Molecular Systems Biology 5; Article number 274; doi:10.1038/msb.2009.31
Citation: ***Molecular Systems Biology*** **5:274**
© 2009 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/09
www.molecularsystemsbiology.com

REPORT

Genome-wide allele- and strand-specific expression profiling

Julien Gagneur[1], Himanshu Sinha[1], Fabiana Perocchi[1], Richard Bourgon[2], Wolfgang Huber[2] and Lars M Steinmetz[1,]*

[1] Gene Expression Unit, European Molecular Biology Laboratory, Heidelberg, Germany and [2] European Bioinformatics Institute, Cambridge, UK
* Corresponding author. European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69117, Germany. Tel.: + 49 6221 387 83 89; Fax: + 49 6221 387 85 18; E-mail: larsms@embl.de

Received 4.2.09; accepted 27.4.09

Recent reports have shown that most of the genome is transcribed and that transcription frequently occurs concurrently on both DNA strands. In diploid genomes, the expression level of each allele conditions the degree to which sequence polymorphisms affect the phenotype. It is thus essential to quantify expression in an allele- and strand-specific manner. Using a custom-designed tiling array and a new computational approach, we piloted measuring allele- and strand-specific expression in yeast. Confident quantitative estimates of allele-specific expression were obtained for about half of the coding and non-coding transcripts of a heterozygous yeast strain, of which 371 transcripts (13%) showed significant allelic differential expression greater than 1.5-fold. The data revealed complex allelic differential expression on opposite strands. Furthermore, combining allele-specific expression with linkage mapping enabled identifying allelic variants that act in *cis* and in *trans* to regulate allelic expression in the heterozygous strain. Our results provide the first high-resolution analysis of differential expression on all four strands of an eukaryotic genome.

Molecular Systems Biology **5**: 274; published online 16 June 2009; doi:10.1038/msb.2009.31
Subject Categories: functional genomics; computational methods
Keywords: allele-specific expression; linkage mapping; microarray analysis; phosphate metabolism; strand-specific expression

Introduction

Genetic variation is the basis of phenotypic variation, and the degree to which this variation is transcribed conditions its impact on phenotype. Up to 90% of the genome of eukaryotic organisms is transcribed (Carninci *et al*, 2005; David *et al*, 2006; Manak *et al*, 2006; Birney *et al*, 2007). Therefore, a significant portion of genetic variation, including that in non-coding sequences, is represented in transcripts. In humans, allelic differential expression (ADE) has been estimated to affect 20–50% of genes (Yan *et al*, 2002; Bray *et al*, 2003; Lo *et al*, 2003; Serre *et al*, 2008). In addition to affecting phenotypic variation, ADE is involved in gene-dosage compensation of sex chromosomes, and imprinting on autosomes (Knight, 2004). Monoallelic expression with random choice between paternal and maternal alleles has also been shown to affect hundreds of autosomal genes and thus to contribute to individual cell variability (Gimelbrant *et al*, 2007).

Recent reports have shown that transcription frequently occurs on both DNA strands (Carninci *et al*, 2005; Katayama *et al*, 2005; David *et al*, 2006; Engström *et al*, 2006). However, so far, genome-wide assessment of allele-specific expression has been carried out using single-nucleotide polymorphism (SNP) arrays (Lo *et al*, 2003; Pant, 2006; Bjornsson *et al*, 2008) and reference genome ORF arrays (Ronald *et al*, 2005a). These studies have not targeted unannotated elements of the genome, nor assessed expression on opposite strands of the same chromosomal position. Therefore, despite the importance of allele-specific expression, the extent of ADE on opposite strands or for non-coding sequences has remained largely unaddressed.

Results and discussion

One tiling array for two genomes

To profile genome-wide allele-specific expression, we designed a high-resolution yeast tiling microarray (David *et al*, 2006; Mancera *et al*, 2008) (Figure 1A) that covers both strands of the genomes of both the laboratory strain S288c (S strain) (Goffeau *et al*, 1996) and the clinical isolate YJM789 (Y strain) (Wei *et al*, 2007). This array allows simultaneous expression

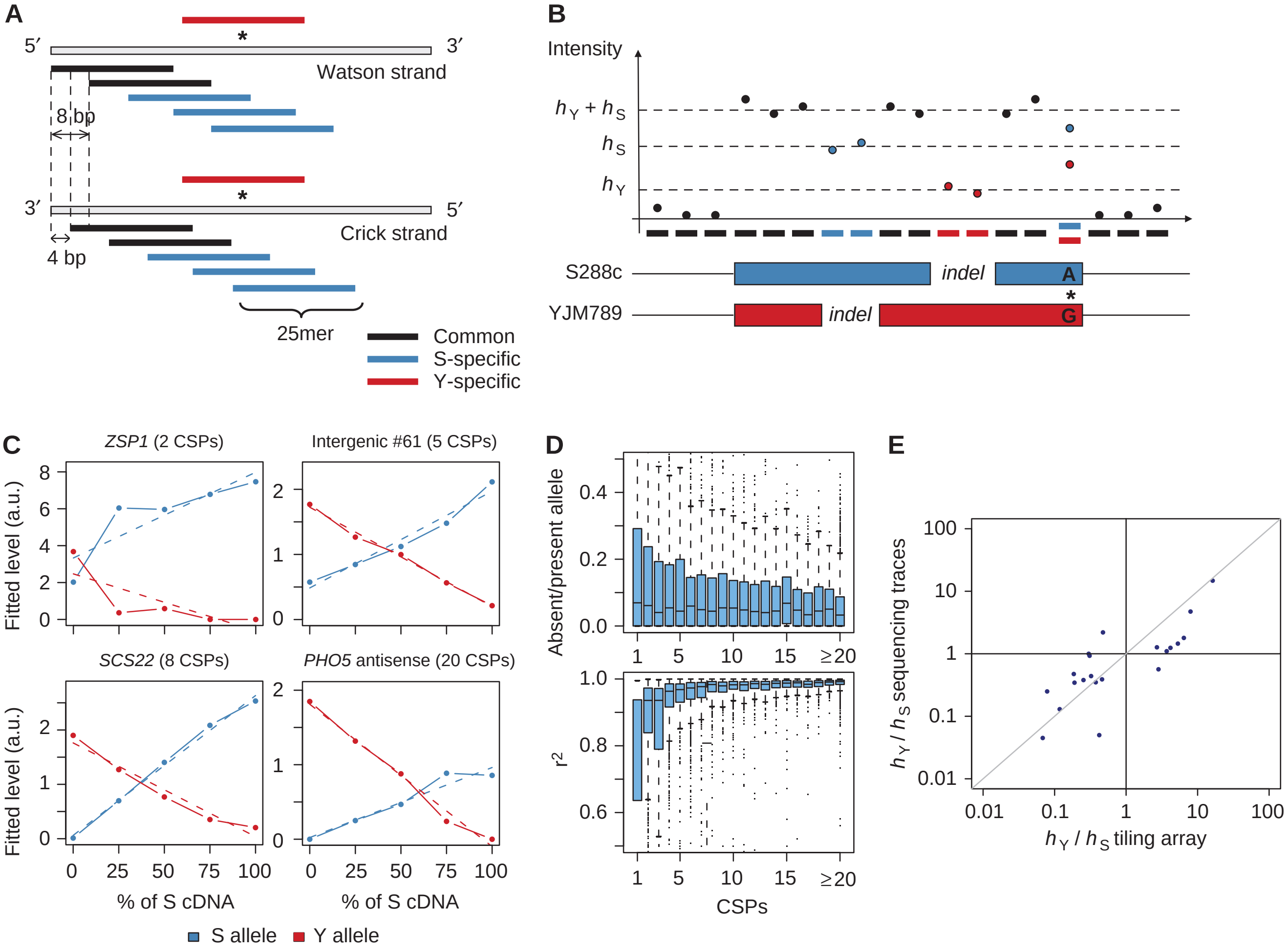

Figure 1 Measuring allelic expression on a tiling array. (**A**) The array contains 25-mer probes (black and blue) that tile both strands of the genome of S288c with a probe offset of 8 bp and a 4-bp shift between the two strands. The array also contains probes (red) complementary to the YJM789 sequence for polymorphic regions, as shown here for a SNP marked by an asterisk. (**B**) Modeling the hybridization intensity. Consider a two-allelic transcript with two indels and one SNP as shown in the lower part. The S allele is at expression level h_S and the Y allele at h_Y. Hybridization intensities of the common probes are ideally expected to be proportional to the sum of the two expression levels. Intensities measured for the probes specific to the S or Y alleles are expected to be proportional to their expression levels, h_S or h_Y, respectively. Owing to cross-hybridization, probes with sequence highly similar to the other allele yield higher intensities (shown here for a SNP). These properties are modeled in equation (1) (Materials and methods). (**C**) Inferred expression level of transcripts in the mixture series. The circles show inferred expression levels for the S allele (blue) and the Y allele (red). Dotted lines mark linear regression. The quality, in terms of both linear behavior and monoallelic calls, improves when moving from *ZSP1* with only two centered specific probes (CSPs) to the antisense of *PHO5* with 20 CSPs. (**D**) Monoallelic calls and linearity of the method. Boxplots of the ratios of inferred expression level of the absent allele over the present allele as a function of the number CSPs (top). In these parental samples, the true value is known to be 0 and the ratio is expected to tend to 0 with increasing CSPs. Boxplots of the r^2 coefficient of the linear fit for expressed alleles as a function of the number CSPs (bottom). Perfect linearity should give r^2 of 1. (**E**) Comparison of allelic expression ratios from tiling array and sequencing traces. For 21 transcripts (see supplementary table VII), allelic expression ratios inferred from tiling array analysis (*X*-axis, log scale) plotted against allelic expression ratios inferred from sequencing traces (*Y*-axis, log scale). The *y*=*x* line (gray) is provided as a reference.

profiling of allelic variants in a heterozygous hybrid strain (designated as Y/S) for coding and non-coding transcripts and in a strand-specific manner. The array tiles both strands of the S genome using 25-mer oligonucleotide probes with 8-bp offset and includes probes matching strain Y at positions of polymorphisms (Figure 1A). Out of the 2.8 million perfect match probes on the array, 86% are common to both genomes, whereas 10 and 4% are specific to S and Y strains, respectively, at insertions, deletions or single-nucleotide polymorphisms.

We hybridized cDNA from the heterozygous Y/S and from the homozygous S and Y strains grown in rich media (YPD). Strand specificity during sample preparation was maintained by inclusion of actinomycin D during reverse transcription to prevent spurious synthesis of second-strand cDNA (Perocchi *et al*, 2007). A segmentation algorithm (Huber *et al*, 2006) was applied to identify transcripts expressed in any of the three strains. In addition to annotated transcripts (e.g., coding genes, tRNAs, snoRNAs), we identified 359 unannotated transcripts, i.e., they do not match any current feature in the SGD database (http://www.yeastgenome.org). Most, if not all, of these unannotated transcripts, are probably non-coding (David *et al*, 2006; Xu *et al*, 2009). The unannotated transcripts consisted of 163 intergenic transcripts and 196 transcripts, which were overlapping annotated genes in antisense orientation. Out of these, 21 intergenic and 21 antisense transcripts were expressed in strain S and not in Y, while 16 intergenic and

35 antisense transcripts were expressed in Y and not in S. Only three unannotated transcripts were specific to the hybrid. Thus, most of the unannotated transcripts (64%) seem to be expressed in both of the evolutionarily distant S and Y strains (Wei *et al*, 2007), suggesting that, despite their low conservation at the sequence level (David *et al*, 2006), the transcription of these unannotated sequences is conserved.

Quantitative estimation of allele-specific expression

Accurate estimation of allele-specific expression was achieved by using both specific and common probes, with the intensities of the latter reflecting the total expression of the two alleles (Figure 1B). One main challenge was accounting for off-target effects. Part of contribution toward hybridization signal of allele-specific probes comes from their cross-hybridization with transcripts of the other allele (Figure 1B). Indeed, in most cases, allele-specific probes have only one nucleotide mismatch with the other allele and show significant hybridization with it. Not accounting for this effect would lead to biased estimation of allele-specific expression levels. This off-target effect was accounted for by modeling the probe intensities as noisy observations of weighted sums of the two allelic levels (equation (1)). The weights represent the affinities of the probe with respect to each allele. They are equal for common probes and can differ for specific probes, none being *a priori* negligible. Hybridizations of genomic DNA yielded estimates of relative affinities by providing a nominally uniform concentration along the genome (David *et al*, 2006). Allele expression levels and probe affinities in our non-linear, heteroscedastic model were inferred using iterative weighted least squares (see Materials and methods). Confidence intervals were obtained by bootstrap re-sampling of the residuals.

Although our tiling array targets both genomes, this is not a prerequisite for the algorithm. The method can incorporate heterozygous genomic DNA if available. It also works for experimental designs that produce cDNA samples from heterozygous strains only or in combinations with homozygous cDNA samples (Supplementary information). Our R package, allelicTxn (available at http://steinmetzlab.embl.de/allelic and in Supplementary information), supports these extensions.

To validate our method, we hybridized cDNA mixtures from homozygous S and Y strains in varying proportions: 0:1, 1:3, 1:1, 3:1 and 1:0. The method was expected to first correctly report monoallelic expression in the 0:1 and 1:0 cDNA samples, and second, to estimate the expression level of each allele in linear relationship with its dilution ratio. As the number of centered specific probes (CSP, probes which interrogate polymorphisms within ±4 bp of their central base, see Materials and methods) per allele increased, the accuracy of monoallelic calls as well as the linearity of the relation between inferred and actual log ratios improved (Figure 1C). When the algorithm was run on the 0:1 cDNA samples, an expression ratio close to 0 (<0.15) could be inferred for more than 83% of the 5404 expressed alleles with at least eight CSPs (Figure 1D, upper panel). In addition, the inferred cDNA levels showed an accurate linear relationship with dilution ratios (linear regression, $r^2>0.90$) for more than 96% of these 5404 alleles (Figure 1D, lower panel).

Furthermore, the sensitivity of the method at different degrees of ADE was evaluated. We considered transcripts with both alleles expressed and with eight CSPs or more. For 81% of the transcripts with more than two-fold difference in expression between the parental strains (142 out of 176), and for 51% (289 out of 570) of transcripts with more than 1.5-fold difference, significant ADE was detected in the 1:1 mixture when using a *P*-value threshold of 0.01. Altogether, these results show that the method can accurately measure allele-specific expression quantitatively for transcripts with a sufficient number of CSPs and detect imbalanced allelic expression levels down to 1.5-fold at a sensitivity of 51%.

Transcriptome profile on all four strands

Applying our method to the three biological replicates of the Y/S heterozygous hybrid, we obtained allele- and strand-specific expression estimates for 5069 transcripts with at least one specific probe (Figure 2A), of which we considered 2914 (57%) to be confident because they had at least eight CSPs. Allelic expression levels for all transcripts including significance estimates for differential expression between alleles are provided in Supplementary Table I and on our website http://steinmetzlab.embl.de/allelic. In total, 454 transcripts showed significant ADE at a false discovery rate (FDR) of 0.05. Among them, 44 transcripts were unannotated (19 antisense and 25 intergenic transcripts). Overall, 371 (82%) of the 454 transcripts showed at least a 1.5-fold difference in allelic expression (Supplementary Figure S1).

For experimental validation, 24 transcripts with significant ADE (FDR <0.05) were selected (15 ORFs, three antisense and six intergenic transcripts). These transcripts spanned a range of expression levels (from the 4th to 82nd percentiles of expressed transcripts) and allelic expression ratios (from 2.1 to 16 fold). We used a method based on sequencing (Ge *et al*, 2005), which estimated allelic expression ratios from relative peak intensities at SNP positions in cDNA-sequence traces. This method yielded informative data for 21 transcripts (see Materials and methods). Allelic ratios inferred by sequencing agreed well over the range of tested ratios with the array-based estimates (Figure 1E, Pearson's correlation=0.8, P-value$=3.1\times10^{-5}$).

Having accurate measurements for allele-specific expression of transcripts on each strand, we compared ADE between strands to identify instances of complex expression regulation. Transcription on opposite strands can mediate regulatory interactions (Hongay *et al*, 2006; Camblong *et al*, 2007; Uhler *et al*, 2007). Expression analysis in the hybrid strain showed 196 pairs of expressed transcripts overlapping on opposite strands (sense–antisense pairs, Supplementary Table II). Out of these, 83 pairs contained 8 CSPs or more for both transcripts and yielded confident estimates of transcript abundance. Among them, 36 showed significant ADE (FDR <0.05, fold change >1.5) for either the sense or the antisense transcript, and two pairs showed significant ADE for both (Figure 2B). The first, *FET4*, has a symmetric allele-specific expression pattern: the Y-alleles for both the antisense and the sense are

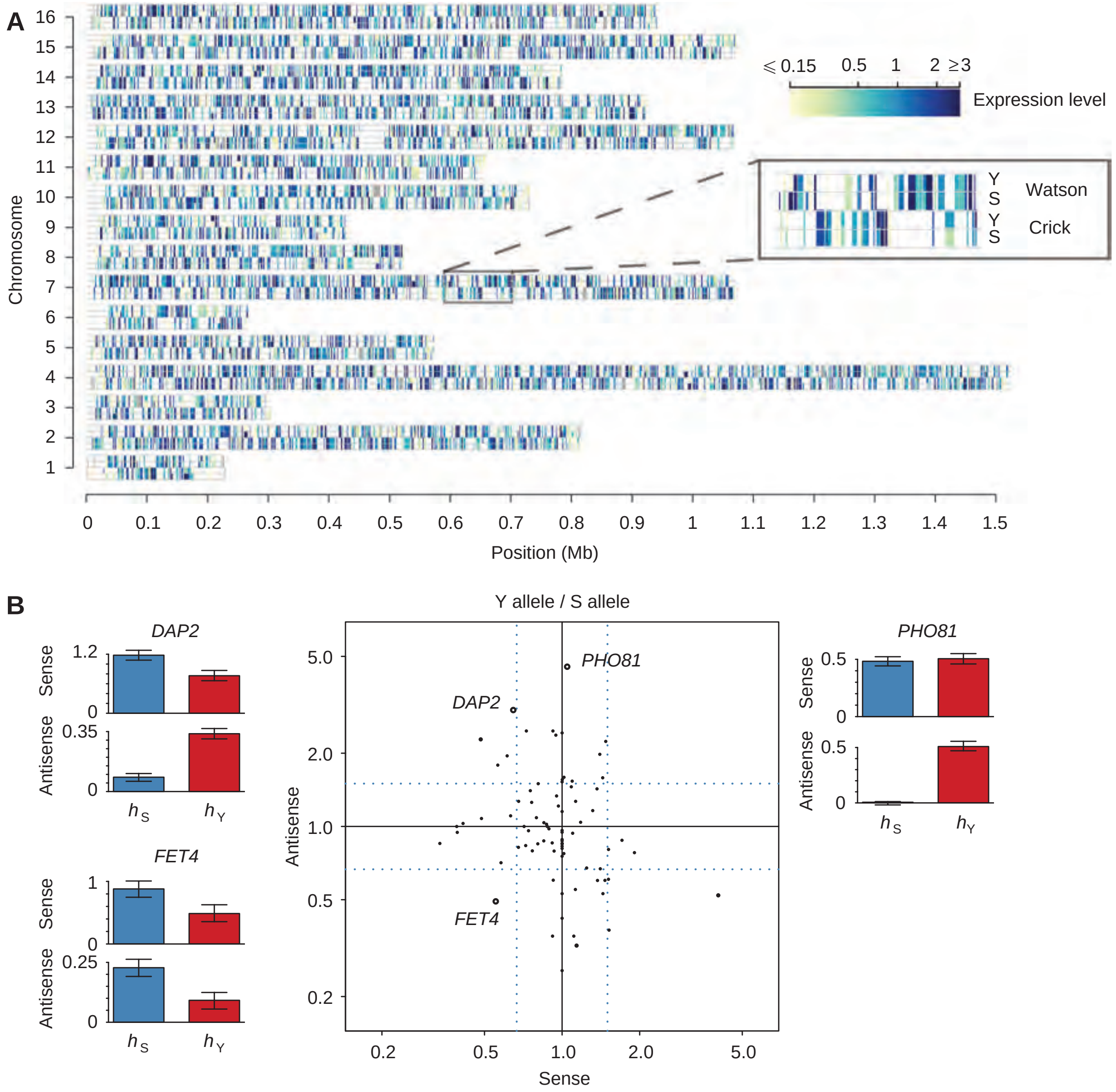

Figure 2 Expression profiles across four strands of a diploid genome. (**A**) Expression levels of all transcripts are shown as colored rectangles positioned with their coordinates on either of the four strands (Y or S, Watson or Crick). One region on chromosome VII is enlarged in the inset. Data shown in this figure are available in supplementary table I. (**B**) Allele-specific expression of sense–antisense transcript pairs. Scatter plot of allelic expression ratios (center panel) for sense (*X*-axis, log scale) versus antisense (*Y*-axis, log scale). Dotted blue lines show 1.5-fold expression differences. Pairs mentioned in the text are labeled and highlighted (bold dots). Allelic expression measurements of three sense–antisense pairs (bar plots) show instances of significant ADE (FDR <0.05) for an anti-correlated pair (*DAP2*), a correlated pair, (*FET4*), and a pair with strong antisense ADE but no difference in sense expression levels (*PHO81*).

less expressed than the S-alleles. The second, *DAP2*, has an anti-symmetric allele-specific expression pattern: one of the two homologous chromosomes expresses strongly the sense transcript and weakly the antisense transcript, whereas the other chromosome shows the opposite pattern. One pair, *PHO81*, showed significant ADE for the antisense transcript whereas the sense transcript showed no strong ADE (95% CI of S-allele level: [0.44, 0.52], and Y-allele: [0.46, 0.55]) (Figure 2B).

Hence, our strand-specific method allows assessing allele-specific expression for transcripts overlapping one another on opposite strands. As such sense–antisense pairs can show asymmetric expression patterns (e.g., one expressed and the other not), the two distinct expression levels would have been confounded if strand specificity had not been taken into account. As most earlier approaches have confounded strandedness, either intentionally through the preparation of double-stranded cDNA or unintentionally through sample preparation artifacts (Perocchi *et al*, 2007), such confounding is a property of existing microarray datasets and is a limitation for their interpretation.

ADE correlates with polymorphism density in promoters

ADE is a consequence of *cis*-regulatory variation, which by definition, acts on the allele of the same chromosome (Knight, 2006; Rockman and Kruglyak, 2006). In yeast, local regulatory polymorphisms have been shown to predominantly consist of *cis*-regulatory polymorphisms and to be enriched in promoter and 3′-UTR regions of transcripts (Ronald *et al*, 2005b). Although a single functional polymorphism might suffice to affect the regulation of a transcript, the higher the density of polymorphisms in a region, the more likely it is that one or more of them have a regulatory impact. To determine whether ADE depends on sequence variation within promoters, we tested the association between ADE and polymorphism density in promoter regions. We measured the degree of differential expression between the two alleles of a transcript by using the ADE coefficient, which ranges between 0 and 1 (Materials and methods). A value of 0 for the coefficient indicates no ADE, whereas 1 indicates monoallelic expression. Across the 2914 confident transcripts, ADE significantly correlated with polymorphism density in promoter regions (defined as the 500-bp interval upstream from the transcription start site) (Kendall's tau test, P-value$=4\times10^{-5}$, Supplementary Figure S2).

A striking example of a region with high ADE lies on chromosome I covering the *DUP240* gene family, which is one of the most polymorphic regions between S288c and YJM789 (Wei *et al*, 2007). Without exception, all *DUP240* genes in this region (*UIP3*, *YAR028W*, *YAR029W*, *PRM9* and *MST28*) showed significant ADE (FDR <0.05, Supplementary Figure S2, inset). These data indicate that sequence variation in the promoter regions is probably a strong contributor to ADE.

Dissecting *cis*- and *trans*-regulatory variations

As opposed to *cis*-regulatory variants, which act on the allele of the same chromosome, *trans*-regulatory variants act on both alleles. The relative contribution of *cis*- versus *trans*-regulation can be assessed by comparing ADE in a hybrid strain to the gene-level differential expression between the homozygous parents (Wittkopp *et al*, 2004). It can be measured as the ratio of *cis*-regulatory divergence to the total regulatory divergence (Materials and methods and Wittkopp *et al*, 2008). Among the 455 transcripts with at least 1.5-fold expression difference between the S and the Y strains and confident ADE estimates, 205 were classified as mainly *trans* (proportion of *cis* effects $<1/3$) and 144 as mainly *cis* (proportion of *cis*-effects $>2/3$), with a median proportion of cis effects being 0.40 (Supplementary Table I and Supplementary Figure S3). Hence, we observed a slight preponderance for *trans*-regulatory effects, similar to a study of 40 differentially expressed genes between BY4741 (an S288c descendant) and RM11-1a in which *trans*-regulation was also reported to have a major contribution (Wang *et al*, 2007).

Although *cis*-acting variants are mostly gene specific, *trans*-acting differences probably affect the level of several downstream genes. To identify which of the transcriptional programs are under the control of *trans*-regulatory variants in the hybrid, we considered transcription factor (TF) target sets and tested them for enrichment in genes differentially expressed between S and Y (FDR <0.05, fold change >1.5) removing the 144 transcripts whose differential expression is mainly attributed to *cis*-effects. Using a comprehensive regulatory network integrating ChIP-chip data and TF binding site predictions (MacIsaac *et al*, 2006), target sets for 17 TFs showed significant enrichment (Fisher's exact test, FDR <0.05, Supplementary Table III). Notably, failing to remove the 144 transcripts with mainly *cis*-effects leads to lower significance levels and to a smaller number of TFs identified (13 instead of 17). Thus, taking allele-specific expression into account increases the power of this analysis. One of the TFs identified was Hap1, an activator of nuclear-encoded mitochondrial genes that is known to be defective in S288c (Gaisne *et al*, 1999); another was Pho4, an activator of the PHO pathway (Oshima, 1997). Analysis of an extensive list of PHO-pathway genes (Supplementary Table IV) showed differential expression between the parental strains for 15 out of 32 genes of this pathway (FDR <0.05, fold change >1.5). All 15 genes, except for the low-affinity phosphate transporter *PHO87*, are highly expressed in S strain and expressed at low levels in both Y and the hybrid strains, reflecting that the PHO pathway is upregulated in the S strain and down-regulated in Y and the hybrid strains.

PHO84-Y allele dominantly represses the PHO pathway in rich media

To identify *trans*-acting factors causative for the differential expression of the PHO pathway, we carried out linkage mapping using a collection of 184 meiotic Y/S segregants genotyped at 55 987 markers (Mancera *et al*, 2008). Resistance to arsenate, a toxic analog of phosphate, was used to assay PHO-pathway activity (Wykoff *et al*, 2007). Profiling of the segregants showed a Mendelian segregation of the arsenate-resistance phenotype (Supplementary Table V). The relative risk factor peaks at a distinct genomic location, centered on *PHO84* (Figure 3A). Two reciprocal hemizygous strains in the hybrid background (Steinmetz *et al*, 2002) were then constructed, in which either the S allele or the Y allele of *PHO84* was deleted. The hemizygous strains carrying only the S allele of *PHO84* recapitulated the S phenotype—being resistant to arsenate (Supplementary Figure S4) and showing high expression of the PHO pathway (Figure 3B)—while the strain carrying only the Y allele of *PHO84* showed the phenotype of the Y and hybrid strains. These results confirm that *PHO84* is the causative *trans*-acting factor, which exerts its effect on both S and Y alleles in the hybrid. *PHO84* encodes a high-affinity inorganic phosphate transporter in the plasma membrane. The non-conservative amino-acid substitution at position 259 from leucine in S to proline in Y (a common variant) has been linked to polychlorinated phenol resistance and is probably essential for protein function (Perlstein *et al*, 2007). This interpretation is consistent with arsenate resistance in the S strains, as these cells would be deficient in arsenate uptake. Moreover, the PHO pathway is upregulated in *PHO84* knockout strains (Wykoff *et al*, 2007) because of positive feedback. Hence, a loss-of-function *PHO84-S* allele also explains the high expression of the PHO-pathway genes.

Conclusion

By assessing expression using a tiling array that contains probes targeting polymorphisms, we were able to estimate

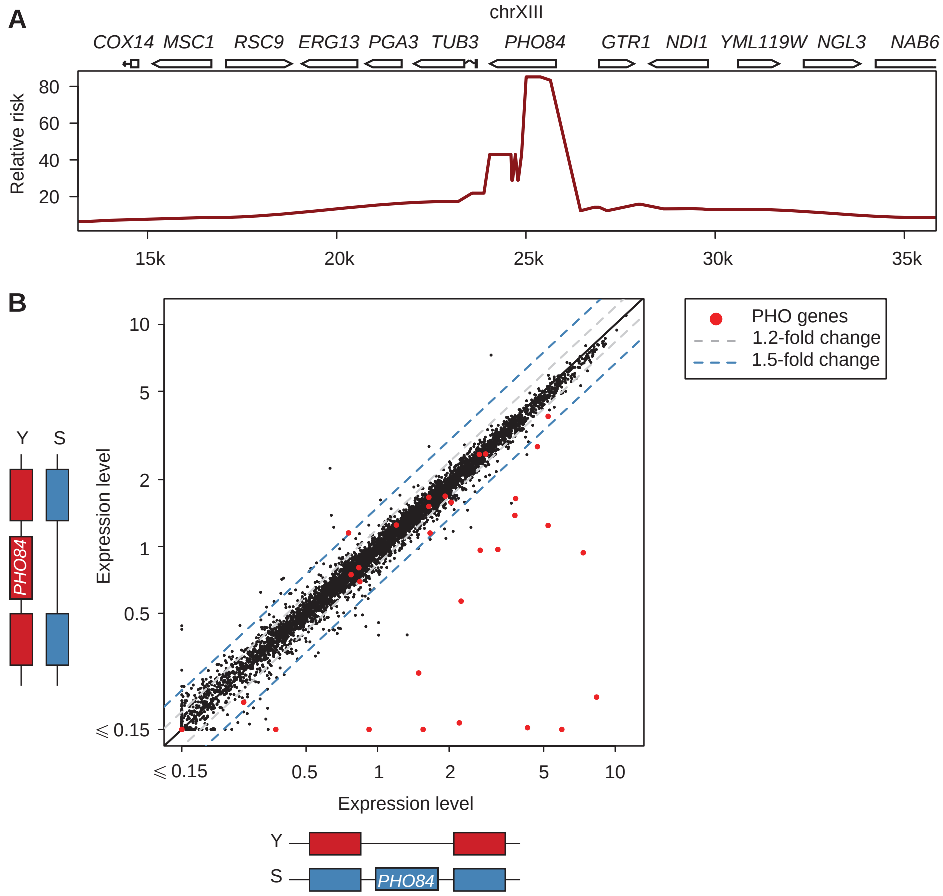

Figure 3 The genetic basis of *trans*-regulation of the PHO pathway in the hybrid. (**A**) Linkage mapping. Relative risk of arsenate resistance for segregants carrying the S allele compared with the Y allele, plotted for markers across a 20-kb region on chromosome XIII around *PHO84*. (**B**) Scatter plot of transcript expression levels for the hybrid strain carrying the *PHO84-S* allele only (*X*-axis, log scale) versus the hybrid strain carrying the *PHO84-Y* allele only (*Y*-axis, log scale). Dotted lines show fold differences of 1.2 (gray) and 1.5 (blue). The majority of the PHO pathway genes (red dots) show differential expression between the two strains showing that the Y allele acts dominantly in the hybrid to repress the PHO pathway.

genome-wide allele-specific and strand-specific expression variation. We have shown using the PHO pathway as an example that integrating allele-specific expression with linkage mapping enables dissecting the genetic variants that act in *cis* and *trans* to regulate allelic expression in a diploid organism. Our computational method is versatile and can be applied to other microarray designs, as long as they contain probes overlapping polymorphic positions of transcripts. Importantly, our method is strand-specific and allows assessing allele-specific expression for transcripts overlapping one another on both strands. As such sense–antisense pairs can show asymmetric expression patterns (e.g., one expressed and the other not), the two distinct expression levels would have been confounded if strand specificity had not been taken into account. Altogether, our data show the importance of assessing transcription on all four strands of a diploid genome. As expression analysis by new sequencing technologies becomes more routine and less expensive, we expect that this expanded view of transcription will become increasingly common.

Materials and methods

Genome sequence and annotation

Sequence and feature files (.gff files) for the S288c genome were obtained from the *Saccharomyces* Genome Database (http://www.yeastgenome.org) on 7 March 2007. The sequence for YJM789 was obtained from Wei *et al* (2007) and aligned to the S288c genome using the procedure described by Wei *et al* (2007).

Microarray data

Microarray data are available at ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/). The cDNA hybridizations are available under accession number E-TABM-569 and the array design is available under A-AFFY-116. We have also used genomic DNA hybridizations from Mancera *et al* (2008) (accession number E-TABM-470). See Supplementary information for details.

Array design

We designed a custom Affymetrix tiling array (product no. 520055) with a total of ~6.5 million probes (25-mers) including perfect match and mismatch probes. The probes tile both strands of the S288c genome at a resolution of 8 bp, with a shift between the strands of 4 bp

(David *et al*, 2006). The array also includes ∼106 000 probes complementary to the YJM789 genome (Wei *et al*, 2007) at positions of polymorphism between the strains. We also added 10 647 negative-control probes of randomly generated sequences with GC content ranging from 2 to 25 GCs.

Yeast strains and sample preparation

Laboratory and clinically derived *S. cerevisiae* strains used in this work were isogenic to S288c and YJM789 and were designated as 'S' and 'Y', respectively. Three independent heterozygous hybrid strains (designated as 'Y/S') were obtained by crossing Y and S strains. Reciprocal hemizygote strains for *PHO84* alleles were constructed by crossing relevant Y and S background strains. Supplementary Table VI lists all strains used in this study.

Total RNA was extracted from yeast cultures grown at 30°C in YPD medium (2% peptone, 1% yeast extract and 2% dextrose) and processed for array hybridizations as described earlier (Perocchi *et al*, 2007). Importantly, to remove reverse transcription artifacts, first-strand cDNA was synthesized in the presence of 6.25 μg/ml actinomycin D. As cDNA is chemically same as DNA, we did not expect any systematic differences between cDNA and genomic DNA labeling.

For making mixture series, cDNA from S and Y strains was mixed in the following proportions, according to mass: 0:1, 1:3, 1:1, 3:1 and 1:0.

Probe filtering and classification

Using the following procedure, we classified each probe as common, S-specific, Y-specific or control. Ungapped alignments of the probes to the S288c genome and the aligned portion of the YJM789 genome were produced using the software exonerate (Slater and Birney, 2005). We considered all perfect matches and near matches (up to two mismatches). A common probe has a unique perfect match to both parental genomes at the same alignment position and no near match. An S-specific probe has a unique perfect match and no further near matches to the S288c genome. It has no perfect match to the YJM789 genome and no near match to the YJM789 genome, except possibly at the same aligned position as its perfect match position in S288c. Y-specific probes were defined analogously. Specific probes whose match overlaps a polymorphism at ±4 bp of its central base were called 'centered specific probes (CSP)'. Finally, we ensured that each negative control probe had neither a perfect nor a near match in either genome.

Normalization and background subtraction

Calibration of intensities between arrays was done using a variant of quantile normalization (Bolstad *et al*, 2003), as follows. The sets of cDNA and genomic DNA (gDNA) hybridizations were treated separately. As specific probes are expected to have different behavior depending on the strain, we restricted the quantile normalization to the set of common probes and used linear interpolation to normalize the intensities of the specific probes.

The background of cDNA hybridizations was subtracted as described earlier (Huber *et al*, 2006). Briefly, probes were binned into 10 groups according to their intensity level in the gDNA hybridizations. For each probe group and for each cDNA hybridization, probes falling outside annotated transcribed regions were used to estimate a background level. This level was then subtracted from the intensities of all probes within the group. To subtract the background of DNA hybridizations, we grouped probes by GC content. For each group and hybridization, we estimated the background level as the 10% trimmed mean of the negative control probes and subtracted it from all probes of the group.

New transcript identification and transcript probe sets

We ran a segmentation algorithm combining heterozygote cDNA hybridizations with parental cDNA hybridizations using the R package 'tilingArray' (Huber *et al*, 2006). Segmentation was carried out on the set of common probes, for which the assumption of a constant level across the transcript can be made. For each chromosome, the segmentation parameter *S* (number of segments) was set so that the average segment size was 1500 bp. Segments corresponding to unannotated transcripts were then categorized as unannotated intergenic or antisense as described earlier (David *et al*, 2006) ('intergenic' were termed 'isolated' in the earlier study). Segments with less than 20 probes were discarded. A subsequent manual inspection discarded six dubious antisense segments and recovered 10.

We subsequently inferred the expression of a transcript from the intensities of its probe set. We defined the probe set of a new transcript as the probes for which the match entirely falls within the boundaries of the segment. We defined the probe set of an annotated transcript as the probes whose match entirely falls within the boundaries of an annotated S288c exon.

Probe intensity model

We modeled y_{ij}, the normalized and background-subtracted intensity of probe i in hybridization j, as

$$y_{ij} = \lambda_{1i}c_{1ij} + \lambda_{2i}c_{2ij} + \varepsilon_{ij} \tag{1}$$

where λ_{1i} and λ_{2i} are the affinities of the probe to its matches in each genome, c_{1ij} and c_{2ij} are the expression levels of the respective complementary sequences in the sample j and ε_{ij} are the errors. The affinities and the expression levels are non-negative real numbers expressed in arbitrary units. For common probes, we have $\lambda_{1i}=\lambda_{2i}$.

We considered five possible types of hybridization samples: genomic DNA (gDNA) of the two homozygous strains S and Y, their cDNA, and cDNA of the heterozygous Y/S. We set $c_{kij}=2$ if sample j is homozygous genomic DNA of genome k. Moreover, we fixed $c_{kij}=0$ if sample j is genomic DNA or cDNA of homozygous strain different from k.

Following Rocke and Durbin (2001), we modeled the variance of the errors ε_{ij} as functions of the expected intensity $I_{ij}=\lambda_{1i}c_{1ij}+\lambda_{2i}c_{2ij}$:

$$\text{var}(\varepsilon_{ij}) = \frac{1}{(\gamma b_j)^2}(1+(a_j+b_jI_{ij}))^2 \tag{2}$$

The coefficients a_j, b_j and γ were inferred using the R package vsn (Huber *et al*, 2002) by treating the cDNA and the gDNA hybridization as two separate groups. We assumed the scaled errors $\varepsilon'_{ij} = \varepsilon_{ij}/\sqrt{\text{var}(\varepsilon_{ij})}$ to be independent and identically distributed and of mean 0.

Least-squares regression

For the cDNA samples of each strain, we assumed a constant level of each allele across one transcript's probe set. The regression proceeds with each transcript separately using probes only of the transcript probe set.

We denoted p_1 and p_2 the nominal expression levels of the alleles in the homozygous strains, h_1 and h_2 the levels of each allele in the Y/S strain. From equation (1), We obtained a set of equations for all hybridizations j and probes i that depend on the hybridization sample types:

$$y_{ij} = \begin{cases} 2\lambda_{1i} + \varepsilon_{ij} & \text{S gDNA} \\ 2\lambda_{2i} + \varepsilon_{ij} & \text{Y gDNA} \\ 2\lambda_{1i}\cdot p_1 + \varepsilon_{ij} & \text{S cDNA} \\ 2\lambda_{2i}\cdot p_2 + \varepsilon_{ij} & \text{Y cDNA} \\ \lambda_{1i}\cdot h_1 + \lambda_{2i}\cdot h_2 + \varepsilon_{ij} & \text{Y/S cDNA} \end{cases} \tag{3}$$

We fitted the model by weighted least squares. More precisely, we searched for a set of affinities and expression levels that minimizes the sum of squared scaled residuals:

$$\min_{\boldsymbol{\lambda},\mathbf{p},\mathbf{h}} F(\boldsymbol{\lambda},\mathbf{p},\mathbf{h}) = \min_{\boldsymbol{\lambda},\mathbf{p},\mathbf{h}} \sum_{i,j} w_{ij}\cdot\varepsilon_{ij}^2 \tag{4}$$

subject to $\boldsymbol{\lambda}\geqslant\mathbf{0}$, $\mathbf{p}\geqslant\mathbf{0}$, $\mathbf{h}\geqslant\mathbf{0}$ and $\lambda_{1i}=\lambda_{2i}$ for common probes, where the weights $w_{ij}=1/\text{var}(\varepsilon_{ij})$ were estimated by using equation (2).

We took advantage of the form of the model for the optimization procedure. Indeed, assuming fixed weights, the cost function is a sum of squared terms bilinear in $\boldsymbol{\lambda}$ and $(\mathbf{p}, \mathbf{h})$. For a given expression-level vector $(\mathbf{p}, \mathbf{h})$, there is a closed-form solution to the unique optimal affinity vector $\boldsymbol{\lambda}$ and vice versa. We thus devised a component-wise optimization algorithm that iteratively optimizes expression levels given affinities and reciprocally, updating the weights at each step using equation (2). We considered that the algorithm had converged, if all fitted expression levels of the last 2 iterations differ by less than a value corresponding to 10% of the background level, and stopped the algorithm if convergence did not occur before the 30th iteration.

Confidence intervals

We estimated confidence intervals per ORF probe set by resampling the scaled residuals with replacement. The regression results in fitted parameters and thus, according to the model, in an estimated intensity $\hat{I}_{ij}$, an estimated weight $\hat{w}_{ij}$ and a scaled residual ε'_{ij} for each observed intensity:

$$y_{ij} = \hat{I}_{ij} + \sqrt{\hat{w}_{ij}}\hat{\varepsilon}'_{ij}$$

We generated new synthetic data as noisy measurements of the fitted intensities: $y^*_{ij} = \hat{I}_{ij} + \sqrt{\hat{w}_{ij}}\hat{\varepsilon}'_{\sigma(ij)}$ where the function σ is a random sampling with replacement of the index pairs *ij*. We repeated this B=999 times and obtained B estimates of the parameters. For all statistics of interest (expression level, allelic differential expression, etc.), 95% equi-tailed confidence intervals were estimated according to the non-parametric basic confidence limit as described in Davison and Hinkley (1997).

P-values and false discovery rates

We estimated significance levels (*P*-values) by simulating data under the null hypothesis for the two following hypotheses:

- H_1: Levels in parent equal: p_1=p_2
- H_2: Levels in hybrid equal: h_1=h_2

We fitted an appropriately constrained model for each probe set and for each hypothesis (p_1=p_2 and h_1=h_2). Similar to the procedure for estimating confidence intervals, we generated B=999 new synthetic data as noisy measurements of those fitted intensities. Here again we sampled scaled residuals of the primary unconstrained fit, because they reflect the true noise better than those of the constrained fits. On each simulated dataset, we performed an unconstrained regression. For each hypothesis respectively, we considered the *T*-statistic.

The *P*-value is then approximated by

$$p = \frac{1 + \#\{t^*_i \geqslant t\}}{B+1}$$

where t is the statistic value for the primary, unconstrained fit and t^*_i, I=1, …, B are the bootstrap statistic values (Davison and Hinkley, 1997).

Treating each hypothesis H_1 and H_2 separately, q-values, i.e. false discovery rates (FDR), were obtained using the R package qvalue (Storey and Tibshirani, 2003) with default parameters.

Sequence validation of differentially expressed transcripts

Quantitative estimates of allelic expression ratios by sequencing were obtained using the method described by Ge *et al* (2005). Primers (Supplementary Table VII) were synthesized such that they spanned multiple SNPs between the two alleles of a transcript. From two independent Y/S strains, XHS768 and XHS769, cDNA was synthesized using random hexamers and PCR was carried out on the resulting cDNA for sequence analysis. PCR products using the same primers on genomic DNA of a Y/S strain, XHS768, was used to provide reference traces in situation of 1:1 allelic concentrations. The resulting sequence traces were analyzed with the software PeakPicker (Ge *et al*, 2005), which estimates allelic expression ratios from relative peak heights at SNP positions. We calculated the allelic ratios of transcripts as the median over all SNPs and traces (Supplementary Table VII). Out of the 24 transcripts tested, one (*HOP1*) did not confirm polymorphic positions in the genomic DNA. Two others (*ICL2* and *YDL237W*) were rejected from further analysis for having ratio estimates derived from less than two SNPs.

ADE coefficient

We defined the ADE coefficient as $(|h_{\mathrm{Y}} - h_{\mathrm{S}}|)/(h_{\mathrm{Y}} + h_{\mathrm{S}})$, where h_{Y} and h_{S} are the expression levels of the Y allele and S allele, respectively in the heterozygote.

Proportion of *cis*- and *trans*-regulatory effects

The ratio of *cis*-regulatory divergence to the total regulatory divergence (Wittkopp *et al*, 2008) is computed as $|C|/(|C| + |T|)$ where C, the *cis*-regulatory effect, is the log ratio of the allelic expression levels in the hybrid and T, the *trans*-regulatory effect, is the difference between the log ratio of the parental gene expression levels and C.

Analysis of *PHO84* reciprocal hemizygote strains hybridizations

The hybridizations of the two *PHO84* reciprocal hemizygote strains were analyzed using the same model as described above. Total transcript expression levels (i.e., $h_{\mathrm{Y}}+h_{\mathrm{S}}$, the sum of the two allele levels for each transcript) were considered for comparison.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Paul McGettigan for work on probe mapping, Wu Wei for providing the YJM789 genome sequence and its alignment to the S288c genome and Sandra Clauder-Münster for technical help.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P *et al* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816

Bjornsson H, Albert T, Ladd-Acosta C, Green R, Rongione M, Middle C, Irizarry R, Broman K, Feinberg A (2008) SNP-specific array-based allele-specific expression analysis. *Genome Res* **18:** 771–779

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185–193

Bray NJ, Buckland PR, Owen MJ, O'Donovan MC (2003) Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet* **113:** 149–153

Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F (2007) Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in *S. cerevisiae*. *Cell* **131:** 706–717

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V *et al* (2005) The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **103:** 5320–5325

Davison AC, Hinkley DV (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press

Engström PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, Kunarso G, Ng EL-C, Batalov S, Wahlestedt C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Wells C, Bajic VB *et al* (2006) Complex loci in human and mouse genomes. *PLoS Genetics* **2:** e47

Gaisne M, Bécam AM, Verdière J, Herbert CJ (1999) A 'natural' mutation in Saccharomyces cerevisiae strains derived from S288c affects the complex regulatory gene HAP1 (CYP1). *Curr Genet* **36:** 195–200

Ge B, Gurd S, Gaudin T, Dore C, Lepage P, Harmsen E, Hudson TJ, Pastinen T (2005) Survey of allelic expression using EST mining. *Genome Res* **15:** 1584–1591

Gimelbrant A, Hutchinson J, Thompson B, Chess A (2007) Widespread monoallelic expression on human autosomes. *Science* **318:** 1136–1140

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, Oliver SG (1996) Life with 6000 genes. *Science* **274:** 546, 563–567

Hongay CF, Grisafi PL, Galitski T, Fink GR (2006) Antisense transcription controls cell fate in Saccharomyces cerevisiae. *Cell* **127:** 735–745

Huber W, Toedling J, Steinmetz L (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22:** 1963–1970

Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Suppl 1)**:** S96–104

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA *et al* (2005) Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564–1566

Knight J (2004) Allele-specific gene expression uncovered. *Trends Genet* **20:** 113–116

Knight J (2006) Analysis of allele-specific gene expression. *Methods Mol Biol* **338:** 153–165

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP (2003) Allelic variation in gene expression is common in the human genome. *Genome Res* **13:** 1855–1862

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7:** 113

Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, Cheng J, Bell I, Ghosh S, Piccolboni A, Gingeras TR (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet* **38:** 1151–1158

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454:** 479–485

Oshima Y (1997) The phosphatase system in Saccharomyces cerevisiae. *Genes Genet Syst* **72:** 323–334

Pant P (2006) Analysis of allelic differential expression in human white blood cells. *Genome Res* **16:** 331–339

Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L (2007) Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat Genet* **39:** 496–502

Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res* **35:** e128

Rocke DM, Durbin B (2001) A model for measurement error for gene expression arrays. *J Comput Biol* **8:** 557–569

Rockman M, Kruglyak L (2006) Genetics of global gene expression. *Nat Rev Genet* **7:** 862–872

Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L (2005a) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res* **15:** 284–291

Ronald J, Brem R, Whittle J, Kruglyak L (2005b) Local regulatory variation in saccharomyces cerevisiae. *PLoS Genet* **1:** e25

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet* **4:** e1000006

Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416:** 326–330

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100:** 9440–9445

Uhler JP, Hertel C, Svejstrup JQ (2007) A role for noncoding transcription in activation of the yeast PHO5 gene. *Proc Natl Acad Sci USA* **104:** 8011–8016

Wang D, Sung H, Wang T, Huang C, Yang P, Chang T, Wang Y, Tseng D, Wu J, Lee T, Shih M, Li W (2007) Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome Res* **17:** 1161–1169

Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, Wilhelmy J, Komp C, Tamse R, Wang X, Jia P, Luedi P, Oefner PJ, David L, Dietrich FS, Li Y *et al* (2007) Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. *Proc Natl Acad Sci USA* **104:** 12825–12830

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **430:** 85–88

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet* **40:** 346–350

Wykoff DD, Rizvi AH, Raser JM, Margolin B, O'Shea EK (2007) Positive feedback regulates switching of phosphate transporters in S. cerevisiae. *Mol Cell* **27:** 1005–1013

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033–1037

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science* **297:** 1143