Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation

Malika Saint[1,2], François Bertaux[1,2,3,*,§], Wenhao Tang[3,*], Xi-Ming Sun[1,2], Laurence Game[1,2], Anna Köferle[4,†], Jürg Bähler[4], Vahid Shahrezaei[3,5] and Samuel Marguerat[1,2,5]

[1] MRC London Institute of Medical Sciences (LMS), Du Cane Road, London W12 0NN, UK

[2] Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, Du Cane Road, London W12 0NN, UK

[3] Department of Mathematics, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK

[4] University College London, Research Department of Genetics, Evolution & Environment and UCL Genetics Institute, London, WC1E 6BT, UK.

[5] Correspondence to v.shahrezaei@imperial.ac.uk or samuel.marguerat@imperial.ac.uk

* These authors contributed equally to this work.

§ Present address: Institut Pasteur, 28 rue du Docteur Roux, 75015 Paris, France.

† Present address: Munich Center for Neurosciences, Ludwig-Maximilian-Universität, BioMedical Center, 82152 Planegg, Germany.

Phenotypic cell-to-cell variability is a fundamental determinant of microbial fitness that contributes to stress adaptation and drug resistance. Gene expression heterogeneity underpins this variability, but is challenging to study genome-wide. Here we examine the transcriptomes of >2000 single fission yeast cells in various environmental conditions by combining imaging, single-cell RNA sequencing (scRNA-seq), and Bayesian true count recovery. We identify sets of highly variable genes during rapid proliferation in constant conditions. By integrating scRNA-seq and cell-size data, we provide unique insights into genes regulated during cell growth and division, including genes whose expression does not scale with cell size. We further analyse the heterogeneity of gene expression during adaptive and acute responses to changing environments. Entry into stationary phase is preceded by a gradual, synchronised adaptation in gene regulation, followed by highly variable gene expression when growth decreases. Conversely, a sudden and acute heat-shock leads to a stronger, coordinated response and adaptation across cells. This analysis reveals that the magnitude of global gene expression heterogeneity is regulated in response to different physiological conditions within populations of a unicellular eukaryote.

Gene expression is tightly regulated at multiple levels, including chromatin structure, transcription, mRNA degradation and translation. This multi-layered process underpins robust and timely expression of single proteins as well as coordinated regulation of entire genetic programmes including dozens of genes. Yet, even in constant environments, expression of specific genes varies between genetically identical cells, leading to cell-to-cell heterogeneity in mRNA numbers and concentrations[1–3]. Cell-to-cell variability in gene expression results from different phenomena. First of all, the random timing of biological reactions makes transcription intrinsically stochastic. This form of variability, also called intrinsic noise, is gene specific and depends on promoter sequence and chromatin states[4,5]. Heterogeneity in quantitative traits such as cell size, growth rate, or concentration of transcription factors also shapes gene expression variability in complex, non-trivial ways. This form of variability is not entirely stochastic and depends on other single-cell attributes that affect biomolecule numbers[6,7]. Furthermore, cells can enter dynamic cellular states characterised by specific gene expression programmes. Examples are progression through the cell cycle or the adoption of distinct metabolic states[8]. Different states co-exist in cell populations or tissues leading to dynamic, yet deterministic, cell-to-cell variability in gene expression. Finally, cells in metazoan tissues belong to different cell types that are important for organ architecture and function. Although reversible and plastic, this form of individuality is not expected to be as dynamic as the transient cellular states.

The development of RNA sequencing protocols supporting the analysis of entire transcriptomes from single cells has been instrumental in describing cell-to-cell variability and phenotypic heterogeneity in multicellular organisms[9]. Because such approaches sample expression levels of many genes in an unbiased manner, they provide insights into the molecular complexity of healthy tissues and tumours, affording better understanding of tissue biology in health and disease[10–12].

Gene-expression variability present in a population of unicellular organisms is conceptually different from heterogeneity in metazoan tissues. Our understanding of its structure and regulation remains superficial as transcriptomic approaches to sampling gene expression in single microbial cells have lagged behind[13]. This is mostly due to the small size and resistant cell walls of microbes[14]. Such approaches are critically needed, because determining the extent of cell-to-cell variability in gene regulation in microbial populations is required to reach a mechanistic understanding of antibiotic resistance, cellular adaptation, or population dynamics and evolution[13,14].

Here, we overcome these limitations and develop an integrated experimental and computational framework to image individual cells of the fission yeast *Schizosaccharomyces pombe*, followed by scRNA-seq analysis and Bayesian true count recovery. Using this approach, we obtain a unique account of heterogeneity in gene expression and cellular states as a function of cell size, growth, and adaptation in this popular model organism.

Imaging and transcriptome analysis of single fission yeast cells

We developed an integrated approach for imaging and isolation of single cells using a tetrad dissection microscope, followed by transcriptome analysis by scRNA-seq (**Fig. 1a**). Datasets combining images and scRNA-seq libraries were generated for 2028 cells across a range of conditions, together with 780 matching control libraries, each obtained from 3 pg of total RNA (ctrRNA, **Supplementary Table 1-3**).

For library preparation, we used a variation of the SCRB-seq protocol[15]. Our approach targets 3'-end cDNA sequences, includes unique molecular identifiers (UMIs), and benefits from optimisations of the Smart-seq2 protocol (**Methods**, **Supplementary Table 9**)[15,16]. We generated 8.2×10^5 mappable sequencing reads/cell, which were clustered around transcription termination sites, and corresponded to 6721.3 unique mRNA molecules on average (**Supplementary Fig. 1a-d**, **Supplementary Table 1-4**). This represents a mean transcriptome coverage $\langle\beta\rangle$ of ~1.5% or ~6%, based on calibration with spike-in controls or single-molecule fluorescence *in situ* hybridisation (smFISH) measurements, respectively (**Supplementary Fig. 1e**).

We could detect 1421.1 genes/cell on average, but as genes with low molecule counts carried little information on true expression levels, we focused this study on a consolidated list of 1011 robustly detected high-confidence genes representing 18.5% and 4.3% of all coding and non-coding genes, respectively (**Methods**, **Fig. 1b**, **Supplementary Fig. 1f-g**). These genes were often highly expressed in cell populations, involved in most cellular processes, and showed constitutive as well as condition-specific regulation (**Fig. 1b**, **Supplementary Fig. 1h**, **Supplementary Table 6**).

The shallow transcriptome coverage inherent to scRNA-seq, together with the high level of amplification required to detect small yeast transcriptomes, made data pre-processing challenging. We used a Bayesian normalisation approach called bayNorm that performs true mRNA count recovery based on cell-specific mRNA capture efficiencies (β) and on gene-specific priors estimated from the data (**Methods, Supplementary Fig. 2a**)[17]. The methodology and performance of this approach is reported in detail elsewhere[17]. Applied to our datasets, bayNorm generated true count distributions highly similar

to measurements from population RNA-seq (**Supplementary Fig. 2b**) and to absolute mRNA counts obtained by smFISH (**Supplementary Fig. 2c**, **Supplementary Table 5**)[17,18]. Finally, true counts correlated with expression levels derived from cell populations[18], while preserving information about cell-to-cell variability in mRNA expression (average $R_{pearson}$=0.78, **Fig. 1c**, **Supplementary Fig. 2b-e**). In summary, we generated the first combined dataset of transcriptomes and microscopy images from single yeast cells, which is analysed in detail below.

Cell-to-cell variability of transcriptome regulation

We first looked for mRNAs with high cell-to-cell variability compared to most genes (Highly Variable Genes, HVG). For this, we used 9 scRNA-seq datasets consisting of 864 unperturbed cells growing exponentially at 2-10x10^6 cell/ml (**Supplementary Table 2**). As reported previously and expected from stochastic gene expression models[19,20], coefficients of variation (σ/m, CV) and means of normalised counts were anti-correlated in cells and ctrRNA (**Supplementary Fig. 3a**). We therefore defined HVGs as mRNAs with CVs significantly higher than this overall trend using simulated data with the Poisson noise as only source of variability as a reference (**Methods, Fig 2a and legend**). We applied this procedure to each dataset normalised separately because it led to CVs closest to smFISH measurements while avoiding batch effects (**Supplementary Fig. 2c**). We identified 411 genes with CV significantly higher than the baseline in at least one dataset, of which 112 were also present in ctrRNAs and were discarded as false positives (FP). This analysis generated a list of 299 high-confidence HVGs (**Supplementary Fig. 3b**, **Supplementary Table 6**). To investigate the specificity of our scRNA-seq approach, we analysed 5 control genes and 8 HVGs covering a range of variability scores and biological functions by smFISH. HVGs showed significantly higher size-corrected Fano factors (scFano), confirming their higher noise levels (**Fig 2b-c**, **Supplementary Fig. 3d**).

Genes periodically expressed during the cell cycle have been identified in synchronised cell populations[21–24]. We hypothesised that these genes would be overrepresented among HVGs, as we sampled asynchronous cells from different cell-cycle stages. Accordingly, 53.3% of the top-500 periodic genes[24] found among high-confidence genes were HVGs (**Supplementary Fig. 3b**, P_{Fisher} = 5.1e^{-9}) as were genes regulated in specific cell cycle phases (**Fig. 2d**). To evaluate our approach's sensitivity, we analysed the top-500 periodic genes that were not HVGs. These showed significant but low amplitude and periodicity in population RNA-seq data, consistent with scRNA-seq being a less sensitive approach (**Supplementary Fig. 4a-c**). Only a minority of periodic genes were false positives (12.5%) confirming the specificity of our experimental and computational protocols (**Supplementary Fig. 3b**, **Supplementary Table 6**). Finally, this analysis demonstrates that periodic gene expression is a single-cell feature of asynchronous populations and not a technical artefact of cell-cycle synchronisation[25].

Notably, most HVGs were not cell-cycle periodic and could not be identified in synchronised cell populations. To characterise these genes, we split them into three categories based on the number of datasets in which they were highly variable (>3: highly pervasive; 2-3: moderately pervasive; 1: lowly pervasive) (**Fig. 2e**, **Supplementary Table 6**). These categories describe how robustly variable each gene is across biological replicates. Importantly, some lowly pervasive genes, like *Isd90,* showed strong

amplitudes of regulation (**Fig. 2b, Supplementary Fig. 3c-d**). Moderately and highly pervasive HVGs were related to mitochondria and heat-shock response. Interestingly, the genes encoding Nmt1 and the associated biosynthetic enzyme Thi2 were among the most pervasively variable HVGs, suggesting widespread heterogeneity in vitamin B1 metabolism. In terms of gene expression regulators, the transcription factor Fil1, which controls the amino acid starvation response, as well as the TATA-associated factor Mot1, a general transcription factor, were pervasively variable (**Fig. 2b**, **Supplementary Fig. 3c-d**)[26,27]. The latter is consistent with TATA-box sequences being associated with variable and noisy genes[5,28–32], and with a role of Mot1 expression variability in this regulation[27]. Lowly pervasive HVGs span diverse functions related to membrane biology and adaptation to external conditions, including genes from the Core Environmental Stress Response (CESR) programme (**also Supplementary Fig. 3b**)[33]. Lowly pervasive variability could result from subtle responses to external fluctuations, consistent with recent budding yeast scRNA-seq data[14].

We finally investigated HVGs association with several cellular and genetic features (**Supplementary Fig. 4d**). Interestingly, budding yeast orthologues of HVGs were highly variable between cells at the protein level. This indicates that the architecture of gene expression variability is at least partially conserved between both yeasts[19]. HVGs were more regulated in response to environmental and genetic perturbations[34,35] suggesting that noisy transcription could underlie rapid adaptation to unpredictable challenges[36]. Variable genes have been reported to evolve rapidly[28,29]. Accordingly, HVGs showed higher evolution rates and non-synonymous/synonymous mutation ratios between fission yeast species[35,37]. Conversely, HVGs showed less negative genetic interactions and co-regulation with other genes[35]. This suggests that variability may be detrimental to large protein complexes or highly connected regulatory networks[30]. Regarding promoter sequence, HVGs were likely to have a canonical TATA box, as observed in other organisms, but showed only moderate enrichments for specific transcription factor binding-sites (**Supplementary Table 10**).

In summary, our analysis defines the functional organisation and pervasiveness of genome-wide gene expression variability in unperturbed fission yeast cells.

Cell size dependence of transcriptome regulation

Single-cell RNA-seq provides a snapshot of gene-expression variability and cell states in a population. Interpreting this information can be greatly facilitated by integrating transcriptomics data with measures of quantitative cellular features[6,19,38–40]. Our combined approach offers such capabilities as it includes microscopy images of each cell matched to their respective transcriptomes (**Fig. 1a**).

We used cell-length measurements from images across all growth datasets to order cells as a function of size, which reflects progression through the cell cycle (**Supplementary Table 3**). We first examined changes in global properties of scRNA-seq measurements as a function of size. Mean cell length during rapid growth was 10.9 µm, consistent with reported data (**Fig. 3a**)[41]. Mean normalised scRNA-seq counts were constant across the size range, consistent with bayNorm returning size-corrected absolute molecule numbers (which are proportional to concentrations; **Fig. 3a**).

In EMM2 medium, fission yeast elongates during G2 for over two thirds of the cell cycle. At mitosis, cells stop growing until cell division, which occurs in G1/S. To validate our image-based classification of scRNA-seq data, we asked whether transcriptome signatures of the M/G1/S phases were apparent in larger cells. As expected, these featured increased transcriptome fractions related to processes specific to G1/S transition and cell-wall biogenesis (**Fig. 3b**). This was also apparent when expression counts were plotted as a function of cell length (**Supplementary Fig. 5a**). Besides cell-cycle signatures, some large cells showed increased transcriptome fractions related to respiratory metabolism, which increases during the reductive building phase of the yeast metabolic cycle (**Fig. 3b**, see below)[8,42].

We then used cell-length measurements to guide our analysis of transcriptional programmes associated with cell proliferation. To do so, we looked for genes differentially expressed between large cells in M/G1/S and small, recently born G2 cells using bayNorm priors specific for each groups (**Fig. 3c**, **Supplementary Fig. 5b-c,e**). We identified 92 genes significantly up-regulated in large cells (**Supplementary Table 6**). Consistent with large cells being in M/G1/S, 28.3% of these were also periodically expressed in synchronised cell populations[21–24]. Twice as many genes (193) were induced in small cells, of which 19.2% are periodic[24]. Importantly, this analysis combined with HVG detection based on ΔCV (**Fig. 2**) retrieved 81.7% of the top-500 periodic genes present in the dataset, with the remaining genes showing no apparent regulation (**Supplementary Fig. 4b-c**). A significant proportion of genes overexpressed in small cells belonged to the stress-response programme ($P_{Fisher,\ one\text{-}sided}$ = 0.02) and/or had hydrolase activity ($P_{Fisher,\ one\text{-}sided}$ = 0.002). Large cells, on the other hand, induced several genes involved in mitochondrial membrane transport. These observations, together with the analysis from **Fig. 3b**, are reminiscent of the yeast metabolic cycle (YMC). We therefore analysed gene signatures of the YMC phases: reductive charging (RC), oxidative (OX), and reductive building (RB). YMC signatures were compartmentalised with cell size and the cell cycle (**Supplementary Fig. 5d**)[42]. RC genes were higher expressed in small cells, while RB genes increased in large cells when DNA replication occurs ($P = 6.1e^{-5}$)[8,42]. This analysis raises the possibility of a YMC, synchronised with the cell cycle, in proliferating asynchronous single fission yeast cells[43,44]. These signatures were not apparent in the HVGs identified in **Fig. 2**, demonstrating the increased sensitivity provided by combining imaging and transcriptomics.

Molecules numbers of most mRNAs increase coordinately (scale) with cell size to maintain concentrations[45,46]. Accordingly, average UMI-corrected raw counts/cell correlated with cell size ($R_{Pearson}$ = 0.17, $P_{Pearson,\ two\text{-}sided} = 6.6e^{-7}$, **Fig. 3a**). Genes that escape this trend have not been characterised globally, yet they could be important in regulating growth and cell-size homoeostasis[46]. To identify genes that escape scaling, we analysed G2 cells between birth and 11 µm in length, beyond which we found strong signatures of the M/G1/S programmes (**Supplementary Fig. 5a**, **Supplementary Table 6**). Seventy-eight genes changed in concentration coordinately with cell size during G2 (**Methods**, $P_{Pearson,\ two\text{-}sided}$ <0.05). Using k-means clustering, we defined 5 small gene clusters, 3 of which increased (Cl1-3) and 1 decreased (Cl5) in concentration (**Fig. 3d**, **Supplementary Table 6**). Cl4 showed significant, but low amplitude, positive regulation with size. We assessed whether these clusters defined one or more cellular states by looking at their correlations between single cells (**Methods**). We found evidence for two states in our datasets. The first was defined by Cl1 and Cl2 that

were positively correlated and contained genes that are also up-regulated during meiotic differentiation (**Fig. 3e**)[47]. Cl3 and Cl5 were anti-correlated and defined a second state containing a small number of genes functioning in carbohydrate metabolism (**Fig. 3e**). Although not significant, this enrichment could hint at a gradual change in cellular energy metabolism coordinated with cell-size. An orthogonal Random Forest approach confirmed that 69 of the 100 genes with the strongest non-linear correlation with cell size were either differentially expressed between big and small cells or escaped scaling (**Supplementary Table 6**). Together, this analysis uncovers gene expression programmes occurring during growth in G2 and escaping coordination with cell size.

Interestingly, 45.5% of HVGs from Fig. 2 were more highly expressed in large or small cells or escaped scaling. Their variability is therefore not stochastic but can be understood in the light of two physiological variables: cell size and cell-cycle stage. This demonstrates the potential of analysing scRNA-seq datasets in the context of quantitative cellular features to understand gene regulation.

Transcriptome heterogeneity within cell populations in response to environmental changes

Defining the impact of environmental signals on gene expression heterogeneity is important to understand how these factors shape population structures and adaptation. We generated a blueprint of 1824 single-cell transcriptional profiles in a series of environmental conditions, including stress response, high cell density, and nutrient depletion (**Supplementary Table 1-2**). To compare and contrast single-cell responses to different environments, we focused on 110 genes that are upregulated as part of the CESR[33]. Transcriptional signatures could be clearly distinguished using principal component analysis (PCA) of bayNorm-normalised counts (**Fig. 4a**). Interestingly, cells growing rapidly in constant conditions occupied a distinct area of the transcriptional space, confirming our previous observation that an exacerbated stress response is not common in single cells during rapid proliferation (**Fig. 2d**). We then examined the specificity of the transcriptional programmes defined by scRNA-seq, focusing on heat-shock and oxidative-stress genes[33]. These signatures singled out cells having experienced the corresponding stresses confirming the specificity of scRNA-seq and the capacity of bayNorm normalisation to correct for experimental batch effects (**Fig. 4b-c**).

We then asked whether the dynamics and heterogeneity of responses differed between perturbations. We first analysed the response of single cells to a gradual change in external conditions. Specifically, we analysed cells growing at densities ranging from $2x10^6$ to $74x10^6$ cells/ml, encompassing rapid proliferation and early stages of stationary phase (**Supplementary Fig. 6a**, **Supplementary Table 1-3**). We observed a progressive increase in CESR mRNAs up to a density of ~$40x10^6$ cells/ml coordinated with a decrease in mRNAs from the translation and the cell-growth programmes (**Supplementary Fig. 6b**)[33]. The concentration of ribosomes is known to increase with growth rate to support higher biosynthetic demand[48]. We therefore wondered whether decreasing concentration of growth-related mRNA would affect growth rate. Surprisingly, growth rates remained constant up to ~$40x10^6$ cells/ml (**Supplementary Fig. 6a**). This indicates that mRNAs of the translation machinery can decrease in concentration in response to environmental changes, independently, and without affecting cell growth. This is consistent with the existence of a free ribosome fraction that buffers growth and

environment[48]. Importantly, only few other mRNA classes showed coordination with cell density, indicating that this behaviour is not ubiquitous (**Supplementary Fig. 6c, left**). Notably, increase in CESR mRNAs concentration was not accompanied, by increased gene-expression noise, nor by appearance of outlier cells having entered a full stress-resistance state (**Fig. 4d-e, Supplementary 6c right**). This result indicates that single cells undergo gradual and synchronised adaptation of gene expression at increasing cell densities.

Strikingly, within the following cell division, we detected a strong and heterogeneous induction of CESR genes (**Fig. 4d-e**), together with a decrease in growth rate (**Supplementary Fig. 6a**). Importantly, exhaustive functional analysis revealed that increased transcriptional heterogeneity was restricted to specific pathways and not a global property of the transcriptomes (**Supplementary Fig. 6c**, **right**). Additional genes showing strong heterogeneous responses were also regulated during meiotic differentiation and growth on glycerol (**Supplementary Fig. 6c middle and right**)[47,49]. Taken together, these data support a model where single cells readjust the balance of the stress- and growth-related transcriptional programmes synchronously as a function of cell density and ahead of changes in growth rate. This is followed, within a single cell cycle, by a substantial, heterogeneous reshuffling of the cellular transcriptome. These findings indicate that entry into stationary phase is a process that increases transcriptional heterogeneity and possibly promotes cell individuality and differentiation.

We then examined the impact of a rapid and severe change in external conditions on gene expression. Cells were briefly switched from 25°C to 37°C in a turbidostat and grown at steady-state at 37°C (adaptation) or 25°C (relaxation). Expression of CESR genes rapidly increased upon temperature switch, and adjusted back to pre-stress levels during both adaptation and relaxation (**Fig. 4f-g**). Strikingly, only a minor increase in transcriptional heterogeneity could be detected during heat shock that did not propagate during adaptation or relaxation, in stark contrast with entry into stationary phase. This suggests that the acute stress response can be synchronous in a cell population and does not lead to phenotypic heterogeneity (**Fig. 4g**). Together, this analysis demonstrates that the level of transcriptional heterogeneity induced by changes in external conditions is variable and regulated, depending on the type and strength of stimulus.

Conclusion

We report an integrated approach to analyse transcriptomes of single yeast cells in combination with phenotypic measurements. We also provide the first account of genome-wide gene expression heterogeneity in fission yeast, during rapid proliferation under constant conditions and in response to environmental changes. In constant conditions, periodic gene expression during the cell cycle is the most robust and pervasive form of heterogeneity. However, G2-specific expression signatures reminiscent of the budding yeast YMC also exist together with genes that escape scaling with cell size. This analysis relied on the ability to order and classify cells based on size, independently of the scRNA-seq data. A setup for quantitative imaging of high-dimensional morphological features coupled to our approach would extend its potential to additional cellular traits such as nuclear size, mitochondrial

numbers, or actin structure. This would enable understanding better the hidden diversity of cellular states that occur during growth and adaptation. We analysed gene expression heterogeneity and its dynamics in response to environmental changes. We observed striking differences between stationary phase entry, a heterogeneous process, and an acute heat-shock response, which appeared more coordinated. This raises the question of whether gene-expression heterogeneity depends on the strength of the challenge and indicates that expression heterogeneity is controlled in a condition-specific manner. Analysis of diverse environmental challenges at the single-cell level will be required to understand the root of this variability. In particular, comparing post-mitotic quiescent cells with proliferating cells would inform on the impact of growth on heterogeneity. Overall, in addition to increasing our understanding of how a single-celled eukaryote functions, the findings reported here highlight the potential of investigating gene regulation as a cause and/or consequence of quantitative cellular phenotypes, such as cell size, genome-wide in single-cells.

Methods

Detailed methods and associated references are available in the online version of the paper.

Acknowledgements

We thank Tom Livermore for his help during the initial part of this project and Claire Stefanelli for her contribution to bayNorm development. We are grateful to Simona Parrinello, Amalia Martínez Segura and Miles Priestman for their input on the manuscript. This research was supported by the UK Medical Research Council, a Leverhulme Research Project Grant (RPG-2014-408), and a Wellcome Trust Senior Investigator Award to J.B. (grant 095598/Z/11/Z). It used the computing resources of the UK Medical Bioinformatics partnership (UK MED-BIO), which is supported by the UK Medical Research Council (grant MR/L01632X/1) and the Imperial College High Performance Computing Service. All figures except 1a and S2a contain original data.

Authors contributions

MS set up the scRNA-seq protocol in yeast, performed all sequencing experiments and part of the computational analysis. WT and FB developed bayNorm and performed most computational analysis together with SM and VS. XMS performed all smFISH and growth experiments. AK developed the first generation of the single-cell isolation and PCR amplification protocol under the supervision of SM and JB. LG assisted with developing the scRNA-seq protocol. SM, VS, and JB supervised this study. SM, VS, JB, WT and MS wrote the paper.

Competing interests

The authors declare no competing interests.

Methods

General comment

RNA-seq analysis of single yeast cells is challenging because their cell wall is resistant to standard lysis conditions that preserve RNA integrity. Moreover, yeast transcriptomes are highly plastic and respond to external conditions within minutes, making cell isolation and manipulation a source of artefacts[14,33]. We have overcome both hurdles by i) snap freezing cells immediately after harvesting, a procedure that fixes both cell morphology and transcriptomes; ii) establishing a protocol for yeast cell lysis at high temperature in conditions that protect RNA integrity, bypassing the need for enzymatic digestion of the cell wall.

Cell culture

$972h^-$ fission yeast cells were cultured with a seeding density of 0.5 x 10^6 cells/ml in all experiments. Standard EMM2 media was used except when indicated otherwise[50]. All conditions are described in detail in **Supplementary Table 2**, and assigned to individual samples in **Supplementary Tables 1 and 3**. *Heat stress*: cells were grown in YE medium at 25°C up to a density of 2-4 x 10^6 cells/ml. The cells were transferred to a water bath maintained at 37°C or 39°C (for datasets, 1712_1, 1712_2, 0408_2). For studying adaptation to heat, cells were transferred post-heat-shock to a turbidostat maintained at a density of 4 x 10^6 cells/ml and a temperature of either 25°C or 37°C (siphon-flow based derivative of the instrument described in ref [51]). *Glycerol growth*: cells were grown in YE medium with 3% glycerol and 0.1% glucose. *Osmotic shock, Oxidative stress*: Cells were cultured in YE medium up to a density of 4 x 10^6 cells/ml. To induce osmotic shock, an equal volume of 2M sorbitol prepared in YE medium was added to the cell culture to a final concentration of 1M for 15 min. To induce oxidative stress, cells were treated with 0.5mM H_2O_2 for 15 or 60 min. *Nitrogen starvation*: Cells were pre-cultured in EMM2 medium with NH_4Cl as nitrogen source up to a density of 2 x 10^6 cells/ml. Cells were harvested by centrifugation, washed twice with EMM medium without nitrogen and re-suspended in medium without nitrogen. Cells were harvested at 6 and 24 hours after starvation. All cultures were snap frozen at the time of harvesting. This treatment kills fission yeast cells and precludes any changes in gene expression during the following isolation steps.

Cell isolation and imaging

Single cells were imaged with a 20x objective on a Singer MSM-400 tetrad-dissection microscope (Singer Instruments), picked into 3µl of QuickExtract™ for RNA extraction solution (Lucigen, Epicenter) in 200µl PCR tubes and immediately snap frozen at -80°C. The use of the QuickExtract™ buffer solution is critical for protecting RNA against degradation during cell lysis. For each ctrRNA sample, 3pg of total RNA isolated from matching cultures by hot phenol extraction were diluted in QuickExtract™ and processed as single cells. Single cell images were analysed using ImageJ. All cells and ctrRNA samples are described in **Supplementary Table 3**.

365 SCRB-seq for yeast library preparation

366 Single cells were lysed at 98°C for 10 min in a PCR machine, and library preparation performed based
367 on[15,16,52], using primer sequences described **Supplementary Table 9**. The protocol was modified as
368 follows. Briefly, oligo dT containing cell indexes and UMIs were added to each well at a final
369 concentration of 1µM. Primers were annealed to the RNA template at 72°C for 3min, and components
370 for reverse transcription added with final concentrations of 100U Superscript II reverse transcriptase
371 (Invitrogen), 10U RNAse inhibitor (Invitrogen), 1x Superscript buffer, 5mM DTT, 1M Betaine (Sigma),
372 1.5mM MgCl2, 1mM of each dNTPs, 1µl of $1/10^6$ dilution of ERCC spikes Set A (NEB), and 1µM RNA-
373 TSO primer. Reverse transcription was carried out at 42°C for 90min, after which the temperature was
374 ramped between 50°C and 42°C for 10 cycles of 2min each. The reaction was heat inactivated at 70°C
375 for 15min and the reaction cooled to 15°C. Each single-cell sample was treated with 20U Exonuclease
376 I (NEB) for 30min at 37°C followed by heat inactivation at 80°C for 20min. Sets of 96 samples were
377 pooled and purified using a PCR purification kit (Qiagen) and eluted in 60µl elution buffer (EB)
378 containing 10mM Tris-Cl, pH 7.5. Samples were treated with 40U of Exonuclease I for 30min at 37°C
379 for a second time followed by heat inactivation at 80°C for 20min. PCR was performed on the pooled
380 sample adding 1x KAPA HiFi buffer, 0.075mM of each dNTPs, 1µM PCR primer, and 1.25U KAPA HiFi
381 enzyme. PCR cycling was done with denaturation at 98°C for 3min, followed by 25 cycles of
382 denaturation, annealing and extension at 98°C, 60°C and 72°C for 20s, 15s and 1min respectively. A
383 final extension of 72°C for 5min was done before cooling the samples at 15°C. Samples were purified
384 using 0.6x Agencourt AMPure XP beads and eluted in 10-15µl nuclease-free 10mM Tris pH 7.5.
385 Libraries were quantified on an Agilent Bioanalyser using an HS-DNA chip to confirm the presence of
386 a clean peak at ~1000bp. Between 1-2ng of PCR library was used for tagmentation using the Illumina
387 Nextera XT kit using a modified I5 primer as described in[15,52] (**Supplementary Table S9**). Between 8
388 and 12 PCR cycles were performed post-tagmentation to amplify the 3' fragments carrying the poly-A
389 tail, the cell barcode and the UMI. The final libraries were purified using Agencourt AMPure XP beads
390 twice at 1x bead concentration and final elution done in elution buffer. Libraries were quantified using
391 on an Agilent Bioanalyser and sequenced.

392

393 Single molecule fluorescence in situ hybridisation (smFISH)

394 Measures of cell size, mRNA number per cell and cellular mRNA concentrations were obtained for 12
395 genes by single molecule fluorescence in situ hybridisation (smFISH) as described in ref [53]. Genes
396 queried are: SPBC16E9.16c (*lsd90*), SPAC27D7.09c, SPCC330.02 (*rhp7*), SPBC725.11c (*php2*),
397 SPBC28F2.12 (*rpb1*), SPBC1826.01c (*mot1*), SPCC1223.11 (*ptc2*), SPBC146.13c (*myo1*),
398 SPAPB1E7.04c ,SPAC328.03 (*tps1*), SPAC2H10.01 and SPCC1739.01. Processed data are provided
399 in **Supplementary Table 7.** Probe sequences are provided in **Supplementary Table 8**. For
400 **Supplementary Fig. 2c**, we used R-code available at:
401 https://stackoverflow.com/questions/35717353/split-violin-plot-with-ggplot2. To calculate size corrected
402 Fano factors (scFano) mRNA counts of each cell were divided by the length of the cell and multiplied

by the average cell size in the experiment. scFano factors were then calculated as the variance over the mean of these normalised counts (σ^2/μ).

Sequencing and Read mapping.

Pools of scRNA-seq libraries were sequenced on an Illumina HiSeq 2500 instrument at the MRC LMS genomics facility. Paired-end reads (100nt) were generated from two pools of 96 samples per sequencing lane. Data was processed using RTA 1.18.64, with default filter and quality settings. The reads were de-multiplexed with bcl2fastq-1.8.4 (CASAVA, allowing 0 mismatches). Read 1 was used to extract cell-specific indexes and Unique Molecular Identifiers (UMI). The corresponding Read 2 was mapped to the fission yeast genome as described in ref [18] . Mapped reads where assigned to fission yeast genes as described in ref [18] using Pombase annotation as of 27/05/2015 and including 5' and 3' UTR sequences. Read1 and Read2 were assigned to specific cell/RNA samples based on cell-specific indexes sequences de-multiplexed using in house Perl scripts. Within each specific cell/RNA sample, reads sharing identical UMI sequences and mapping to the same gene were collapsed.

UMI Correction

Unique Molecular Identifiers (UMI) are short random DNA sequences, typically 6-8nt in length, which are appended to every single cDNA molecule during SCRB-seq for yeast sequencing library preparation. In SCRB-seq for yeast, UMIs are part of the first-strand reverse transcription primer[15]. UMIs are commonly used to remove PCR amplification biases but importantly have been recognized to be themselves prone to sequencing errors and biases[54,55]. These lead, for a given gene, to an enrichment in the fraction of UMIs with small sequence distances, also called Hamming distances, that are higher than expected by chance[54]. This phenomenon is present in our data and leads to an overestimation of the library diversity. To correct for this bias, we developed an original network-based method which removes recursively, at each genomic locus, reads associated to UMIs that differ by a distance of 1 nucleotide (Hamming distance = 1) from the UMIs with the highest abundance. Our method is identical to the adjacency method introduced and implemented recently in UMI-tools[54]. Applying our UMI error correction method removes ~ 30% of the raw reads pool (**Supplementary Fig. 1b**). For dataset descriptions, statistics and raw counts see **Supplementary Tables 1-4**.

Average gene (Supplementary Figure 1d)

Average profiles were obtained from raw UMI-corrected counts for 10 cells using the deeptools package[56]. The function "`computeMatrix scale-regions`" and default bin size of 10nt and flanking regions of 300nt were used.

438 Selection of high-confidence genes

439 Genes representing >0.16% of the total molecules detected in the transcriptome of at least one cell
440 across all datasets were included in the high-confidence gene-set used in this study. This empirical filter
441 leads to a list of 1011 genes with varied functions and regulation patterns across conditions (**Fig. 1b**,
442 **Supplementary Fig. 1f-h**, **Supplementary Table 6**). Importantly this approach included in the high-
443 confidence list genes with high expression in a small number of cells. This would not have been possible
444 using a single cut-off on mean expression values across the dataset. This filtering protocol removed
445 mostly lowly expressed genes with very high fraction of cells with zero counts (dropouts) for which
446 detection becomes mostly stochastic (**Supplementary Fig. 1f**). Accordingly, the mean expression
447 levels of the discarded genes after removal of zero values was 1.23 molecules/cell significantly lower
448 than high-confidence genes (mean = 5.8 molecules/cell, $P_{willcox\ one\text{-}sided} = 0$, **Supplementary Fig. 1g**).
449 Together, this analysis confirms the low information content of the discarded genes and validates our
450 filtering approach.

451

452 Estimation of the mean capture efficiency using spike-ins and smFISH

453 We define the capture efficiency β_i of a cell i to be the probability of observing (sequencing) any one of
454 the cell's original mRNA molecules. The mean capture efficiency $\langle\beta\rangle$ is the mean of the capture
455 efficiencies across all cells. As for any given gene, the observed UMI-corrected counts per cell are lower
456 than the original number of mRNA molecules present in a cell, capture efficiencies range between 0
457 and 1. Spike-in controls can be used to estimate β_i and mean capture efficiency $\langle\beta\rangle$[57]. To do this, we
458 divided the total number of spike-in molecules observed within each cell by the corresponding
459 theoretical number of input spike-in molecules. The mean of these ratios across all the cells is 0.015.
460 We believe this number is an underestimate of the true capture efficiency, given recent absolute
461 estimates of average mRNA counts in fission yeast populations[18]. Consistent with our observations, it
462 has been reported recently that spike-ins have a lower capture efficiency than mRNAs[58]. An alternative
463 way to estimate mean capture efficiency relies on estimates of absolute mRNA molecules numbers per
464 cell obtained by smFISH. Using 12 different genes, we fit a linear regression between the mean
465 expression of UMI corrected sequencing counts and the mean of the corresponding smFISH counts.
466 Using this approach, the mean capture efficiency is estimated to be the coefficient of variable, which is
467 about 0.06 (**Supplementary Fig. 1e**). In summary, the mean capture estimates obtained from spike-in
468 controls and smFISH measurements are very different. We chose to use the geometric mean of the two
469 estimates, leading to a mean capture efficiency $\langle\beta\rangle$ of 0.03. This estimate is one of the parameters for
470 our Bayesian data normalization protocol described below (bayNorm)[17]. We note that, within this range,
471 our biological conclusions are not overly sensitive to specific values of $\langle\beta\rangle$. The dependence of bayNorm
472 normalization on the choice of $\langle\beta\rangle$ is systematically explored elsewhere[17].

473

474 Estimation of capture efficiencies of single cells

Single cells' capture efficiencies (β_i) are proportional to cell-specific global scaling factors (s_i) that are commonly used in normalisation of single cell RNA-seq data (see for example[59]):

$$\beta_i = \langle\beta\rangle \times s_i/\langle s\rangle$$

where the constant of proportionality is related to the mean capture efficiency $\langle\beta\rangle$. The scaling factors (s_i) can be estimated using spike-in controls[57], or alternatively from the data directly. Simple estimates of s_i are the total number of molecules detected per cell (total count, TC), or the mean of the number of molecules of a subset of genes detected in each cell. Popular bulk RNA-seq methods such as DESeq are designed to compute global scaling factors s_i[60,61]. Unfortunately, these methods are not applicable to scRNA-seq datasets because of the high frequency of drop-outs present in the data (drop-outs: the proportion of genes with zero counts across cells)[59]. Alternative methods have been developed specifically for scRNA-seq (for example see[62,63]). We carefully assessed several existing methods for estimation of scaling factors, and settled for estimations of s_i based on the mean of UMI-corrected counts of a carefully chosen subset of genes in each given cell. The rationale behind this choice is the following: we argued that genes with i) high drop-out rates, ii) high variability in ctrRNA control experiments (showing technical variability), or, iii) those in the tail of the mean expression distribution (which have disproportionally high contribution to the total count) are not suitable for scaling factor estimation. Specifically, we used a list of 768 genes for global scaling factors estimation that were selected as follows: i) Genes with dropout rate > 70% were excluded (zero UMI-corrected counts in more than 70% of the cells across all datasets, 202 genes). ii) The top 20 higher expressed genes across datasets after TC normalisation were removed. iii) Genes with significantly high technical variability in ctrRNA controls were removed. To do this, we called highly variable genes in 11 ctrRNA datasets using the Bioconductor package scran and excluded 21 genes that were noisy in at least 5 of the 11 datasets[63,64]. Interestingly, the procedure described above produced the highest correlation between single-cell capture efficiencies and cell sizes (0.1781 for this procedure, 0.149 for the method proposed in[63] and 0.0568 for the spike-in estimates).

Normalisation using bayNorm

Single-cell RNA-seq data are commonly normalised by dividing raw counts by the global scaling factor s_i estimated for each cell[59]. We have recently developed bayNorm an alternative Bayesian approach to scRNA-seq normalisation that also provides simultaneous imputation for the drop-outs[17]. In this approach, given the raw count x_{ij} observed in the j^{th} cell for the i^{th} gene, and given the capture efficiency β_j of the j^{th} cell, we estimate the posterior distribution of the expected number of mRNAs x_{ij}^0 that were present originally in the cell. We found that a reasonable choice for the likelihood of observed counts x_{ij} is a Binomial distribution with size x_{ij}^0 and probability β_j[17]. In addition, we assume that the prior for x_{ij}^0 follows a Negative Binomial distribution (NB) with mean μ_i and size factor ϕ_i, with the following parameterisation:

512 $$\sigma^2 = \mu + (\mu)^2/\phi.$$

513

514 Using the Bayes rule, the posterior distribution of the i^{th} gene in the j^{th} cell can be expressed as:

515

516 $$\underbrace{\Pr(x_{ij}^0 | x_{ij}, \mu_i, \phi_i, \beta_j)}_{\text{Posterior}} = \frac{\overbrace{\Pr(x_{ij}|x_{ij}^0, \beta_j)}^{\text{Likelihood: Binomial}} \times \overbrace{\Pr(x_{ij}^0|\mu_i, \phi_i)}^{\text{Piror:NB}}}{\underbrace{\Pr(x_{ij}|\mu_i, \phi_i, \beta_j)}_{\text{Marginal likelihood}}}$$

517

518 Outputs of the bayNorm normalisation procedure are either samples from the posterior distribution
519 described above or its maximum *a posteriori* estimate (**Supplementary Fig. 2a**). In addition to raw
520 RNA-seq counts, bayNorm normalisation requires as inputs prior distributions of the parameters μ_i and
521 φ_i for each gene. In bayNorm, these are estimated from the scRNA-seq data directly using an Empirical
522 Bayes approach (see ref [17] for details). Prior estimation can be done using data from all cells across all
523 datasets irrespective of experimental conditions. We refer to this procedure as "Global" normalisation.
524 Alternatively, if cells can be split in different groups based on experimental conditions or phenotypic
525 information for instance, prior parameters μ_i and φ_i can also be estimated within each group
526 independently. We refer to this procedure as "Local" normalisation. On one hand, the use of global
527 priors based on Empirical Bayes reduces the technical batch effects that occur between different
528 experiments. On the other hand, using local priors for different groups of cells enhances the resolution
529 and sensitivity of differential expression analysis between these groups. The flexibility of prior parameter
530 estimation allows for heterogeneous cell populations to be accounted for[65]. Bayesian normalisation, as
531 implemented in bayNorm, has several additional advantages over widely used normalisation
532 approaches that rely on dividing molecules numbers by global scaling factors (see also[62]). First,
533 bayNorm also provides imputation by replacing a large proportion of zero counts by non-zero values,
534 reducing strongly the fraction of drop-outs in the normalised data (from 42.27% to 3.56% in the cell
535 datasets for high-confidence genes). Second, bayNorm effectively corrects for the experimental batch
536 dependent variation in average capture efficiency and reduces batch-specific biases performing
537 similarly to SCnorm but without the need for multiple expression-dependent scaling factors[17,62]. Also,
538 use of global priors as explained above can further reduce batch effects. Third, bayNorm normalisation
539 preserves the uncertainty present in the data particularly for cells with low coverage, reducing false
540 discovery rates in differential expression analysis. Finally, bayNorm produces mRNA distributions and
541 noise estimates close to the state-of-the-art smFISH measurements (**Supplementary Fig. 2c,**
542 **Supplementary Fig. 3c-d**) and averaged transcriptome structures close to high-quality population
543 estimates (**Fig. 1c, Supplementary Fig. 2b,e**).

544

Mean capture efficiencies, prior distributions, posterior distribution and point estimate datasets used in this study

Mean capture efficiency was set at 0.03 throughout the manuscript except for **Supplementary Fig. 2c and 3c** where a capture efficiency of 0.06 calculated from smFISH data was used. Prior distributions where generated as follows: **Fig. 1c**, **3a-b**, **3d-3e**, **4, and Supplementary Fig. 2b**, **2d-e**, **3c**, **4c**, **5a-c**, **6b-c** data were normalised using "global" priors obtained from all cells in the dataset in order to correct for batch effects. **Supplementary Fig. 2c** used priors calculated from rapidly growing cells. **Fig. 2a**, **2d-e**, **Supplementary Fig. 3a-c** data were normalised using "local" priors estimated within each individual dataset to exclude any residual contribution of batch differences to HVG calls. **Figure 3c** and **Supplementary Fig. 5d-e** data were normalised using "local" priors estimated independently for sets of either large (13-16μm) or small (8-10μm) cells to maximize sensitivity of DE analysis.

Detection of noisy or highly variable genes (Fig. 2)

Expression of a given gene can vary among cells of a population. Cell-to-cell variability in gene expression, also called noise, is defined as the coefficient of variation $(\sigma/\mu)^2$, where σ and μ are the standard deviation and the mean of expression scores across cells respectively. A number of modelling and experimental studies have shown that gene expression noise is inversely correlated with mean gene expression calculated across cells[19,20,66,67]. Genes with particularly high cell-to-cell variability are called noisy or highly variable genes (HVGs) and are defined as having significantly higher noise than most genes of similar means (**Fig. 2a**). Identifying HVGs from scRNA-seq experiments is challenging due to the strong technical noise present in the data and several teams have addressed this problem[64,66–68]. The general consensus is to decompose the total noise observed into its technical and biological components. To do this, the dependence of technical noise to the mean is measured and used to infer potential additional biological noise present for each gene (**Fig. 2a, Supplementary Fig. 3a**). Here, we have developed an original method for HVG detection based on bayNorm normalised data and sets of computed synthetic control RNA-seq data (synRNA) to estimate noise floors. synRNA are generated as follows: Given a scRNA-seq dataset with a raw count matrix x_{ij} and a vector of estimated cell-specific capture efficiencies β_j, we produced a set of synRNA data x_{ij}^{syn} with similar mean expressions and capture efficiencies as the real experimental data but with no biological variability above what is expected from the Poisson distribution. Poisson noise is the minimal amount of noise expected if no additional biological variability is present. To do this, we first generate a gene expression dataset x_{ij}^{Poisson} sampled from a Poisson distribution with mean expression obtained from raw count matrix x_{ij} and capture efficiencies β_j:

$$x_{ij}^{\text{Poisson}} = \text{Poisson}(\lambda_i), \text{ where } \lambda_i = < \frac{x_{ij}}{\sum_i x_{ij}} > < (\sum_i x_{ij})/\beta_j >$$

Both means above are calculated across cells (index j). In a final step, we used binomial downsampling to generate a x_{ij}^{syn} dataset from x_{ij}^{Poisson} simulating the effect of partial RNA capture during the scRNA-seq procedure:

$$x_{ij}^{\text{syn}} = \text{Binom}(x_{ij}^{\text{Poisson}}, \beta_j)$$

Finally, x_{ij}^{syn} data are normalised with bayNorm using prior parameters estimated from the original raw data (i.e. identical to those used for normalisation of the cells data). To identify highly variable genes, a local regression between noise and mean expression of all genes from the normalized synRNA dataset is calculated and compared to the noise levels observed in the corresponding normalised experimental datasets (log-log, **Fig. 2a**). To call genes with noise levels significantly above the synRNA fitted line, we use an approach similar to the one proposed in ref [64] and based on an adaptation of the gene expression variation model[68]. Briefly, vertical differences (only considering positive residuals) between noise levels in the experimental dataset and the fitted line were calculated (illustrated as ΔCV on **Fig. 2a**). The differences were normalised by dividing by the residuals from the regression, which follow a normal distribution. Most genes were assumed not to deviate from the centre of the distribution significantly. The centre was found by the kernel density of the normalized differences. Next, a normal distribution was fitted using differences which were below that centre. P-values were then extracted based on the normal distribution and adjusted using the Benjamini and Hochberg procedure. Noisy genes were called independently for each batch of 96 cells or ctrRNAs generated in this study after normalisation with bayNorm using priors estimated within each dataset ("local" priors). For each gene in each dataset, noise and mean values across cells where calculated using pooled expression scores of 100 samples of the bayNorm posterior distribution per cell. Using this design, gene variability was assessed in 100 bootstrapped versions of each dataset. Genes were called noisy if they had an FDR < 0.1 in 85 or more bootstrap samples. Genes called in at least one rapid growth cell dataset (2502_1, 2502_3, 2502_5, 2502_7, 2502_9, 1904_1, 0109_3, 1711_1, 1711_2), and none of the ctrRNA datasets (2502_2, 2502_4, 2502_6, 2502_8, 2502_1, 1904_2, 0408_5), were called HVGs and are discussed further in this study (**Supplementary Table 6**).

Functional analysis of HVG (Fig. S4d)

Levels of quantitative variables describing a series of genes features were compared between HVGs (including or not the top-500 periodic genes) and all other genes from the high confidence set using one-sided Wilcoxon tests. Features were obtained from four studies and are listed below[19,34,35,37]. Labels were adjusted to be self-explanatory and a detailed description of each feature is available in the original publications.

1) Features as in the additional file 2 from Koch et al[35]. Label from Figure 4d are listed with labels from the original additional file 2 between parentheses. "Yeast conservation" (Yeast.conservation), "dN/dS"

(dN.dS), “Multifunctionality (associated GO terms)” (Multifunctionality), “Disordered domains [%]” (Disorder), “Number of physical protein interactions” (PPI.degree), “Fitness” (SM.fitness.defect), “Copy number (paralogues number)” (Copy.number), “Codon Adaptation Index (CAI)” (CAI), “Codon unsage bias (Nc)” (Nc), “Number of co-expressed genes” (Co.expression.degree), “Protein length” (Protein.length), “Expression level (RNA)” (Expression.level), “Expression variation (RNA, Koch et al)” (Expression.variation), “Number of protein domains” (Num.of.domains), “Number of single protein domains” (Num.of.unique.domains), “Broad conservation” (Broad.conservation), “Negative genetic interaction degrees” (Observed.GI.degree).

2) From Rhind et al[37]: “Evolutionary rates (Rhind et al)” (“Rate” values from table S30).

3) From Pancaldi et al[34]: “Expression variation (RNA, Pancaldi et al)” (Between condition variability score, Table S1).

4) From Newman et al[19]: “Cell-to-cell variability (YEPD, Newman et al)” (DM values in YEPD), “Cell-to-cell variability (SD, Newman et al)” (DM values in SD).

Transcriptome fractions of functional categories (Fig. 3b)

To assign cells to particular functional categories, we calculated z-scores for the sums of the counts of each category within each cell. Cells were assigned to a given category if the category |z-score| was > 1.2 in more than 70 out of 100 bayNorm posterior samples. Categories with assigned cells significantly larger or smaller than the whole population are shown on **Fig. 3b** ($P_{Wilcox} < 0.05$).

Differential Gene Expression analysis (Fig. 3c, Supplementary Fig. 5d-e)

Several Differential Expression (DE) analysis methods, tailored for scRNA-seq analysis have been published. A recent comparative analysis[69] and our own experience identifies MAST[70] as a reliable method. Therefore, in this study, we used the MAST package with method = “glm”, the “ebayes” option enabled, and considering adjusted P-values from the continuous part of the hurdle model utilized in MAST (multiple testing adjustment method: Benjamini and Hochberg)[70]. DE detection was run independently on 100 bayNorm posterior distribution samples for **Fig. 3**. Genes called DE in > 90% of the posterior distributions were considered differentially expressed. Log_2 ratios are the mean of the Log_2 ratios from each posterior distribution. An additional cut off on log_2 ratios (>0.2 or <-0.2) was used on **Fig. 3**. **Figure 3** DE analysis used two sets of cells, either large (13-16µm) or small (8-10µm).

For this analysis, “large” or “small” cells sets were normalised by bayNorm using different local priors specific for each set. In order to demonstrate that our DE analysis reflected gene expression differences related to different cell sizes and is not an artefact of the use of local priors, we performed the following experiment. Two sets of 50 cells were selected from the “large” or “small” cells sets and normalised using bayNorm and local priors. This subsampling experiment was repeated 20 times. In parallel, the labels of the “big” and “small” cells were randomised in advance before subsampling and normalisation

as above. Both groups were then used for DE. As expected, this second randomised set showed barely any genes with robust and reproducible differential expression validating our approach (**Supplementary Fig. 5e**).

Random forest model

To identify genes that have non-linear correlation of expression levels with cell size, we built a random forest model of cell size given gene expression levels. We chose a subset of normal cells and applied a filtering criterion such that cells with total counts less than 100,000 or greater than 300000 were removed. In addition, we filtered out cells with length <6µm or >25µm. We then applied a Random Forest model described in[71] on the filtered dataset. Genes were ranked according to the importance statistic "%IncMSE" returned by the model (**Supplementary Table 6**).

YMC analysis and validation (Fig. S5d)

Gene signatures of the three proposed YMC phases: OX (oxidative), RB (reductive building) and RC (reductive charging) were obtained from ref[42]. Fission yeast orthologues were identified for each signature and DE ratio between large and small cells from **Fig. 3c** plotted for each lists (**Supplementary Fig. 5d**). To add statistical support to the observed expression patterns, we used the following rationale. **Supplementary Fig. 5d** shows that big/small DE ratios increase from the RC to OX and RB phases. This pattern leads to a positive correlation between YMC cycle steps and DE ratio. We compared the slope of the regression line between DE ratios and steps of the YMC and compared it to 1000 datasets where the same ratios were randomised between YMC steps, or where ratios were randomly sampled from the whole dataset. The R squared values from our data where significantly higher than those from the permutation (z-scores, $p < 10^{-4}$). The same was true when using Pearson correlations. Together, this analysis confirms that the observed pattern is unlikely to have arisen by chance.

Promoter sequence analysis

Promoter sequences of HVGs (position relative to TSS, -300 to +100) were analysed using genes that where neither HVGs nor false positives as a reference set. We used the tool CentriMo[72] available as part of the MEME software suite to identify known motifs enriched in these promoters based on the YEASTRACT lists of budding yeast motifs (**Supplementary Table 10**).

Data availability

All raw RNA-seq datasets are available in ArrayExpress accession number E-MTAB-6825. Cell size measurements and all smFISH data are available as Supplementary Material. The bayNorm package in available from Bioconductor https://bioconductor.org/packages/release/bioc/html/bayNorm.html.

References

1. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6,** 451–64 (2005).

2. Shahrezaei, V. & Swain, P. S. The stochastic nature of biochemical networks. *Curr. Opin. Biotechnol.* **19,** 369–74 (2008).

3. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135,** 216–26 (2008).

4. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–6 (2002).

5. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.* **10,** 443–456 (2009).

6. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell* **163,** 1596–1610 (2015).

7. Shahrezaei, V. & Marguerat, S. Connecting growth with gene expression: of noise and numbers. *Curr. Opin. Microbiol.* **25,** 127–135 (2015).

8. Mellor, J. The molecular basis of metabolic cycles and their relationship to circadian rhythms. *Nat. Struct. Mol. Biol.* **23,** 1035–1044 (2016).

9. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13,** 599–604 (2018).

10. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science (80-.).* **358,** 58–63 (2017).

11. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat. Rev. Cancer* **17,** 557–569 (2017).

12. Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14,** e8046 (2018).

13. Saliba, A.-E., C Santos, S. & Vogel, J. New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.* **35,** 78–87 (2017).

14. 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).

15. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. *Characterization of directed differentiation by high-throughput single-cell RNA-Seq*. As preprint: *bioRxiv* (Cold Spring Harbor Labs Journals, 2014). doi:10.1101/003236

16. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–81 (2014).

17. Tang, W. *et al.* bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. As preprint: *bioRxiv* 384586 (2018). doi:10.1101/384586

18. Marguerat, S. *et al.* Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **151,** 671–683 (2012).

19. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441,** 840–6 (2006).

20. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38,** 636–643 (2006).

21. Rustici, G. *et al.* Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*

731 **36,** 809–17 (2004).

732 22. Peng, X. *et al.* Identification of Cell Cycle-regulated Genes in Fission Yeast. *Mol. Biol. Cell* **16,**
733 1026–1042 (2005).

734 23. Oliva, A. *et al.* The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol.* **3,**
735 e225 (2005).

736 24. Marguerat, S. *et al.* The more the merrier: comparative analysis of microarray studies on cell
737 cycle-regulated genes in fission yeast. *Yeast* **23,** 261–277 (2006).

738 25. Cooper, S. On a heuristic point of view concerning the expression of numerous genes during
739 the cell cycle. *IUBMB Life* **64,** 10–17 (2012).

740 26. Duncan, C. D. S., Rodríguez-López, M., Ruis, P., Bähler, J. & Mata, J. General amino acid
741 control in fission yeast is regulated by a nonconserved transcription factor, with functions
742 analogous to Gcn4/Atf4. *Proc. Natl. Acad. Sci.* **115,** E1829–E1838 (2018).

743 27. Ravarani, C. N. J., Chalancon, G., Breker, M., de Groot, N. S. & Babu, M. M. Affinity and
744 competition for TBP are molecular determinants of gene expression noise. *Nat. Commun.* **7,**
745 10417 (2016).

746 28. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations
747 in gene expression. *Nat. Genet.* **38,** 830–834 (2006).

748 29. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J. & Hartl, D. L. Genetic Properties
749 Influencing the Evolvability of Gene Expression. *Science (80-.).* **317,** 118–121 (2007).

750 30. Lehner, B. Selection to minimise noise in living systems and its implications for the evolution of
751 gene expression. *Mol. Syst. Biol.* **4,** 170 (2008).

752 31. Blake, W. J., KÆrn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression.
753 *Nature* **422,** 633–637 (2003).

754 32. Weinberger, L. *et al.* Expression noise and acetylation profiles distinguish HDAC functions.
755 *Mol. Cell* **47,** 193–202 (2012).

756 33. Chen, D. *et al.* Global transcriptional responses of fission yeast to environmental stress. *Mol.*
757 *Biol. Cell* **14,** 214–229 (2003).

758 34. Pancaldi, V., Schubert, F. & Bähler, J. Meta-analysis of genome regulation and expression
759 variability across hundreds of environmental and genetic perturbations in fission yeast. *Mol.*
760 *Biosyst.* **6,** 543–52 (2010).

761 35. Koch, E. N. *et al.* Conserved rules govern genetic interaction degree across species. *Genome*
762 *Biol.* **13,** R57 (2012).

763 36. López-Maury, L., Marguerat, S. & Bähler, J. Tuning gene expression to changing
764 environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* **9,** 583–593
765 (2008).

766 37. Rhind, N. *et al.* Comparative functional genomics of the fission yeasts. *Science* **332,** 930–6
767 (2011).

768 38. Lane, K. *et al.* Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to
769 Dynamic Patterns of NF-κB Activation. *Cell Syst.* **4,** 458–469.e5 (2017).

770 39. Cadwell, C. R. *et al.* Electrophysiological, transcriptomic and morphologic profiling of single
771 neurons using Patch-seq. *Nat. Biotechnol.* **34,** 199–203 (2016).

772 40. Nichterwitz, S. *et al.* Laser capture microscopy coupled with Smart-seq2 for precise spatial
773 transcriptomic profiling. *Nat. Commun.* **7,** 12139 (2016).

774 41. Turner, J. J., Ewald, J. C. & Skotheim, J. M. Cell size control in yeast. *Curr. Biol.* **22,** R350-9
775 (2012).

776 42. Kuang, Z. *et al.* High-temporal-resolution view of transcription and chromatin states across

distinct metabolic states in budding yeast. *Nat. Struct. Mol. Biol.* **21,** 854–863 (2014).

43. Silverman, S. J. *et al.* Metabolic cycling in single yeast cells from unsynchronized steady-state populations limited on glucose or phosphate. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6946–51 (2010).

44. Slavov, N., Airoldi, E. M., van Oudenaarden, A. & Botstein, D. A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Mol. Biol. Cell* **23,** 1986–97 (2012).

45. Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends Genet.* **28,** 560–5 (2012).

46. Schmoller, K. M. & Skotheim, J. M. The Biosynthetic Basis of Cell Size Control. *Trends Cell Biol.* (2015). doi:10.1016/j.tcb.2015.10.006

47. Mata, J., Lyne, R., Burns, G. & Bähler, J. The transcriptional program of meiosis and sporulation in fission yeast. *Nat. Genet.* **32,** 143–7 (2002).

48. Metzl-Raz, E. *et al.* Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *Elife* **6,** (2017).

49. Malecki, M. *et al.* Functional and regulatory profiling of energy metabolism in fission yeast. *Genome Biol.* **17,** 240 (2016).

50. Moreno, S., Klar, A. & Nurse, P. Molecular genetic analysis of fission yeast Schizosaccharomyces pombe. *Methods Enzymol.* **194,** 795–823 (1991).

51. Takahashi, C. N., Miller, A. W., Ekness, F., Dunham, M. J. & Klavins, E. A Low Cost, Customizable Turbidostat for Use in Synthetic Circuit Characterization. *ACS Synth. Biol.* **4,** 32–38 (2015).

52. Semrau, S. *et al.* Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.* **8,** 1096 (2017).

53. Keifenheim, D. *et al.* Size-Dependent Expression of the Mitotic Activator Cdc25 Suggests a Mechanism of Size Control in Fission Yeast. *Curr. Biol.* **27,** 1491–1497.e4 (2017).

54. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27,** 491–499 (2017).

55. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–6 (2014).

56. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42,** W187–W191 (2014).

57. Lun, A. T. L., Calero-Nieto, F. J., Haim-Vilmovsky, L., Göttgens, B. & Marioni, J. C. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* **27,** 1795–1806 (2017).

58. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14,** 381–387 (2017).

59. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14,** 565–571 (2017).

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

61. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

62. Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14,** 584–586 (2017).

63. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA

sequencing data with many zero counts. *Genome Biol.* **17,** 75 (2016).

64. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5,** 2122 (2016).

65. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative Single-cell Transcriptomics. *Brief. Funct. Genomics* (2018). doi:10.1093/bfgp/ely009

66. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).

67. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).

68. Chen, H.-I. H., Jin, Y., Huang, Y. & Chen, Y. Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics* **17,** 508 (2016).

69. Jaakkola, M. K., Seyednasrollah, F., Mehmood, A. & Elo, L. L. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* **18,** bbw057 (2016).

70. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16,** 278 (2015).

71. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

72. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* **40,** e128 (2012).

Figure legends

Figure 1: *Imaging and transcriptome analysis of single fission yeast cells*

(a) Experimental and analysis pipelines. Batches of 96 single cells were imaged and isolated using a MSM-400 dissection microscope (Singer Instruments). Single-cell RNA-sequencing data were generated using SCRB-seq for yeast, normalised using bayNorm[17], and used for functional analysis. **(b)** Overall coverage of scRNA-seq datasets and selection of high-confidence genes. The highest raw count of each coding and non-coding gene observed across 21 datasets (n=2028 cells) is plotted as a function of the number of cells in which it was detected. All genes (blue, n=6646 genes) and the high-confidence genes used for all further analysis in this study (red, n=1011 genes) are shown. High-confidence genes were defined as genes that represented >0.16% of the transcriptome of at least one cell. **(c)** Gene expression levels in single cells. Normalised scRNA-seq counts for high-confidence genes are plotted as a function of population average mRNA copies/cell data from Marguerat et al[18]. Normalised counts in single cells (blue, n=2,050,308 measurements), and average counts across cells (red) are shown. $R_{pearson}$ = 0.61 for average counts, (n=1011 genes) and 0.48 when including low-confidence genes.

Figure 2: *Cell-to-cell variability of fission yeast transcriptome*

(a) Identification of highly variable genes (HVG). Coefficients of variation of normalised counts are plotted against their respective mean expression for: all filtered genes (grey, mostly hidden, n=1001 genes), genes called variable (red, n=125 genes), or simulated synthetic controls (blue, synRNA, n=1001 genes; **Methods**). Genes are called variable if their ΔCV is significantly higher than the distribution of ΔCVs from synthetic controls of similar mean expression using z-scores and assuming normality ($P < 0.1$ in >= 85% of bootstrapped samples; **Methods**). The dotted line represents a Loess fit to synRNAs and the ΔCV of an example variable gene is highlighted by an arrow. **(b)** Validation by smFISH of mRNA called variable from scRNA-seq data. Representative smFISH images are shown for low-variability control *rpb1* mRNA and three mRNAs with different levels of variability (*lsd90*, *mot1*, and SPAC27D7.09c, see also **Supplementary Fig. 2c, 3c-d**). Scale bars representing 5μm and size-corrected Fano factors (scFano) are indicated on each plot. **(c)** Boxplot showing size-corrected Fano factors measured by smFISH for HVGs (n=8 genes) or control genes (n=5 genes). P values for a one-sided Wilcoxon test are shown above the figure. Boxplots represent median, interquartile range, and most extreme data points that are not more than 1.5 times the interquartile range. **(d)** Functional analysis of variable genes from 9 datasets of 96 cells (n=864 cells) during rapid proliferation. Significance of overlap of variable mRNAs in each dataset with selected functional categories is shown ($-\log_{10}$ of $P_{Fisher\ one\text{-}sided}$ for overlap of HVGs from each dataset and each category, p-values are corrected for multiple testing using Benjamini and Hochberg, number of tests = 14 categories/dataset, as shown on the figure). False positive mRNAs called from total RNA control experiments have been filtered out (**Supplementary Fig. 3b**). Note that some categories are more pervasively variable across datasets

than others. **(e)** Functional analysis of non-periodic HVGs. HVGs that not among the top 500 most cell-cycle periodic genes in cell populations were sorted into three categories with “low”, “moderate”, or “high” pervasive variable expression (n = 235 genes). Selected distinctive gene functions are shown in the plot. Boxplots represent median, interquartile range, and most extreme data points that are not more than 1.5 times the interquartile range.

Figure 3: ***Cell-size dependence of fission yeast transcriptome***

(a) Cell length distribution across 864 cells during rapid proliferation and global characteristics of the corresponding transcriptomes. Average raw expression scores (blue) or average bayNorm normalised expression scores (green) are shown for cell length bins of 1µm. Note the positive correlation of raw scores with cell size that is lost after normalisation. **(b)** Single cells were assigned to functional categories based on their relative transcriptome signatures. Boxplots show cells assigned to categories associated with cell sizes significantly smaller (blue, $P_{Wilcox\ one\text{-}sided} < 0.05$) or larger (red, , $P_{Wilcox\ one\text{-}sided} < 0.05$) than the overall population (green). Boxplots are overlaid onto the cell size frequency histogram shown in (a). The vertical line marks the average cell length in the dataset. Boxplots represent median, interquartile range, and most extreme data points that are not more than 1.5 times the interquartile range. **(c)** Differential expression (DE) analysis between large (13-16µm, n=292 cells) and small (8-10µm, n=281 cells) cells performed using the MAST package[70]. Number of bootstrap iterations showing significant DE call are plotted for each gene ($P_{MAST} <0.05$, total iterations = 100) as a function of MAST $\log_2$ DE ratios. Genes significantly induced in small and large cells are highlighted in blue and red, respectively (cut-off: number of significant iterations > 90 and absolute $\log_2$ ratio > 0.2, **Methods**, **Supplementary Table 6**). Selected functional categories significantly enriched in either list are shown along with enrichment p-values ($P_{Fisher\ one\text{-}side}$, p-values are corrected for multiple testing using Benjamini and Hochberg). **(d)** Transcripts that change in concentration during the G2 growth phase (non-scaling genes, NSG). Average bayNorm expression scores were computed in bins of 1µm for cells shorter than 11µm, normalised to the smallest size bin and used for k-means clustering (n=414 cells, **Supplementary Fig. 5a**). Only genes with significant linear correlation with cell size were included in this analysis (n=78 genes, $P_{Pearson\ two\text{-}sided} <0.05$). Boxplots represent median and interquartile range. **(e)** Co-regulation of NSG clusters. Pearson correlations between clusters from (d) are shown.

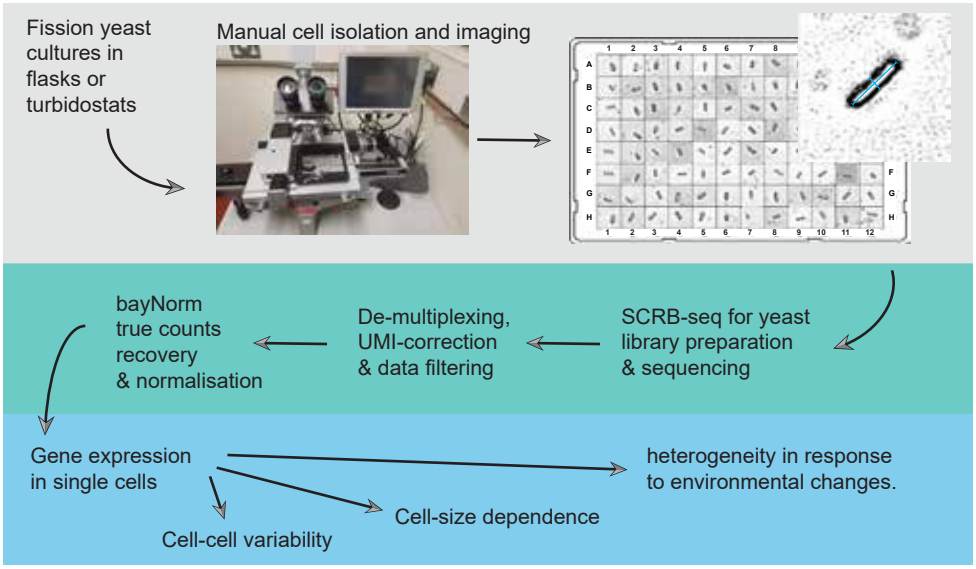
Figure 4: ***Gene-expression heterogeneity of fission yeast populations in response to environmental changes***

(a) Single cells show distinct stress signatures of gene expression in response to different external conditions. PCA analysis of normalised gene expression scores for the core environmental stress response genes (CESR). A total of 1824 cells growing in different external conditions and including cells from Fig 1-3 were analysed (**Supplementary Table 1-4**). Each condition is colour coded as per the legend on the right, and larger groups are circled and annotated (n = 1824 cells). **(b)** As in (a) showing genes specific for the heat-shock response (n = 1824 cells)[33]. **(c)** As in (a), showing genes specific for

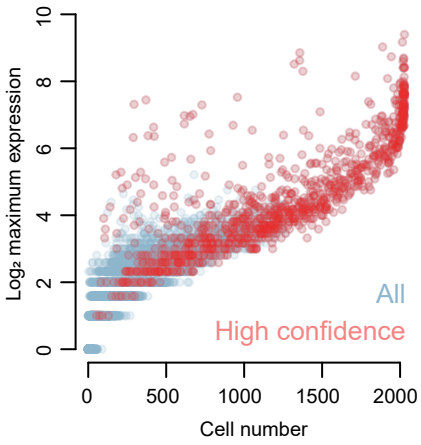
the oxidative-stress response (n = 1824 cells)[33]. **(d)** Heterogeneity in CESR gene expression during rapid proliferation and entry into stationary phase. Average CESR gene expression per cell plotted as a function of cell size. Cell density is colour coded as per the legend on the right (n = 1056 cells). **(e)** Related to (d) showing between cell coefficients of variation in CESR gene expression (main panel) and average expression (insert). Note the strong increase in average CESR expression per cell accompanied by increased expression heterogeneity occurring at higher cell densities (n = 1056 cells). Boxplots represent median, interquartile range, and most extreme data points that are not more than 1.5 times the interquartile range. **(f)** Heterogeneity in CESR gene expression during acute response and adaptation to heat shock. Average CESR gene expression per cell plotted as a function of cell size. Conditions are colour coded as per legend on the right (n = 576 cells). **(g)** Related to (f) showing between cell coefficients of variation in CESR gene expression (main panel) and average expression (insert) (n = 576 cells). Note the acute increase in average expression per cell of heat-shock genes and the lack of increase in expression heterogeneity during acute and adaptive responses. Boxplots represent median, interquartile range, and most extreme data points that are not more than 1.5 times the interquartile range.

Figure 1, Saint et al.

a



b



c

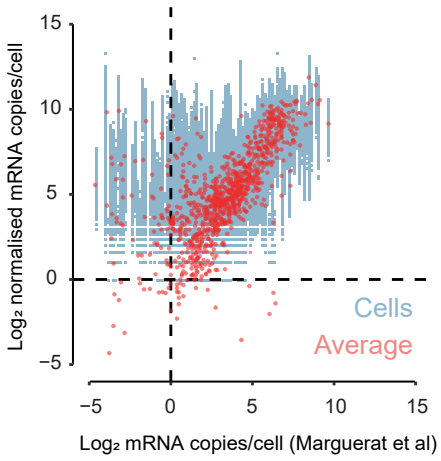


Figure 2, Saint et al.

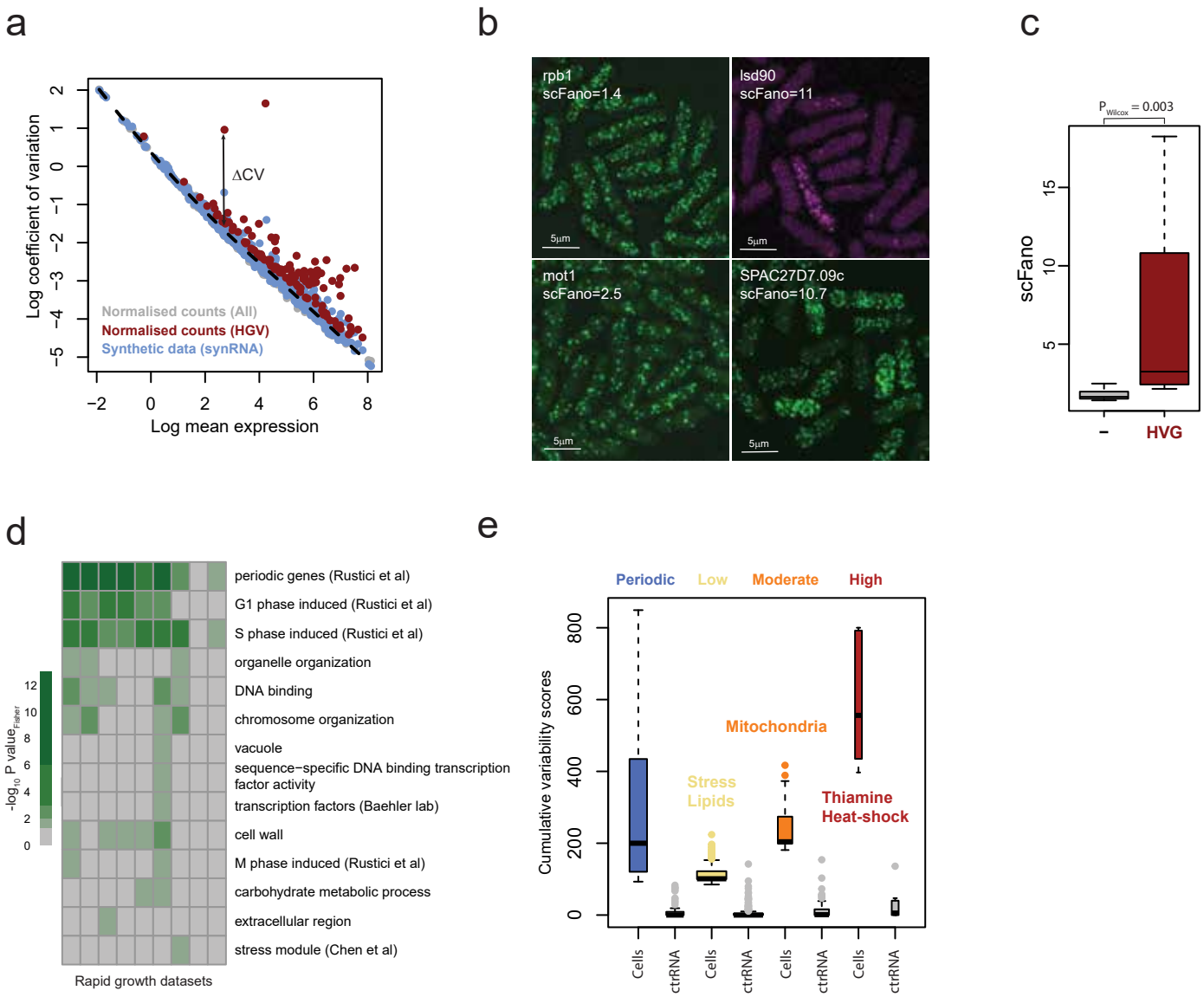


Figure 3, Saint et al.

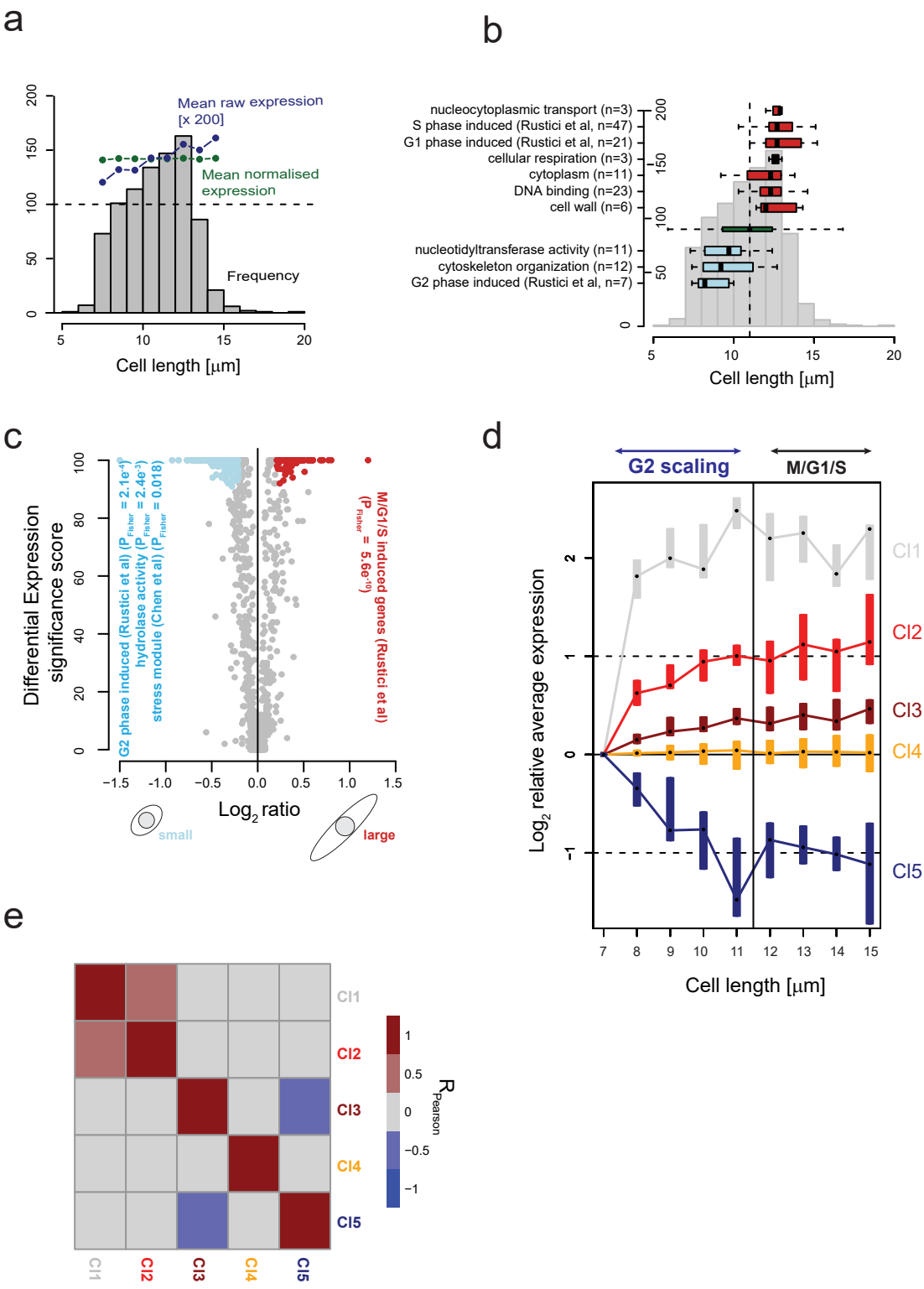
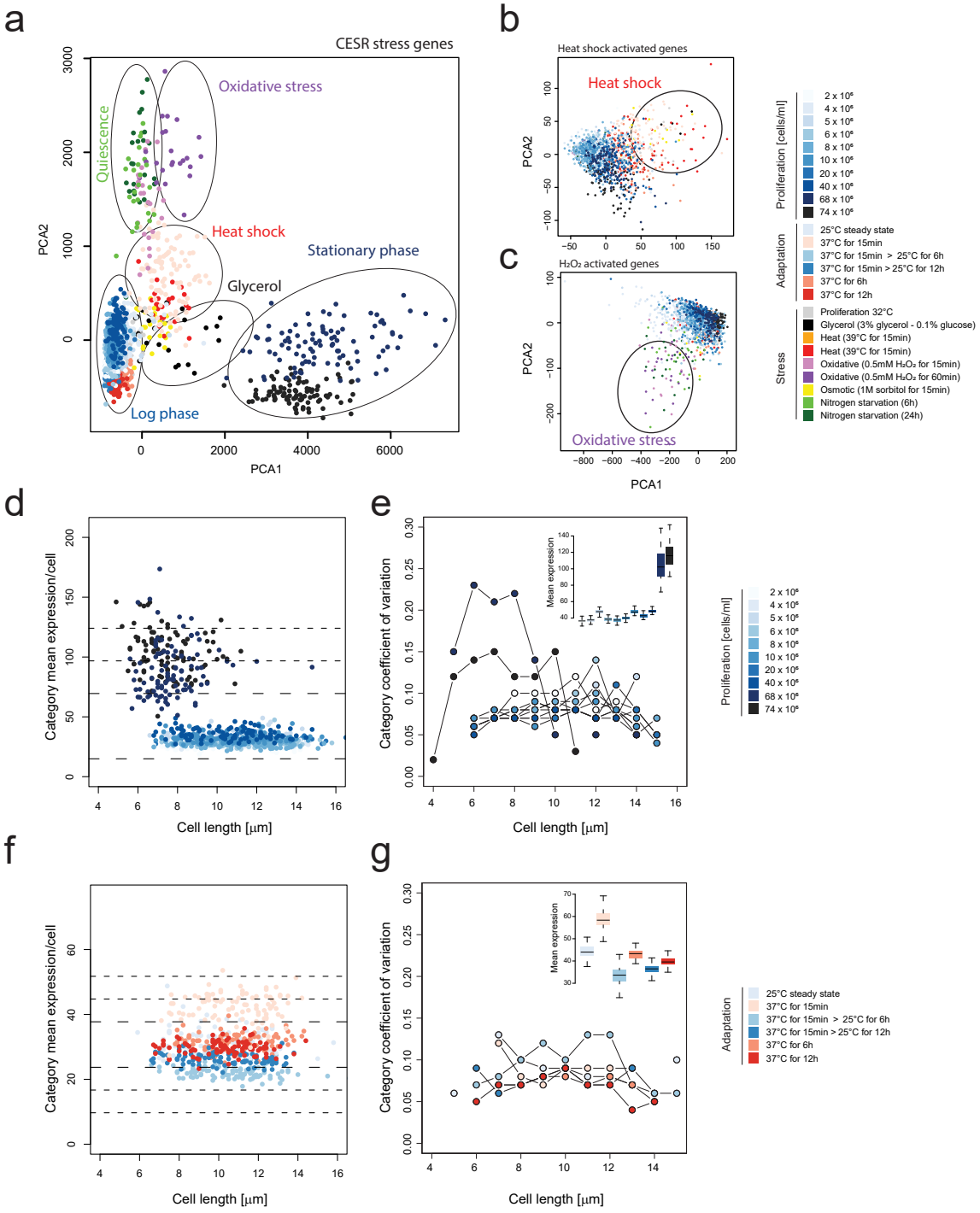


Figure 4, Saint et al.



Supplementary figures and references

Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation

Malika Saint[1,2], François Bertaux[1,2,3,*,§], Wenhao Tang[3,*], Xi-Ming Sun[1,2], Laurence Game[1,2], Anna Köferle[4,†], Jürg Bähler[4], Vahid Shahrezaei[3,5] and Samuel Marguerat[1,2,5]

[1] MRC London Institute of Medical Sciences (LMS), Du Cane Road, London W12 0NN, UK

[2] Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, Du Cane Road, London W12 0NN, UK

[3] Department of Mathematics, Faculty of Natural Sciences, Imperial College London, London SW7 2AZ, UK

[4] University College London, Research Department of Genetics, Evolution & Environment and UCL Genetics Institute, London, WC1E 6BT, UK.

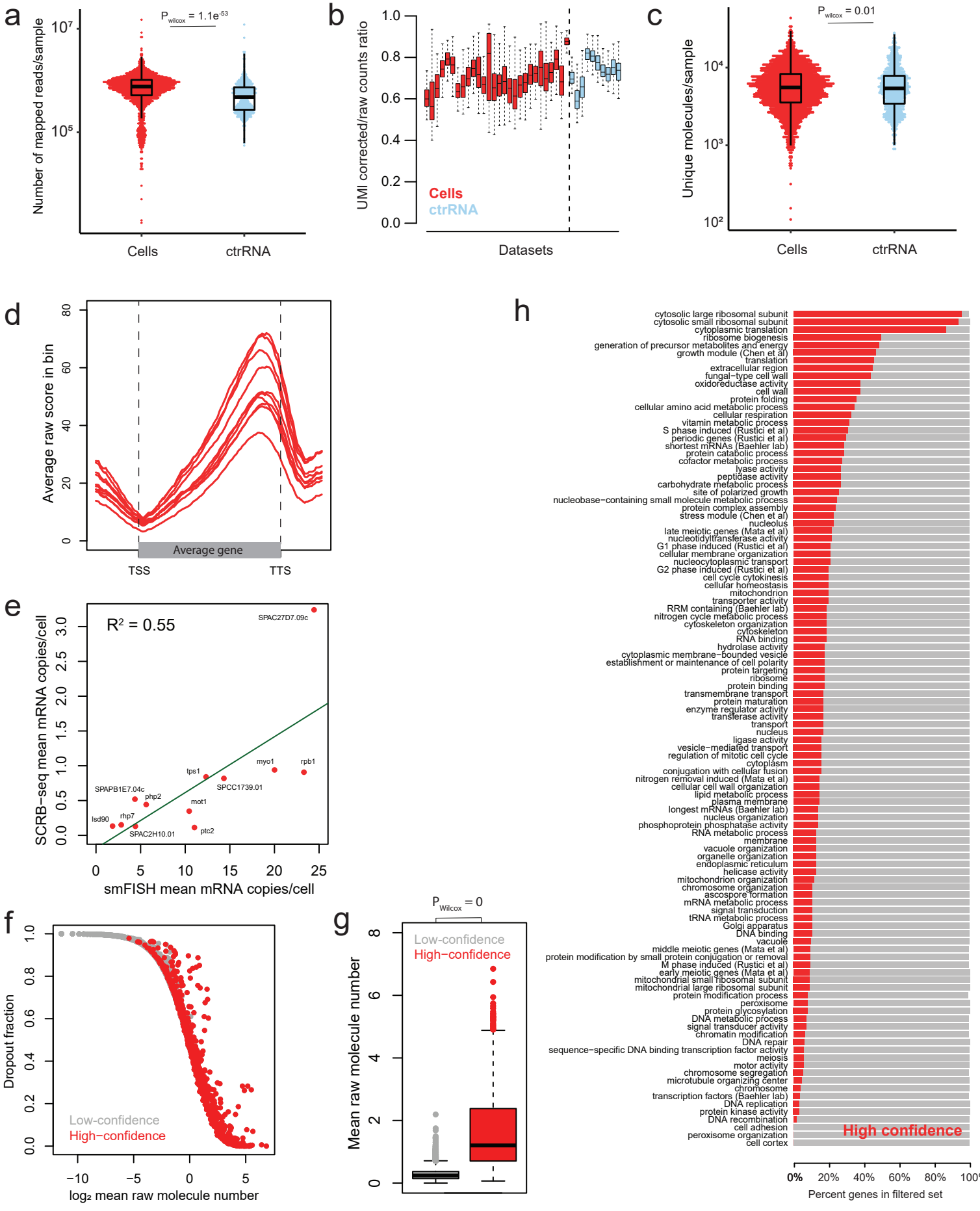
[5] Correspondence to v.shahrezaei@imperial.ac.uk or samuel.marguerat@imperial.ac.uk

* These authors contributed equally to this work.

§ Present address: Institut Pasteur, 28 rue du Docteur Roux, 75015 Paris, France.

† Present address: Munich Center for Neurosciences, Ludwig-Maximilian-Universität, BioMedical Center, 82152 Planegg, Germany.

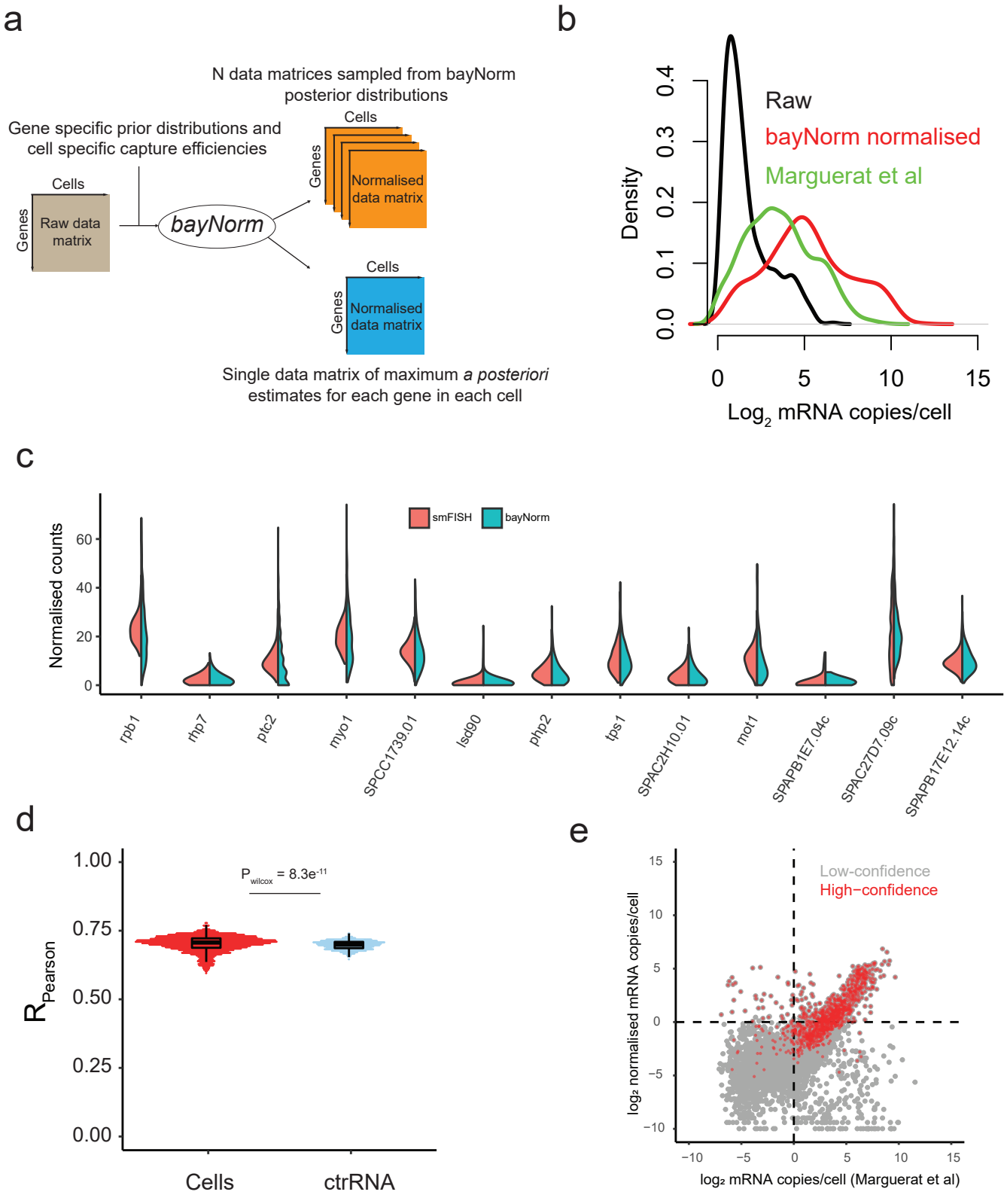
Figure S1, Saint et al.



Supplementary Figure 1: Supplement to *Imaging and transcriptome analysis of single fission yeast cells I*

(a) Distribution of numbers of mapped reads per sample in cells (red, n=2028 cells) and ctrRNA (blue, n=780 samples) datasets. P-value of a one-sided Wilcoxon test is shown. Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(b)** Fraction of molecules that passed filter after correcting for UMI errors (Hamming distance < 2, n = 3672 samples). Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(c)** Distribution of unique mRNA molecule numbers per sample after UMI filtering in cells (red, n=2028 cells) and ctrRNA (blue, n=780 samples) datasets. P-value of a one-sided Wilcoxon test is shown. Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(d)** Average gene coverage plot for 10 cells (dataset 2501_1, n=10 cells). Profiles were obtained using the 'deeptools' package[1] using the default 10nt bin size and 300nt flanking regions. **(e)** Comparison of mean UMI-corrected sequencing counts, or molecule numbers, and average absolute mRNA counts defined by smFISH for 12 genes (common names labelled on the plot). Linear regression line is shown in green and its R^2 coefficient is shown in the top left corner. For bayNorm data, n=864 cells. For smFISH data, n=194 cells (rpb1, lsd90), n=106 cells (rhp7, mot1), n=224 cells (ptc2, php2, SPAC27D7.09c), n=199 cells (SPBC146.13c, SPAPB1E7.04c, SPAPB17E12.14c), n=207 cells (SPCC1739.01, SPAC328.03, SPAC2H10.01). **(f)** Fraction of cells with mRNA molecule number = 0 (dropout) as a function of the mean molecule number of each gene. High-confidence genes (red, n=1011 genes) and low-confidence genes, which were filtered out (grey, n=6037 genes) are shown. **(g)** Boxplot showing the mean mRNA molecule numbers after removing zero values. High-confidence genes (right, red, n=1011 genes) and low-confidence genes, which were filtered out (left, grey, n=6037 genes) are shown. P-value of a one-sided Wilcoxon test is shown. Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(h)** Functional analysis of high-confidence gene dataset used in this study. The proportion of each gene category present in the high-confidence dataset is indicated in red. Mean coverage across categories is 19.8% (n=1011 genes).

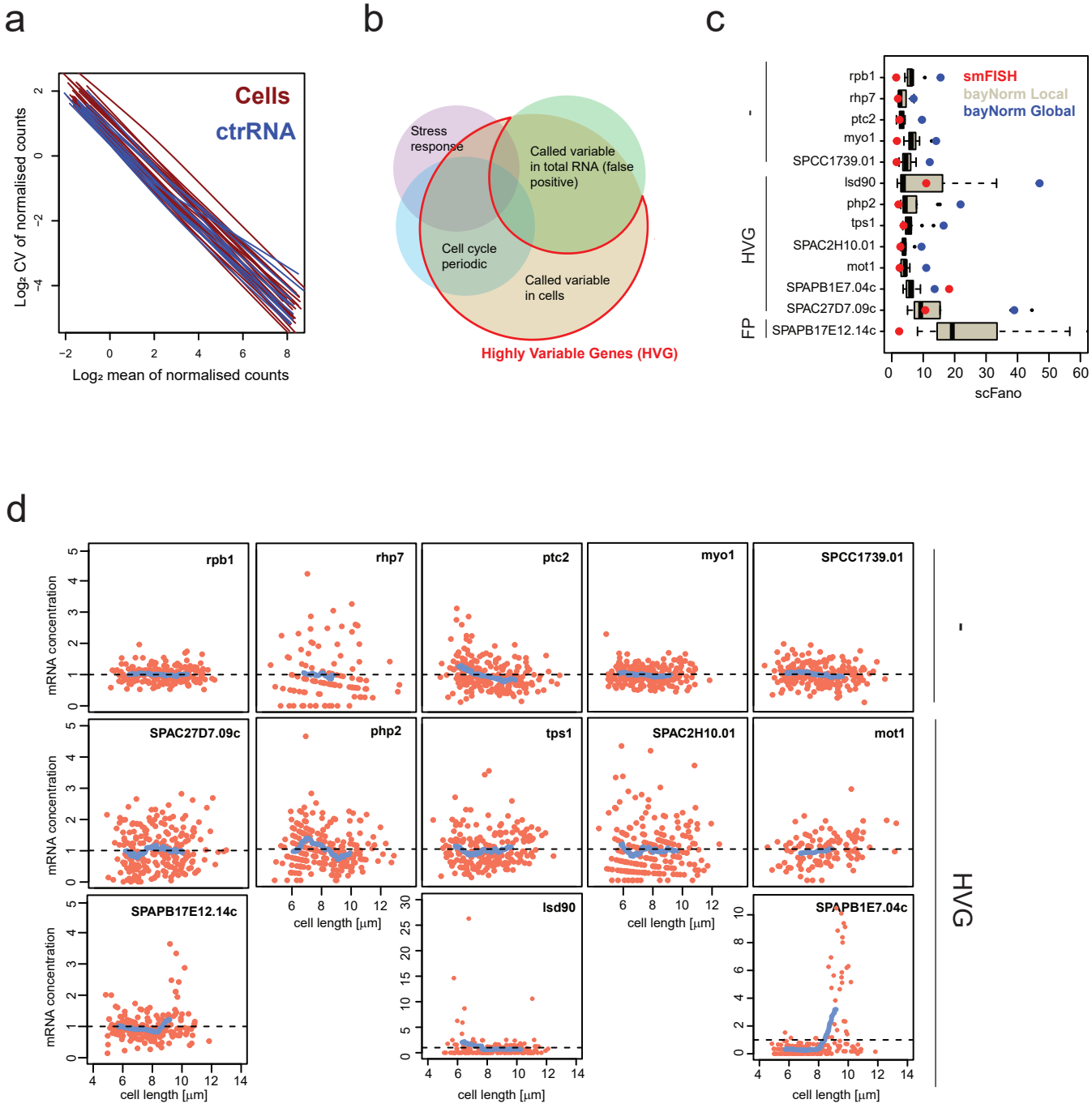
Figure S2, Saint et al.



Supplementary Figure 2: Supplement to *Imaging and transcriptome analysis of single fission yeast cells II*

(a) Cartoon of bayNorm normalisation procedure. **(b)** Comparison of raw and bayNorm-normalised scRNA-seq counts averaged across cells with population estimates from Marguerat et al[2] (n=1011 genes). The shift between the green and red curves depends on the choice of $\langle\beta\rangle$. **(c)** Violin plots showing smFISH molecule numbers (left half, blue) and bayNorm normalised scores (right half, red) for 13 genes using $\langle\beta\rangle = 0.06$. Data were median centred to allow for cross-comparison. Note the similar count distribution between both approaches. For bayNorm data, n=864 cells. For smFISH data n=194 cells (rpb1, lsd90), n=106 cells (rhp7, mot1), n=224 cells (ptc2, php2, SPAC27D7.09c), n=199 cells (SPBC146.13c, SPAPB1E7.04c, SPAPB17E12.14c), n=207 cells (SPCC1739.01, SPAC328.03, SPAC2H10.01). **(d)** Pearson correlation between bayNorm-normalised scRNA-seq counts (molecule numbers/cell) in individual cells (red, n=2028 cells) or ctrRNA (blue, n=780 samples) datasets, and population average mRNA copies/cell data from Marguerat et al[2]. P-value of a one-sided Wilcoxon test is shown. Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(e)** Average gene expression levels in single cells. Average bayNorm-normalised scRNA-seq counts are plotted as a function of population average mRNA copies/cell data from Marguerat et al[2]. High-confidence genes (red, n=1011 genes) and low-confidence genes, which were filtered out (grey, n=6037 genes) are shown. $R_{pearson} = 0.6$ for high-confidence genes and $R_{pearson} = 0.48$ for all fission yeast genes.

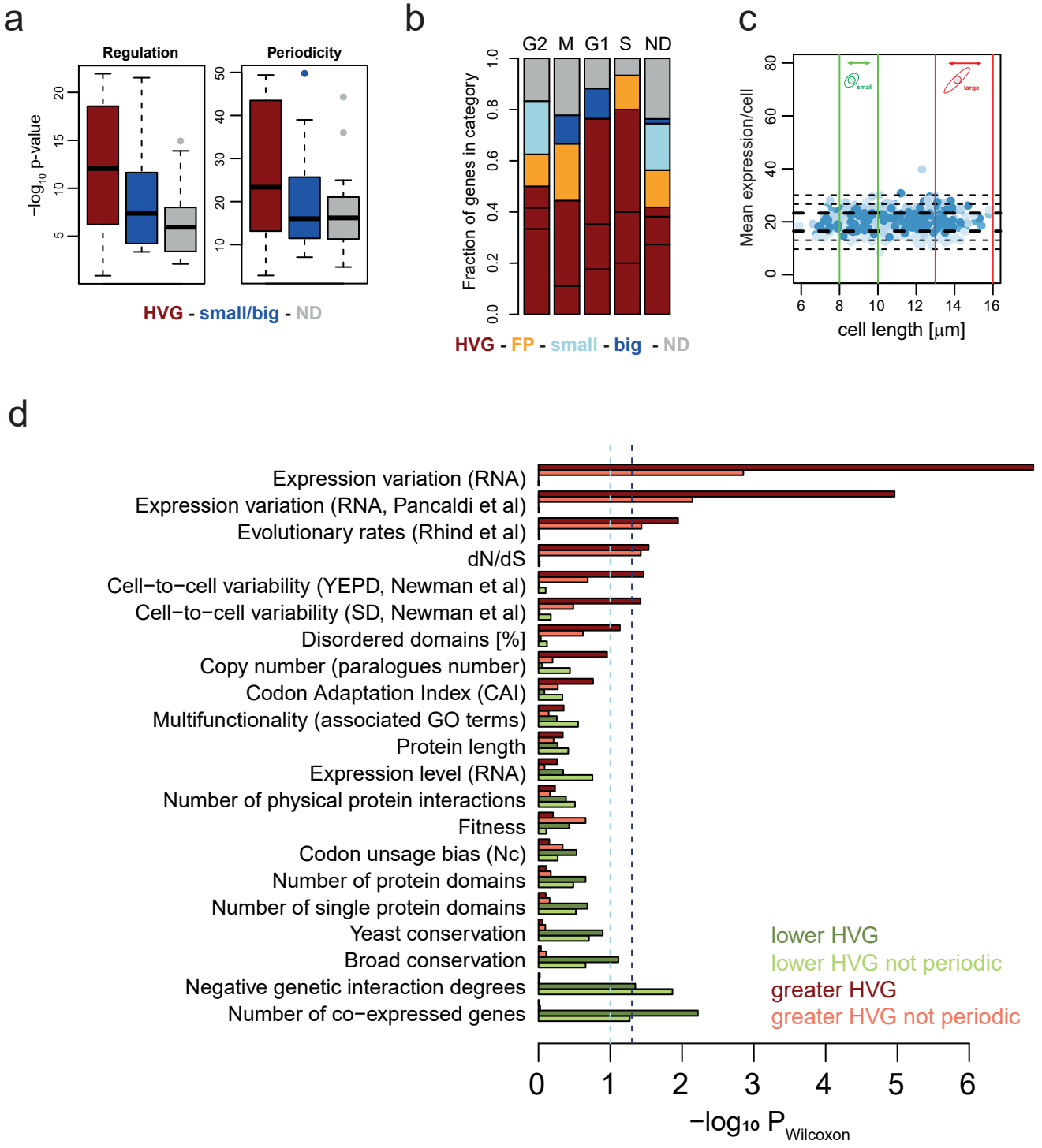
Figure S3, Saint et al.



Supplementary Figure 3: Supplement to *Cell-to-cell variability of the fission yeast transcriptome I*

(a) Relation between mean expression and coefficient of variation (CV) for cell (red, n=28 datasets) and ctrRNA datasets (blue, n=11 datasets). Loess fits are plotted. Note the strong inverse relationship across all expression levels. **(b)** Identification of highly variable mRNAs. Variable mRNAs were called from 9 datasets of 96 cells, and 7 sets of total ctrRNA controls to define false positives. Genes called variable in at least 1 cell dataset, but never in ctrRNAs, constitute the HVG list used throughout this study (red crescent-shaped area, **Supplementary Table 6**). Note that mRNAs described as cell-cycle regulated or part of the core environmental stress response (CESR) show very low levels of false positives, thus validating the approach. **(c)** Size-corrected Fano factors of 13 genes measured by smFISH (red dot), by scRNA-seq normalised using priors local to each of the 9 datasets used for calling HVGs (grey boxplots), or by scRNA-seq normalised using global priors estimated across all the cells form the 9 datasets used for calling HVGs (blue dots). Note that the global normalisation tends to overestimate variability (**Methods**). Label on the left marks control genes (-), highly variable genes (HVG) or false positives (FP) according to the analysis from **Fig. 2**. scFano factors were obtained from n=864 cells for bayNorm data. For smFISH data, scFano factors were obtained from n=194 cells (rpb1, lsd90), n=106 cells (rhp7, mot1), n=224 cells (ptc2, php2, SPAC27D7.09c), n=199 cells (SPBC146.13c, SPAPB1E7.04c, SPAPB17E12.14c), n=207 cells (SPCC1739.01, SPAC328.03, SPAC2H10.01). Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(d)** smFISH measurements of mRNA concentrations/cell plotted as a function of cell length. Moving averages are shown in blue (for cell numbers and data see **Supplementary Table 7**). Label on the right marks control genes (-) or highly variable genes (HVG). The rhp7 genes was lowly express in our smFISH dataset (2.8 copies/cell on average) which is compatible with its higher variability in concentration.

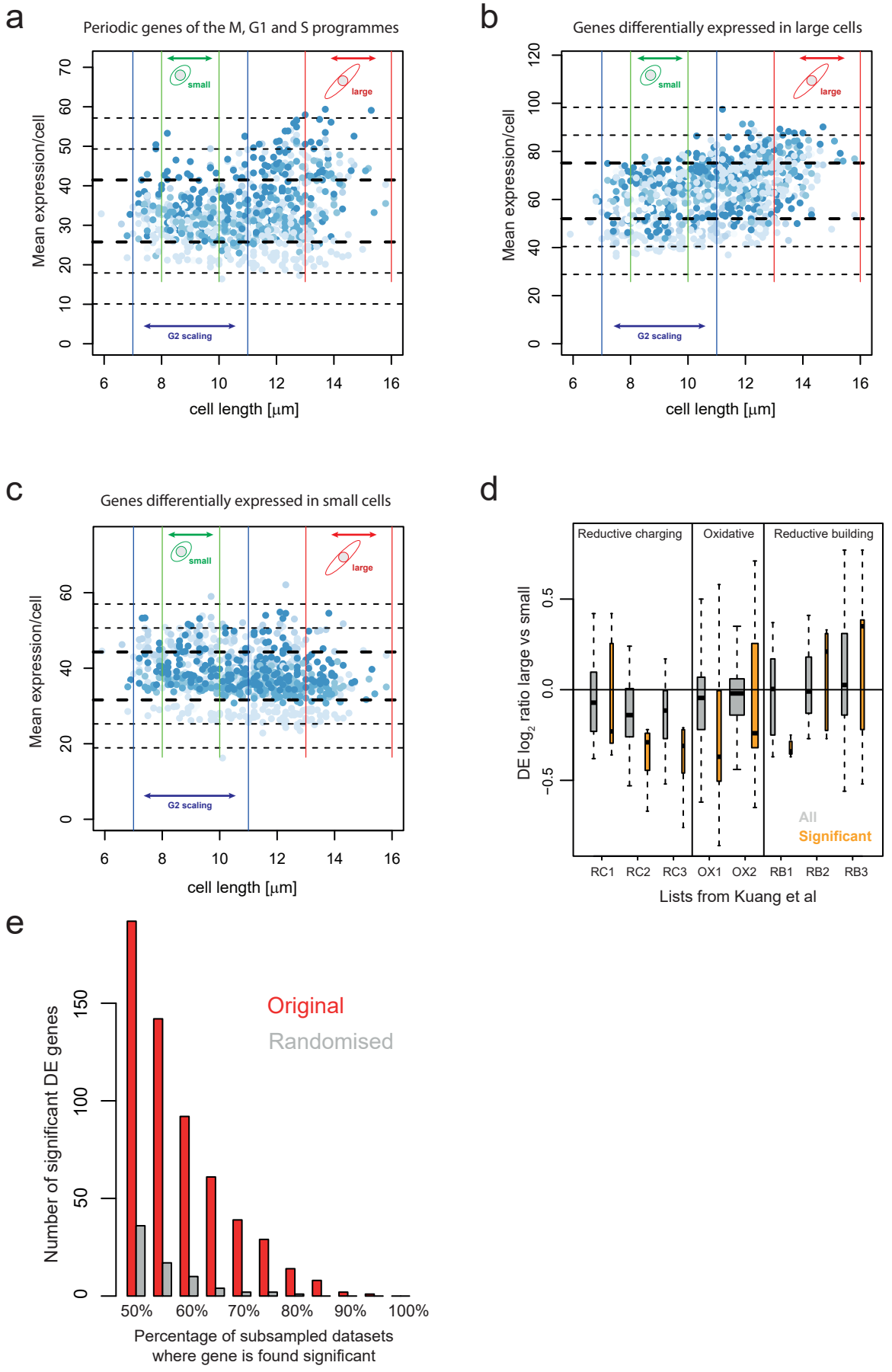
Figure S4, Saint et al.



Supplementary Figure 4: Supplement to *Cell-to-cell variability of the fission yeast transcriptome II*

(a) Level of periodic regulation of the top-500 most periodic genes from Rustici et al[3] present in the high-confidence filtered set. P-values for the level of regulation (amplitude) and periodicity as defined in Cyclebase (www.cyclebase.org)[4] are shown for genes classified as HVG (dark red, see **Fig 2, Supplementary Fig. 3**), regulated in small or large cells (blue, see **Fig 3 and Supplementary Fig. 5**) or not classified (grey). Statistical tests are described in ref[4]. Boxplots represent median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(b)** Detectability of cell-cycle periodic genes in scRNA-seq data. Fraction of the top-500 most periodic genes from Rustici et al[3] present in the high-confidence filtered set that could be detected by different computational approaches. Genes were either: called HVG (dark red, see **Fig 2, Supplementary Fig. 3**), called false positives (orange, see **Fig 2, Supplementary Fig. 3**), not HVG with increased expression in small cells (light blue, see **Fig. 3 and Supplementary Fig. 5**), not HVG with increased expression in large cells (dark blue, see **Fig. 3 and Supplementary Fig. 5**), not called by any approach (grey). Periodic genes are split in five categories according to Rustici et al[3], depending on their peak expression phase (G2, M, G1, S). ND denotes genes not assigned to a specific phase. **(c)** Mean expression per cell of the top-500 most periodic genes from Rustici et al[3] present in the high-confidence filtered set that were never called variable in this study (in grey on panel (a)) as a function of cell length. Expression levels in rapidly proliferating cells are shown (n = 864 cells). Green and red bars mark the size ranges of the small and large cells used for differential expression analysis respectively (see **Fig 3 and Supplementary Fig. 5**). **(d)** Characterisation of quantitative cellular or genetic features of HVGs (n=299 genes). Levels of different features were compared between HVGs and all other genes. P-values for one-sided Wilcoxon test are shown (alternative “greater” or “less” in red and green respectively). Gene list included the top-500 periodic genes (dark colours) or not (light colours). See **Methods** for a description of the features and their origin.

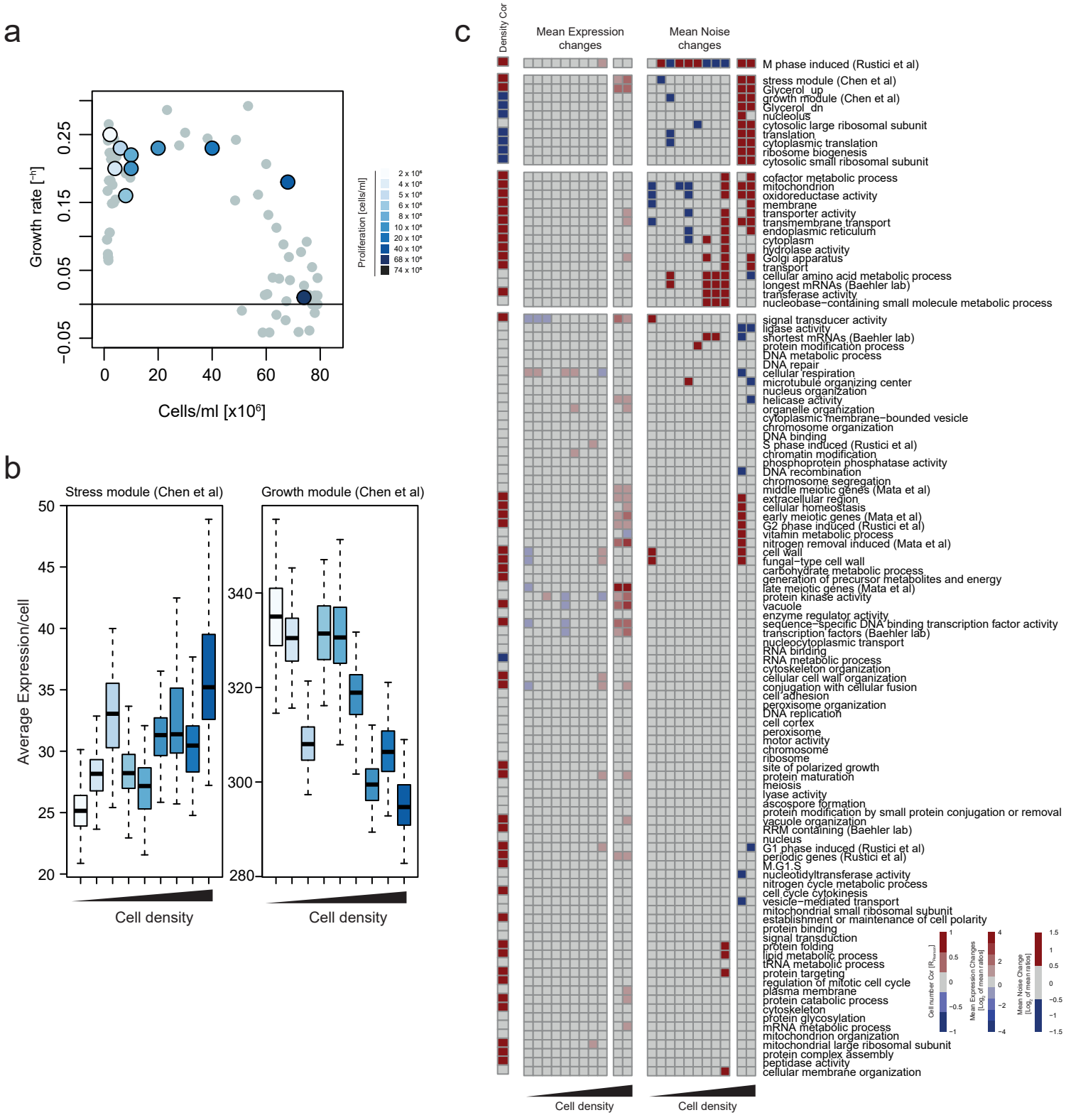
Figure S5, Saint et al.



Supplementary Figure 5: Supplement to *Cell-size dependence of fission yeast transcriptome*

(a) Mean expression per cell of genes associated with the M, G1 and S phases of the cell cycle[3] as a function of cell length. Fission yeast cells during rapid proliferation are shown (n = 864 cells). The size ranges of small (green bars) and large (red bars) cells used for differential analysis are shown. Blue bars show the size range used for defining non-scaling genes (NSG). Conditions are colour coded as on **Fig. 4** legend. **(b)** As in (a) for genes significantly up-regulated in large cells late in the cell cycle compared to small cells in G2. **(c)** As in (a) for genes significantly up-regulated in small cells in G2 compared to large cells late in the cell cycle. **(d)** Signatures of the Yeast Metabolic Cycle (YMC) in single fission yeast cells in asynchronous cultures (n=573 cells). Differential gene expression $\log_2$ ratios between large and small cells were obtained with MAST[5] and are shown for sets of genes participating in the YMC. Eight gene lists containing genes from the three phases of the YMC (top) were obtained from Kuang et al[6] (**Supplementary Table 6**). For each list, ratios of all genes are shown in grey and of genes significantly regulated in orange (Number of significant iterations > 90 and absolute $\log_2$ ratio > 0.2, width of the boxes are proportional to number of genes, **Methods**, **Supplementary Table 6)**. The slope of the regression line between DE ratios and steps of the YMC is positive and significantly different from randomised data (z-scores, $p < 10^{-4}$, **Methods**). Boxplots represent, median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(e)** Validation of DE analysis. MAST DE analysis was performed between 20 datasets of 50 “large” or “small” cells form **Fig. 3c** (red), or between 20 datasets of 50 cells of randomised sizes. Each dataset was normalised individually using bayNorm and local priors. The number of genes called significant in at least 50% and up to at least 100% the 20 datasets is shown. Note the very low amount of DE genes detected in the randomised datasets in grey.

Figure S6, Saint et al.



Supplementary Figure 6: Supplement to *Gene-expression heterogeneity of fission yeast populations in response to environmental changes*

(a) Related to **Figure 4d-e**. Population growth rate plotted as a function of cell density for scRNA-seq datasets. Conditions are colour coded as per legend on the right. Grey dots represent measurements from three additional independent growth experiments. Note the constant growth rate up to a cell density of >40x10^6 cells/ml. **(b)** Average expression per cell for genes from the stress response (left) and growth (right) programmes[7]. Boxes are colour-coded according to cell density as in (a). (n=864 cells). Boxplots represent, median, interquartile range and most extreme data points that are not more than 1.5 times the interquartile range. **(c)** Functional analysis of gene expression changes during growth and entry into stationary phase. Left (Density Cor): Correlation of mean expression levels of gene categories with cell densities between 2-40x10^6 cells/ml. Note that some categories increase in concentration co-ordinately with cell density, including the stress response programme, while other decrease, e.g. components of the ribosome, reminiscent of the P and R proteome fractions described in microorganisms[8]. Middle: changes in average expression of functional categories with cell density. For each category, the mean expression of genes from the category in each of the 11 cell-density datasets from **Fig. 4d-e** is compared to the other 10 datasets (10 ratios per density dataset). Log_2 of the mean of the 10 ratios is shown for each density dataset. Right: as in middle but using noise measurements for each category and dataset. This analysis demonstrates that expression heterogeneity occurs during entry into stationary phase for specific categories only.

Suppementary references

1. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42,** W187–W191 (2014).
2. Marguerat, S. *et al.* Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **151,** 671–683 (2012).
3. Rustici, G. *et al.* Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.* **36,** 809–17 (2004).
4. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43,** D1140–D1144 (2015).
5. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16,** 278 (2015).
6. Kuang, Z. *et al.* High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. *Nat. Struct. Mol. Biol.* **21,** 854–863 (2014).
7. Chen, D. *et al.* Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell* **14,** 214–229 (2003).
8. Scott, M., Klumpp, S., Mateescu, E. M. & Hwa, T. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Mol. Syst. Biol.* **10,** 747 (2014).