Published in final edited form as:
Nat Microbiol. 2019 April ; 4(4): 683–692. doi:10.1038/s41564-018-0346-9.

Sensitive high-throughput single-cell RNA-Seq reveals within-clonal transcript-correlations in yeast populations

Mariona Nadal-Ribelles[#1,2,3,4], **Saiful Islam**[#1,2], **Wu Wei**[#1,2,5], **Pablo Latorre**[#3,4], **Michelle Nguyen**[1,2], **Eulàlia de Nadal**[3,4], **Francesc Posas**[3,4], and **Lars M. Steinmetz**[1,2,6,*]

[1]Department of Genetics, Stanford University, School of Medicine, California, USA

[2]Stanford Genome Technology Center, Stanford University, California, USA

[3]Departament de Ciències Experimentals i de la Salut, Cell Signaling Research Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain

[4]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain

[5]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

[6]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany

[#] These authors contributed equally to this work.

Abstract

Single-cell RNA-seq has revealed extensive cellular heterogeneity within many organisms, but few methods have been developed for microbial clonal populations. The yeast genome displays unusually dense transcript spacing, with interleaved and overlapping transcription from both strands, resulting in a minuscule but complex pool of RNA protected by a resilient cell wall. Here, we have developed a sensitive, scalable, and inexpensive yeast single-cell RNA-seq (yscRNA-seq) method that digitally counts transcript start sites in a strand- and isoform-specific manner. YscRNA-seq detects the expression of low-abundant, non-coding RNAs, and at least half of the protein-coding genome in each cell. Within clonal cells, we observed a negative correlation for the expression of sense/antisense pairs, while paralogs and divergent transcripts co-express. Combining yscRNA-seq with index sorting, we uncovered a linear relationship between cell size

[*]Corresponding author.

Code availability

Custom code generated in this study can be downloaded from *steinmetzlab.embl.de/yscRNASeq*

Author contribution

M.N.R., S.I., W.W. and L.M.S. conceived the project. M.N.R., S.I., W.W., P.L., M.N. developed the protocol and performed the analyses. M.N.R., S.I., M.N. performed the experiments. W.W. and P.L. performed the computational analyses. M.N.R., S.I., P.L., W.W., E.N., F.P., L.M.S. participated in experimental design, data analysis and writing the manuscript. E.N., F.P. and L.M.S. supervised the work. All authors read and edited the manuscript.

Competing interests

The authors declare no competing financial interests.

and RNA content. Although we detected an average of ~3.5 molecules/gene, the number of expressed isoforms are restricted at the single-cell level. Remarkably, the expression of metabolic genes is highly variable, while their stochastic expression primes cells for increased fitness towards the corresponding environmental challenge. These findings suggest that functional transcript diversity acts as a mechanism for providing a selective advantage to individual cells within otherwise transcriptionally heterogeneous populations.

Eukaryotic transcription is pervasive, and results from the stochastic process of gene expression leading to cell-to-cell heterogeneity. Advances in sequencing technologies and library preparation have made single cell RNA-seq (scRNA-seq) accessible to several tissues and cell lines, but few methods have been successful for unicellular microorganisms1.

The development of yeast-specific scRNA-seq has been hampered by the intrinsic nature of yeast. First, its small cell size (2-5 μm) results in a minute amount of RNA per cell, which is estimated to be at least 10 times lower than mammalian cells where scRNA-seq has been pioneered2 (1 pg *versus* 10 pg, respectively3). Second, the cell wall poses a barrier for single-cell RNA isolation, and therefore standard RNA extraction procedures are incompatible with efficient scRNA-seq library preparation. Third, the yeast transcriptional landscape is composed of bidirectional and overlapping transcripts embedded in a dense genome—this presents a challenge for RNA-seq, and requires stranded-libraries to capture complex genome architectures. As such, scRNA-seq has only recently been able to be applied to microorganisms like yeast, using low-throughput and labor-intensive methods in combination with non-stranded library preparation4,5.

S. cerevisiae is the only organism for which transcript isoforms have been mapped in bulk at both 5'- and 3'-ends by transcript isoform profiling (TIF-seq)6, however cell-to-cell heterogeneity has mainly been studied on a case-by-case basis. Yeast populations show extensive isoform heterogeneity that contribute to phenotypic diversity6,7. Whether this variability results from the co-expression of several isoforms or from a cell-specific selection has not been investigated; this would require resolving individual cells from a homogenous population. The strongest limitation to yeast-specific scRNA-seq is the lack of strand-specific transcript isoform methods that are sensitive enough to globally assess the transcriptome of single yeast cells. This absence underscores the need for the development of novel technologies.

Here, we set out to develop a high-throughput single-cell RNA-seq method for yeast that integrates indexed cell sorting for prior phenotyping, is inexpensive (approximately US$12 per cell), and is strand-specific. By applying yscRNA-seq, we quantitatively characterized the extent to which isogenic single-cells deviate in gene expression, and measured the stochastic expression of highly variable genes that can result in fitness variation within microbial populations.

Results

To measure absolute gene expression and transcription start site (TSS) usage in individual yeast cells, we performed unbiased index sorting of single cells from exponentially-growing yeast cultures in rich media (YPD) using 96 well plates containing absolute ethanol for fixation and RNA preservation. For each well, we measured the forward scatter (FSC) as a proxy for cell size by fluorescence-activated flow cytometry (FACS; Supplementary Figure 1a). Following cell sorting, we applied yscRNA-seq to 285 individual yeast cells (2 plates for BY4741 and 1 plate for YJM789; see Methods). After ethanol evaporation, cells were lysed in buffer containing zymolyase, and 5000 molecules of external RNA control consortium (ERCC) transcripts. A 5'-biotinylated template-switching oligo (TSO) containing P5 and a unique molecular identifier (UMI) was used to generate the first strand. Full-length dscDNA libraries were amplified with limited numbers of PCR cycles, and size distribution was validated by Bioanalyzer profiling (Supplementary Figure 1b). Cell-specific adapters were introduced by tagmentation using homemade Tn58, preloaded with adapters (see Methods). This greatly reduced the cost-per-cell to US$12 (Supplementary Table 1). Tagmented libraries (96 samples) were pooled, and strand-specific libraries were eluted by removing the biotinylated strand with streptavidin beads (Figure 1a). Size distribution was assessed before sequencing (Supplementary Figure 1c).

A first-sequencing read (*read1*) was used to obtain the UMI and the gene identity, while a shorter read (*index1*) was used to retrieve cellular barcodes. After preprocessing and aligning the reads (see Methods), we applied a filter to consider only UMIs with at least 3 reads for analysis (Supplementary Figure 1d). Additionally, we restricted the analysis to cells with more than 500K uniquely-mapped reads and over 1000 expressed genes (≥ 1 UMI-gene), which removed low-quality cells with a higher ratio of mitochondrial RNA (Supplementary Figure 1e,f). Overall, this resulted in 127 cells for BY4741 out of 190, and 48 out of 95 for YJM789.

To measure the quantitative efficiency of yscRNA-seq, we compared the number of detected molecules against expected unique ERCC molecules, and observed a linear correlation across the entire dynamic range (Figure 1b) with a 15.5-25.8% efficiency (an average of 1290 ERCC molecules out of 5000 spiked-in, 775 molecules after filtering, respectively). To gauge the 5'-end detection accuracy of yscRNA-seq, we aligned reads to the reference TSS, determined by TIF-seq6 and from ERCC annotation. Our method faithfully identified TSS boundaries (Supplementary Figure 1g).

Next, we benchmarked the sensitivity of yscRNA-seq relative to other scRNA-seq methods with a previously established metric that calculates the number of ERCC molecules required to reach 50% detection probability9,10. Indeed, we find that yscRNA-seq is among the most sensitive methods, with a median 50% detection probability requiring only 4.7 molecules (Figure 1c). Similarly, we compared yscRNA-seq to the only available yeast scRNA-seq dataset from Gasch *et al.*4, which used *Fluidigm*'s C1 System. Overall, yscRNA-seq yielded a higher number of genes per cell (3399 versus 2392) and a good genome-wide correlation despite scRNA-seq from Gasch *et al.* not being strand-specific (Spearman Correlation 0.71; Figure 1d, Supplementary Figure 1h).

Given the quantitative nature of yscRNA-seq, we assessed the concordance between RNA abundance (estimated in bulk by competitive PCR from Miura *et al.*3) and our BY4741 yscRNA-seq dataset. We observed a linear correlation genome-wide (Spearman Correlation

0.76; Figure 1e), however when we performed the same analysis after correcting for ERCC efficiency (15.5%), the predicted number of molecules per gene across the entire transcriptome increased3 (Supplementary Figure 1i). In both cases, the increase in RNA abundance can be explained by the fact that our method provides an unbiased transcriptome quantification of both coding and noncoding transcripts (ncRNAs), the latter of which had not previously been considered. Our data also suggest that ERCC correction alone may overestimate expression, possibly due to differences in reverse transcription efficiencies9,11.

To confirm that yscRNA-seq transcriptomes resemble those obtained in bulk, we compared the sum of all BY4741 libraries to our previously-published tiling array data from the same background and condition12. Interestingly, the comparison with tiling array data resulted in the highest correlation to other methods (Spearman correlation 0.83) due to the strand-specific nature of both approaches (Supplementary Figure 1j).

Overall, yscRNA-seq quantitatively recapitulated gene expression from two independent studies and performed within the most-efficient scRNA-seq protocols—the highest for yeast. One explanation for the higher sensitivity is that libraries are directly generated from live-sorted cells through sequential reactions, which reduces sample loss and handling to a minimum.

To validate that a single yeast cell was present in each well, and that there was no well-to-well contamination, we applied yscRNA-seq to a randomly-sorted plate from a mixture of two yeast strains (BY4741 and YJM789) at a 1:1 ratio. YJM789 is a less-commonly used haploid *S. cerevisiae* strain isolated from an HIV patient13. At the genomic level, YJM789 differs from the reference strain in approximately 60,000 single nucleotide polymorphisms (SNPs) and 6,000 indels13. We plotted the allele frequency of reads mapping to each genotype, and considered cell doublets if more than 1% of the reads mapped to more than one genotype. Reassuringly, the frequency of doublets (2 or more cells/reaction) was less than 1%, thus confirming that yscRNA-seq libraries originated from individual cells (Figure 1f, Supplementary Figure 1k).

For a single wild-type *S. cerevisiae* cell, the majority of expressed genes belonged to open reading frames (3072 ORFs), which encompassed 90.5% of the detected transcripts. Meanwhile, ncRNAs—cryptic unstable transcripts (CUTs), stable unannotated transcripts (SUTs) and others—only represented 9.5% of the detected RNA pool (Figure 2a, Supplementary Table 2). The ratio of ORFs/ncRNAs resembled those reported from bulk studies14–16, but the distribution of detected ORFs was narrower than for ncRNAs. This may be due to lower abundance and stability of ncRNAs relative to mRNAs. Interestingly, we obtained similar means of expression for YJM789, indicating that the number of expressed genes per cell is similar across strains (Supplementary Table 2). Moreover, when we analyzed the expression matrix for all genes in each individual cell, only 252 of 7272 genes (or 8 ORFs) were not detected in any of the 127 cells (Supplementary Table 3). Overall, we observed a good genome-wide expression correlation within individual cells or across all samples (Pearson correlation 0.66 to 0.83; Supplementary Figure 2a,b). Altogether, these data suggest that at least half of the yeast genome is expressed in a complex pattern within each individual cell.

To assess the complexity of yscRNA-seq libraries, we downsampled reads to fixed depths from one representative sample (1,027,599 reads after QC and filtering). We found that 250,000 reads recapitulated 90% of the total number of genes detected indicating that even at low sequencing depth yscRNA-seq captures a large fraction of the transcriptome (Supplementary Figure 2c), which can further lower the cost/cell.

To assess genome-wide transcriptional architectures by exploiting the sensitivity and strand-specific nature of yscRNA-seq, we analyzed the correlation between paralogous genes (those that arose from whole-genome duplication17), sense-antisense pairs (SAPs), and divergent transcripts12 (Figure 2b). The correlation was only computed for gene pairs in which at least one gene in the pair was detected in ≥3 cells. Expression of SAPs showed a strong negative correlation compared to random gene pairs (one-tailed Wilcoxon test, p-value 1.45e-19). This is consistent with previous data showing anticorrelated expression of SAPs from bulk samples18.

To explore the potential of yscRNA-seq, we investigated *GAL1*, which is one of the most studied SAPs. In bulk samples under glucose growth conditions, the *GAL1* sense-transcript is repressed and a ncRNA (*CUT445),* which originates from a region of the 3'-UTR that is constantly expressed in glucose and galactose that overlaps the *GAL1* TSS19,20 For the majority of our single cells in YPD, we could neither detect the sense nor the antisense transcript—remarkably, for those where we could, the expression pattern was mutually exclusive (Supplementary Figure 2d). In contrast, when we investigated the expression of gene pairs that originate from bidirectional promoters and paralogs, we observed a positive correlation for both groups (one-tailed Wilcoxon test, p-value 2.63e-09 and 3.6e-09, respectively Figure 2b). These results indicate that transcription events reported in bulk reproduce within single cells, while the bimodal expression for the *GAL1-CUT445* SAP is unresolvable in bulk assays.

One of the main layers of transcriptional complexity arises from transcript isoform diversity, where an average of 26 isoforms per gene have been reported in yeast by TIF-seq6. To characterize the TSS variability in single cells, we sought to identify TSSs applying a similar criteria used for TIF-seq6, in which the highest-expressed position defined the major isoform. Secondary isoforms were iteratively assigned if reads occurred outside a 15 nt window centered from the previous isoform (see Methods). To make the data comparable across both studies, we assessed TSS diversity by considering only TSS positions with ≥2 UMIs in yscRNA-seq, and ≥2 reads in TIF-seq6, for the same background and condition. From the sum of isoforms per gene across all cells, we detected a linear correlation between TSS number/gene (Spearman correlation 0.63; Figure 2c). On average, each gene expressed 3.46 unique molecules (Supplementary Figure 2e), however the mean number of TSSs per gene across all cells was 1.19 (Figure 2d). Indeed, when we clustered cells based on the TSS variability of a representative gene (*YLL014W* with a mean of 3.09 UMIs and 1.25 TSSs of a total of 8 used-TSS-positions in our dataset), the majority of cells expressed only one isoform (coinciding with the bulk TSSs detected by TIF-seq6 at position-0), 32 cells co-expressed a second isoform from a different TSS, and in some rare cases three isoforms were detected (Figure 2e). As a control, the sum of the most-expressed isoforms of all BY4741 cells revealed a linear correlation to the expression of the second most-expressed isoform (Spearman correlation 0.93). Furthermore, while the major isoform dominated the expression, the second isoform was still abundant within the population (Supplementary Figure 2f). Overall, our results suggest that while a main isoform is preferred in an individual cell, TSS diversity is achieved through the expression of select isoforms between cells.

To understand how TSS variability in single cells leads to the observed population isoform diversity, we assessed each single cell's contribution by computing the cumulative sum of unique TSS isoforms/cell, comparing the result to the total number of TIF-seq6 5'-TSS isoforms (a total of 148,000 5'-isoforms with ≥2 reads, Figure 2f). We observed a logarithmic increase of isoforms, that starts to plateau at ~100 cells without reaching saturation. These data support the notion that the abundant TSS diversity seen in populations is composed of the heterogeneous expression of limited gene isoforms per cell. As such, a large number of single cells would be needed to reconstruct isoform diversity.

To investigate the capacities of yscRNA-seq to resolve clonal yeast populations, we applied t-distributed Stochastic Neighbor Embedding (tSNE), using the entire transcriptomes for all datasets. Indeed, tSNE analysis revealed two distinct strain-specific clusters (BY4741 and YJM789, Figure 3a), validating yscRNA-seq as a sensitive method for distinguishing different yeast strains.

Co-regulation of cell size with transcriptome size has been extensively investigated in yeast, and has traditionally been approached through the use of mutants that uncouple cell division and growth, or by arresting the cell cycle but not cell growth21–23. The quantitative nature of yscRNA-seq, together with estimated cell size (by FSC index sorting), provides a tool to link transcriptome to phenotype. Moreover, the presence of ERCC spike-ins allows libraries to be normalized against technical noise rather than total number of reads24,25. We found that the absolute number of RNA molecules increased linearly with cell size for both genetic backgrounds (Figure 3b). Comparing unique ERCC molecules against cell size (or against total RNA molecules) confirmed that the observed changes in transcriptome size exceeded technical noise (Supplementary Figure 3a,b).

Reverse transcription and template switching are major sources of technical noise26, while intrinsic biological noise can originate from the stochastic nature of transcription and by the regulation of mRNA stability. To identify variable genes, we took advantage of the presence of ERCC spike-ins24, and classified genes as variable if their squared coefficient-of-variation (CV^2) was higher than the technical and expected biological variance using a 10% false discovery rate (CV^2>0.25)24. For exponentially-growing wild type BY4741, we identified 400 variable RNAs (FDR <0.01; Figure 3c). Of these, 100 were ncRNAs (38 CUTs and 62 SUTs), while the remaining 300 variable genes were protein-coding. Remarkably, 50% of the detected ncRNAs were highly variable at the single-cell level, probably due to their intrinsic properties. Because UMI incorporation in yscRNA-seq occurs by template-switching at the 5'-end, we ruled out gene-length-dependent effects of the observed CV^2 (Supplementary Figure 3c). Therefore, transcriptome composition was heterogeneous, despite most cells expressing a relatively similar number of genes.

To characterize the nature of variable ORFs in BY4741, we performed KEGG pathway-enrichment analysis. As expected, variable genes were enriched in cell-cycle-periodic transcripts16,27. Surprisingly, the largest groups of variable genes belonged to metabolic pathways such as galactose and glycolysis/gluconeogenesis, among others (Figure 3d). For YJM789, we detected 567 variable genes (FDR<0.01), of which 124 overlapped between YJM789 and BY4741 and were significantly enriched for metabolic processes (Supplementary Figure 3d). These results suggest that variable genes are partially strain-dependent, while several genes might be intrinsically variable.

To understand variability within the BY4741 population, we applied tSNE analysis. We identified two distinct clusters enriched for previously annotated cell-cycle-regulated genes16,27 (Figure 3e, Supplementary Figure 3e). We then phase-ordered the expression of cluster-specific genes based on previously annotated expression peaks16,27. *Cluster 1* contained cells that expressed early (M/G1) and late (G2/M) cell cycle genes, while *Cluster 2* was enriched for cells expressing G1, S and late S/G2 genes (Figure 3f). To rule out the contribution of batch effects, we projected both replicates of BY4741 onto the tSNE and confirmed that batch effects are negligible (Supplementary Figure 3f). Our data suggests that, while phase-specific expression exists, the yeast cell cycle segregates cells into two subpopulations with distinct transcription signatures: those corresponding to cell cycle entry/exit, and those corresponding to progression through G1/S/G2.

Since the majority of variable genes were related to metabolic processes such as sugar and nucleotides metabolism, we assessed whether this variability was cell cycle dependent. We overlaid the expression of variable galactose genes (*GAL3, EMI2, GLK1, HXK1*) onto the tSNE (Figure 4a), as well as representative examples of the remaining KEGG-enriched pathways (Supplementary Figure 4a). We did not observe cell cycle segregation, suggesting that the transcriptional heterogeneity of these genes is cell-cycle-independent. Galactose-induced genes are tightly repressed in the presence of glucose (YPD), and have been used as a model to understand transcriptional noise and regulation19,28,29. Interestingly for BY4741, we detected the expression of at least one variable gene belonging to the galactose pathway in virtually all cells (Figure 4b). Despite over 80% of cells exponentially growing in glucose displaying no expression of *GAL3,* 20% of cells had more than 4 unique *GAL3* mRNAs and 1,5% more than 10 mRNAs — an expression level above the expected noise.

To understand whether the expression of galactose-variable genes was coordinated, we analyzed the pair-wise expression of all galactose metabolism genes. We did not observe any strong correlations among gene pairs, suggesting their uncoordinated expression-regulation in single cells grown in YPD (Figure 4c). We hypothesized that heterogeneity of variable galactose-related genes during growth in glucose could provide a selective advantage if cells were to face a rapid change in carbon source. To assess this possibility, we tagged the endogenous *GAL3* gene with GFP at the C-terminus, sorting two populations of 10,000 cells from YPD based on GFP intensity (top 2% and a random sort control). Then, we then followed their growth in YPD or YPGal and measured Gal3-GFP fluorescence under both conditions (Supplementary Figure 4b). While all populations had a similar growth rate in YPD, the top 2% of cells expressing Gal3 in YPD, displayed an increased fitness in galactose (Figure 4d). These results indicate that stochastic expression variation can have a functional consequence, and can be used as a bet-hedging strategy.

Discussion

Here we report the development of yscRNA-seq, a method for quantitatively profiling gene expression and TSS variation in a strand-specific, cost-effective manner. YscRNA-seq enables cell profiling with minute amounts of RNA, and represents the most sensitive, inexpensive yeast scRNA-seq method to date. In addition, FACS-based single-cell isolation allows correlating cell parameters, such as cell size or reporter abundance, to the transcriptome. YscRNA-seq provides a high-throughput approach for linking phenotype to transcriptome by means of a single experiment.

Pervasive transcription in yeast contributes to complex transcriptional architectures that require strand-specific methods to be resolved at the single cell level[6,12,14]. By applying yscRNA-seq, we identified genome-wide co-expression of paralogous and divergent gene pairs within single cells. In contrast, SAPs showed a stronger anticorrelation of expression in single cells than previously reported in bulk samples[18,30], establishing yscRNA-seq as a tool to resolve transcriptional architectures masked in bulk experiments.

By including UMIs, we were able to digitally count gene expression and TSS usage. We detected about half of the genome as simultaneously and heterogeneously expressed in single cells. Additionally, the cumulative expression of approximately 100 cells virtually covered the entire transcriptome (>99% of ORFs). On average, each gene expressed 3.46 transcript molecules per cell but primarily from one TSS isoform. This suggests that a limited number of TSS isoforms per cell composes the massive isoform heterogeneity observed in bulk samples. These findings challenge the assumption of one mRNA molecule per gene per cell, and suggest rather the use of one isoform per gene per cell.

By recording cell size during sorting, we observed a positive correlation between total RNA molecules and cell size. Our single cell analysis revealed a high degree of heterogeneity in gene expression between cells that affects at least 5.4-7.7% of genes for both strains (11.7-16% of expressed genes/cell). As expected, these genes included cell cycle-regulated transcripts, clustered in only two sets with distinct signatures (mitotic entry/exit and G1/S/G2 progression) rather than phase-specific clusters. This suggests that cell cycle signatures are not as pronounced if they are measured directly from unsynchronized single cells.

Surprisingly, we found an even larger number of variable genes related to metabolic processes. One such process is galactose metabolism, whose early transcriptional activation in the presence of both glucose and galactose has been previously reported, and proposed to shorten the adaptation to diauxic shift[31,32]. Indeed, we tested the functional consequences of stochastic expression for one of the highly variable genes (i.e. Gal3-GFP), and found that this variability provides a fitness advantage upon change in carbon source. This observation suggests that, while the process of transcription is inherently noisy, stochastic gene expression has functional implications for cellular fitness at the single cell level.

Methods

Strains

Wild type *S. cerevisiae* BY4741 (S288C) and YJM789 strains were used for yscRNA-seq libraries. BY4741 (MATa his3-D1 leu2-D0 met15-D0 ura3-D0) and its derivative ySMN235 (*GAL3*-GFP::hphNT1) were obtained by genomic integration using a PCR-based strategy33.

Cell growth and isolation

BY4741 or YJM789 were pre-inoculated in rich media and grown overnight until OD_{660}=1. The next morning, cells were diluted to OD_{660}= 0.05 and grown for at least two full divisions. Prior to sorting, cells were diluted to OD_{660}= 0.01 and sonicated for 3 pulses of 0.5 seconds to remove aggregates (Sonicator Branson M1800). Then, cells were filtered by passing samples through cell strainer snap-cap tubes (BD Falcon). Live single cells were index-sorted (BD InFlux) into 96 well plates containing 5µl of absolute ethanol. Gates on the pulse-width were stringent, in order to remove potential doublets. Forward and Side scatter (FSC and SSC respectively) values were recorded for each well, FSC was used as an approximation of cell size. For each plate, well-*H12* was not sorted, and served as negative control.

For experiments with mixed strains of BY4741 and YJM789 (Figure 1f, Supplementary Figure 1k), cells were grown independently as described above, except that strains were mixed at a 1:1 ratio prior to random sort (unindexed) into a 96 well plate.

yscRNA-seq library preparation

Sorted 96 well plates were left in a sterile hood to allow complete ethanol evaporation. Then, the cell wall was digested with a lysis buffer containing zymolyase (1 U/µl) and 5000 molecules of External RNA Control Consortium (ERCC). Cell lysis was performed for 10 minutes at 37 °C followed by 3 minutes at 72 °C in 5 µl. Plates were immediately placed on ice and 5 µl of reverse transcription reaction mix (RT-buffer and *Invitrogen* Super Script II) was added to synthesize first strand cDNA from a 5'-Biotinylated oligo dT primer (Bio-AATGATACGGCGACCACCGATCGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT). The oligo dT primer contains a PvuI site between the P5 sequence and the dT stretch that is used to remove 3'-ends. UMI incorporation was done in the same reaction by using a template-switching oligonucleotide consisting of RNA bases (Bio-AAUGAUACGGCGACCACCGAUNNNNNNGGG).

The TSO primer sequence (Bio-GAATGATACGGCGACCACCGAT) contains P5 followed by six random nucleotides (UMI) and 3 GGGs used for template-switching. Full-length library amplification was performed with the UMI-PCR primer, which anneals to a common region (P5 sequence) present in both the oligo dT and TSO region to minimize the formation of primer-dimers. This reaction was performed for a limited number of cycles (21 cycles) with Advantage 2 Polymerase mix (Clontech). The resulting full-length dscDNA libraries were diluted (1:20) and subjected to qPCR to validate the presence or absence of cDNA library, using primers for a highly expressed housekeeping gene (*TDH3.Fw:* TCGTCAAGTTGGTCTCCTGG and *TDH3.Rev:* GGCAACGTGTTCAACCAAGT). Full-length dscDNA libraries were purified using Agencourt Ampure beads XP (Beckman Coulter), following the manufacturer's instructions (0.6:1 ratio), and eluted in 15 μl of Elution Buffer (Qiagen). DscDNA library size was validated by Bioanalyzer (Supplementary Figure 1b). To incorporate a cellular barcode and the Illumina P7 adapter, dscDNA libraries were tagmented with our homemade Tn58 and preloaded with cell-specific barcodes as previously described34. Briefly, a plate containing 96 adapters (STRT-TN5-1 to 96) was mixed with TN5-U (PHO-CTGTCTCTTATACACATCTGACGC) 50 μM each. Primers were annealed by denaturing at 95 °C for 3 minutes and cooled down to room temperature. Annealed primers were loaded into Tn5 to make a 10X transposome stock in 80% Glycerol, and incubated for 1 hour at 37°C (50 μM transposase, 50 μM of 96 different annealed adapters). Loaded Tn5 plates can be stored for up to a month at -20°C. Tagmentation was performed by mixing the eluted dscDNA libraries with 2 μl of transposase stock (3-6 ng dscDNA, transposase 1X, 2X TAPS, 10% DMF in 20 μl) for 5 minutes at 55 °C. The tragmentation reaction was inactivated by incubation at 85 °C for 5 minutes. Tagmented libraries were captured using streptavidin beads—a 1:20 dilution of Dynabeads MyOne Streptavidin C1 beads (Invitrogen) was used, pre-washed twice with 2xBTW buffer (10 mM Trizma HCl pH 7.5, 1 mM EDTA, 2M NaCl, 0.02% Tween). To each reaction, 20 μl of pre-washed MyOne Streptavidin C1 beads were added to each well and incubated for 5 minutes at room temperature. Then, all 96 reactions were pooled together and washed once with TNT buffer (20 mM Tris-Cl pH7.5, 50 mM NaCl, 0.02% Tween 20), once with PB Buffer (Quiagen) and three times with TNT buffer. Libraries were then resuspended in PvuI digestion buffer (NEB, PvuI 0.4 U/μl, incubated for 1 hour at 37 °C) to remove the 3'-ends. After 3'-end removal, streptavidin beads were washed twice with TNT buffer and resuspended in 15 μl of nuclease-free water. To generate strand-specific libraries, samples were briefly denatured at 70 °C for 10 minutes to release the non-biotinylated strand. Then, beads were quickly bound to magnets, to remove the biotinylated strand and single-strand cDNA (sscDNA) was recovered from the supernatant. The resulting TSS-enriched sscDNA (purified with Ampure beads as above, at a 0.8:1 ratio) and libraries were quantified using the KAPA library quantification kit (Illumina). Between 8-16 pmol were hybridized and sequenced using a HiSeq Rapid Run to a sequencing depth of 1-2 million reads/cell.

Single stranded cDNA library sequencing

Sequencing was performed using one plate per lane in an Illumina HiSeq 2500. Libraries were loaded without denaturation in the absence of PhiX. Custom primers with Locked Nucleic Acids (LNA, Exiqon) were spiked-in, following manufacturer's instructions. The *read1* primer is located upstream of the UMI and was used to obtain a 68 bp read containing the UMI sequence, followed by the TSS of the gene of interest. A second custom index primer (*index1*) was used to determine the identity of the cellular barcode (8 bp length). LNA primers (0.5 μM) were spiked-in following manufacturer's instructions for the *read1* and *index1* primers.

UMI_PCR_read1: +GAATGA+TACGGCG+ACCA +CCGA+T

Index1: CTGT+CT+CTT+ATA+CA+CA+TCTGA+CG+C

The "+" indicates the position of the LNA. A detailed step-by-step protocol of yscRNA-seq can be found at *protocols.io*.

Read pre-processing, alignment and filtering

To process the sequencing data (fastq), we used a custom script written in Java to trim out the first 6 nt corresponding to UMI sequences. Poly-G's following the UMI nucleotides were also trimmed out. Only reads containing up to 14 G's were kept for analysis. These reads were aligned to a reference sequence combining the *S. cerevisiae* genome (Saccer3, SGD R64 version, www.yeastgenome.org) and ERCC control sequences using Novocraft (http://www.novocraft.com) with default parameters. Of note, ERCC sequences provided by the manufacturer do not contain the restriction sites used to clone the synthetic ERCCs downstream of the T7 promoter. Due to the 5'-specific nature of yscRNA-seq, these nucleotides are sequenced from *read1* and cause misalignments to the reference ERCC annotation. We generated a modified the ERCC reference sequence by including the remaining T7 promoter nucleotides in the reference annotation. Updated ERCC sequences can be accessed from the Gene Expression Omnibus (see [Data Availability]). Unique mapped reads with a quality score >30 and no soft clips in the 5'-end were kept for downstream analysis35.

Molecule counting and TSS identification

For each UMI the total number of reads was counted, and UMIs supported with less than 3 reads were removed. Reads that contained the same UMI sequence and mapped to the same 5′-end location in the genome, were included in the number of reads supporting this UMI.

To define TSS isoforms, we used a similar approach as previously reported (TIF-seq)6. Briefly, we defined a first set of initial TSSs by considering the position in the gene where the overlap with the first nucleotide of the reads occurs. Then, we counted the number of UMIs and reads supporting these TSSs. Next, for each gene, we ordered all the TSSs present in all cells (n=127) by expression. Then, secondary isoforms were called by iterating through the ordered set of TSSs in the following way: if the TSS was within a ±7 nt range that had not been defined as a confident TSS in previous iterations, its position was used to define a new TSS. Following this criteria, the most expressed TSSs were the first to be considered as confident TSS; if the TSS was within a ±7 nt range of a previously-defined TSS, we assigned its UMIs and reads to the confident TSS to which it overlaps. In the case a TSS overlapped with two defined TSS, the number of UMIs and reads were divided and assigned to both TSSs. Finally, we considered only the confident TSS isoforms that were supported by at least two UMIs. The aforementioned computation was used to define the total confident TSS isoforms per gene. To identify TSS isoforms per cell, we iterated through each gene/cell and assigned each TSS to the most expressed TSS isoform at gene level.

Data availability

All data generated in this study has been uploaded at Gene Expression Omnibus (GEO) under accession number GSE122392.

Variable genes calling

We inferred variable genes as previously reported24. Briefly, we tested the null hypothesis that the biological squared coefficient of variation of a gene is smaller or equal than a chosen minimum ($CV^2 \leq 0.25$). After estimating the coefficients (see Supplementary Note 6 of Brennecke *et al.*, 2013 for an in-depth explanation of the statistical modeling), p-values for each gene were obtained from the cumulative distribution function of the chi-squared-distribution using the *pchisq()* function of the *stats* R package. Finally, we selected genes with an FDR<0.01 after multiple testing correction.

Correlations

Correlation values from plots generated by the *LSD* package were generated automatically using the *comparisonplot()* function with cor=T. Spearman's rho values are displayed on the bottom right corner of the plots unless otherwise stated. Gene-pair expression correlation (normalized by total number of RNA) was computed using the *cor.test()* function from the *stats* R package with default parameters. Cell-to-cell genome-wide and galactose metabolism variable genes correlation in gene expression was computed using the *cor()* function from the *stats* R package with default parameters.

Marker genes

We identified cluster specific genes using the *FindAllMarkers()* function from the *Seurat* package in R. We only tested genes detected in at least 25% of the cells of one of the clusters. Among these genes, we tested those that differ a minimum in the fraction of detection between the two clusters. For testing, we used a likelihood-ratio test for single cell gene expression that considers changes in the fraction of cells expressing a gene together with differences in the quantity of this gene among cells (McDavid, A. *et al.* 2013)36 (Parameters: *min.pct*=0.25, *min.diff.pct*=0.25, *only.pos*=T, *test.use*="bimod").

Normalization

Median ratio normalization: Normalization of raw UMI counts was done using the *DESeq2* package (v.1.18.1) in R (Love *et al.*, 2014)37. The *estimateSizeFactorsForMatrix()* function was used to compute cell-wise size factors for yeast RNA and for ERCC separately. Then, these size factors were applied to the raw UMI counts to obtain normalized expression values.

ERCC normalization: To estimate the total number of molecules per gene (as represented in Supplementary Figure 1i), the detection efficiency of the libraries was computed using ERCC data. We estimated the efficiency of detection as the average number of detected ERCCs per cell over the initial number of spiked-in ERCCs (5000 molecules). Then, we divided each gene's raw UMI count by this value.

Sensitivity of the method

To evaluate the sensitivity of yscRNA-seq, we measured the 50% detection probability of ERCC molecules. This measure, as proposed in Svensson *et al.*, 201710, is obtained through a binomial logistic regression model. In this model, the detection of ERCC molecules is considered as a function of the initial number of ERCC molecules available in the sequencing reaction. This model estimates the number of molecules required to achieve a 50% detection probability of ERCC.

For the binomial logistic regression, we used the *glm(family=binomial(logit))* function from the *stats* R package. The ERCC sensitivity computation pipeline was adapted from Bagnoli *et al*., 201811. We compared our 50% detection probability against the datasets from Svensson *et al* 201710 and Bagnoli *et al*., 201811.

Growth curves

ySMN235 strains were grown in YPD to mid-exponential log phase and subjected to sorting as described above. Two groups of 10,000 cells (top 2%, or random) were sorted based on Gal3-GFP intensity in 1 ml of YP media. Cells were immediately grown in duplicate in YPD (2% glucose) or YPGal (2%) for 50 hours in a Synergy H1 plate reader (30 °C, orbital shaking). Growth was measured by optical density at OD_{660} every 1.5 hours.

Data visualization

Data were visualized using R programming language and the packages *ggplot2* (v2.2.1), *plotly* (v4.8.0), *LSD* (v4.0-0), *Seurat* (v2.3.1), *geneplotter* (v1.56.0), *ggpubr* (v0.1.6.999), *VennDiagram* (v1.6.20), *cowplot* (v0.9.2) and *Superheat* (v1.0.0). Color palettes used were taken from the *viridis* (v0.3.0), *wesanderson* (v0.3.6) and *RColorBrewer* (v1.1-2) packages. tSNE visualizations were generated with *Seurat*38 using the default Bames-Hut implementation from *Rtsne*. We used PCA as a dimensionality reduction (dimensions 1 to 10) and default parameters (perplexity=30, theta=0.5 and Euclidian distances computation) for all the tSNE plots. *FlowJo* was used to analyze and generate FACS data.

Statistical information

To test for differences in the correlation of expression of pairs of genes against random pairs of genes, we used the one-tailed Wilcoxon signed-rank test with the *wilcox.test()* function from the *stats* R package. Functional enrichment analysis was performed with *gProfileR* (v0.6.6), which uses a hypergeometric test (Fisher's exact test) with default g:SCS multiple testing correction39.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Sten Linnarsson for kindly providing reagents during the initial tests with yscRNA-seq. We thank the Protein Expression and Purification Core Facility at EMBL, Bianca Hennig and Lars Velten for kindly providing in-house purified Tn5. We thank René Böttcher for fruitful discussions. We thank Derek Caetano-Anollés for editing and refining the manuscript. M.N.R. was a recipient of an EMBO long-term fellowship (Stanford University) and later of a Maria de Maeztu Postdoctoral Fellowship (Doctores Banco de Santander-María de Maeztu at Universitat Pompeu Fabra). P.L. is a recipient of a FI Predoctoral Fellowship (Generalitat de Catalunya). This work was supported by the National Institutes of Health (NIH) and a European Research Council Advanced Investigator Grant (AdG-294542) to L.M.S. and the National Key Research and Development Program of China (2017YFC0908405) to W.W. The study also was supported by grants from the Spanish Ministry of Economy and Competitiveness (BFU2015-64437-P and FEDER, BFU2014-52125-REDT, and BFU2014-51672-

REDC to F.P.; BFU2017-85152-P and FEDER to E.N.), the Catalan Government (2017 SGR 799), the Fundación Botín, and by the Banco Santander through its Santander Universities Global Division to F.P. and the Unidad de Excelencia Maria de Maeztu, MDM-2014-0370. F.P. is recipient of an ICREA Acadèmia (Generalitat de Catalunya).

References

1. Picelli S. Single-cell RNA-sequencing: The future of genome biology is now. RNA Biol. 2017; 14:637–650. [PubMed: 27442339]
2. Tang F, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6:377–382. [PubMed: 19349980]
3. Miura F, et al. Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. BMC Genomics. 2008; 9:574. [PubMed: 19040753]
4. Gasch AP, et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. PLoS Biol. 2017; 15
5. Saint M, Bertaux F, Tang W, Sun X-M, Game L. Single-cell phenotyping and RNA sequencing reveal novel patterns of gene expression heterogeneity and regulation during growth and stress adaptation in a unicellular eukaryote. bioRxiv. 2018; doi: 10.1101/306795
6. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature. 2013; 497:127–31. [PubMed: 23615609]
7. Pelechano V, Wei W, Steinmetz LM. Widespread co-translational RNA decay reveals ribosome dynamics. Cell. 2015; 161:1400–1412. [PubMed: 26046441]
8. Hennig BP, et al. Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. G3 (Bethesda, Md.); Genes|Genomes|Genetics. 2018; 8:79–89.
9. Svensson V, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017; 14:381–387. [PubMed: 28263961]
10. Bagnoli JW, et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat Commun. 2018; 9
11. Grün D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014; 11:637–640. [PubMed: 24747814]
12. Xu Z, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009; 457:1033–7. [PubMed: 19169243]
13. Wei W, et al. Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proc Natl Acad Sci U S A. 2007; 104:12825–12830. [PubMed: 17652520]
14. David L, et al. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci. 2006; 103:5320–5325. [PubMed: 16569694]
15. Pelechano V, Wei W, Jakob P, Steinmetz LM. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. Nat Protoc. 2014; 9:1740–59. [PubMed: 24967623]
16. Granovskaia MV, et al. High-resolution transcription atlas of the mitotic cell cycle in budding yeast. Genome Biol. 2010; 11:R24. [PubMed: 20193063]
17. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 2004; 428:617–624. [PubMed: 15004568]
18. Xu Z, et al. Antisense expression increases gene expression variability and locus interdependency. Mol Syst Biol. 2011; 7:468. [PubMed: 21326235]
19. Lenstra TL, Coulon A, Chow CC, Larson DR. Single-Molecule Imaging Reveals a Switch between Spurious and Functional ncRNA Transcription. Mol Cell. 2015; 60:597–610. [PubMed: 26549684]
20. Murray SC, et al. Sense and antisense transcription are associated with distinct chromatin architectures across genes. Nucleic Acids Res. 2015; 43:7823–7837. [PubMed: 26130720]
21. Aldea M, Jenkins K, Csikász-Nagy A. Growth Rate as a Direct Regulator of the Start Network to Set Cell Size. Front Cell Dev Biol. 2017; 5:57. [PubMed: 28603712]
22. Turner JJ, Ewald JC, Skotheim JM. Cell Size Control in Yeast. Curr Biol. 2012; doi: 10.1016/j.cub.2012.02.041

23. Schmoller KM, Turner JJ, Kõivomägi M, Skotheim JM. Dilution of the cell cycle inhibitor Whi5 controls budding-yeast cell size. Nature. 2015; 526:268–72. [PubMed: 26390151]

24. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013; 10:1093–5. [PubMed: 24056876]

25. Velten L, et al. Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol. 2017; 19:271–281. [PubMed: 28319093]

26. Zajac P, Islam S, Hochgerner H, Lönnerberg P, Linnarsson S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. PLoS One. 2013; 8

27. Spellman PT, et al. Comprehensive identification of cell cycle – regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998; 9:3273–3297. [PubMed: 9843569]

28. Teste M-A, Duquenne M, François JM, Parrou J-L. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in Saccharomyces cerevisiae. BMC Mol Biol. 2009; 10:99. [PubMed: 19874630]

29. Houser JR, et al. An improved short-lived fluorescent protein transcriptional reporter for Saccharomyces cerevisiae. Yeast. 2012; 29:519–530. [PubMed: 23172645]

30. Huber F, et al. Protein Abundance Control by Non-coding Antisense Transcription. CellReports. 2016; 15:2625–2636.

31. Venturelli OS, Zuleta I, Murray RM, El-Samad H. Population Diversification in a Yeast Metabolic Program Promotes Anticipation of Environmental Shifts. PLoS Biol. 2015; 13:1002042.

32. Wang J, et al. Natural Variation in Preparation for Nutrient Depletion Reveals a Cost–Benefit Tradeoff. PLoS Biol. 2015; 13

33. Janke C, et al. A versatile toolbox for PCR-based tagging of yeast genes: New fluorescent proteins, more markers and promoter substitution cassettes. Yeast. 2004; 21:947–962. [PubMed: 15334558]

34. Islam S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2013; 11:163–166. [PubMed: 24363023]

35. Pelechano V, Wei W, Steinmetz LM. Genome-wide quantification of 5’-phosphorylated mRNA degradation intermediates for analysis of ribosome dynamics. Nat Protoc. 2016; 11:359–76. [PubMed: 26820793]

36. McDavid A, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics. 2013; 29:461–467. [PubMed: 23267174]

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. [PubMed: 25516281]

38. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018; 36:411–420. [PubMed: 29608179]

39. Ri Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res Web Serv. 2007; 35:193–200.

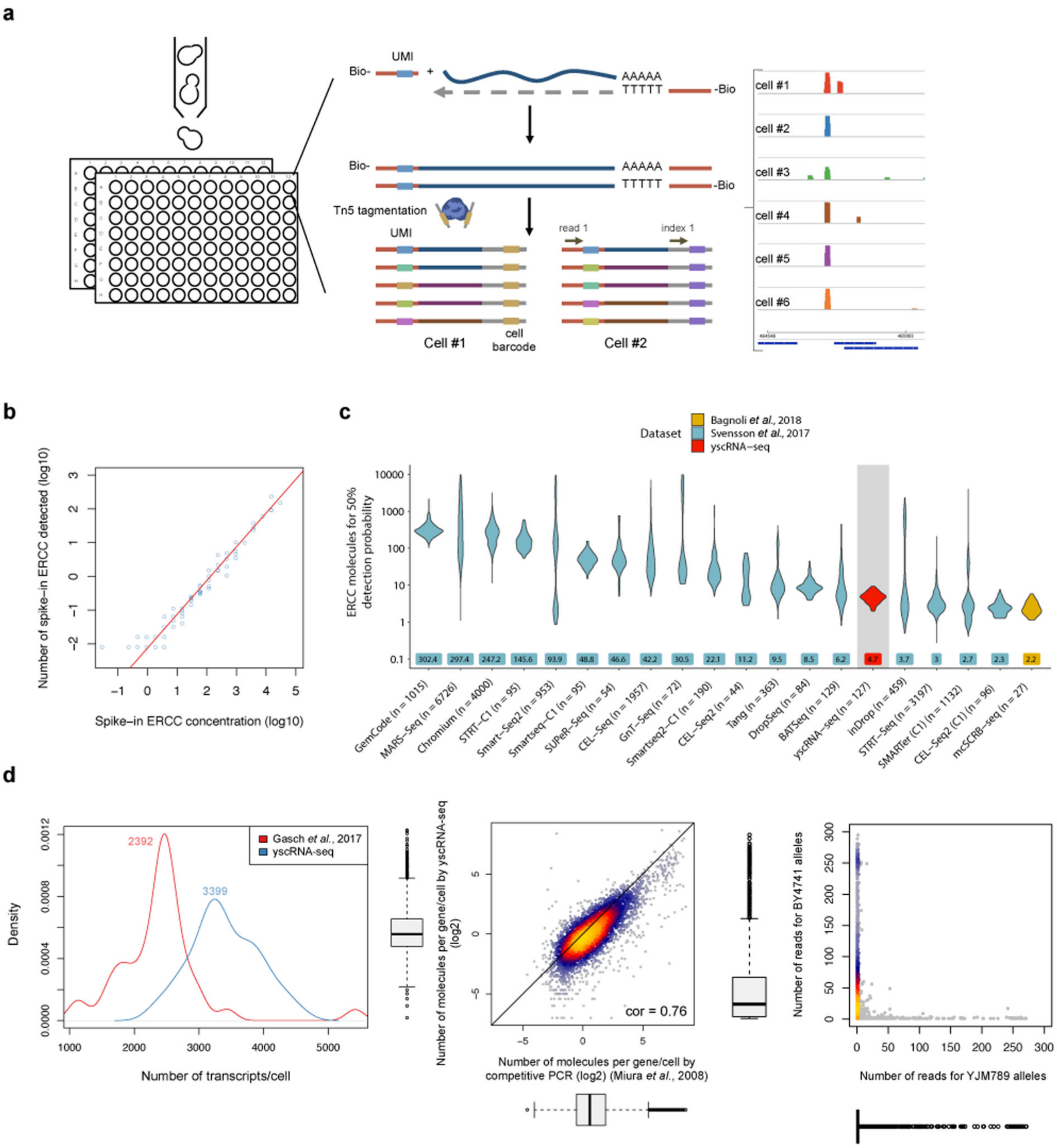

Figure 1. Absolute transcriptome quantification of single yeast cells by using yscRNA-seq

(a) Schematic representation of the yscRNA-seq workflow and representative example. Full-length cDNA libraries were generated from oligo dT and Unique Molecule Identifiers (UMI)-containing template-switching oligonucleotides (TSO). Following second-strand synthesis, double-stranded cDNA (dscDNA) libraries were tagmented with Tn5-loaded cell-specific barcodes. Single-strand cDNA (sscDNA) was obtained by briefly denaturing dscDNA bound to streptavidin beads, followed by high-throughput sequencing with custom primers. The right panel represents a histogram of reads at the *SCC4* locus for individual cells (rows). ORFs are indicated as blue boxes. **(b)** ERCC reads were mapped to the provided sequence from the manufacturer with a modification on the 5'-end to include the upstream restriction site used for ERCC expression (see Methods). The plot shows the correlation of unique ERCC molecules ($\log_{10}$) detected by yscRNA-seq and the expected spike-in concentration provided by the manufacturer. **(c)** Comparison of yscRNA-seq sensitivity to other scRNA-seq methods. The violin plots show the probability density function across several scRNA-seq methods. For each approach, the distribution of the 50% detection probability9,10 of ERCC molecules is plotted for the indicated number of cells (*n,* parentheses). Highlighted rectangles show the corresponding median value for each distribution. **(d)** Distribution of the total number of detected transcripts per cell by yscRNA-seq or Gasch *et al.*, 20174. The mean of the distribution for each method is shown. **(e)** Spearman correlation of detected number of molecules ($\log_2$) by competitive PCR3 (x-axis) versus mean UMI/gene from yscRNA-seq (y-axis) (n=3211). Correlation value is shown in the bottom right corner. **(f)** Comparative analysis of yscRNA-seq library-reads containing YJM789 SNP (x-axis) or BY4741 (y-axis) from unindexed random sorted cells of a 1:1 mixed sample containing YJM789 and BY4741 for a 96 well plate.

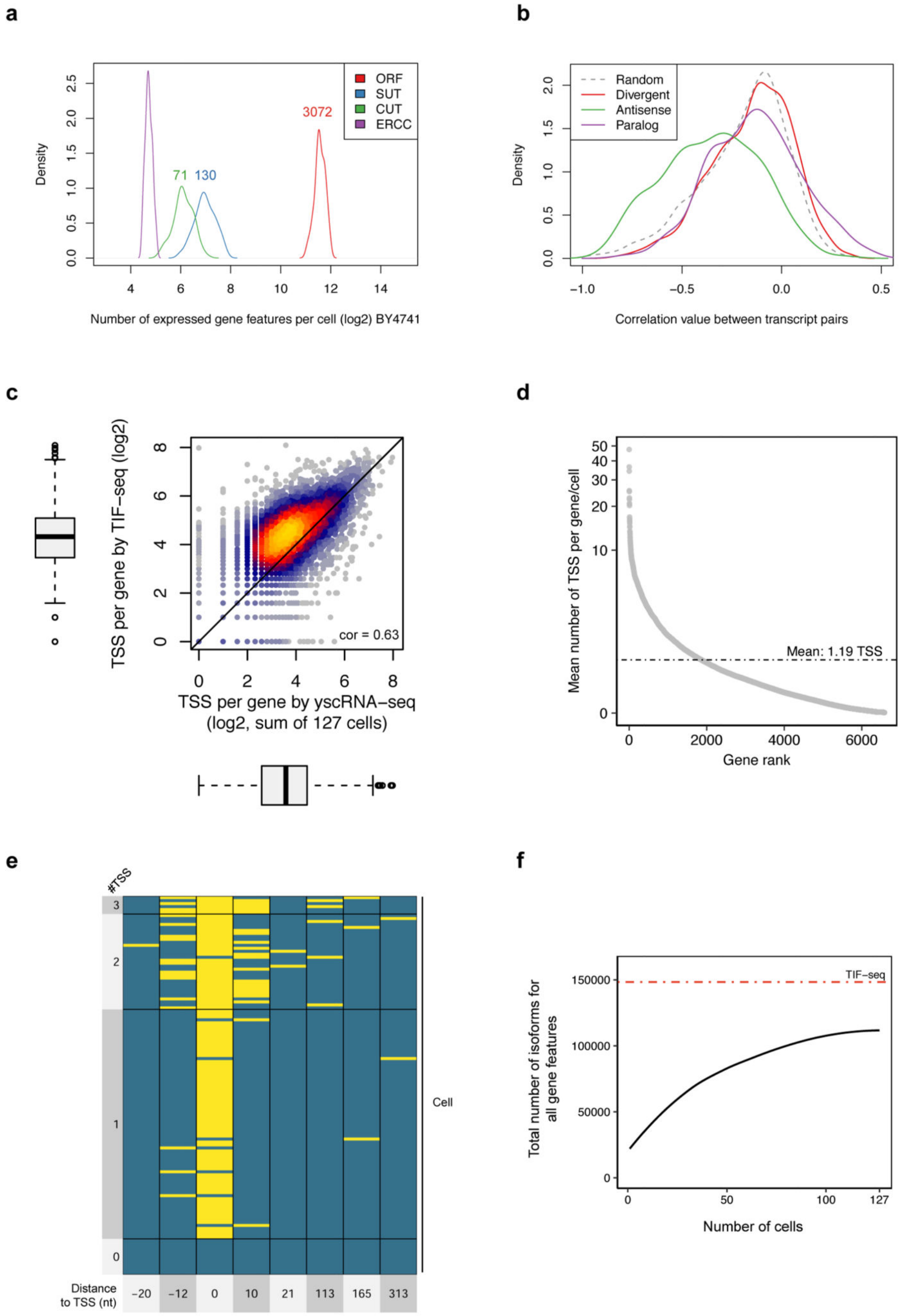

Figure 2. yscRNA-seq as a tool to quantitatively profile transcriptional architectures and TSS variation.

(a) Average distribution of all RNA species counted from yscRNA-seq libraries prepared from 127 BY4741 cells exponentially grown in YPD. Gene feature annotation was done based on Xu *et al.*, 200912. Density represents the distribution of genes expressed (≥1 UMI). The mean of the distribution for each gene feature is shown. **(b)** Distribution of genome-wide Pearson correlation values between transcripts in divergent orientation (red line, n=2555, one-tailed Wilcoxon test p-value 2.63e-09), annotated ORFs with antisense transcripts originating from their 3’-UTR (green line, n=202, one-tailed Wilcoxon p-value 1.45 e-19), paralogs (purple line, n=370) or a subset of random genes (grey dashed line, n=1846, one-tailed Wilcoxon test p-value 3.6 e-09). Classification of bidirectional and sense-antisense transcripts was obtained from Xu *et al.*, 200912 and paralogs from Kellis *et al.*, 200417. **(c)** Spearman correlation of the number of TSS isoforms per gene identified by yscRNA-seq and TIF-seq6. Detected yscRNA TSSs (x-axis) were obtained by pooling all yscRNA-seq libraries for all BY4741 cells (n=127) and comparing them to TIF-seq6. Correlation value is shown in the bottom right corner. **(d)** Ranked mean number of TSSs. Each dot represents the mean number of TSSs detected per gene per cell for BY4741 (127 cells) for genes detected in at least one cell (252/7272 genes were not detected). Mean of all TSSs is indicated with a dotted line. **(e)** Representative example of TSS usage in single cells. The binary heatmap represents the position (x-axis) and the usage (yellow) of the identified TSSs over *YLL014W* across all BY4741 yscRNA-seq libraries obtained from two biological replicates. A TSS was considered as ‘used’ if at least 2 UMIs/position were detected. Each row represents an individual cell, and each column represents the position to the TSS identified by TIF-seq6. **(f)** Cumulative sum of the number of TSSs used across BY4741 (y-axis) in comparison to the total number of 5’-TSS observed by TIF-seq6 (y-axis, approximately 148,000 5’-isoforms).

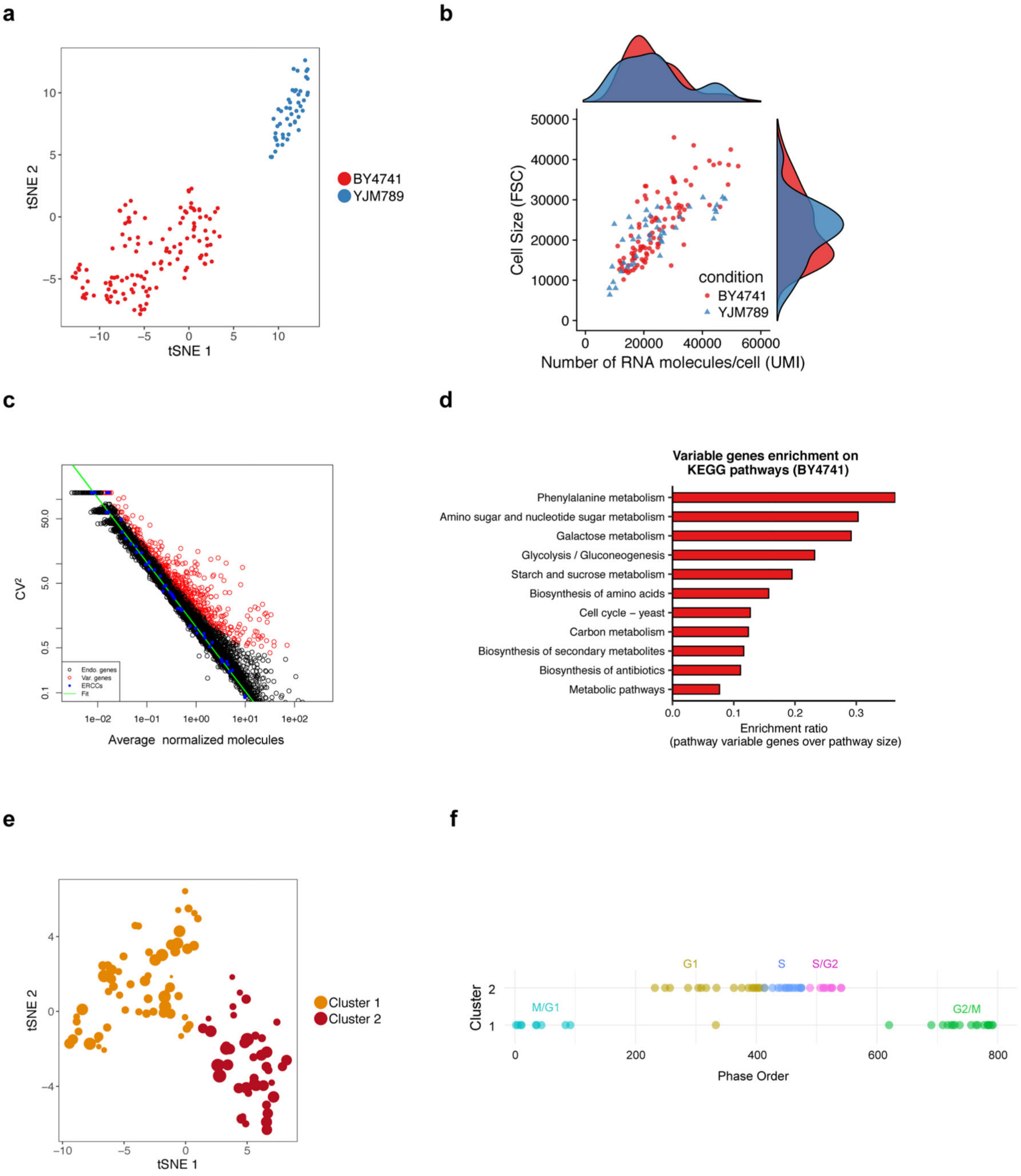

Figure 3. yscRNA-seq reveals a high-resolution map of the transcriptional heterogeneity within clonal yeast populations.

(a) tSNE plot using whole transcriptome data obtained by yscRNA-seq for the indicated strains (n=127 BY4741 and n=48 YJM789). **(b)** Correlation of cell size and transcriptome abundance. Correlation of number of RNA molecules/cell by yscRNA-seq (x-axis) with cell size (y-axis). Cell size units represent the FSC-value determined during index sorting for the indicated yeast backgrounds. **(c)** Detection of variable genes. Squared coefficient of variation (CV^2) of normalized UMI counts versus mean normalized expression (UMI) for all genes are shown in black across BY4741 (n=127). Variable genes are highlighted in red, and ERCCs are shown in blue. Inference of variable genes was done by taking into account technical noise from ERCC spike-ins as previously reported24. Genes that deviate from technical noise for whose biological coefficient of variation is more than our chosen minimum (CV^2>0.25) were considered as variable. Variable genes are highlighted in red, ERCCs are shown in blue and solid green line represents the technical noise fit. **(d)** KEGG-pathway enrichment of highly variable genes identified in (c). Categories are ranked based on effect size (total number of enriched genes over total number of annotated genes). **(e)** Clustering of 127 BY4741 cells based on genome-wide expression matrix. Each dot corresponds to a cell, with dot size representing cell size (FSC-value). Clusters were generated with *Seurat* using 0.6 as the value for resolution parameter. **(f)** Cell cycle stage of clusters generated in (e). Each dot represents the expression of cluster-specific genes ordered on the x-axis, based on their peak expression through cell cycle phases, determined by Spellman *et al*27.

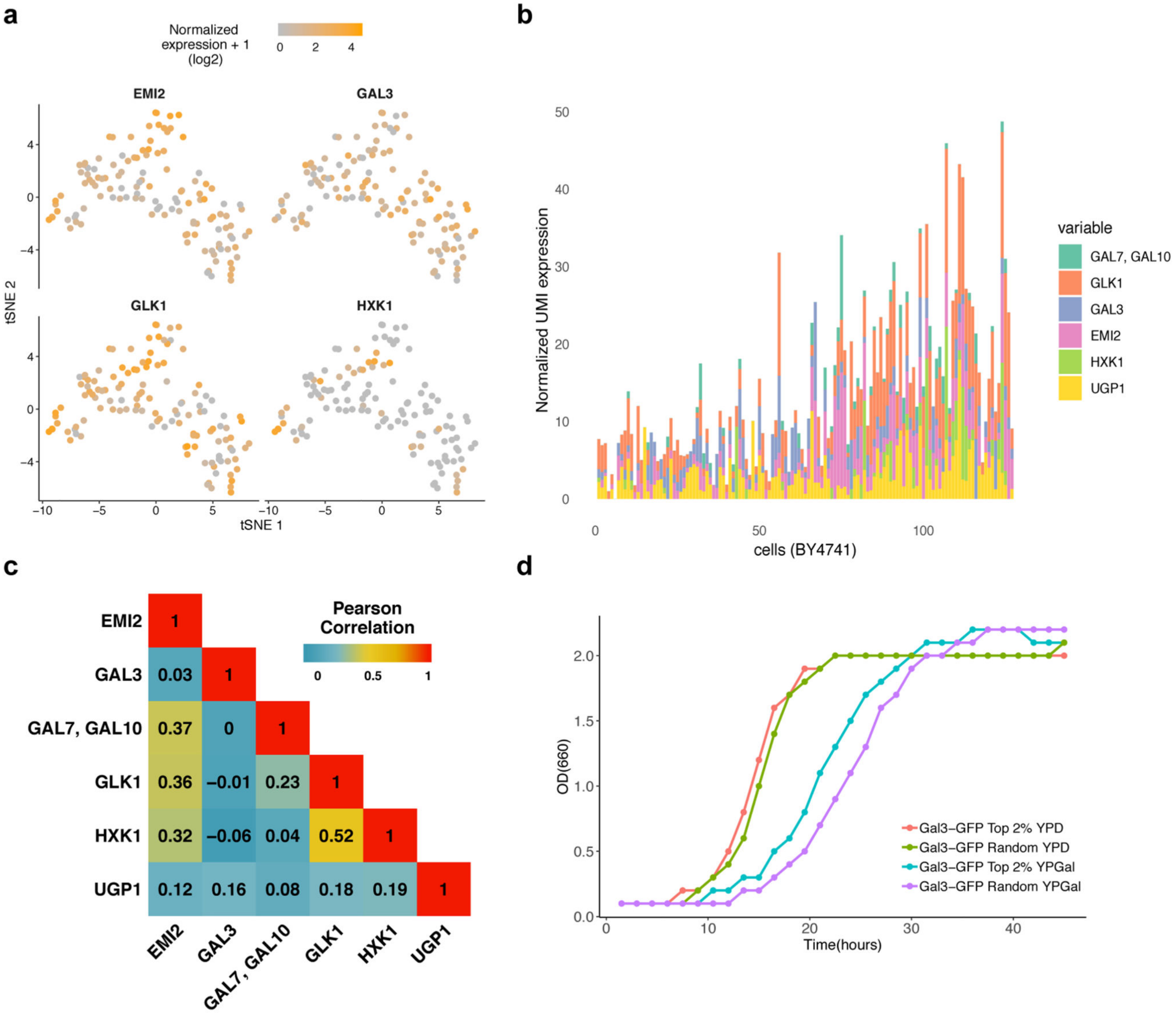

Figure 4. Functional consequences of stochastic gene expression.

(a) Overlay over the tSNE plot from Figure 3e of four representative variable genes from the galactose metabolism KEGG category. Each dot represents cells in the BY4741 dataset (127 cells); the color scale represents the normalized UMI expression, $\log_2$(UMI+1). **(b)** Expression of all variable galactose-related genes. Each stacked bar represents the normalized UMI expression for the indicated genes for each BY4741 cell. Bars are ordered based on their sorted position in the plate. **(c)** Heatmap representing the pair-wise Pearson correlation of variable galactose-related genes using the normalized number of unique molecules per cell ($\log_2$) across the BY4741 dataset (n=127). Warmer colors indicate a stronger correlation and colder colors represent a weaker correlation. **(d)** ySMN235 (BY4741 Gal3-GFP) was exponentially grown in YPD. Two groups of 10,000 cells were sorted based on their GFP intensity in YPD. Groups containing Top-2% GFP (Gal3-GFP Top-2%) or a random sort (Gal3-GFP Random) were sorted by FACS into YP media with no sugar, and quickly inoculated into YPD or YPGal. Growth curves over 48 hours were performed with a Synergy H1 plate reader. Data were obtained by measuring optical density (OD660) every 1.5 hours.