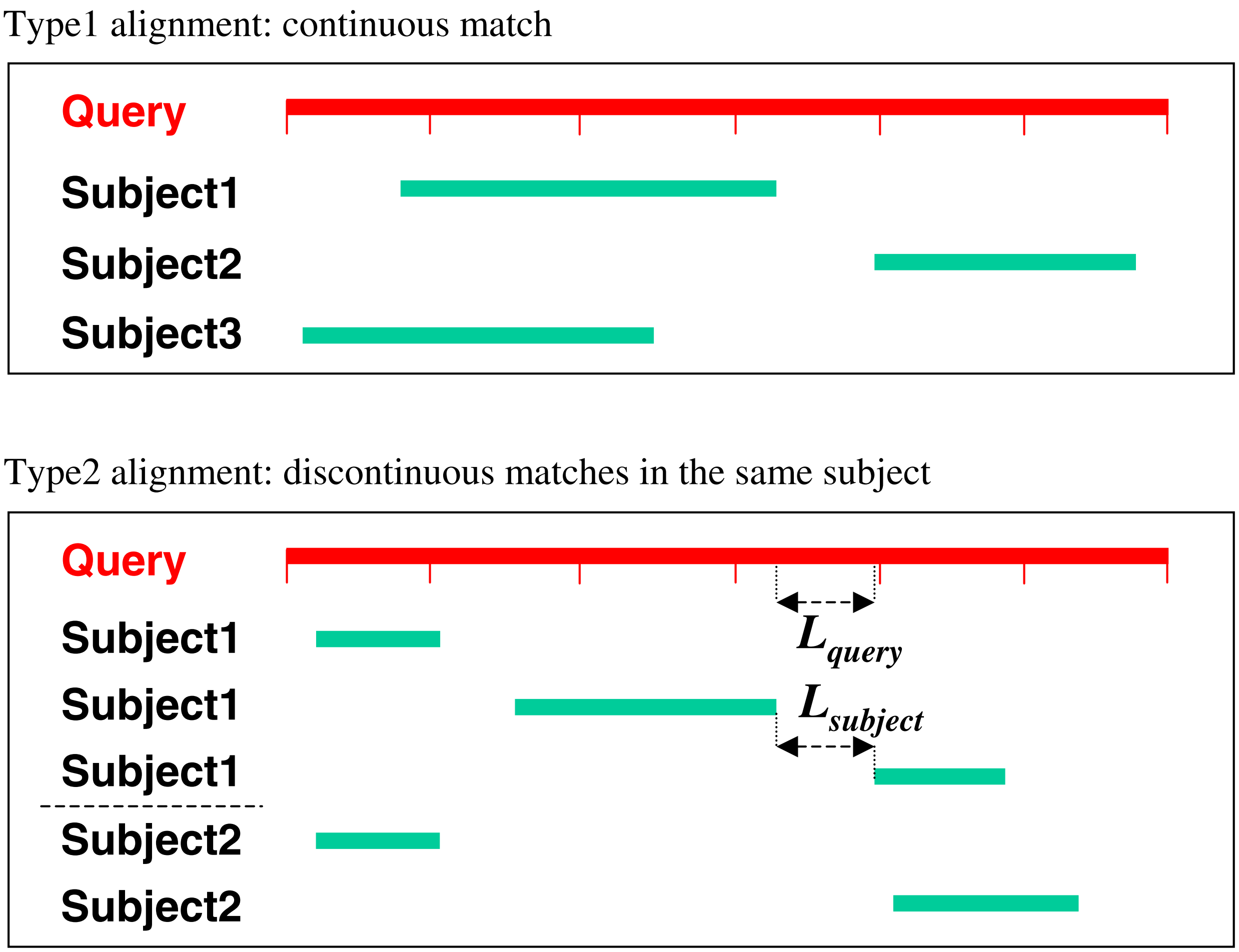

Figure 1
Possible types of blast alignments between a probe (query) and the *S. cerevisiae* genome (subject). (1) The query sequence has a continuous match in one or several different subject sequences (Type1 alignment); (2) the query sequence has discontinuous matches aigainst several regions in the same subject sequence (Type2 alignment). $L_{query}$and $L_{subject}$ are the distances (in bp) between two successive blocks of alignments in query and subject sequences, respectively.

exceeding cross-hybridization parameters or (2) from a discontinuous series of short similarity regions (within the same subject) appearing as several alignments which all together exceed the cross-hybridization parameters (Figure 1 and "Methods"). This last situation is of critical importance because a previous study [29] suggests that short regions of sequence identity spread over the entire length of a probe may result in cross-hybridization artefacts in cDNA microarrays. This type of alignments was found in 17.6% and 5.1% of cases when comparing entire CDS to other coding sequences and intergenic regions, respectively. Such cases were treated as described in "Methods".

In order to predict cross-hybridizing regions in each CDS, three parameters were combined. The alignment size (*AS*) threshold was set to 50 nt, based on predicted melting temperatures [30] with an average GC content of 39%, as in the *S. cerevisiae* genome. The percentage of identity (*ID*) was set to 70% based on previous studies in *Escherichia coli* [31] (52% GC content) and in *Arabidopsis thaliana* [32] (40% GC content). Finally, a longest contiguous perfect match segment (*LC*) of 30 nt was considered sufficient for cross-hybridization.

Using above criteria, 4523 (78%) of yeast CDS are not expected to show cross-hybridization with any other transcript from the yeast genome. In such cases, the corresponding probes could be easily designed in the preferred region (as described below). For the remaining 1280 CDS, cross-hybridizations are expected with other CDS (841), with intergenic regions (191), or with both simultane-

Table 1: Total number of *Saccharomyces cerevisiae* protein coding genes and corresponding probes

	No CH	*CH with other CDS only*	*CH with IR only*	*CH with both CDS & IR*	*Total*
Natural CDS	4523	841	191	248	5803
Designed probes					
PCR products	4972	0	0	0	4972
Oligonucleotides	688	65	9	69	831

CH: cross-hybridization; ***CDS***: coding sequences; ***IR***: intergenic region

ously (248) (Table 1). In such cases, designed probes were selected to avoid the cross-hybridizing region(s) as far as possible (see below).

Probe design (made of PCR products (4972 cases) or oligonucleotides (831 cases)) predicts that 5660 CDS (97.5% of the total) should be free of cross-hybridization. This is a gain of 1137 CDS compared to full-length probes often used in other arrays. For example, HXT2 (*YMR011w*), a high-affinity hexose transporter is known to share high amino acid sequence identity with other members of the same family [33]. At the nucleic acid level, we found that *YMR011w* has potential cross-hybridizing regions with all the other 15 members of the hexose family: HXT1; HXT3-12; HXT14-16 as well as with *YLR081w*, the gene encoding the GAL2 protein. In a previous study, the HXT2 transcript signal could not be determined because of high sequence similarity with the other hexose transporter genes [34]. In this work, our designed probe for HXT2 does not show cross-hybridization with members of the same family (see also Additional file: 7).

No non cross-hybridizing probes could be designed for the 143 remaining CDS (2.5% of the total). Among them, 9 cross-hybridize with intergenic sequences and the 134 others with CDS only or with both CDS and intergenic sequences (Table 1). The complete list of such CDS as well as their putative cross-hybridizing sequences is given by the additional data file 2.

Designed primers and corresponding PCR products

From the previously defined non cross-hybridizing regions, 4972 pairs of primers were designed to generate PCR products as probes (see additional data file 3). The average GC content of the designed primers is 47.8%. As far as possible, reverse primers were selected for their proximity to the 3'-end of each gene. This feature was chosen because, in most experiments, the hybridizing target will be prepared by reverse transcription with oligo(dT)-primed mRNA. Distances between the reverse primer and the stop codon is less than 100 nt for 62% of the CDS and less than 300 nt for 94% of the CDS. Reverse primers were selected at greater distances from the stop codon in 316 remaining cases, due to the presence of putative cross-hybridizing regions near the 3'-end of these CDS. The forward or reverse primers were selected in the 5'- or 3'-untranslated regions (UTR) in 37 CDS because their non cross-hybridizing regions were shorter than 300 nt and located in the 5- or 3' regions of the CDS (see additional data file 4). Attention was placed not to overlap the next CDS.

Minimal size of the PCR products was limited to 300 bp because it was previously reported that small genes (<300 nt) often showed lower intensity signals in hybridization experiment [31]. Average size of our PCR product set is 521 bp. The GC content of the designed PCR probes range from 20 to 62 % (with an average of 40.6) compared to 20 to 59% (with an average of 40%) for entire yeast CDS. PCR amplification of yeast DNA using the above primer pairs showed good quality products of 98.8% of experiments.

Long oligonucleotides as probes

For the 831 CDS in which non cross-hybridizing regions are shorter than 300 nt, we decided to use synthetic 71-mer oligonucleotides as probes (see additional data file 5). Using this strategy, we reduced the number of cross-hybridizing probes by 688, leaving only 143 CDS in which cross-hybridization is unavoidable. Oligonucleotide probes were selected within the last 800 nt of each CDS. They have a melting temperature close to 70°C (as determined using the *melting* program [35] with sodium concentration and nucleic acid concentration in excess set to 0.825 M and 0.02 M, respectively). Note that 199 selected oligonucleotides partially overlap the 5'- or the 3'-UTR (mainly for short CDS). For each of these oligonucleotides, the sequence in 5' or 3'-UTR is 41 nt and does not overlap the next CDS.

Sensitivity and specificity of our microarrays

We have determined the sensitivity and specificity of our microarrays from results of hybridization experiments described in the next section. Under our experimental conditions, 82–92% of all probe spots produce detectable transcript signals i.e. showing Cy3 or Cy5 intensities higher than the 95% upper limit of the distribution of empty spot signals. This is indicative of the high sensitiv-

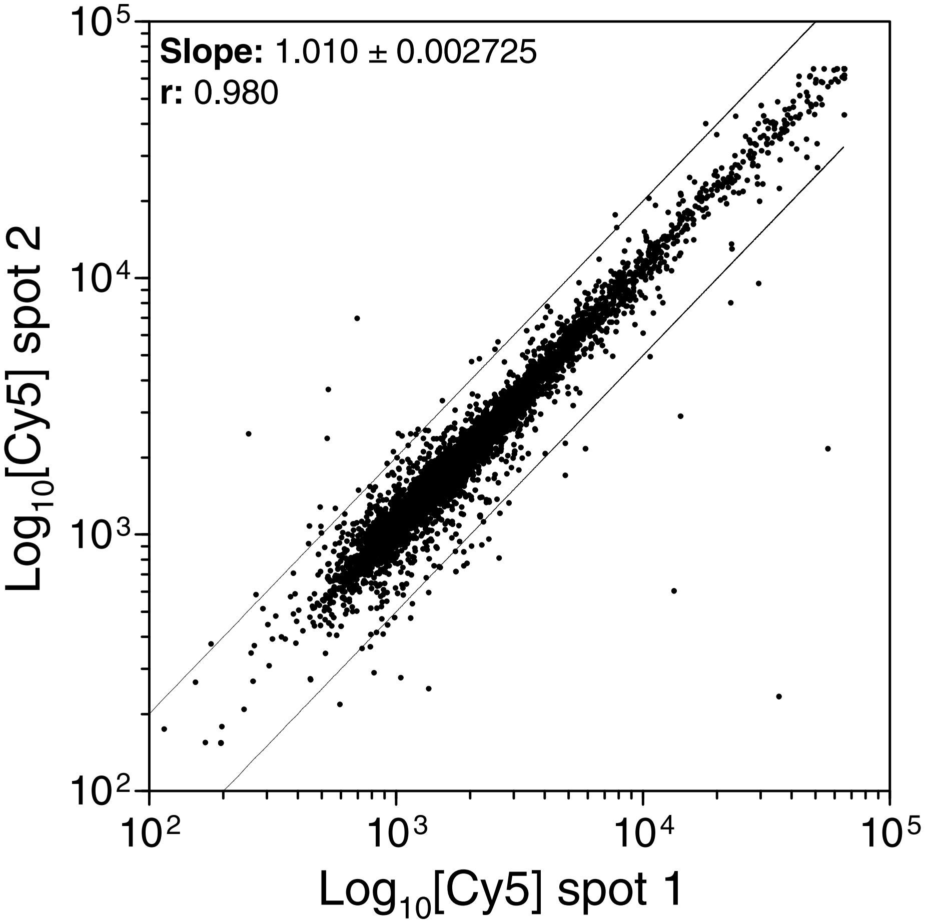

Figure 2

Comparison of transcript level measurements between duplicated probe spots within a single array from Batch616. A scatter plot of Cy5 intensities at duplicated probe spots from a single hybridization is shown. cDNA targets were prepared from total RNA isolated from wild-type BY4742 (Cy3 labelled) and Δ*ydr225w* (Cy5 labelled) strain as described in "Methods". The 2 × limits are shown.

ity of our system. An example (for Cy5 channel) of the dynamic range and reproducibility of the duplicate measurements of an array is shown in Figure 2. As can be seen, transcript intensities can be measured over a 3-log range and the reproducibility of measurements of the two duplicates is high (slope close to 1 with r value of 0.98). Similar results were obtained for the Cy3 channel (data not shown).

We also compared hybridization results using arrays from 3 different batches (Batch800, Batch400, and Batch616) successively prepared in this order. Batch400 gave higher values for background signals than Batch800 and Batch616 (Table 2). In our most recent experiments, which give the best transcript abundance, Batch616 was used.

We have examined the effects of probe size, GC content, and position relative to stop codon on signal strength (case of Cy5 channel). Results of hybridization experiment with Δ*ymr011w* strain are shown by Figure 3. The probe size (300–700 bp) has no significant effect on signal although the running average of the intensities ranges from 2039 to 4145. The same is essentially true for the distance between the probe and the stop codon (except for very high figures (>1000 nt)). On the contrary, the GC probe composition shows some influence on intensities, especially for lower GC as could be expected. As can be seen, the signal intensity increases with the GC probe content with a maximun at 44 to 46% GC. Note, however, the large distribution of values over the entire composition range. Similar results were obtained for the Cy3 channel (data not shown). Moreover, the figure observed in

Table 2: Comparison of hybridization signals between different batches (Batch800, Batch400, and Batch616) of arrays.

	Total number of arrays	*Median*		*Mean*		
		Background	*Probe*	*Background*	*Probe*	*TA*
Cy5 Channel						
Batch800	2	50	536	57	1379	24.0
Batch400	5	314	681	335	1993	5.9
Batch616	13	54	1570	72	3761	52.2
Cy3 channel						
Batch800	2	56	506	62	971	15.6
Batch400	5	249	1047	293	3092	10.5
Batch616	13	75	1144	92	2627	28.5

For all arrays of a given batch, the mean intensity value for each gene was first calculated considering all intensity measurements above 95% of the background threshold. Distribution of intensities were computed for all genes. Results are described by their medians and means. The transcript abundance (TA) is the mean intensity of the probe in the batch over mean intensity of the background signal.

Δ*ymr011w* hybridization is the same in other hybridization experiments (data not shown).

Within two independent experiments involving BY4742 (Cy3-labelled) and Δ*ydr225w* (Cy5-labelled) cDNA, we compared the oligonucleotide signals to the PCR probe signals (Table 3). The hybridization signals of oligonucleotides tend to be higher than those of PCR products (e.g. respective mean values 3512 and 3093 for Cy5 channel, and 2280 and 2194 for Cy3 channel in experiment B). The same is not true for median values. Nevertheless, signals of oligonucleotide spots remain highly significant in comparison to background signals (more than 20 fold in Cy5 channel and more than 6.5 fold in Cy3 channel). In our experimental conditions, sensitivity (94%) of the oligonucleotide probe spots was similar than the one observed for the PCR probe spots (95%). In separate experiments which involved PCR and oligonucleotide probes with extensions in the 5'- or 3'-UTR, we found that more than 87% of the probe spots produce a detectable signal with signal ratios over background ranging from 4.8 to 9.5.

Experimental validation of predicted cross-hybridization profiles

In order to test experimentally our cross-hybridization predictions, we have used the collection of yeast deletion mutants. We have first selected a set of probes representative of the various situations encountered with regard to potential cross-hybridizations (i.e. different alignment size and different percentage of identity between the tested probe and the potential cross-hybridizing gene) (*Part I* in Table 4 [Additional file: 6] & Table 5 [Additional file: 7, *part II*]). Eleven deletants were selected to represent cases in which the probe shows potential cross-hybridization with only one other coding sequence and the 2 other deletants (*YML113w* and *YMR011w*) to represent cases with potential cross-hybridization to several other coding sequences. One deletant (*YEL033w*) was selected for its predicted cross-hybridization with an intergenic sequence and a last one (*YNL143c*) has putative cross-hybridizing regions with one coding sequence and 14 intergenic sequences. Putative cross-hybridizing regions range from 31 to 494 nt with identities higher than 79%. The last 3 deletants (*YBR025c*, *YML072c*, and *YLR109w*) were selected as negative controls (no putative cross-hybridization).

For each of the 18 different hybridization experiments, intensities were normalized as described in "Methods" and the Cy5/Cy3 ratios were calculated for all the probes. For convenience, values are presented as a normal or a flip ratio (FR)(see "Methods"). Figure 4 describes the theoretical results expected: a ratio close to 1 or -1 (for a flip ratio) (with yellow color) should be observed at the spot corresponding to the cross-hybridizing gene (used as control) whereas the color at the spot corresponding to the tested gene (deleted gene in mutant) should be either yellow (in case of cross-hybridization) or green (in case of non cross-hybridization). Results are given in Tables 4 (see Additional file: 6, *Part II*) & 5 (see Additional file 7, *Part II*). For each spot, we determined the transcript abundance (given by TA5 and TA3) by comparing the hybridization signal at the spot with the average background signal of the array. In order to allow comparisons between the 18 experiments, we also calculated the mean of the Cy5/Cy3 ratios (given by Control ratio) for the tested gene in the 17 experiments in which the tested gene is not deleted. As shown, all values are close to 1 or -1, indicating the homogeneity and reproducibility of experiments (*Part III* in

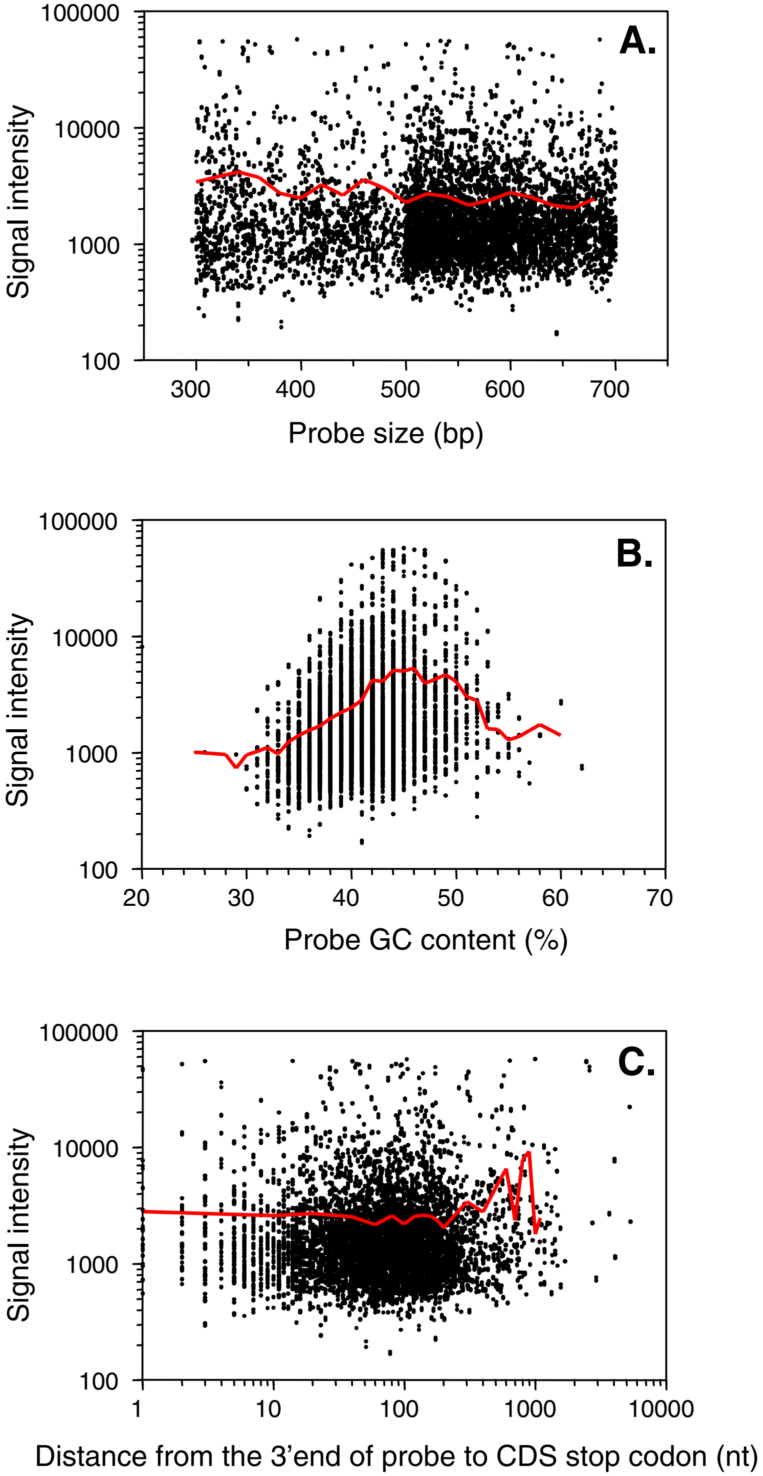

Figure 3

Effects of probe size, GC probe content and position relative to stop codon on signal intensities. Cy3- and Cy5-labelled cDNA targets were prepared from the total RNA isolated from the wild-type strain BY4742 and Δ*ymr011w*, respectively. Hybridization was performed as described in "Methods" with microarray from Batch616. For the Cy5 channel, normalized signals of each spot were computed as a function of probe size (panel A), GC probe content (panel B) and the distance from 3'end of the probe to stop codon of the CDS (panel C). Panel A includes PCR probes only, panels B and C include both PCR and oligonucleotide probes. Red curves represent running averages of the signal intensities.

Table 3: Comparison of hybridization signals between oligonucleotide and PCR probes.

Exp. A	*Total number of spots*	*Median*		*Mean*		
		Background	*Probes*	*Background*	*Probes*	*TA*
Oligonucleotide						
Cy5 channel	1718	162	1616	166	4561	27.5
Cy3 channel	1718	299	1535	301	3760	12.5
PCR	6					
Cy5 channel	9962	155	1619	157	2181	13.9
Cy3 channel	9962	291	1666	295	2123	7.2

Exp. B	*Total number of spots*	*Median*		*Mean*		
		Background	*Probes*	*Background*	*Probes*	*TA*
Oligonucleotide						
Cy5 channel	1576	177	845	175	3512	20.0
Cy3 channel	1576	339	822	350	2280	6.5
PCR						
Cy5 channel	9936	172	2047	170	3093	18.2
Cy3 channel	9936	346	1600	355	2194	6.2

Intensities and corresponding background values were computed separately for oligonucleotide and PCR probes. Experiments were done twice (Exp. **A** and **B**) using microarrays from batch J100B. Cy3- and Cy5-labelled cDNA targets were prepared from the total RNA isolated from the wild-type strain BY4742 and Δ*ydr225w*, respectively. For each channel, the distribution of intensities for all genes was computed for oligonucleotide or PCR probe. Results are given by median and mean of the intensities. The transcript abundance (TA) is the mean intensity of probe over mean intensity of background.

Table 4 (see Additional file: 6) & Table 5 (see Additional file: 7)).

As expected, our three control genes *YBR025c*, *YML072c*, *YLR109w* (botton lines in Table 4 - see Additional file: 6) predicted not to cross-hybridize, show highly negative flip ratios (-100 to -33.3) at the tested gene spots. Correspondingly, the transcript abundance in the mutant (TA5) is significantly lower than in the wild-type (TA3). For the probes predicted to cross-hybridize with only one other *S. cerevisiae* gene, cross-hybridization was observed for *YMR170c*, *YMR169c*, *YDR225w*, *YGR148c*, *YGL147c*, *YBR145w*, and *YNL143c* (flip ratios at the tested spots range from -2.1 to -1.2) but not for *YOL103w*, *YDR497c*, *YGR086c*, *YJR148w* and *YLR058c* (flip ratios at the tested spots range from -50 to -9). Note that, with the exception of *YGR148c* (missing spot of the putative cross-hybridizing gene *YGL031c* on our array) and *YNL143c* (hybridization signal of the putative cross-hybridizing gene *YGR069w* below the background signal), ratios at the cross-hybridizing spots are close to 1 or -1, as expected.

The difference between the first 7 genes (that cross-hybridize) and the last 5 (that do not cross-hybridize) may have several causes. First, the transcript abundance for the potential cross-hybridizing gene may, of course, influence the signal at the tested spots. In our experiments, however, low and high levels of transcript abundance from the potential cross-hybridizing gene (given by TA3 and TA5) were observed in each category, suggesting that this factor has limited importance. Similarly the alignment size (*AS*) does not seem to play a major role (67 to 348 nt for the probes that do not cross-hybridize compared to 31 to 494 nt for those that do). In contrast, the *LC* value seems to play the most important part in our results. If one excepts *YBR145w*, *LC* values for the 6 tested probes which cross-hybridize, range from 31 to 187 nt, and for the 5 tested probes which do not cross-hybridize, from 15 to 24 nt. The cross-hybridization observed for *YBR145w* (with a *LC* value of 14 nt) is probably due to the high transcript abundance of the cross-hybridizing gene *YMR303c* (TA5 of 338).

Results of experiments with the probes (*YML113w* and *YMR011w*) having putative cross-hybridization with several other yeast genes (Table 5 - see Additional file: 7) are compatible with the above conclusions in the sense that *YMR011w* (whose *LC* values range from 11 to 20) do not show cross-hybridization (flip ratio of -16.6) while *YML113w* does (*LC* values from 11 to 32 with a flip ratio of -1.5).

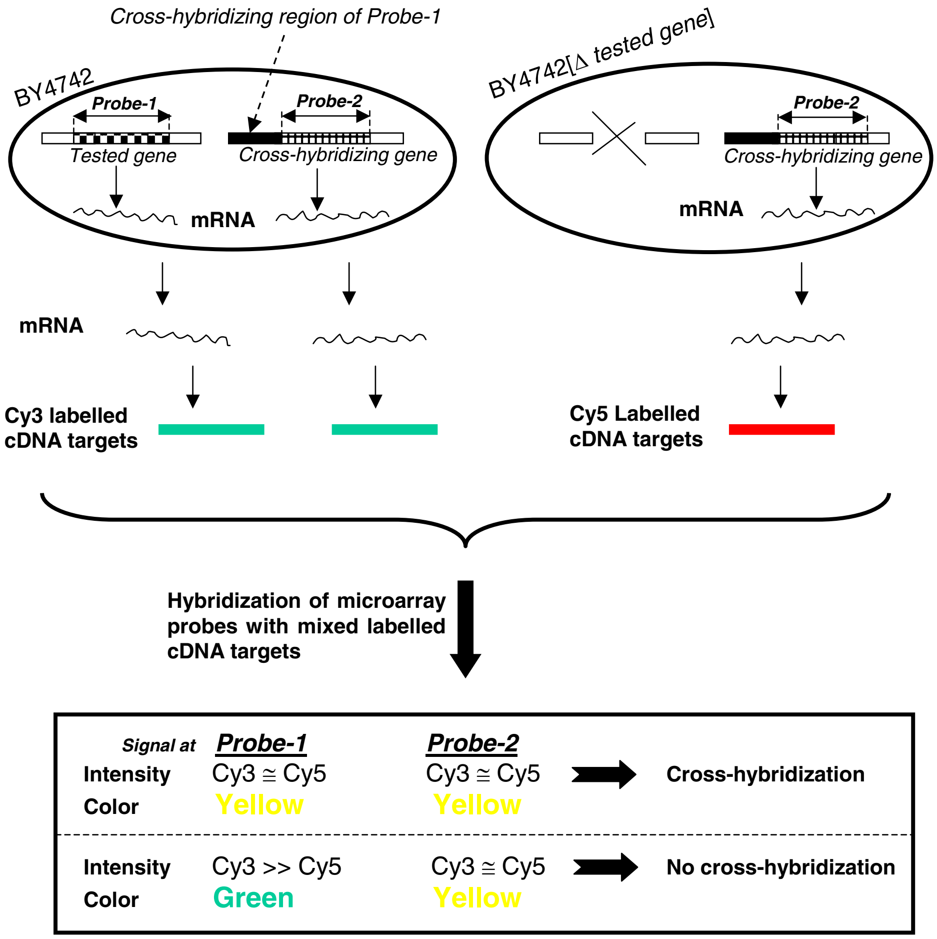

Figure 4
Protocol strategy used in the array experiments.

The cross-hybridization observed for *YNL143c* (*LC* of 31 nt) cannot be due to its potential cross-hybridizing gene (*YGR069w*) because its transcript level is undetectable. However, the *YNL143c* probe also shares 31–34 nt in common with several intergenic regions with might be transcribed since some of them are close to the next CDS. This result suggests that transcripts from intergenic regions can lead to cross-hybridization, thus validating our concern to examine cross-hybridization of probe with intergenic sequences. For the probe *YEL033w*, however, the flip ratio of -16.6, indicates an absence of cross-hybridization with a possible transcript from the potential cross-hybridizing intergenic region (position 92686–92717 on the chromosome 4) that is only 45 nt distant from the next CDS.

In summary, of all parameters that may interfere with cross-hybridization results, it seems that *LC* plays the most important part. From our data, two different sequences with *LC* ≤ 24 should not significantly cross-hybridize in regular experiments while sequences with *LC* > 24 do. Effect of the transcript abundance of the cross-hybridizing gene may of course play some role, in particular when *AS* and *LC* are low (i.e. *AS* of 53 with *LC* of 14). Considering all our results, we predict that a total of 5660 probes in our microarray design should be specific for the selected gene.

Validation of the OliD program

Following the above results, we developed a program (*OliD*) that allows the automatic design of oligonucleotides probe of various size, GC content, or melting temperature, depending on user-defined input parameters (see "Methods"). This program has been successfully tested on CDS from *S. cerevisiae* (data not shown). From an entry file of 5607 CDS (without intron) and the specified parameters (oligonucleotide length of 65–70 nt, Tm range of 70–90°C; *AS*, *ID* and *LC* of 50, 70, and 24, respectively) the *OliD* program successfully designed "unique" oligonucleotides for 5332 CDS (95.1%). Because of their higher similarity with other sequence regions of the yeast genome, 253 CDS (4.5%) failed to be represented by a "unique" oligonucleotide probe. In such cases, the program selects an oligonucleotide with the minimum number of cross-hybridizing regions and displays the location of the cross-hybridizing regions. For the remaining 22 CDS (0.4%), no oligonucleotide could be designed because of their short length or excessive AT-richness.

Discussion

In this work, we designed and experimentally tested a novel hybridization array for all the 5803 predicted coding sequences (CDS) of *S. cerevisiae*. Our design differs from currently used yeast arrays by the list of probes selected and by our concern to minimize artefactual cross-hybridization between related sequences.

The prediction of protein coding genes in *S. cerevisiae* is not as simple as it may seem. During the last 7 years, since the publication of the original sequence [8], corrections have been introduced in several yeast chromosome sequences. Although most of them are only minor ones (except for the chromosome 4), the complete sequence of chromosome 3 has been entirely re-sequenced (MIPS, http://mips.gsf.de). In addition, the original sequence annotation, essentially based on CDS size, resulted in some degree of over-prediction and under-prediction, due to that all partially overlapping CDS were considered and that short coding sequences were avoided, respectively. In this work, we used the first extensive comparative sequencing data on different yeast species to address both problems. A number of novel CDS whose protein products are conserved in other yeast species have been included, mostly short ones [12,13]. At the same time, a number of questionable CDS (mostly partially overlapping ones) have been eliminated, resulting in a total of 5803 CDS. Note that a total of 98 partially overlapping CDS are still included in this set, although we believe that in most cases, only one of them will correspond to an actual gene. For those partially overlapping CDS, our probe design procedure automatically removed the overlap.

While this work was in progress, other publications reported the re-annotation of the *S. cerevisiae* genome, based on theoretical, comparative or experimental studies. The likelihood of predicted CDS was assessed using methods such as the three-dimensional space curve (Z curve) or the dinucleotide composition, resulting in 5645 CDS or 5300–5400 CDS, respectively [10,11]. The first figure is close to 5651 CDS, an estimation based on statistical and comparison analysis with partial sequence data from a number of other yeast species [36]. Similarly, comparisons with the finished *S. pombe* genome resulted in an estimation of a maximum of 5804 real CDS for *S. cerevisiae*, and it was suggested that 370 others, defined as totally spurious CDS, should be disregarded [14]. Interestingly, 306 of those CDS were also eliminated in our work. This indicates a good correlation of results despite the fact that our approach took into account sequence comparisons within a less diverged phylogenetic group (the hemiascomycetes) than *S. pombe* (an euascomycete). Similarly, 317 of the unlikely CDS resulting from Z curve analysis are also eliminated in our work. A number of experimental methods were recently used to reveal the presence on non-annotated short CDS. Using an original transposon-mediated gene trapping associated with expression analysis and similarity searching, Kumar *and coll.* [15] identified 137 previously non-annotated CDS. Most of those CDS, however, are short ones that are included in antiparallel orientation into other annotated CDS or are in subtelomeric locations, and are not in the present work. By a combination of expression profiling and mass spectrometry of yeast proteins, 62 novel genes were proposed [16]. This work also provides independent experimental support for 13 previously identified genes [12,13] that are already included in our set of yeast genes. These new non-annotated genes, treated in our annotation procedure (as described before), revealed that 17 novel non-annotated CDS not considered yet in the present work could be considered as real genes.

The second major concern of our work was to reduce cross-hybridization between related yeast sequences. Such cross-hybridizations can result from (*i*) duplications of genes or short repeated sequences (such as di- or trinucleotide arrays) which can be predicted by sequence analysis, and from (*ii*) the existence of transcribed sequences (pseudo-genes, disabled CDS or relics) in intergenic regions that may resemble to actual gene sequences [37,38]. In addition, natural transcripts contain 5'- and 3'-UTR which, except in a few cases, have not been experimentally determined and therefore correspond to "intergenic" region in databases. These are the reasons why we have compared the sequences of all predicted CDS with the entire yeast genome and not only with other coding sequences.

According to our probe design and its experimental validation using mutants from the systematic deletion collection, we propose here a complete yeast array in which each of the 5660 probes is specific for a unique target gene. This represents more than 97% of the actual yeast CDS. Our experiments showed that, of all sequence similarity parameters, the *LC* value plays the most important role to determine cross-hybridization. No cross-hybridization was observed for *LC* < = 24. If we use *LC* > 24 as a limit, 5579 probes (96% of the total) are expected to be specific for a unique target gene. In an independent work, cross-hybridization was reported between sequences sharing more than 15 contiguous identical nucleotides [39]. Using this more stringent criterion, a total of 5454 (94% of the total) of our designed probes are still expected to be specific for a unique target gene. As expected, other parameters than *LC* nevertheless plays some role in cross-hybridization. For example, a sequence of 162 nt with 81% of nucleotide identity was found sufficient to favour cross-hybridization in our experimental conditions, an observation which is consistent with earlier studies [31,32,39]. The expression level (transcript abundance) of the putative cross-hybridizing sequences is also a determinant. In the case of highly expressed gene such as *YMR303c*, cross-hybridization was observed with the *YBR145w* probe although the two sequences share only 53 nt segment with 86% identity and a *LC* of 14.

Despite our probe design protocol, cross-hybridization could not be eliminated for 143 CDS because they are part of highly conserved gene families. Only 5 of them are parts of ancient duplicated chromosome blocks as previously described [40]. For the remaining 650 CDS pairs in such blocks, our design successfully discriminated each CDS from its partner(s). One example is provided by the well-studied yeast gene family encoding the alcohol dehydrogenases (ADH)[41]. From the five distinct ADH genes present in *S. cerevisiae*, *ADH1* (*YOL086c*) and *ADH2* (*YMR303c*) show cross-hybridization with other members of the family. Using our probe, it is now possible to study individually *ADH3* (*YMR083w*), *ADH4* (*YGL256w*) and *ADH5* (*YBR145w*) without cross-hybridizations with the other family members.

In 236 cases (37 PCR products and 199 oligonucleotides), our probes include extensions into 5'- or 3' intergenic region flanking the CDS. Hybridization with those probes has been experimentally verified. Although 5'- and 3'-UTR have not been determined for all yeast genes, it was previously predicted that in 85–95% of cases, poly(A) sites are located between 55 and 145 nt after the stop codon of the CDS [42], a figure which is perfectly consistent with the fact that the average distance between two converging CDS is 326 bp [43]. Except in two cases (*YBR191wa* and *YLR149ca*), our probes are entirely included into the predicted 3'-UTR regions. It is, therefore, unlikely that they will hybridize with transcript from the neighbouring gene.

Compared to other yeast arrays, our design is unique by the combination of PCR products and long oligonucleotide probes on the same array. The high success rate of PCR reactions is attributable to our carefully designed primers, in which attention was focussed on the uniqueness in the entire yeast genome sequence. Thus, in most cases, the reverse primers designed (without their common 5'-tag) can also be used to prime the synthesis of the first strand cDNA target. Although the mRNA priming with specific reverse primers will shorten the average length of cDNA, this priming method should increase the sensitivity of detecting rare mRNAs in samples because there is no more priming competition between rare mRNAs and abundant mRNA as it can be observed with oligo(dT) primers. Therefore, the use of a mixture of oligo(dT) and specific reverse primers should provide uniform coverage of cDNA target length and should increase the detection of rare mRNAs.

Another explanation for the absence of cross-hybridization could be that the cross-hybridizing region lies upstream of a long CDS such that this region is under-represented in our labelled reverse transcriptase-generated cDNA. This, however, seems unlikely because the distance from the probe to the stop codon of the corresponding gene ranges from 18 to 246 nt (for probes showing cross-hybridization) and from 38 to 240 nt (for probes not showing cross-hybridization).

While this work was in progress, two programs that select oligonucleotides as microarray probes were reported. In OligoPicker program [23], the cross-hybridization approach is based on finding the occurrences of contiguous perfect matches. In our approach, we also emphasize that contiguous base pairing is one of the most important parameters during the cross-hybridization procedure. However, we focus on the longest perfect match, which was experimentally verified as a leading parameter for cross-hybridization. In the same study [23], the secondary structure of the selected oligonucleotides was assessed by calculation of regions that may self anneal, whereas in our study, prediction of the secondary structures was calculated from the free energy of the oligonucleotide molecule [44]. Our design follows the same strategy as in the OligoArray algorithm [22], except that it (1) first determines the non cross-hybridizing regions within the sequence entry (instead of scanning the sequence to check for oligonucleotide specificity); (2) can computes different oligonucleotide lengths within the same run; and (3) finds the best oligonucleotide as configured by the user. In addition, the *OliD* program displays the location of cross-hybridizing regions as this information might be useful during the microarray hybridization analysis.

Conclusions

We have described a novel yeast microarray that combines a refined annotation of *S. cerevisiae* genome and an experimentally validated method that predicts cross-hybridizing regions between two related sequences. In our approach, we emphasize that contiguous base pairing is one of the most important parameters for cross-hybridization, in particular the longest contiguous perfect match (*LC*). Our yeast microarray experiments show a suitable dynamic range, good reproducibility and sensitivity. Since oligonucleotide-based microarrays are now preferentially set up, we developed an oligonucleotide design program, *OliD*, which allows a user-defined input parameters and automatic design of long oligonucleotides. Performances of the *OliD* program have been successfully evaluated on a well-established data set.

Methods

DNA sequences and databases

DNA sequences of the 16 *S. cerevisiae* chromosomes were downloaded from MIPS on February 2001. A set of 5803 protein coding genes (see "Results") was used in this work and referred to here as the GSC database. The GSC database includes CDS originally described at MIPS plus novel short CDS discovered in Génolevures [13] or other comparative sequence analysis [12]. All remaining sequences from *S. cerevisiae* were considered as "intergenic" sequences and called GSCIG database. The list of intron-containing CDS was compiled from MIPS, YIDB [45] and ARES [46] databases. Position and length of trinucleotide repeats relative to *S. cerevisiae* CDSs were taken from [47].

Design of DNA probes

Our probe design involves 3 steps as illustrated by Figure 5.

Step1. Detection of non cross-hybridizing regions within coding sequences

Each sequence (here the entire CDS as query) was first compared to our GSC database using *blastn* program (versus 2.1.2) [48] without low complexity filter, in order to identify potential cross-hybridizing regions. Self-matching scores were ignored. The two possible types of alignments are described in Figure 1. Type2 alignments were considered as putative cross-hybridizing regions if the following two conditions were simultaneously met: (*i*) the successive alignment blocks between the query and subject sequences have the same strand orientation. (*ii*) L_{query} and $L_{subject}$ distances are less than 500 nt and more than -7 nt (cases of partial overlaps). The threshold value of 500 nt was chosen because it corresponds to the average size of probes to be designed. Overlapping alignment threshold was arbitrary set at 7 nt.

Potential cross-hybridizing regions were identified when alignments meet one of the following two conditions: (i) $AS > 50$ and $ID > 70$; or (ii) $LC > 30$, where *AS*, *ID*, and *LC* are the size, the percentage of identity, and the longest contiguous perfect match segment in each alignment, respectively. Note that for Type2 alignments, *AS* and *ID* were re-calculated globally from extreme coordinates of blocks. The remaining part of each CDS (without potential cross-hybridizing regions) was retained from subsequent analysis.

After elimination of introns and trinucleotide repeats, all segments larger than 72 nt were extracted and compared against our GSCIG database using the same procedure as above to eliminate potential cross-hybridization with intergenic regions. After this step, a list of non cross-hybridizing regions was obtained for each CDS, among which our probes were selected. When possible, preference was given to sub-segments of 800 nt in length selected by their proximity to the 3'-end of the CDS. In some cases, sequences were extended to the neighbouring intergenic regions. Based on the size of available non cross-hybridizing regions two distinct routes were followed (see Figure 5): *Step 2* or*Step 3*.

Step 2. Design of primer pairs

The selected non cross-hybridizing regions were subjected to *primer3* program [49] with the following parameters: primer length: 20 nt, primer melting temperature range: 56–63°C, and optimum PCR product size: 300–700 bp. *Primer3* was allowed to return up to 10 pairs of primers as candidates (*forward primer* or *reverse primer* designate the oligonucleotide localized near the start or the stop codon of the CDS, respectively). Each candidate primer sequence was first compared to the 16 yeast chromosome sequences using *blastn* program without low-complexity filter and *W* (word size) set to 14. Primers were considered as "unique" if no match of 14 nt was found or if matches > = 14 nt leaves mismatches at the 3'-end. For each CDS, preference was given to "unique" primers, or those exhibiting minimal size match. Finally, the best primer pair for each CDS was selected based on the following preferences: (*i*) an optimal probe length of 500 bp; (*ii*) at least one of the two primers is "unique"; and (*iii*) position of reverse primer is as close as possible to the 3'-end of the open reading frame. In our design, 97.2% of primers were "unique".

Step3. Design of long oligonucleotide probes

For CDS in which the remaining non cross-hybridizing regions are shorter than 300 nt, no PCR probe was designed. Instead, 71-mer synthetic oligonucleotides were synthesized, that met the following criteria: GC content between 34–46 (optimum 39%); absence of monotonous hexanucleotides (TTTTTT, CCCCCC, GGGGGG and AAAAAA); and absence of secondary structures (as determined using the *mfold 2.3* program [44], with temperature set to 68°C).

Oligonucleotide design tool

The *OliD* program, follows the design strategy described above, except that monotonous nucleotide sequences are

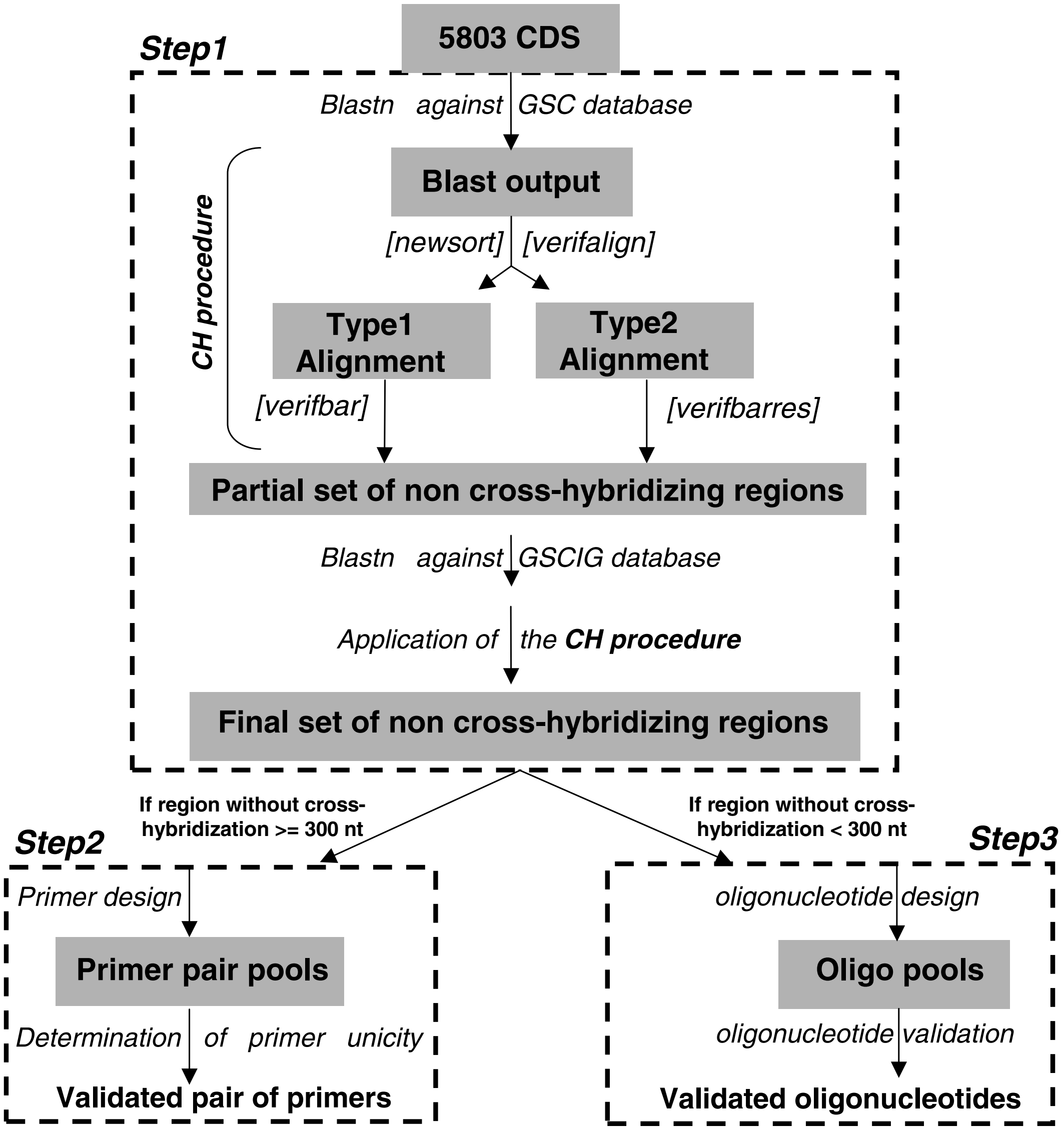
5803 CDS
Step1
Blastn against GSC database
Blast output
CH procedure
[newsort] [verifalign]
Type1 Alignment
Type2 Alignment
[verifbar]
[verifbarres]
Partial set of non cross-hybridizing regions
Blastn against GSCIG database
Application of the CH procedure
Final set of non cross-hybridizing regions
If region without cross-hybridization >= 300 nt
If region without cross-hybridization < 300 nt
Step2
Step3
Primer design
Primer pair pools
Determination of primer unicity
Validated pair of primers
oligonucleotide design
Oligo pools
oligonucleotide validation
Validated oligonucleotides

Figure 5

Schematic representation of the design procedure. The main scripts (written in *perl* and *sh shell* programming languages) used in this work are in brackets []. [*Newsort*] allows the rearrangement of the blast output table in a positional order along the query and subject sequences. [*Verifalign*] detects the different types of alignments (as described in Figure 1). [*Verifbarres*] and [*Verifbar*] allow the detection of potential cross-hybridization regions in Type2 and Type1 alignments, respectively. All scripts are available upon request.

not excluded. The input sequences are individual exon sequences in fasta format. The *OliD* command line allows the user to define several input parameters such as size, GC content, and melting temperature of the designed oligonucleotides; the cross-hybridization parameters (alignment size, percentage of identity and longest contiguous match); the maximal distance from the 3'end of the sequence; the number of oligonucleotides per input sequence and the minimal distance between two adjacent oligonucleotides (if two or more oligonucleotides are selected).

Briefly, the program first compares each input sequence against the genome sequence and predicts all non cross-hybridizing regions within the input sequence according to selected parameters. From these regions, all possible oligonucleotides that fulfill the user-defined parameters are listed. Then, an oligonucleotide score is calculated as the sum of individual score parameters (size, GC content, and Tm), $[W_{size}/|D|_{size} + W_{gc}/|D|_{gc} + W_{Tm}/|D|_{Tm}]$, where W is the user-defined weight of each parameter, and D is the difference between the optimum value defined by the user and the value found for the selected oligonucleotide. $|D|$ is arbitrary set to 1 if $|D| < 1$. If the oligonucleotide is not located within the user-defined position on the entry sequence, a penalty score is calculated as [P(EntryLen – PosMax – PosOligo)] where P is the weight position (-100000 by default); EntryLen, the length of the sequence entry; PosMax (user-defined parameter), the maximal distance from the 3'end of which the oligonucleotide should be designed; and PosOligo, the end position of the selected oligonucleotide. The penalty score is also added to the oligonucleotide score. The list of oligonucleotides is first sorted by their number of cross-hybridizing regions, then by their scores. Finally, the best oligonucleotides are chosen. Sequences and user-defined parameters are automatically transferred to intermediate programs (*blast*, *melting*, *mfold*, etc...) and the results are saved in two output files. The first one is a tabulated spreadsheet table containing the oligonucleotide sequences as well as its position on the entry sequence and the cross-hybridizing regions found. The second one contains rejected sequences with the reasons of rejection. *OliD* is written in python 2.2 (with the use of biopython modules) and runs under all platforms supporting *python*, *blast* and *mfold* programs. *OliD* is available upon request.

Microarray preparation

Primer pairs for each of the PCR products selected as probes were synthesized at 40 nmoles scale using standard cyanoethyl phosphoramidite chemistry method. The forward and reverse primers are 34–35 nt long (20 nt specific and a common 5'-tag sequence: forward primer-tag: 5'-CGACGCCCGCTGATA; reverse primer-tag: 5'-GTCCGGGAGCCATG). Synthesis of the 71-mer oligonucleotides was done using the same chemistry but the 5'-end is modified by the aryl amino-group NH_2- $[CH_2]_6$-. All primers were synthesized by Eurogentec s.a.

PCR amplification reactions were done in two rounds. In the first round, one hundred-microliter reactions were performed using each primer pair with the following reagents: 20 ng of *S. cerevisiae* DNA template, 50 pmoles of each primer, 0.25 units of Taq polymerase (Eurogentec s.a.), 1 × PCR buffer (Eurogentec s.a.), and 1.8 mM $MgCl_2$. Thermalcycling was carried out in Gene Amp PCR system (Applied Biosystems) with 5 min denaturation step at 95°C, followed by 40 cycles of 1 min at 95°C, 30 sec at 55°C, 1 min at 72°C, and a final cycle at 72°C for 10 min. All PCR reactions were analyzed by agarose gel electrophoresis. Reactions that failed to amplify or show multiple bands were repeated using different conditions (e.g. higher annealing temperatures) to favour amplification of the desired product. The second round of PCR amplification follows the same protocol as the first round except for the use of 5 ng of first round PCR products as templates, 50 pmoles each of forward (with an amino-group NH_2-$[CH_2]_{12}$- at the 5'-end of the oligonucleotide) and reverse tag-primers, and the annealing temperature of 45°C. PCR reactions were purified with 96-well Millipore MultiScreen-FB filters (Millipore), then eluted in a final volume of 40 µl $NaPO_4$ 20 mM, pH 8.0, and stored at -20°C until use.

Using the robot SDDC-2 (Engineering Services Inc., ESI), DNAs (PCR products, 71-mer oligonucleotides, and control genes) were deposited in close duplicates (1 fmol DNA of PCR product and 10 fmol DNA of long oligonucleotide) onto microarray aldehyde glass slides with 200-µm spacing between neighbouring spot centers. Subsequent microarray treatments were performed as recommended by the manufacturer (Telechem International). The microarray view is described in Figure 6. Note that batches 800, 400, and 616 (prepared successively in this order from November 2001 to February 2002) used in our hybridization experiments differ from the final ones described above, since the probes corresponding to the genes *YBR145w*, *YDR225w*, *YEL033w*, *YGL147c*, *YGR148c*, *YJR148w*, *YMR011w*, *YMR169c*, *YMR170c*, and *YNL143c* are present in the form of PCR products instead of 71-mer oligonucleotides.

Yeast strains

All yeast strains used in this study were provided by the Euroscarf collection of *S. cerevisiae* deletions http://www.uni-frankfurt.de/fb15/mikro/euroscarf. These includes the parental strain BY4742 (*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*)[50] as well as the corresponding deletants in which each of the following genes was replaced by the KANMX4 cassette: *YBR025c*, *YBR145w*, *YDR225w*,

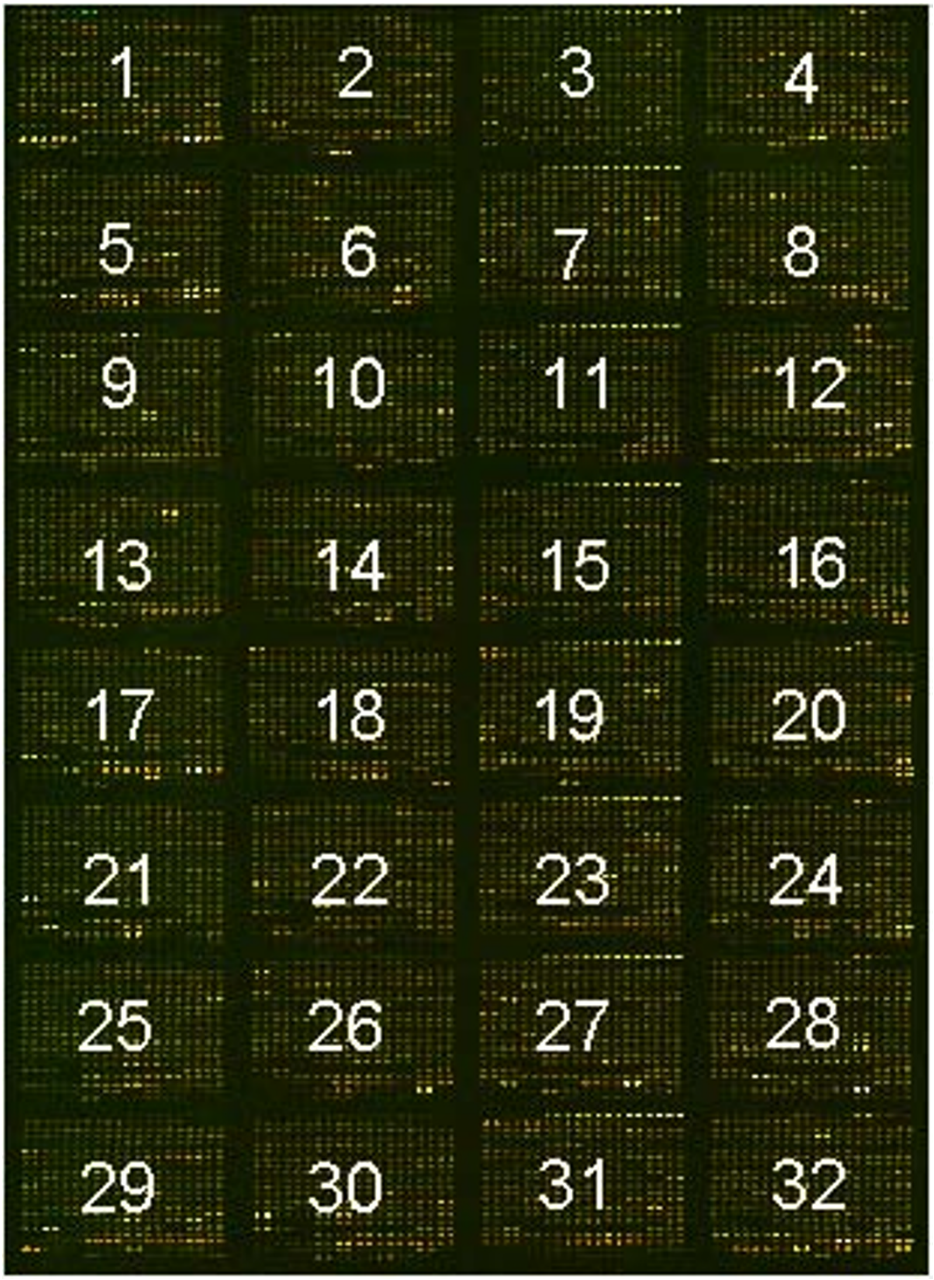

Figure 6
A view of the microarray. The array is composed of 32 grids of 420 spots each. A total of *ca.* 200 empty spots are distributed through the array for background controls. Probe spots are deposited in close duplicates. A set of PCR products and synthetic oligonucleotides was selected as controls. These include scorecard kits (Amersham Biosciences) in grids 4, 8, 12, 16, 20, 24, 28, and 32; serial dilutions of the signal normalization luciferase gene (row 1 in grids 3, 7, 11, 15, 19, 23, 27, 31) for which control RNA spike (Promega) can be obtained; 10 long oligonucleotides covering *YKL182w* (6153 bp) and 9 long oligonucleotides covering *YLR310c* (4767 bp) ORFs as controls for reverse transcription efficiency; 10 intergenic regions, 10 intronic sequences, and mitochondrial genes, 20 non-monotonous trinucleotide repeats (72-mer oligonucleotides) and 4 serial dilutions (in grid 1, 4, 29 and 32) of the total genomic DNA from the wild-type strain S288c; 3 *E. coli* genes (*tuf*A, *ace*F, *kdt*A) as negative controls; the LexA binding domain, the LacZ 5' and 3'end regions, the Pho4 binding domain, the Gal4 binding domain, the GFP, the TAP and GST as commonly used tags or reporter genes; *Leu1*, *His5*, and *Ura4* from *S. pombe*, the human*CBF2* and*c-myc* genes, and the *kan*^{R} gene as heterologous genes and markers. Printing buffer was deposited on the empty spots. The array shown results from hybridization with cDNA targets of total RNA isolated from wild-type BY4742 (Cy3-labelled) and *Δydr225w* (Cy5-labelled) strain.

YDR497c, *YEL033w*, *YGL147c*, *YGR086c*, *YGR148c*, *YJR148w*, *YLR058c*, *YLR109w*, *YML072c*, *YML113w*, *YMR011w*, *YMR169c*, *YMR170c*, *YNL143c*, and *YOL103w*. The correct replacement of the target gene by the KANMX4 cassette in each deletion strain was confirmed by PCR amplification. All the deletant strains have normal growth phenotype.

mRNA preparation, cDNA synthesis and hybridization

For each experiment, freshly grown yeast cells were inoculated into 400-ml glucose-rich liquid medium, grown to mid log phase (OD_{600} = 0.5), and total RNA was purified as described by [51]. Hybridization experiments were performed as described in Figure 4. For each experiment, about 10 μg of total RNA were converted to labelled cDNA targets using an indirect labelling kit (Amersham-Pharmacia-Biotech). All experiments were performed using the same RNA preparation from the wild-type strain BY4742 as reference. The amount of cDNA as well as the incorporation of Cy3 and Cy5 dyes into cDNA targets were quantified by measuring the absorbance of each sample at 260 nm, 550 nm and 650 nm, respectively. For each experiment, labelled-Cy3 and -Cy5 cDNA targets were then combined in equal amount, vacuum dried (Speed-vac centrifugation), and re-solubilized in 50 μl hybridization buffer composed of Dig Easy Hyb solution (Roche diagnostics) with 0.5 mg/ml of salmon sperm DNA. After prehybridization in 50 μl hybridization buffer for 2 h at 42°C, the hybridization was performed with CyDye-labelled cDNA solution at 42°C for approximately 16 h. Following hybridization, slides were washed twice in 0.1 × SSC, 0.1% SDS for 5 min and finally in 0.1 × SSC for 30s. Then, slides were immediately dried by centrifugation (2 min at 400 × g).

Array data processing

Hybridized microarrays were scanned using a Genepix 4000A fluorescence reader (Axon) with a resolution of 10 μm. Signal quantification for each probe on the microarray was performed with Genepix image acquisition software (Axon). For each experiment, intensity signals of the *ca.* 200 empty spots were measured and the 95% upper confidence interval of these figures was used as the lower limit for significant measurements. Intensity signals of duplicated DNA spots were processed independently. Since the dye bias appears to be dependent on grid location in the array, a "within-print-tip-group" normalization method [52] was applied. This procedure corrects the differential rates of incorporation of Cy3 and Cy5 so that the log2-transformed ratio for each array grid is close to 0. Fluorescence ratios were computed based on hybridization signals normalized with background corrections. Ratio below 1.0 was inverted, multiplied by -1 for symmetry to give the flip ratio [FR].

Abbreviations

CDS, coding sequence

TA3 or TA5, transcript abundance for Cy3- or Cy5-labelled cDNA targets, respectively

FR, flip ratio

AS, alignment size

ID, percentage of identity

LC, longest contiguous perfect match segment

cDNA, complementary DNA

Authors' contributions

ET conceived the study and carried out the design procedure, the cross-hybridization experiments and all the programming steps of the *OliD* program, as a postdoctoral fellow in the group of BD. FT participated to the programming steps of the PCR probe design. LB carried out the microarray construction and preliminary hybridization tests to assess quality and reproducibiliy of the yeast microarrays. BD supervised the study and provided valuable input for data analysis and biological interpretations. All authors read and approved the final manuscript.

Additional material

Additional File 1

List of non-coding CDS in Saccharomyces cerevisiae. *CDS were named according to MIPS nomenclature, except 4 coding sequences that were so far disregarded in MIPS databases.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S1.doc]

Additional File 2

List of CDS sharing cross-hybridizing sequences regions with other CDS or with intergenic (IG) sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S2.xls]

Additional File 3

List of designed primers and corresponding PCR products. The table contains all the designed primers (left and right primers) used for the amplification of the PCR products, as well as their features (position along the CDS, the nucleotide sequence, the GC content, the probe length and the CDS cross-hybridization information before and after this study).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S3.xls]

Additional File 4

PCR Probes with forward and reverse primers designed in the 5'- and 3'-UTR regions, respectively.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S4.doc]

Additional File 5

List of oligonucleotide probes. The table contains all the designed oligonucleotides used as probes, as well as their features (position along the CDS, the nucleotide sequence, the GC content, and the CDS cross-hybridization information before and after this study).

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S5.xls]

Additional file 6

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S6.doc]

Additional file 7

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-38-S7.doc]

Acknowledgements

We wish to thank O. Sismero, L. Frangeul and E. Deveaud for technical discussions and help during the microarray hybridizations and the programming steps of *OliD* program; H.W Mewes, M. Muensterkoetter and U. Gueldener for useful discussions about the re-annotation of the yeast genome; D. Marechal, M. Bolotin-Fukuhara and all GARNISH members for comments and useful discussions; and C. Jacq for the critical evaluation of this manuscript. This work was supported by the European Commission under contracts "CYGD" (QLRI-CT1999-01333) and "GARNISH" (QLK3-2000-00174). BD is a member of Institut Universitaire de France.

References

1. Duggan DJ, Bittner M, Chen Y, Meltzer P and Trent JM: **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21:**10-14.
2. Lipshutz RJ, Fodor SP, Gingeras TR and Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21:**20-24.
3. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH and Linsley PS: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat Biotechnol* 2001, **19:**342-347.
4. DeRisi JL, Iyer VR and Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
5. Hauser NC, Vingron M, Scheideler M, Krems B, Hellmuth K, Entian KD and Hoheisel JD: **Transcriptional profiling on all open reading frames of Saccharomyces cerevisiae.** *Yeast* 1998, **14:**1209-1221.
6. yMGV: **Yeast Microarray Global viewer.** 2003 [http://www.transcriptome.ens.fr/ymgv/].
7. SMD: **Stanford Microarray Databases.** 2003 [http://genome-www5.Stanford.EDU/microarray/SMD/index.shtml].
8. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H and Oliver SG: **Life with 6000 genes.** *Science* 1996, **274:**563-567.
9. Harrison PM, Kumar A, Lang N, Snyder M and Gerstein M: **A question of size: the eukaryotic proteome and the problems in defining it.** *Nucleic Acids Res* 2002, **30:**1083-1090.
10. Mackiewicz P, Kowalczuk M, Mackiewicz D, Nowicka A, Dudkiewicz M, Laszkiewicz A, Dudek MR and Cebrat S: **How many protein-coding genes are there in the Saccharomyces cerevisiae genome?** *Yeast* 2002, **19:**619-629.
11. Zhang CT and Wang J: **Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve.** *Nucleic Acids Res* 2000, **28:**2804-2814.
12. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH and Johnston M: **Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11:**1175-1186.
13. Blandin G, Durrens P, Tekaia F, Aigle M, Bolotin-Fukuhara M, Bon E, Casaregola S, de Montigny J, Gaillardin C, Lepingle A, Llorente B, Malpertuy A, Neuveglise C, Ozier-Kalogeropoulos O, Perrin A, Potier S, Souciet J, Talla E, Toffano-Nioche C, Wesolowski-Louvel M, Marck C and Dujon B: **Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited.** *FEBS Lett* 2000, **487:**31-36.
14. Wood V, Rutherford KM, Ivens A, Rajandream M-A and Barrell B: **A Re-annotation of the Saccharomyces cerevisiae Genome.** *Comp Funct Genom* 2001, **2:**143-154.
15. Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, Miller P, Gerstein MB and Snyder M: **An integrated approach for finding overlooked genes in yeast.** *Nat Biotechnol* 2002, **20:**58-63.
16. Oshiro G, Wodicka LM, Washburn MP, Yates JR 3rd, Lockhart DJ and Winzeler EA: **Parallel Identification of New Genes in Saccharomyces cerevisiae.** *Genome Res* 2002, **12:**1210-1220.
17. Eisen MB, Spellman PT, Brown PO and Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.
18. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci U S A* 1999, **96:**2907-2912.
19. Brown CS, Goodwin PC and Sorger PK: **Image metrics in the statistical analysis of DNA microarray data.** *Proc Natl Acad Sci U S A* 2001, **98:**8944-8949.
20. Nielsen HB and Knudsen S: **Avoiding cross hybridization by choosing nonredundant targets on cDNA arrays.** *Bioinformatics* 2002, **18:**321-322.
21. Raddatz G, Dehio M, Meyer TF and Dehio C: **PrimeArray: genome-scale primer design for DNA-microarray construction.** *Bioinformatics* 2001, **17:**98-99.
22. Rouillard JM, Herbert CJ and Zuker M: **OligoArray: genome-scale oligonucleotide design for microarrays.** *Bioinformatics* 2002, **18:**486-487.
23. Wang X and Seed B: **Selection of oligonucleotide probes for protein coding sequences.** *Bioinformatics* 2003, **19:**796-802.
24. Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F and Zollner A: **Overview of the yeast genome.** *Nature* 1997, **387:**7-65.
25. Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, Ballesta JP, Banrevi A, Bolle PA, Bolotin-Fukuhara M and Bossier P *et al.*: **Complete DNA sequence of yeast chromosome XI.** *Nature* 1994, **369:**371-378.
26. Sharp PM and Li WH: **The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.
27. Bussey H, Kaback DB, Zhong W, Vo DT, Clark MW, Fortin N, Hall J, Ouellette BF, Keng T and Barton AB *et al.*: **The nucleotide sequence of chromosome I from Saccharomyces cerevisiae.** *Proc Natl Acad Sci U S A* 1995, **92:**3809-3813.
28. Johnston M, Andrews S, Brinkman R, Cooper J, Ding H, Dover J, Du Z, Favello A, Fulton L and Gattung S *et al.*: **Complete nucleotide sequence of Saccharomyces cerevisiae chromosome VIII.** *Science* 1994, **265:**2077-2082.

29. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE and Davis RW: **Discovery and analysis of inflammatory disease-related genes using cDNA microarrays.** *Proc Natl Acad Sci U S A* 1997, **94:**2150-2155.
30. Bolton ET and McCarthy BJ: **A general method for the isolation of RNA complementary to DNA.** *Proc. Natl. Acad. Sci* 1962, **48:**1390.
31. Richmond CS, Glasner JD, Mau R, Jin H and Blattner FR: **Genome-wide expression profiling in Escherichia coli K-12.** *Nucleic Acids Res* 1999, **27:**3821-3835.
32. Girke T, Todd J, Ruuska S, White J, Benning C and Ohlrogge J: **Microarray analysis of developing Arabidopsis seeds.** *Plant Physiol* 2000, **124:**1570-1581.
33. Bisson LF, Coons DM, Kruckeberg AL and Lewis DA: **Yeast sugar transporters.** *Crit Rev Biochem Mol Biol* 1993, **28:**259-308.
34. DeRisi J, van den Hazel B, Marc P, Balzi E, Brown P, Jacq C and Goffeau A: **Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants.** *FEBS Lett* 2000, **470:**156-160.
35. Le Novere N: **MELTING, computing the melting temperature of nucleic acid duplex.** *Bioinformatics* 2001, **17:**1226-1227.
36. Malpertuy A, Tekaia F, Casaregola S, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P, de Montigny J, Durrens P, Gaillardin C, Lepingle A, Llorente B, Neuveglise C, Ozier-Kalogeropoulos O, Potier S, Saurin W, Toffano-Nioche C, Wesolowski-Louvel M, Wincker P, Weissenbach J, Souciet J and Dujon B: **Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes.** *FEBS Lett* 2000, **487:**113-121.
37. Harrison P, Kumar A, Lan N, Echols N, Snyder M and Gerstein M: **A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution.** *J Mol Biol* 2002, **316:**409-419.
38. Fischer G, Neuveglise C, Durrens P, Gaillardin C and Dujon B: **Evolution of gene order in the genomes of two related yeast species.** *Genome Res* 2001, **11:**2009-2019.
39. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD and Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.** *Nucleic Acids Res* 2000, **28:**4552-4557.
40. Wolfe KH and Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387:**708-713.
41. Jornvall H, Persson B and Jeffery J: **Characteristics of alcohol/polyol dehydrogenases. The zinc-containing long-chain alcohol dehydrogenases.** *Eur J Biochem* 1987, **167:**195-201.
42. van Helden J, del Olmo M and Perez-Ortin JE: **Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals.** *Nucleic Acids Res* 2000, **28:**1000-1010.
43. Dujon B: **The yeast genome project: what did we learn?** *Trends Genet* 1996, **12:**263-270.
44. Walter AE, Turner DH, Kim J, Lyttle MH, Muller P, Mathews DH and Zuker M: **Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding.** *Proc Natl Acad Sci U S A* 1994, **91:**9218-9222.
45. Lopez PJ and Seraphin B: **YIDB: the Yeast Intron DataBase.** *Nucleic Acids Res* 2000, **28:**85-86.
46. Grate L and Ares M Jr: **Searching yeast intron data at Ares lab Web site.** *Methods Enzymol* 2002, **350:**380-392.
47. Richard GF and Dujon B: **Trinucleotide repeats in yeast.** *Res Microbiol* 1997, **148:**731-744.
48. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
49. Rozen S and Skaletsky HJ: **Primer3.** 1998 [http://www-genome.wi.mit.edu/genome_software/other/primer3.html].
50. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P and Boeke JD: **Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications.** *Yeast* 1998, **14:**115-132.
51. Ausubel F, Brent R, Kingston E, Moore D, Seidman J, Smith J and Struhl K eds: *Current protocols in molecular biology* 1993.
52. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.