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

Mariona Nadal-Ribelles, Saiful Islam, Wu Wei, Pablo Latorre, Michelle Nguyen, Eulàlia de Nadal, Francesc Posas and Lars M. Steinmetz

Single-cell RNA sequencing 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 that is 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-abundance, noncoding RNAs and at least half of the protein-coding genome in each cell. In clonal cells, we observed a negative correlation for the expression of sense-antisense pairs, whereas paralogs and divergent transcripts co-expressed. By combining yscRNA-seq with index sorting, we uncovered a linear relationship between cell size and RNA content. Although we detected an average of ~3.5 molecules per gene, the number of expressed isoforms is restricted at the single-cell level. Remarkably, the expression of metabolic genes is highly variable, whereas their stochastic expression primes cells for increased fitness towards the corresponding environmental challenge. These findings suggest that functional transcript diversity acts as a mechanism that provides a selective advantage to individual cells within otherwise transcriptionally heterogeneous populations.
Articles

sorting, we applied yscRNA-seq to 285 individual yeast cells (two plates for strain BY4741 and one plate for YJM789; see Methods). After ethanol evaporation, the cells were lysed in a buffer containing zymolyase and 5,000 molecules of external RNA control consortium (ERCC) transcripts. A 5′-biotinylated template-switching oligonucleotide (TSO) containing P5 and a unique molecular identifier (UMI) was used to generate the first strand. Full-length double-stranded cDNA (dscDNA) libraries were amplified with a limited number of PCR cycles and size distribution was validated by Bioanalyzer profiling (Supplementary Fig. 1b). Cell-specific adapters were introduced by tagmentation using homemade TN5 (ref. 4) preloaded with adapters (see Methods). This substantially 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 (Fig. 1a). The size distribution was assessed before sequencing (Supplementary Fig. 1c).

A first-sequencing read (read1) was used to obtain the UMI and the gene identity, whereas a shorter read (read2) was used to retrieve cellular barcodes. After pre-processing and alignment of the reads (see Methods), we applied a filter to consider only UMIs with at least three reads for analysis (Supplementary Fig. 1d). In addition, we restricted the analysis to cells with more than 500,000 uniquely-mapped reads and over 1,000 expressed genes (≥ 1 UMI gene), which removed low-quality cells with a higher ratio of mitochondrial RNA (Supplementary Fig. 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 the expected unique ERCC molecules and observed a linear correlation across the entire dynamic range (Fig. 1b) with a 15.5–25.8% efficiency (an average of 1,290 ERCC molecules out of 5,000 spiked-in and 775 molecules after filtering). 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 consistently identified the TSS boundaries (Supplementary Fig. 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 a 50% detection probability13. We found that yscRNA-seq was among the most sensitive methods, with only a median of 4.7 molecules required for a 50% detection probability (Fig. 1c). Similarly, we compared yscRNA-seq to the only available yeast scRNA-seq dataset from Gasch et al.4, which used the C1 System from Fluidigm. Overall, yscRNA-seq yielded a higher number of genes per cell (3,399 versus 2,392) and a good genome-wide correlation despite the scRNA-seq from Gasch et al.4 not being strand-specific (Spearman’s correlation = 0.71; Fig. 1d and Supplementary Fig. 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’s correlation = 0.76; Fig. 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 increased1 (Supplementary Fig. 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 efficiencies13,14.

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 condition15. Interestingly, the comparison with the tiling array data resulted in the highest correlation to other methods (Spearman’s correlation = 0.83) due to the strand-specific nature of both approaches (Supplementary Fig. 1j).

Overall, yscRNA-seq quantitatively recapitulated gene expression from two independent studies and performed within the range of 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 whether 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 that was isolated from an HIV patient16. At the genomic level, YJM789 differs from the reference strain in approximately 60,000 single nucleotide polymorphisms and 6,000 indels17. We plotted the allele frequency of reads that mapped 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 (two or more cells per reaction) was lower than 1%, thus confirming that yscRNA-seq libraries originated from individual cells (Fig. 1f and Supplementary Fig. 1k).

For a single wild-type S. cerevisiae cell, the majority of expressed genes belonged to open reading frames (ORFs; n = 3,072), 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 (Fig. 1a and Supplementary Table 2). The ratio of ORFs to ncRNAs was similar to those reported from bulk studies14–16, but the distribution of the 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 analysed the expression matrix for all genes in each individual cell, only 252 of 7,272 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 in individual cells or across all samples (Pearson’s correlation coefficient = 0.66–0.83; Supplementary Fig. 2a,b). Together, these data suggest that at least half of the yeast genome is expressed in a complex pattern in each individual cell.

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

To assess the genome-wide transcriptional architectures by exploiting the sensitivity and strand-specific nature of yscRNA-seq, we analysed the correlation between paralogous genes (those that arose from whole-genome duplication17), sense–antisense pairs (SAPs) and divergent transcripts18 (Fig. 2b). The correlation was only computed for gene pairs in which at least one gene in the pair was detected in ≥3 cells. The expression of SAPs showed a strong negative correlation compared to random gene pairs (P = 1.45 × 10−19, one-tailed Wilcoxon test). This is consistent with previous data that show anticorrelated expression of SAPs from bulk samples18.

To explore the potential of yscRNA-seq, we investigated GALI, which is one of the most studied SAPs. In bulk samples under glucose growth conditions, the GALI sense transcript is repressed and a ncRNA (CUT445), which originates from the 3′-untranslated region (3′ UTR) that is constantly expressed in glucose and galactose that overlaps the GALI TSS19,20. For the majority of our single cells
Fig. 1 | Absolute transcriptome quantification of single yeast cells using yscRNA-seq. 

a, Schematic of the yscRNA-seq workflow and a representative example. Full-length cDNA libraries were generated from biotinylated (Bio) oligo(dT)- and UMI-containing TSOs. Following second-strand synthesis, dscDNA libraries were tagmented with Tn5-loaded cell-specific barcodes. Single-stranded cDNA was obtained by briefly denaturing dscDNA bound to streptavidin beads, which was followed by high-throughput sequencing with custom primers. The right panel shows representative histograms of reads at the SCC4 locus of individual cells (rows). ORFs are represented by blue boxes. 

b, ERCC reads were mapped to the sequence provided by the manufacturer with a modification at the 5′ end to include the upstream restriction site used for ERCC expression (see Methods). The plot shows the correlation between the unique ERCC molecules 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). The highlighted rectangles show the corresponding median value for each distribution. 

d, Distribution of the total number of detected transcripts per cell determined by yscRNA-seq and Gasch and colleagues4. The mean of the distribution for each method is shown. 

e, Spearman’s correlation between the number of molecules detected by competitive PCR3 and the mean UMI per gene from yscRNA-seq (n = 3,211). 

f, Comparative analysis of yscRNA-seq library-reads containing YJM789 single nucleotide polymorphism or BY4741 from unindexed random sorted cells of a 1:1 mixed sample containing YJM789 and BY4741 for a 96-well plate. For e, f, warmer colours indicate a stronger correlation and colder colours represent a weaker correlation. Boxplots represent the distribution of the data for the corresponding axes respectively.
Fig. 2 | yscRNA-seq as a tool to quantitatively profile transcriptional architectures and TSS variation. **a**, Distribution of all RNA species counted from yscRNA-seq libraries prepared from 127 BY4741 cells that were exponentially grown in YPD medium. The gene feature annotation was conducted based on Xu and colleagues12. Density represents the distribution of the expressed genes (≥1 UMI). The mean of the distribution for each gene feature is shown; SUTs (stable unannotated transcripts) and CUTs (cryptic unstable transcripts). **b**, Distribution of genome-wide Pearson’s correlation values between transcripts in divergent orientations (n = 2,555; P = 2.63 × 10−9, one-tailed Wilcoxon test), annotated ORFs with antisense transcripts originating from their 3’-untranslated region (n = 202; P = 1.45 × 10−14, one-tailed Wilcoxon test), paralogs (n = 370; P = 3.6 × 10−9, one-tailed Wilcoxon test) and a subset of random genes (n = 1,846). The classification of the bidirectional and sense-antisense transcripts was obtained from Xu et al.12 and the paralogs from Kellis and colleagues17. **c**, Spearman’s correlation between the number of TSS isoforms per gene that were identified by yscRNA-seq and by TIF-seq6. The TSSs detected by yscRNA (x axis) were obtained by pooling all yscRNA-seq libraries for all BY4741 cells (n = 127) and comparing them to TIF-seq6. **d**, Ranked mean number of TSSs. Each dot represents the mean number of TSSs detected per gene per cell for BY4741 (n = 127 cells) for genes that were detected in at least one cell (252 of 7,272 genes were not detected). The mean of all TSSs is indicated by a dashed 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 that were obtained from two biological replicates. A TSS was considered to be ‘used’ if at least two UMIs per position were detected. Each row represents an individual cell and each column represents the position relative to the TSS identified by TIF-seq6. **f**, Cumulative sum of the number of TSSs used (y axis) across BY4741 (x axis) in comparison to the total number of 5’ TSSs observed by TIF-seq6 (red dashed lines; approximately 148,000 5’ isoforms).
cultured in YPD media, we could neither detect the sense nor the antisense transcript—remarkably, for those in which we could, the expression pattern was mutually exclusive (Supplementary Fig. 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 ($P = 2.63 \times 10^{-9}$ and $3.6 \times 10^{-9}$, respectively, one-tailed Wilcoxon test; Fig. 2b). These results indicate that transcription events reported in bulk reproduce in single cells, whereas the bimodal expression for the GAL1-CUT445 SAP is unresolved in bulk assays.

One of the main layers of transcriptional complexity arises from transcript isoform diversity, where an average of 26 isoforms per gene has been reported in yeast by TIF-seq. To characterize the TSS variability in single cells, we sought to identify TSSs by applying a similar criterion to that used for TIF-seq, in which the highest-expressed position defined the major isoform. Secondary isoforms were iteratively assigned if reads occurred outside a centered 15 nt window centred from the previous isoform (see Methods). To make the data comparable across both studies, we assessed the diversity of TSSs by considering only TSS positions with ≥2 UMI s in yscRNA-seq and ≥2 reads in TIF-seq for the same background and condition. From the sum of isoforms per gene across all cells, we detected a linear correlation between the TSS numbers per gene determined by TIF-seq and yscRNA-seq (Spearman’s correlation = 0.63; Fig. 2c). On average, each gene expressed 3.46 UMI s (Supplementary Fig. 2e); however, the mean number of TSSs per gene across all cells was 1.19 (Fig. 2d). When we clustered cells on the basis of the TSS variability of a representative gene (YL0144W with a mean of 3.09 UMI s 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-seq at Position 0), 32 cells co-expressed a second isoform from a different TSS and in some rare cases three isoforms were detected (Fig. 2e). As a control, the sum of the most-expressed isoforms of all BY4741 cells revealed a linear correlation with the expression of the second most-expressed isoform (Spearman’s correlation = 0.93). Furthermore, although the major isoform dominated the expression, the second isoform was still abundant within the population (Supplementary Fig. 2f). Overall, our results suggest that although a main isoform is preferred in an individual cell, TSS diversity between cells is achieved through the expression of select isoforms.

To understand how TSS variability in single cells leads to the observed population isoform diversity, we assessed the contribution of each cell by computing the cumulative sum of unique TSS isoforms per cell, comparing the result to the total number of TIF-seq 5′-TSS isoforms (a total of 148,000 5′ isoforms with ≥2 reads; Fig. 2f). We observed a logarithmic increase in isoforms that begins to plateau at ~100 cells without reaching saturation. These data support the idea 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 neighbour embedding (tSNE) analysis using the entire transcriptomes for all datasets. Analysis using tSNE revealed two distinct strain-specific clusters (BY4741 and YJM789; Fig. 3a), thus validating yscRNA-seq as a sensitive method for distinguishing different yeast strains.

The 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 growth. We observed a positive correlation for both groups ($P = 2.63 \times 10^{-9}$ and $3.6 \times 10^{-9}$, respectively, one-tailed Wilcoxon test; Fig. 2b). These results indicate that transcription events reported in bulk reproduce in single cells, whereas the bimodal expression for the GAL1-CUT445 SAP is unresolved in bulk assays.

To understand variability within the BY4741 population, we applied tSNE analysis. We identified two distinct clusters that were enriched for previously annotated cell-cycle-regulated genes (Fig. 3c and Supplementary Fig. 3e). We then phase-ordered the expression of cluster-specific genes based on previously annotated expression peaks. Cluster 1 contained cells that expressed early (M/G1) and late (G2/M) cell-cycle genes, whereas Cluster 2 was enriched for cells expressing G1, S and late S/G2 genes (Fig. 3f).

To rule out the contribution of batch effects, we projected both BY4741 replicates onto the tSNE and confirmed that the batch effects are negligible (Supplementary Fig. 3f). Our data suggests that, although phase-specific expression exists, the yeast cell cycle segregates cells into two subpopulations with distinct transcription signatures: those that correspond to cell-cycle entry/exit and those that correspond to progression through G1/S/G2.

As the majority of variable genes were related to metabolic processes, such as sugar and nucleotide metabolism, we assessed whether this variability was cell-cycle dependent. We overlaid the expression of variable galactose genes (GAL3, EM12, GLK1 and HXX1) onto the tSNE (Fig. 4a), as well as representative examples of the remaining KEGG-enriched pathways (Supplementary Fig. 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 medium) and have been used as a model to understand transcriptional noise and regulation. Interestingly, for BY4741, we detected the expression of at least one variable gene that belongs to the galactose pathway in virtually all cells (Fig. 4b). Although more than 80% of the cells growing exponentially in glucose displayed no expression of GAL3, 20% of cells had more than four unique GAL3 mRNAs and 1.5% had more than ten mRNAs—a expression level that is above the expected noise.

Reverse transcription and template switching are major sources of technical noise, whereas intrinsic biological noise can originate from the stochastic nature of transcription and the regulation of mRNA stability. To identify variable genes, we took advantage of the presence of ERCC spike-ins and classified genes as variable if their squared coefficient of variation (CV2) was higher than the technical and expected biological variance using a 10% false discovery rate (FDR; CV2 > 0.25). For exponentially growing wild-type BY4741, we identified 400 variable RNAs (FDR < 0.01; Fig. 3c). Of these, 100 were ncRNAs (38 CUTs and 62 SUTs), whereas 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 the incorporation of UMIs in yscRNA-seq occurs by template switching at the 5′ end, we ruled out gene-length-dependent effects of the observed CV2 (Supplementary Fig. 3c). Therefore, the 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 transcripts. Surprisingly, the largest groups of variable genes belonged to metabolic pathways, such as galactose and glycolysis/glucogenesis, among others (Fig. 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 Fig. 3d). These results suggest that variable genes are partially strain-dependent, whereas several genes might be intrinsically variable.

To understand variability within the BY4741 population, we applied tSNE analysis. We identified two distinct clusters that were enriched for previously annotated cell-cycle-regulated genes (Fig. 3c and Supplementary Fig. 3e). We then phase-ordered the expression of cluster-specific genes based on previously annotated expression peaks. Cluster 1 contained cells that expressed early (M/G1) and late (G2/M) cell-cycle genes, whereas Cluster 2 was enriched for cells expressing G1, S and late S/G2 genes (Fig. 3f).

To rule out the contribution of batch effects, we projected both BY4741 replicates onto the tSNE and confirmed that the batch effects are negligible (Supplementary Fig. 3f). Our data suggests that, although phase-specific expression exists, the yeast cell cycle segregates cells into two subpopulations with distinct transcription signatures: those that correspond to cell-cycle entry/exit and those that correspond to progression through G1/S/G2.

As the majority of variable genes were related to metabolic processes, such as sugar and nucleotide metabolism, we assessed whether this variability was cell-cycle dependent. We overlaid the expression of variable galactose genes (GAL3, EM12, GLK1 and HXX1) onto the tSNE (Fig. 4a), as well as representative examples of the remaining KEGG-enriched pathways (Supplementary Fig. 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 medium) and have been used as a model to understand transcriptional noise and regulation. Interestingly, for BY4741, we detected the expression of at least one variable gene that belongs to the galactose pathway in virtually all cells (Fig. 4b). Although more than 80% of the cells growing exponentially in glucose displayed no expression of GAL3, 20% of cells had more than four unique GAL3 mRNAs and 1.5% had more than ten mRNAs—an expression level that is above the expected noise.
To understand whether the expression of galactose-variable genes was coordinated, we analysed the pairwise expression of all galactose-metabolism variable genes. We did not observe any strong correlations between gene pairs, which suggests uncoordinated regulation of expression in single cells cultured in YPD (Fig. 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 green fluorescent protein (GFP) at the C-terminus, sorting two populations of 10,000 cells from YPD on the basis of GFP intensity (top 2% and a random-sort control). We then followed their growth in YPD or

Fig. 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 between cell size and transcriptome abundance. Correlation of number of RNA molecules (UMI) per cell by yscRNA-seq with cell size. The FSC values determined during index sorting for the indicated yeast backgrounds represent the cell sizes. For each axis the density distribution for each indicated strain is shown. c, Detection of variable genes. The CV² values across BY4741 (n = 127 cells) of the normalized UMI counts versus mean normalized expression (UMI) for endogenous genes, variable genes and ERCCs are shown. Inference of variable genes was performed by taking into account the technical noise from ERCC spike-ins as previously reported. Genes were considered to be variable when they deviated from technical noise and the biological coefficient of variation was more than our chosen minimum (CV² > 0.25). The solid green line represents the technical noise fit. d, KEGG-pathway enrichment of highly variable genes that were identified in c. The categories are ranked on the basis of effect size (total number of enriched genes over total number of pathway annotated genes). e, Clustering of 127 BY4741 cells based on the genome-wide expression matrix. Each dot corresponds to a cell, with dot size representing cell size (FSC value). The clusters were generated with Seurat using 0.6 as the value for the resolution parameter. f, Cell-cycle stage of the 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, as determined by Spellman and colleagues.
YPGal and measured Gal3–GFP fluorescence for both conditions (Supplementary Fig. 4b). Although all populations had a similar growth rate in YPD, the top 2% of cells expressing Gal3 in YPD displayed an increased fitness in galactose (Fig. 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 the quantitative profiling of gene expression and TSS variation in a strand-specific, cost-effective manner. YscRNA-seq enables cell profiling using minute amounts of RNA and represents a sensitive and inexpensive approach. In addition, FACS-based single-cell isolation allows cell parameters, such as cell size or reporter abundance, to be correlated to the transcriptome. YscRNA-seq provides a high-throughput approach to link phenotype to the 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 in single cells. In contrast, SAPs demonstrated a stronger anticorrelation of expression in single cells than previously reported in bulk samples\(^{18,30}\), thus establishing yscRNA-seq as a tool to resolve transcriptional architectures that are 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. In addition, the cumulative expression of approximately 100 cells virtually covered the entire transcriptome (>99% of the ORFs). On average, each gene expressed 3.46 transcript molecules per cell but primarily from one TSS isoform. These findings challenge the assumption of...
one mRNA molecule per gene per cell and instead suggest 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 per cell). As expected, these genes included cell-cycle-regulated transcripts that were clustered in only two sets with distinct signatures (mitotic entry/exit and G1/S/G2 progression) rather than phase-specific clusters. This suggests that the 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 shift31,32. Indeed, we tested the functional consequences of stochastic expression for one of the highly variable genes (that is, GAL3–GFP) and found that this variability provides a fitness advantage following a change in the carbon source. This observation suggests that, although 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 YM789 strains were used for yscRNA-seq libraries. BY4741 (MATa his3-D1 leu2-D0 met15-I D0 ura3-D0) and its derivative ySMN235 (GAL3–GFP–phhp1NT1) were obtained by genetic integration using a PCR-based strategy33.

Cell growth and isolation. Cells (BY4741 or YM789) were pre-inoculated in rich media and grown overnight until they reached OD600 = 1. The next morning, cells were collected at OD600 = 0.05 and grown for at least two full divisions. Before sorting, cells were diluted to OD600 = 0.01 and sonicated for three pulses of 0.5 s to remove aggregates (Sonicator Branson M1800). The cells were then filtered by passing the samples through cell strainer snap-cap tubes (BD Falcon). Live single cells were indexed (Sorted 96-well plates) into 96-well plates containing 5 µl absolute ethanol. The gates on the pulse-width were stringent to remove potential doublets. FSC and side scatter 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 a negative control.

For experiments using mixed strains of BY4741 and YM789 (Fig. 1f and Supplementary Fig. 1k), cells were grown independently as described above, except that strains were mixed at a 1:1 ratio before random sorting (unindexed) into 96-well plates.

yscRNA-seq library preparation. Sorted 96-well plates were left in a sterile hood to allow complete ethanol evaporation. The cell wall was then digested with a lysis buffer containing zymolyase (1 µl 1−1) and 5,000 molecules of ERCC. Cell lysis was performed for 10 min at 37 ºC, followed by 3 min at 72 ºC in 5 µl. The plates were immediately placed on ice and 5 µl reverse transcription reaction mix (RT-buffer and Invitrogen SuperScript II) was added to synthesize first strand cDNA from a 5′-biotinylated oligo(dT) primer (Bio-AATGATACGGCGACCACAGCTTGTAGGTCA TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT). The oligo(dT) primer contains a PvuI adapter, dscDNA libraries were tagmented with our homemade Tn5 (ref. 8) and (Supplementary Fig. 1b). To incorporate a cellular barcode and the Illumina P7 primer, the dscDNA library size was validated using a Bioanalyzer following the manufacturer's instructions (0.6:1 ratio) and eluted in 15 µl nuclease-free water. To generate strand-specific libraries, the samples were briefly denatured at 70 ºC for 10 min to remove the non-biotinylated strand. The strands were then quickly bound to magnets to remove the biotinylated strand and single-stranded cDNA was recovered from the supernatant. The resulting TSS-enriched single-stranded cDNA (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 and 16 pmol were hybridized and sequenced using a HiSeq Rapid Run to a sequencing depth of 1–2 x 106 reads per 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 (Exiqon) were spiked-in following the manufacturer’s instructions. The read1 primer is located upstream of the UMI and was used to obtain a 68bp 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). Locked nucleic acid primers (0.5 µl) were spiked-in following the manufacturer’s instructions for the read1 and index1 primers. UMI PCR-read1: +GAATAG+TACCCGGCCACCCA+CGGGG+T Index1: CTG+CT+CTT+ATT+CA+CA+CTTGA+CG+GA The + indicates the position of the locked nucleic acid. A detailed step-by-step protocol of yscRNA-seq can be found at protocols.io (https://doi.org/10.17594/protocols.io.t29eqh6).

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-Gs following the UMI nucleotides were also kept for analysis until up to 14 Gs were kept for analysis. Only reads that were mapped to the genome were kept for analysis. Reads that were not mapped to the genome were removed. 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. Notably, the ERCC sequences provided by the manufacturer do not contain the restriction sites used in the yscRNA-seq protocol of yscRNA-seq can be found at protocols.io (https://doi.org/10.17594/protocols.io.t29eqh6).

Molecule counting and TSS identification. For each UMI, the total number of reads was counted and UMIs supported with fewer than three 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 an approach similar to one previously reported (TIF-seq). 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. We then counted the number of UMIs and reads supporting these TSSs. Next, for each gene, we ordered all of the TSSs present in all cells (n = 127) according to expression. Then, subsequent TSSs were identified by considering the ordered set of TSSs in the following way: if the TSS was within a range of ±7 nt that had not been defined as a confident TSS in previous iterations, its position was used to define a new TSS. Following this criterion, the most-expressed TSSs were the first to be considered as confident TSSs; 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 overlapped. If a TSS was positioned within ±7 nt range of the number of UMIs and reads were divided and assigned to both TSSs. Finally, we only considered the confident TSS isoforms that were supported by at least two UMIs. The
Correlations. Correlation values from plots generated by the LSD package were generated automatically using the compareplot() function with cor=r.T. The Spearman's rho values are displayed on the bottom right corner of the plots unless otherwise stated. Gene-pair expression correlation (normalized to the 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 that were 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 ratio test (binomial) that considers changes in the fraction of cells expressing a gene together with differences in the quantity of this gene among cells (McDavid et al.,21; Parameters: min.pct=0.25, min.diff.pct=0.25, only.pos=T, test.use = 'binom').

Normalization. Median ratio normalization. Normalization of raw UMI counts was done using the DESeq2 package (v1.18.1) in R (Love and colleagues22). The estimateSizeFactorsForMatrix() function was used to compute cell-wise size factors for yeast RNA and for ERCC separately. These size factors were then 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 Fig. 1f), 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 (5,000 molecules). We then divided the raw UMI count of each gene 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.31, 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(formula=binomial(logit)) function from the stats R package. The ERCC sensitivity computation pipeline was adapted from Bagnoli and colleagues31. We compared our 50% detection probability against the Svensson et al.31 and Bagnoli et al.31 datasets.

Growth curves. The ySMN235 strains were grown in YPD to the mid-exponential log phase and subjected to sorting as described above. Two groups of 10,000 cells (top 2% and random) were sorted based on Ga3–GFP intensity in 1 ml VP media. The cells were immediately grown in duplicate in YPD (2% glucose) or YPGal (2%) for 5h in a Synergy H1 plate reader (30°C, orbital shaking). Growth was measured by optical density at OD600 every 1.5h.

Data visualization. Data were visualized using the R programming language and the packages ggplot2 (v2.2.1), plotly (v4.8.0), LFD (v4.0.0), Seurat (v2.3.1), geneplotter (v1.56.0), ggupp (v0.16.899), VennDiagram (v1.6.20), cowplot and Superheat (v1.0.0). The colour palettes used were taken from the viridis (v0.3.0), wesanderson (v0.3.6) and RColorBrewer (v1.1-2) packages. The TSNE visualizations were generated with Seurat31 using the default Barnes–Hut implementation from Rtse. We used principal component analysis (PCA) as a dimensionality reduction (dimensions 1 to 10) and default parameters (perplexity=30, theta=0.5 and Euclidian distances computation) for all of the TSNE plots. FlowJo was used to analyse and generate FACS data.

Statistical analysis. To test for differences in the correlation of expression of pairs of genes against random pairs of genes, we used the one-sided 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 correction31.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

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

Data availability

All data generated in this study has been uploaded to Gene Expression Omnibus under accession number GSE123392.

Received: 2 July 2018; Accepted: 7 December 2018;
Published online: 4 February 2019

References

1. Picelli, S. Single-cell RNA-sequencing: the future of genome biology is now. RNA Biol. 14, 637–650 (2017).
2. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382 (2009).
3. Miura, F. et al. Absolute quantification of the budding yeast transcriptome by qPCR from single-cell. Nucleic Acids Res. 37, 377–382 (2009).
4. Gasch, A. P. et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. PLoS Biol. 15, e2004050 (2017).
5. Saint, M., Berluxa, F., Tang, W., Sun, X.-M. & Game, L. Single-cell phenotyping and RNA sequencing reveal novel patterns of gene expression heterogeneity and regulatory heterogeneity in yeast. Nat. Commun. 9, 1400–1412 (2018).
6. Hennig, B. P. et al. Large-scale low-cost NGS library preparation using a robust TruSeq purification and tagmentation protocol. G3 8, 79–89 (2018).
7. Svensson, V. et al. Power analysis of single-cell RNA-seq sequencing experiments. Nat. Methods 14, 381–387 (2017).
8. Bagnoli, J. W. et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 11, 637–640 (2014).
9. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. Nature 457, 1033–1037 (2009).
10. Wei, W. et al. Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YM789. Proc. Natl Acad. Sci. USA 104, 12825–12830 (2007).
11. David, L. et al. A high-resolution map of transcription in the yeast genome. Proc. Natl Acad. Sci. USA 103, 5320–5325 (2006).
12. Pelecanos, V., Wei, W., Jakob, P. & Steinmetz, L. M. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. Nat. Protoc. 9, 1750–1759 (2014).
13. Granovskai, M. V. et al. High-resolution transcription atlas of the mitotic cell cycle in budding yeast. Genome Biol. 11, R24 (2010).
14. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428, 617–624 (2004).
15. Xu, Z. et al. Antisense expression increases gene expression variability and regulatory heterogeneity in yeast responding to stress. Nat. Protoc. 9, 1740–1759 (2014).
16. Schmoller, K. M., Turner, J. J., Kõivomägi, M. & Skotheim, J. M. Dilution of transcriptional locus interdependency. Mol. Cell 60, 981–990 (2015).
17. Turner, J. J., Ewald, J. C. & Skotheim, J. M. Cell size control in yeast. Nat. Rev. Mol. Cell Biol. 15, 289–297 (2014).
18. Xu, Z. et al. Antisense expression increases gene expression variability and locus interdependency. Mol. Biol. Cell 27, 271–281 (2017).
19. Lenstra, T. L., Coulon, A., Chow, C. C. & Larson, D. R. Single-molecule imaging reveals a switch between spurious and functional ncRNA transcription. Mol. Cell 60, 597–610 (2015).
20. Murray, S. C. et al. Sense and antisense transcription are associated with distinct chromatin architectures across genes. Nucleic Acids Res. 43, 7823–7837 (2015).
21. Aldea, M., Jenkins, K. & Csikász-Nagy, A. Growth rate as a direct regulator of cell size control in yeast. PLoS Biol. 8, e2004050 (2010).
22. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095 (2013).
23. Veltén, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. 19, 271–281 (2017).
26. Zajac, P., Islam, S., Hochgerner, H., Löwnerberg, P. & Linnarsson, S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS ONE* **8**, e65270 (2013).

27. Spellman, P. T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).

28. Teste, M.-A., Duquenne, M., François, J. M. & Parrou, J.-L. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*. *BMC Mol. Biol.* **10**, 99 (2009).

29. Houser, J. R. et al. An improved short-lived fluorescent protein transcriptional reporter for *Saccharomyces cerevisiae*. *Yeas* **29**, 519–530 (2012).

30. Huber, F. et al. Protein abundance control by non-coding antisense transcription. *Cell Rep.* **15**, 2625–2636 (2016).

31. Venturelli, O. S., Zuleta, I., Murray, R. M. & El-Samad, H. Population diversification in a yeast metabolic program promotes anticipation of environmental shifts. *PLoS Biol.* **13**, e1002042 (2015).

32. Wang, J. et al. Natural variation in preparation for nutrient depletion reveals a cost–benefit tradeoff. *PLoS Biol.* **13**, e1002041 (2015).

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* **21**, 947–962 (2004).

34. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2013).

35. Pelechano, V., Wei, W. & Steinmetz, L. M. Genome-wide quantification of 5′-phosphorylated mRNA degradation intermediates for analysis of ribosome dynamics. *Nat. Protoc.* **11**, 359–376 (2016).

36. McDavid, A. et al. Data exploration, quality control and testing in single-cell transcriptomic data across different conditions, technologies, and experiments. *Bioinformatics* **29**, 461–467 (2013).

37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Biostatistics* **15**, 550 (2014).

38. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

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.* **35**, 193–200 (2007).

Acknowledgements
The authors would like to thank S. Linnarsson for providing reagents during the initial tests with yscRNA-seq. We thank the Protein Expression and Purification Core Facility at EMBL, B. Hennig and L. Velten for providing in-house purified Tn5. We thank R. Bottcher for fruitful discussions. We thank D. 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-Maria 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 and a European Research Council Advanced Investigator Grant (grant no. AdG-294542 to L.M.S.) and the National Key Research and Development Program of China (grant no. 2017YFC0908405 to W.W.). The study was also supported by grants from the Spanish Ministry of Economy and Competitiveness (grant nos BFU2015-64437-P, FEDER, BFU2014-52125-REDT and BFU2014-51672-REDC to F.P.; BFU2017-85152-P and FEDER to E.d.N.), the Catalan Government (grant no. 2017 SGR 799), the Fundación Botín, the Banco Santander through its Santander Universities Global Division to F.P. and the Unidad de Excelencia María de Maeztu, grant no. MDM-2014-0370. F.P. is a recipient of an ICREA Académia (Generalitat de Catalunya).

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

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41564-018-0346-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.M.S.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Software and code

Policy information about availability of computer code

- Data collection: no software was used
- Data analysis: custom script: Java (UMI debarcoding), R (processing, analysis and visualization)
  - Comercial: novocraft (aligner), FlowJo (FACS)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated in this study has been uploaded at Gene Expression Omnibus (GEO). The accession code will be public at the time of acceptance (GSE122392)
Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No calculations were done to determine sample size. The total number of cells per each strain is indicated. |
|-------------|---------------------------------------------------------------------------------------------------------|
| Data exclusions | Samples that did not meet our quality criteria for number of reads and number of detected genes were discarded from the analysis. Criteria for exclusion is explained in detail in the manuscript |
| Replication | Data for BY4741 represents two biological replicates, YJM789 dataset represents one biological replicate. Growth curves represent the mean of two biological replicates each one of them with two technical replicates |
| Randomization | Randomization of samples does not apply to the study since clonal yeast populations are used |
| Blinding | Blinding was not possible because it is not relevant to the present study |

Reporting for specific materials, systems and methods

Materials & experimental systems

| n/a | Involved in the study |
|-----|------------------------|
|     | Unique biological materials |
|     | Antibodies |
|     | Eukaryotic cell lines |
|     | Palaeontology |
|     | Animals and other organisms |
|     | Human research participants |

Methods

| n/a | Involved in the study |
|-----|------------------------|
|     | ChiP-seq |
|     | Flow cytometry |
|     | MRI-based neuroimaging |

Flow Cytometry

Plots

- Confirm that:
  - The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
  - The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
  - All plots are contour plots with outliers or pseudocolor plots.
  - A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

In short, over-night grown BY4741 and YJM789 were diluted to OD 0.05 grown for at least two cell divisions. Prior to sorting cell were diluted to OD 0.01 and briefly sonicated for 3 pulses 0.5 seconds to remove aggregates and cells were filtered by passing samples through cell strainer snap cap. Live single cells were index-sorted on the forward versus side scatter (FSC and SSC respectively).

Instrument

For single cell index sorting, cells were sorted with FACS InFlux (BD Instruments). Cell sorting for growth was done with FACSAriaII (BD Instruments)

Software

FACS data for GFP measurements was done with FlowJo
Cell population abundance: For fluorescence experiments 10000 cells of each group were sorted or measured. For single cell experiments 1 single cell was sorted into 96 well plates.

Gating strategy: For single cell sorting gates on the FSC and SSC for the entire population were drawn and a second stringent pulse width was done to eliminate non single cell sorting. For fluorescence Gal3 experiments gating strategy was either random or top 2% based on the fluorescence.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.