Refining the resolution of the yeast genotype-phenotype map using single-cell RNA-sequencing

Authors: Arnaud N'Guessan[1], Wen Yuan Tong[2,¥], Hamed Heydari[3,4], Alex N Nguyen Ba[1,2]

1. Department of Cell and Systems Biology, University of Toronto, Ramsay Wright Laboratories, 25 Harbord St, M5S3G5, Toronto, Ontario, Canada.

2. Department of Biology, University of Toronto at Mississauga, 3359 Mississauga Rd, L5L 1C5, Mississauga, Ontario, Canada

3. Department of Molecular Genetics, University of Toronto, Medical Science Building, Room 4386, 1 King's College Cir, M5S1A8, Toronto, Ontario, Canada.

4. Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Room 230, M5S3E1, Toronto, Ontario, Canada.

¥. Present address: Center of Molecular and Cellular Oncology, Yale University, 300 George Street Suite 6400, New Haven, CT

ABSTRACT

Genotype-phenotype mapping (GPM) or the association of trait variation to genetic variation has been a long-lasting problem in biology. The existing approaches to this problem allowed researchers to partially understand within- and between-species variation as well as the emergence or evolution of phenotypes. However, traditional GPM methods typically ignore the transcriptome or have low statistical power due to challenges related to dataset scale. Thus, it is not clear to what extent selection modulates transcriptomes and whether cis- or trans-regulatory elements are more important. To overcome these challenges, we leveraged the cost efficiency and scalability of single-cell RNA sequencing (scRNA-seq) by collecting data from 18,233 yeast cells from 4,489 F2 segregants derived from an F1 cross between the laboratory strain BY4741 and the vineyard strain RM11-1a. More precisely, we performed eQTL mapping with the scRNA-seq data to identify single-cell eQTL (sc-eQTL) and transcriptome variation patterns associated with fitness variation inferred from the segregant bulk fitness assay. Due to the larger scale of our dataset and its multidimensionality, we could recapitulate results from decades of work in GPM from yeast bulk assays while revealing new associations between phenotypic and transcriptomic variations at a broad scale. We evaluated the strength of the association between phenotype variation and expression variation, revealed new hotspots of gene expression regulation associated to trait variation, revealed new gene function with high expression heritability and highlighted the larger aggregate effect of trans-regulation compared to cis-regulation Altogether these results suggest that integrating large-scale scRNA-seq data into GPM improves our understanding of trait variation in the context of transcriptomic regulation.

INTRODUCTION

The process by which DNA encodes proteins via transcription and translation has been studied for decades to make sense of organisms' phenotypes. However, being able to explain organisms' phenotypes from their genetic material, i.e. genotype-phenotype mapping (GPM), has been a long-lasting problem with important applications (1,2). Indeed, making sense of genetic variation at the phenotypic level enables the understanding of trait variation between and within species as well as the emergence and evolution of phenotypes (3). For instance, reverse genetics approaches, e.g. gene knockout or transgenic technologies, and forward genetics approaches like GWAS and QTL mapping helped in determining the function of multiple genes and the effects of mutations on growth in different environments (4). However, reverse genetics approaches typically fail to account for natural variation and forward genetics approaches like QTL mapping typically focus on genetic and phenotypic variation so they cannot highlight selection on the transcriptome.

An essential characteristic of this problem is the multi-layered organization of the GPM. Indeed, GPM is not strictly restricted to the direct association between genotypes and phenotypes. This association is better resolved and complemented by understanding the intermediary transcriptome layer, e.g. cell mechanisms at the transcriptomic level are involved in diseases and pathogenicity (2,5–8). However, it is not clear to what extent transcriptomic changes relate to phenotypic changes or selection. Pioneering work from Mary-Claire King and Allan Charles Wilson set the tone for investigating this question by proposing that variations in morphological and behavioral traits arise more often through gene expression regulation than evolution at the protein-coding level (9). François Jacob then postulated an essay that stemmed from this theory in which he highlights how evolution acts as a tinkerer that works from already available material, i.e. through regulation of gene expression, to create new adaptations (10). This constituted the core of the evolutionary developmental biology which matured into the still-debated claim that new adaptations mainly emerge through cis-regulation of gene expression, i.e. through noncoding DNA regulating a neighbor gene contrarily to trans-regulators acting on distant genes (11–14). This debate has been reinforced by the technical difficulties and complexity of assessing the evolution and outcome of mutations in non-coding regions (11,12). For example, although human genome-wide association studies (GWAS) loci are mostly identified in noncoding regions and have well-characterized cis-regulatory effects, their trans-regulatory effects are usually challenging to detect (15–17). Indeed, small effect sizes, multiple testing corrections and high false-positive rates for trans-regulation impose statistical and computational challenges for its detection (18,19). Advances in sequencing technologies helped to partially solve these issues, particularly in the context of transcriptome analyses of the model organism *Saccharomyces cerevisiae*. For instance, Brem et al (2002) used microarray technology to relate the gene expression profiles of 40 yeast segregants from a lab (BY) and natural vineyard strain (RM) to their genetic markers (20). They found that cis-acting modulation is the main mechanism for regulating gene expression. Nearly two decades later, by greatly increasing statistical power, Albert and collaborators (2018) found that most of the expression variation arises through trans-regulation using non-multiplexed RNA-seq to analyze 5,720 genes in 1,012 yeast segregants generated by a crossing between RM and BY (21). The analysis method they used, i.e. expression quantitative trait loci (eQTL) mapping, consists in correlating allele frequencies to gene expression levels to find the loci modulating expression.

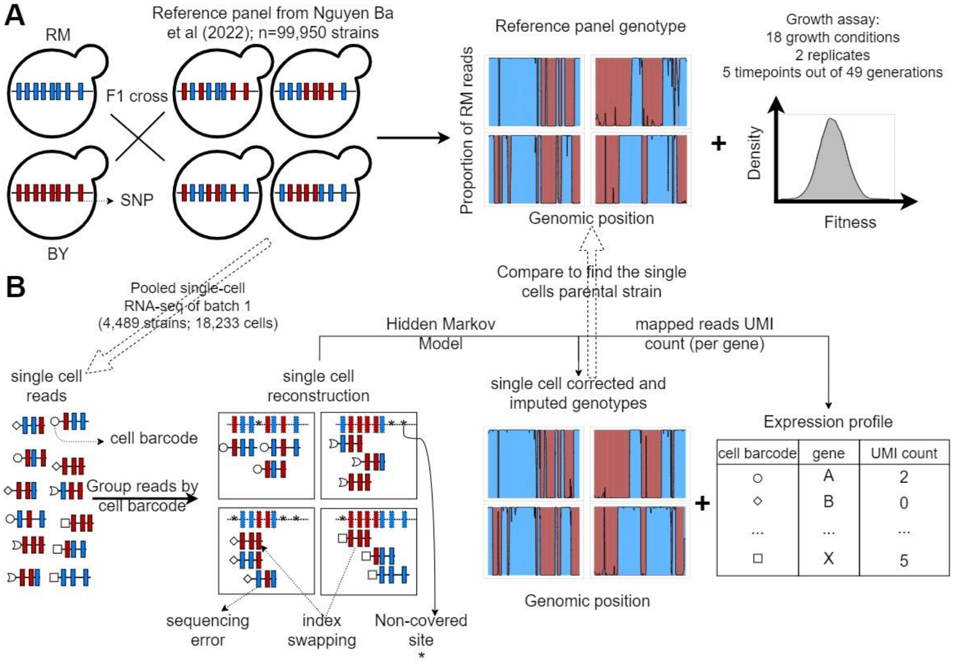
Although eQTL mapping is a traditional GPM analysis that accounts for the transcriptomic layer, it is typically realized through non-multiplexed RNA-seq which tends to have low statistical power due to challenges with experimental scale and confounding factors (22,23). Thus, eQTL mapping traditionally cannot identify significant low-effect regulatory mutations that are important for understanding the genetic bases of complex traits and diseases (24,25). Furthermore, most eQTL studies only assess the average transcriptomic profile of bulk populations without being able to capture the profile of rare cell lineages within a population. This is a critical limitation in heterogenous populations such as cancer or microbial populations where rare lineages can drive relapse or drug resistance (26).

Here, we sought to circumvent the challenges of non-multiplexed bulk RNA-seq imposed by the scale and population heterogeneity by performing eQTL mapping through single-cell RNA sequencing (scRNA-seq) of a pool of ~4,500 well-characterized F2 segregants derived from a yeast cross (21,27,28). In the same way that combinatorial indexing/barcoding and multiplexing enable the collection of large-scale fitness and genotype data (29), we hypothesized that scRNA-seq could help us collect both genotype and expression data on a large pool of segregants. We employ several strategies to overcome previous obstacles of eQTL mapping studies: i) we pool cells from thousands of F2 segregants during the growth step and perform a single scRNA-seq run on the culture to account for environmental effects. The expression profiles of these lineages cannot be captured cost-effectively with bulk RNA-seq as pooling segregants in a single bulk assay would not capture individual strains' expression profiles. Moreover, ii) from the exome sequencing data of single cells, we take advantage of the reference panel to validate that we accurately infer the genotype of each cell from extremely low number of reads mapping to polymorphic sites per cell (effectively ~0.2x coverage). Finally, iii) we leverage the presence of thousands of high-coverage cells to train an unsupervised learning model to correct (denoise) and infer an imputed expression profile of poorly covered cells.

Using this approach, we integrated the resulting transcriptomic data from growth in rich media with a pre-existing yeast GPM to reveal new GPM features at a broad scale. For instance, we estimated the heritability of the transcriptome and the extent to which the transcriptome is associated with fitness. This allowed us to reveal that a negligible portion of expression variation is related to the studied phenotype variation independently of genotypic variation. This implies that almost all the phenotypic variation related to expression is also modulated by mutations. This differentiates our work from previous bulk RNA-seq studies where expression is systematically controlled for differences in lineage growth or phenotypes due to an assumption that a considerable amount of phenotypic variation is related to expression independently of genomic mutations (21). We also show that the increased scale of our scRNA-seq dataset enables single-cell eQTL mapping (sc-eQTL) and identifies more hotspots of expression regulation containing previously identified QTL. We also exploit the identified sc-eQTL to analyze the patterns of cis- and trans-regulation in the GPM.

Our single-cell RNA-seq approach is consistent with yeast GPM results from non-multiplexed assays

We initially aimed to show that performing scRNA-seq at a large scale can generate data that are consistent with non-multiplexed DNA and RNA sequencing. For this purpose, we analyzed a dataset of thousands of yeast lineages generated by Nguyen Ba and collaborators (2022) (29). To understand the yeast GPM, they collected fitness and genotype data from ~100,000 F2 segregants derived from an F1 cross between a laboratory strain of yeast (BY) and a natural vineyard strain (RM) (**Figure 1A**).


A
RM
BY
F1 cross
SNP
Reference panel from Nguyen Ba et al (2022); n=99,950 strains
Reference panel genotype
Proportion of RM reads
Genomic position
+
Growth assay:
18 growth conditions
2 replicates
5 timepoints out of 49 generations
Density
Fitness
B
Pooled single-cell RNA-seq of batch 1 (4,489 strains; 18,233 cells)
Compare to find the single cells parental strain
Hidden Markov Model
mapped reads UMI count (per gene)
single cell reads
cell barcode
Group reads by cell barcode
single cell reconstruction
single cell corrected and imputed genotypes
Genomic position
+
Expression profile
sequencing error
index swapping
Non-covered site
*

cell barcode	gene	UMI count
○	A	2
◇	B	0
...	...	...
□	X	5

Figure 1 F2 yeast segregants datasets. A) Reference panel and experiments from the barcoded bulk sequencing previously described in (29). The 99,995 F2 yeast segregants in the reference panel were derived from an F1 cross between a laboratory strain of yeast (BY) and a natural vineyard strain (RM). Thus, they only have 2 possible alleles at each of the 41,594 polymorphic sites. The lineage barcodes enabled fitness estimation from competition assays in 18 environments recapitulating the adaptation to temperature gradients, the ability to process different sources of carbon and the resistance to antifungal compounds. B) Pooled scRNA-seq dataset from a single batch. We performed scRNA-seq of the first batch of F2 segregants (n=4,489) to obtain genotypes that are similar to the reference panel and cell's expression profiles. The F2 segregant barcodes are typically not expressed so single cells have their own cell barcode which is included in all amplified reads of the same droplet. Non-covered sites, sequencing errors and the presence of reads in the wrong library (index swapping) are corrected for using the HMM described in Figure S1.

Using this approach named barcoded bulk QTL mapping or BB-QTL mapping, they revealed the complex polygenic and pleiotropic nature of phenotypes as well as an unprecedented number of pairwise epistatic interactions. To integrate transcriptomic data into this GPM, we performed scRNA-seq using the 10X Genomics Chromium microfluidics platform (30). Other scRNA-seq methods like Smart-seq2 offer a better breadth of coverage per cell (31). However, due to the higher throughput or number of cells generated by 10x chromium and the different coverage profiles across cells from the same F2 segregant, it is more advantageous to use 10x Chromium to capture transcriptomic variation and reconstruct the reference genomes from single cells (31,32). The throughput of 10x Genomics is typically in the order of 10^4 cells which limits us to a few thousand lineages per library if we want to obtain a reasonable number of cells per lineage. This is why we focus on a single batch of 4,489 F2 segregants. This method allowed us to obtain both genotype and expression profiles from 18,233 cells of the first batch of F2 segregants (**Figure 1B**). The F2 segregant barcodes are typically not expressed so single cells have their own cell barcode which is included in all amplified reads of the same droplet. Although this short-read scRNA-seq is a high-throughput approach (**Table 1**), it comes with challenges like gene dropouts and low sequencing depth in some cells caused by technical artifacts of PCR amplification (25,26).

Metric	Value
Library size (number of reads)	479,781,081
Median number of covered genes per cell	1,311
Median number of UMIs per cell	5,158
Doublet rate	0.10
Average number of cell barcodes per F2 segregant	4.83
Polymorphic sites mean breadth of coverage	0.03

Table 1 scRNA-seq dataset features. The doublets were defined as scRNA-seq barcodes with high coverage (>=third quartile) without significant genotype association to any F2 segregants and for which the 2 most similar F2 segregants do not share genotypic similarity ($R^2 < 0.1$). The breadth of coverage is defined as the proportion of BY/RM polymorphic sites covered in a single cell.

To overcome these challenges, the unique molecular identifiers (UMIs) of the 10X Genomics platform provide a control for PCR amplification biases by quantifying gene expression from unique transcribed molecule counts instead of sequencing read counts (31). However, UMIs still do not correct for dropouts or missing data in the expression profile and this could obscure potential associations between genotype and transcription (see below). We addressed this issue by fitting an imputation model called DISCERN to the expression data (33). DISCERN is a neural network that learns how to reconstruct the expression profile of high-coverage cells after embedding it in a lower dimension. This eliminates the gene dropouts and denoises the expression profile. The fact that the reconstructed expression is highly but not perfectly correlated to the original expression (mean R^2 =50.0%, median R^2 =52.9%; R^2 s.d. = 10.1%) and the high variance explained by the first principal component of the imputed expression (99.6%) is consistent with an effective denoising. We take advantage of this denoised transcriptome for variance partitioning (as described later).

In addition, Hidden Markov Models (HMMs) can infer accurate genotype data even at sequencing depths as low as 0.1x (29). Nguyen Ba and collaborators (2022) designed an HMM to infer the segregants genotypes from the observed reads at low depth of DNA sequencing by accounting for sequencing error rate, recombination rate, and index swapping rate (29). As there are only two ancestral lineages, there are only two possible alleles for the strains at each of the 41,594 polymorphic sites. Thus, the genotype of the segregants can be represented by the frequency of only one of the parental alleles, which is RM in the dataset. Applying this model to low-coverage segregants yielded genotypes that are significantly similar to high-coverage replicates (29). We sought to use a similar model to infer genotypes from scRNA-seq data but we anticipated that some of these parameters may differ due to increased error rate of the reverse-transcriptase, increased index swapping due to pooled reaction, etc (**Figure S1**). In Nguyen Ba et al, those rates were heuristically determined, but here we estimated these from the read mapping data and found that re-estimated parameters from data increased the proportion of recovered strains in the single cell data from 58.6% to 72.0%.

After adapting the HMM to the scRNA-seq data, we sought to validate that the resulting cell genotypes relate well to their corresponding strain in the reference panel obtained by non-multiplexed DNA sequencing strategies. Ideally, each single-cell barcode (from 10x Genomics Chromium) should be associated with a single cell and a cell should have a clear match with a unique strain in the reference panel. However, several factors can obscure these associations, e.g. a single-cell droplet containing cells from 2 different strains, a low-coverage cell, uncertainty in the allele of the reference genotype, etc. Thus, we designed an approach to assign cells to the correct reference panel strain (see **Methods**). This approach relies on two metrics of similarity between the cells and the strains' genotypes, i.e. the expected distance between them, which should be minimized for the best match, and the relatedness (R^2). The statistical significance of the relatedness between single cells and reference lineages was determined by a permutation test (**Figure S2**). From the read mapping alone, we obtained a mean R^2 of 0.59 (σ = 0.19 and median = 0.64), which was significantly improved after applying our HMM to correct for misidentified alleles and imputing data in low-coverage sites using recombination probability. Indeed, the single-cell HMM genotypes yield a mean R^2 of 0.73 (σ = 0.18 and median = 0.81; **Figure 2A**). We found that the distribution of relatedness after HMM was still left-skewed, with many cells statistically significantly assigned to a reference genotype despite having what appeared to be low relatedness. Upon investigation, it was found that these could be explained by genotyping uncertainty either in the single-cell and/or in the reference panel genotype (**s**) (**Table S1**).

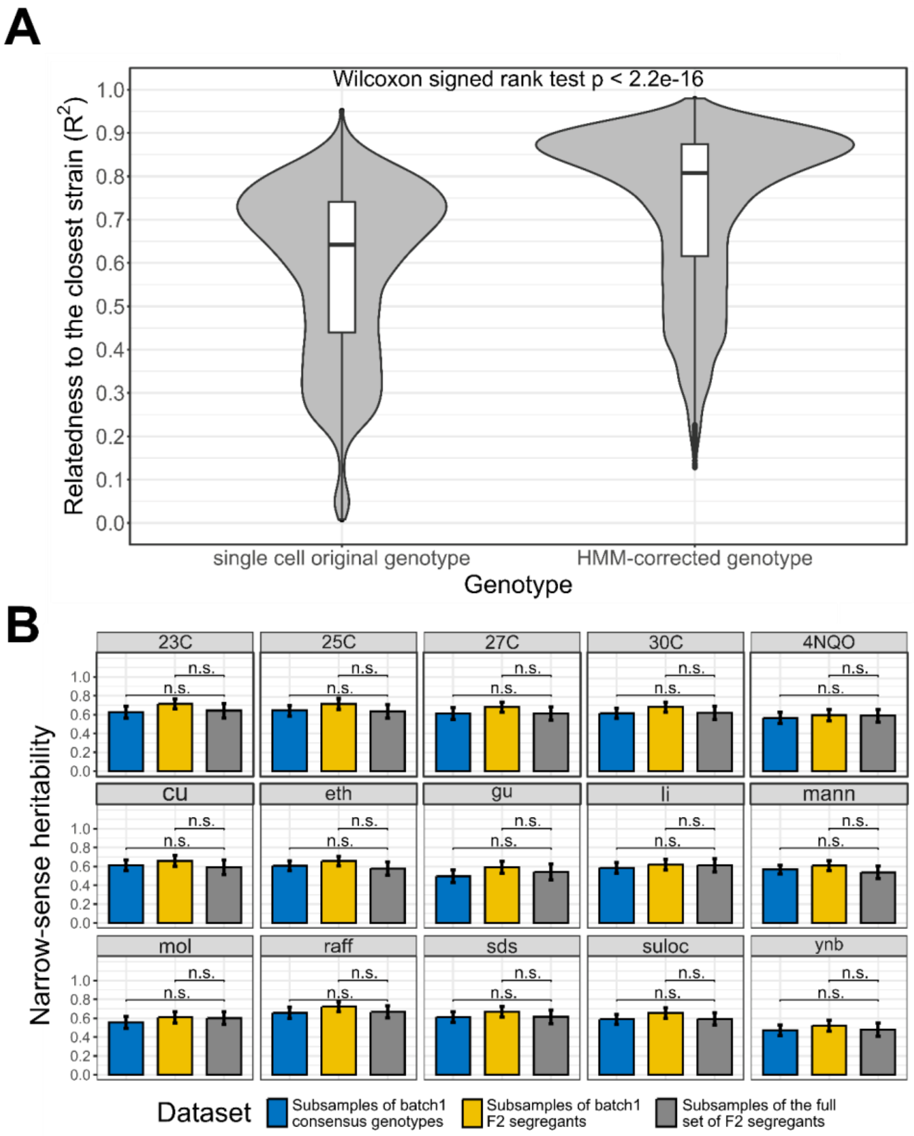


Figure 2 Single-cell RNA-seq data recapitulate bulk DNA and RNA assays results. A) Effect of the HMM on the relatedness between single-cell genotypes and their closest reference lineage.

The single-cell original genotype represents the genotype of the cells before the correction with the HMM. The relatedness to the closest lineage in batch 1 has been measured with the adjusted R^2. To control for genotype uncertainty, only the 13,069 barcodes with a significant lineage assignment (lineage-barcode genotype correlation $FDR<0.05$) and a reference lineage with lower uncertainty than the single cell HMM are selected, which represents 72.2% of the barcodes. We then rounded the genotypes to remove the uncertainty during the comparison. Wilcoxon signed test p-value is indicated above the violin plots. B) Narrow-sense heritability measured with scRNA-seq consensus genotypes and non-multiplexed DNA sequencing. Three datasets are compared through a t-test on the bootstrap narrow-sense heritabilities. The consensus genotypes of the batch 1 F2 segregants obtained from the associated scRNA-seq data are represented in blue, the batch 1 F2 segregant genotypes in the reference dataset (bulk sequencing) are represented in yellow and the whole reference dataset of F2 segregants is represented in grey. For each dataset, the narrow-sense heritability has been measured from 500 random subsamples of 1000 F2 segregants. The bars represent the mean narrow-sense heritability, and the error bar length represents the 95% CI. The results are illustrated for the 15 environmental conditions where a minority of batch 1 F2 segregants have missing fitness data. The 23C-30C represents the temperature for the competition assay in YPD media while the other phenotypes represent growth on YNB, molasses (mol), mannose (Mann) or raffinose (raff) and chemical resistance to copper sulfate (Cu), ethanol (eth), guanidinium chloride (gu), lithium acetate (Li), Sodium dodecyl sulfate (SDS) and suloctidil (suloc) (29).

To further establish that the genotyping obtained from scRNA-seq data was comparable to previous non-multiplexed genotyping of the reference genotype panel, we estimated the contribution of genetic variation to the phenotypic variation, i.e. fitness heritability. Nguyen Ba and collaborators (2022) estimated the narrow- and broad-sense heritabilities of complex phenotypes associated with temperature gradient, carbon source and chemical resistance for which RM and BY segregants exhibit a significant level of diversity (29). We used our lineage assignment to that panel to obtain fitness but used our single-cell genotyping to perform this association (**Methods**). Encouragingly, all GCTA-REML estimates of narrow-sense heritability in the scRNA-seq batch 1 dataset (blue in **Figure 2B**) are similar to the estimates from Nguyen Ba and collaborators (2022) whether or not we focus on batch 1 (yellow and grey rectangles in **Figure 2B**).

Although the variance partitioning is consistent with previous studies, it only provides a broad view of the genotype-phenotype map as it does not allow to identify the loci that significantly explain phenotype variation. If the genotypes obtained by scRNA-seq were of high quality, then we would expect that a QTL mapping model from scRNA-seq would yield a similar model to non-multiplexed DNA sequencing data. To achieve this, we used a cross-validated stepwise forward linear regression on the strain fitness and consensus genotypes data from single cells that shared the same lineage assignment (**Methods**). Performing the QTL mapping on the batch 1 scRNA-seq dataset enabled the identification of 29 QTL compared to 31 QTL identified with the bulk barcoded approach (**Tables S2 and S3**) (29). These QTL were largely similar as shown by the non-significant difference between the effect sizes (Wilcoxon signed rank test *p* = 0.29) and by a model similarity metric (29) that considers the recombination distance between matched QTL, the similarity of the effect sizes and the allele frequencies (**Methods**). Using this approach, we estimated that the similarity score between the batch 1 single cells QTL and the batch 1 BB-QTL is 86.2% while each model respectively had a similarity score of 78.7% and 78.2% with the full BB-QTL mapping performed on the 99,950 F2 segregants (29) (**Figure S4**). The QTL identified from the scRNA-seq dataset also recapitulated several important biological features of the reference panel such as an enrichment of non-synonymous and disordered region QTL (29) (**Figure S5**).

Integrating scRNA-seq data to an existing GPM highlights selection on the transcriptome

Having shown that our scRNA-seq allows us to relate the cells to the corresponding F2 segregant fitness data via the genotype relatedness, we next sought to highlight new associations within the BY/RM GPM. Selection is often highlighted at the genotype level through convergent evolution, an increase in allele frequency within a population, or population genetics metric (34–36). However, the central dogma of molecular biology and the evolution tinkering model both entail that phenotype variation should be linked to transcriptomic variation. As our dataset included all these variables, we sought to evaluate the expression-genotype association strength and provide a variance partitioning framework to evaluate the association between the transcriptome and trait variation (**Methods**). The GCTA-REML variance partitioning model used to estimate trait heritability in the previous section can also be modified to include gene expression as the response variable and cell genotypes as the only random effect (**Methods**). This enables the quantification of expression heritability, i.e. the variance of expression explained by genotype. Using this approach on our large-scale dataset, we estimated that genotype explains 75.7% of expression variance, which is slightly higher than previous estimates from non-multiplexed eQTL mapping studies. Indeed, Albert and collaborators (2018) estimated that genotype explains 70% of expression variance using a lower-power dataset of 5,720 genes in 1,012 yeast segregants generated by the same parental strains (RM and BY).

In addition to expression heritability, the multidimensionality of our scRNA-seq dataset enables the evaluation of the association between traits and expression, which we demonstrate with the 30C phenotype (**Figure 3**).

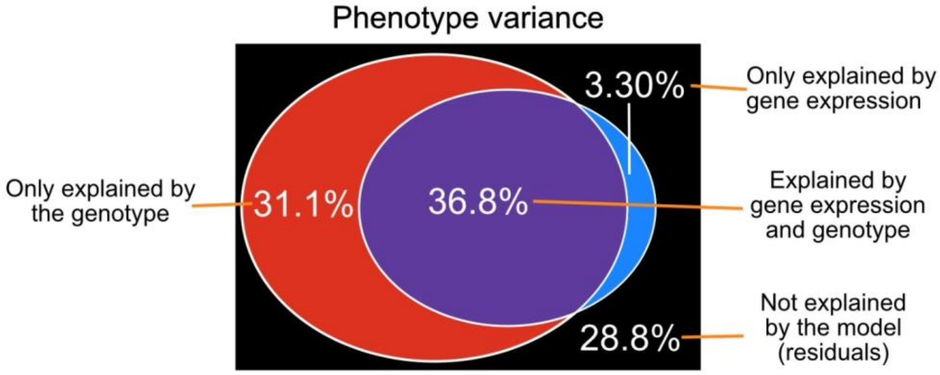


Figure 3 Variance partitioning of the 30C phenotype from scRNA-seq data. The percentages represent the proportion of fitness variance (whole rectangle area) explained by the components and their overlaps. The ellipse area represents the phenotype variance explained by genotype or expression variation. The black area of the rectangle represents the residual of the model while the other colored areas represent the shared and exclusive components explaining fitness variation. The size of the shared component area (purple) is obtained by subtracting the variance explained by the model in Equation 4 to the one in Equation 2 or in Equation 3 as it is a shared component of trait variation explained by genotype and expression variation. The size of the red genotype-exclusive component is obtained by subtracting the size of the shared component from the variance explained by the model in Equation 2. The size of the blue expression-exclusive component is obtained by subtracting the size of the shared component from the variance explained by the model in Equation 3. Finally, the size of the residuals is obtained by subtracting 1 by the sum of the size of all the defined components (red + blue + purple).

Commented [AN1]: Refer to the equation

The components of this variance partitioning all relate to at least one biological phenomenon. Indeed, the portion of trait variation explained exclusively by the genotype variation (red in **Figure 3**) represents the effect of mutations on fitness via several biological phenomena such as protein stability, enzymatic function etc, independent of expression level. For the 30C phenotype, this component explains 31.1% of the fitness variation in the BY/RM background which is slightly lower than the 36.8% explained by the shared component between phenotype, genotype and expression variations (purple in **Figure 3**). The latter represents the association between selection (fitness) and the transcriptome either through loci influencing fitness via expression directly or through loci affecting expression via an effect on cell fitness (indirectly) (37,38). Its considerable association with fitness variation thus supports the evolution tinkering model. As for the phenotype variation explained exclusively by gene expression (blue in **Figure 3**), it could represent epigenetics and stochastic gene expression, which weakly explain variations in the 30C phenotype.

Although this model accurately estimates the narrow-sense heritability of 30C, the residuals still represent 28.8% of fitness variation. This could be explained by unmeasured factors like high-order epistasis. However, the broad-sense heritability of this phenotype is similar to the narrow-sense heritability, suggesting that the residuals are mostly not explained by genotype and expression (29). Nguyen Ba et al. (2022) also estimated that epistasis only explained around 5% of fitness (29). Other factors like mitochondrial mutations, protein-protein interactions, or other protein properties could play a role in explaining the residual of the model and are good candidates to extend this analysis. These results suggest that a single run of scRNA-seq on a single batch of F2 yeast segregants converges with bulk DNA sequencing results while revealing previously hidden components of the GPM.

Revealing hidden components of the yeast GPM with scRNA-seq

Our integrative scRNA-seq approach is not limited to enabling the quantification of the association between transcriptomic changes and trait variation. Indeed, the same approach we used to identify QTL can be used to detect loci regulating gene expression which can reveal the cell mechanisms underlying trait variation through transcriptomic changes. We thus modified the QTL mapping framework such that the response variable is the level of expression of a single gene in the single cells (**Methods**). Because our cells are not synchronized, differences in the cell cycle could create transcriptomic variation that could obscure the association with genotypic variation. To control for this, we averaged out the effect of cell cycle by reconstructing the consensus expression and genotype profiles of each segregant significantly associated to the cells. This approach is a cost-efficient way to perform eQTL mapping from the expression profile and genotype of cells from thousands of lineages in a multiplexed way (sc-eQTL mapping).

Consistent with yeast non-multiplexed eQTL results, the genes with the highest expression heritability are enriched in functions related to carbohydrate catabolic process (GO:0016052) and cellular biosynthetic process (translation GO:0006412, organelle assembly GO:0070925, ribosome biogenesis GO:0042254 and gene expression GO:0010467) (Fisher’s exact test FDR<0.05; **Methods**). In both datasets, these genes are also highly expressed, which reflects the positive correlation between expression heritability and expression levels ($R^2 = 0.66$ and $p < 2.2e\text{-}16$). Although some of these highly expressed genes have important functions modulated by mutations, their higher heritability could also be explained by higher statistical power to measure heritability when expression counts are higher. Moreover, genes with the lowest expression heritability observed in the RM/BY background, which we defined as the bottom 10% expression heritability, are enriched in functions related to the cell cycle biological process (GO:0007049, Fisher’s exact test FDR<0.05) which is consistent with bulk RNA-seq eQTL (21,39). This does not suggest that genes with this function are not important for variation in gene expression at the whole-genome scale but rather that mutations observed in the RM/BY background are not correlated with variation in the expression of these specific genes. This could reflect conservation for genes involved in the cell cycle biological process (40,41).

Because of the increased scale of our collection, our approach is more powered to estimate the gene expression heritability. Indeed, we detected a median of 21 eQTL per gene (mean =34.3, s.d. = 29.7) which is almost 4 times higher than what has been detected by previous non-multiplexed RNA-seq (21). We were also able to detect new overrepresented biological processes, i.e. DNA metabolic process (GO:0006259) and the response to nutrient levels (GO:0031667), for which the variation of expression levels is weakly associated to the genetic variation observed across the F2 RM/BY segregants.

The functional enrichment analysis using scRNA-seq data revealed new associations between expression heritability and biological processes in the RM/BY genetic background. However, while it suggests that many eQTL are also QTL, it cannot accurately point to the specific loci involved in trait variation and cannot address whether mutations on regulatory hubs have stronger effects on traits. To investigate this, we mapped the QTL to hotspots of gene regulation (or regulatory hubs), which we defined as 25 kb genomic windows that were repeatedly identified in the eQTL mapping procedure (for different genes). This was done to acknowledge the uncertainty in the exact position of the eQTL due to linkage disequilibrium and power. We then ranked the 30C QTL identified by Nguyen Ba and collaborators (2022) based on their absolute effect size and correlated it to the rank of the eQTL hotspots based on the number of regulated genes. This resulted in a positive correlation (Spearman $\rho = 0.33$ and $p = 5.21\text{e-}5$), suggesting that larger effects on the regulatory network translate into larger trait variation. Indeed, we observed that some previously reported high-effect-size QTL genes are located in eQTL hotspots, e.g. MKT1, HAP1, and IRA2 (**Figure 4A**).

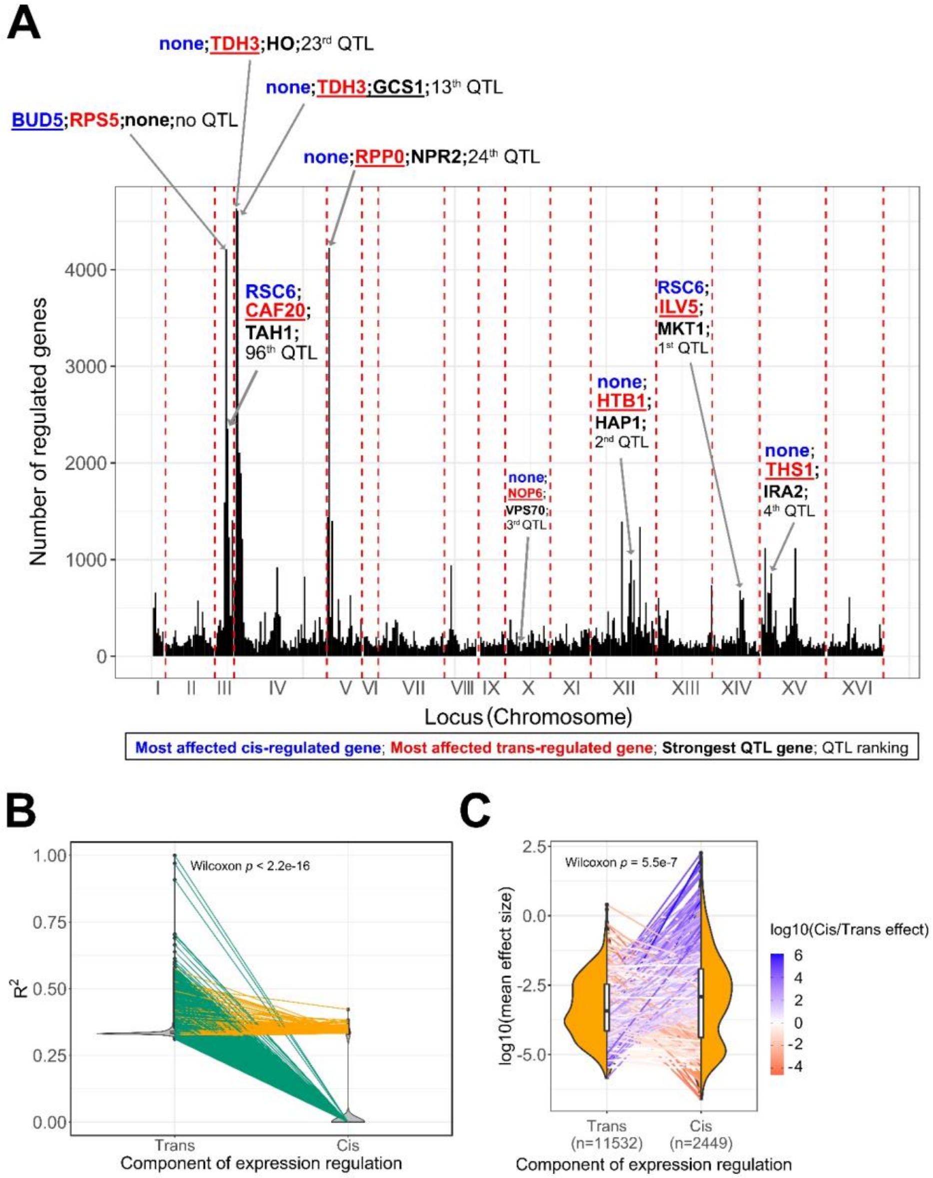


Figure 4: eQTL features underlying trait variation across the BY/RM segregants. A) Mapping of the 30C QTL in the eQTL hotspots. We represent the hotspots of expression regulation as genomic windows (25 kb) to acknowledge the uncertainty around the real position of the eQTL

due to linkage disequilibrium. We annotated the 5 top eQTL hotspots and the eQTL hotspots in which the top additive QTL identified by the BB-QTL mapping of the 30C phenotype are located. In these regions, we represented the most affected trans-regulated genes in red, the most affect cis-regulated gene in blue and the genes of the top QTL in black. The double quotation characters represent the absence of such genes in the associated region. We also represented the rank of the QTL in the set of 159 QTL of the 30C phenotype. B) Partitioning of the expression heritability or explained variance (R^2) among cis- and trans-eQTL. Each pair of points connected by a line represents a gene. Green lines represent the genes that are only have trans-eQTL and orange lines represent the genes that have both trans- and cis-eQTL. C) Comparison of the mean effect size between cis- and trans-eQTL. Each pair of points connected by a line represents a gene. The ratio of the average effect size between cis- and trans-eQTL is represented by the line color. The sample size of each eQTL category is represented in the x-axis. This is the number of trans-eQTL and cis-eQTL used for calculating the average effect sizes per gene not the number of points per distribution.

Performing this rank test on individual genes also yielded the result that eQTL effect is correlated with the fitness effect for 35.1% of the genes (permutation test $p < 0.05$, see **Methods**). Although this correlation does not apply to most genes, it reveals potential regulatory mechanisms explaining the importance of the strongest growth loci or QTL. For instance, MKT1, i.e. the strongest growth loci, is part of a regulation hotspot affecting genes that are important for yeast growth like ENP1 which is involved in RNA processing and HXT6 which is involved in glucose uptake (42–44). Among the strongest growth loci, VPS70 is part of a hotspot of regulation that strongly affects the expression of RSF2, a zinc-finger protein regulating glycerol-based growth and respiration (45). Furthermore, the highest peak for expression regulation contains important growth loci in chromosome IV around the mating type loci. This suggests the presence of cells with different mating types in the dataset which we confirmed from the read mapping to Mat-a and Mat-α genes. This is consistent with previous budding yeast eQTL mapping and is also expected because the mating types in yeast express sets of genes that are “turned off” in other mating types (20,21,46). This peak of expression regulation is also responsible for regulating TDH3 which is involved in glycolysis and glucogenesis and can have an important effect on fitness (47).

These hotspots suggest that expression differences in BY/RM would predominantly be due to mutations in trans-regulatory elements. To test this, we partitioned the variation in gene expression between cis- and trans-regulatory loci for each gene (see Methods). This analysis revealed that all the genes are affected by at least one polymorphic trans-regulatory locus and that these polymorphic trans-regulatory loci explain most of that gene’s expression (**Figure 4B**). It is well known that mutations in promoters and nearby enhancers can influence gene expression (48,49). Indeed, we identified many genes that contained an allele in a cis-regulatory element that strongly explain that gene’s expression variation (n=750 genes out of 6088, **Figure 4B**). As expected, mutations in cis-regulatory elements were of stronger effect size than trans-eQTL individually, but the cumulative aggregate effect of all trans-eQTL acting on that gene was comparable to the few cis-eQTL they had (**Figure 4C**). This can be explained by the fact that there are more opportunities for mutations to arise in trans-regulatory elements. Finally, we found that trans-eQTL have two times higher odds of affecting cell fitness than cis-eQTL (χ^2 $p = 0.01$).

Furthermore, by repeating our eQTL hotspot analysis with Albert et al. (2018) data, we observed a non-significant association between eQTL hotspot and QTL (χ^2 $p = 0.50$). In that bulk RNA-seq dataset, QTL have less odds of being in eQTL hotspots compared to other loci (od ratio = 0.72) while in this scRNA-seq dataset, QTL have twice more odds of being in eQTL hotspots. This could be explained by the smaller sample size of that bulk RNA-seq dataset and by the fact that they control for differences in growth across lineages. Taken together, these results suggest that the link between the genetic basis of transcription variation across RM/BY segregants and fitness could only be revealed by integrating large-scale transcriptomic data to an existing GPM, which scRNA-seq facilitates.

CONCLUSION

By leveraging the scalability of scRNA-seq, we obtained thousands of transcriptomes from a reference pool of strains in a single experiment. This enabled the analysis of the association between genotype, transcriptome, and phenotype at an unprecedented scale. Questions surrounding transcriptomic variation and phenotypic variation have been at the center of many previous quantitative genetics studies (19–21,27,46,50–53). For instance, how strong is the connection between the transcriptome and trait variation? Although most GWAS loci are located at noncoding regions and are well-characterized for cis-regulation effects, is trans-regulation more important for phenotypic and expression variation? What are the gene functions that are the most heritable and modulable at the expression level? These ideas and discoveries all support the fact that researchers can gain valuable insight into the evolution of traits by integrating the transcriptome in GPM analyses, which can translate into fundamental knowledge or other important applications where phenotypes evolve.

In this study, we took advantage of a previously characterized BY/RM cross where the genetic basis of growth in various environments was examined in detail (29). By integrating transcriptomic data in this genotype-phenotype map, we revealed how transcriptomic components are involved in trait variation at a broad scale. Similar to a previous study, which obtained transcriptomes by individual strain sequencing, we found that gene expression is highly heritable. Further, our study design also allowed us to conclude that gene expression contributes to a significant portion of the phenotypic variation in this strain collection.

This finding is corroborated by our findings that most eQTL detected in our study were previously shown to be QTL. This is perhaps not surprising given that QTL in this cross were previously inferred to be in regulatory genes, but this provides a more mechanistic view of the effect of an allele on phenotype. Indeed, we find a bias for trans-regulation for generating transcription innovation where the cumulative effect of trans-eQTL on gene expression are significant. That is not to say that cis-regulatory alleles are dispensable as cis-regulatory alleles often have a large effect on gene expression. This is similar to studies on human transcriptomes where trans-eQTL are cumulatively more impactful and enriched in sets of complex trait loci (19,51–53). This genome-wide view of the genetic basis of transcriptional variation has consequences for the evolution of phenotypes, as the target size afforded by trans-eQTL is far larger than cis-eQTL.

Thus, adaptation to small and fluctuating environmental changes may proceed preferentially through allelic changes or recombination of many small-effect trans-eQTL, but large expression changes are likely to require some cis-eQTL.

In this study, we leveraged the availability of genotype and phenotype data on our pool of strains. This was obtained by liquid handling robotics and pooled competitive growth assay with barcode sequencing. While this was performed on a very large scale, it was essentially obtained by brute-force and through approaches that are not necessarily applicable to other systems. While it is clear from our results that single-cell sequencing can achieve the same genotype quality as single-reaction genotyping, it is much harder to obtain phenotype data from scRNA-seq. Thus, our framework might not be readily translatable to other systems where similar studies on the GPM are desirable. However, two observations from this cross can be used to suggest an experimental approach. First, while epistasis is important, it contributes to a relatively small portion of the phenotypic variance. Second, transcriptomic variation contributes little to the missing heritability. Thus, it may be possible to use predicted fitness instead of observed fitness and recapitulate essentially similar results as this study. Indeed, the phenotype under study has a negligible component of its variance explained by transcription independent of the genotype, but this feature was unknown before the study. This alternative approach can only relate expression to the heritable component of traits, which would have missed this variance explained by transcription independent of the genotype and explains why we did not employ it. Predicted fitness could be obtained from bulk-segregant analysis where the additive effect of loci can be inferred from whole-genome sequencing (28,54). However, it is not clear if the proposed experimental approach is generalizable in any other context.

However, despite the study's limitation on generalizability, our scRNA-seq framework helps bridge the understanding of how genetic variation influences transcriptomic variation at a broad scale. Our framework relies on identifying the genome of single cells from the transcriptome, which is going to be possible from low-coverage sequencing when genetic variation within the pool is high (such as this cross, microbiome sequencing, or cancer cells with extensive copy number variation), and from low cell diversity with sufficient transcriptomic variation such that aggregation of single-cells with similar transcriptomes can afford pseudo-high coverage sequencing. Although we restricted the scope of this work to broad-scale transcriptomic and genomic patterns related to trait variation, we demonstrated that it is possible to integrate large-scale transcriptome data into an existing GPM, which can be exploited in the future for more fine-scale analysis like co-expression analysis and highlighting how specific trans-regulatory patterns within the cell network lead to trait variation. Thus, integrating genotype, transcriptome, and phenotype using scRNA-seq data can be particularly efficient for developing a better understanding of other important traits or diseases.

MATERIAL AND METHODS

Yeast strains and segregants

We analyzed cells from a single batch (batch 1) of 4,489 F2 segregants obtained from a F1 cross between the yeast laboratory strain BY4741 and the vineyard strain RM11-1a generated in a previous study (29). These strains have been selected to generate this collection of segregants because they exhibit differences in multiple phenotypes including the adaptation to temperature, the ability to process different sources of carbon and the ability to resist antifungal compounds. Therefore, the genetic variation observed across the segregants can be correlated to the differences in growth rate observed in the 18 environments recapitulating these phenotypes in the Nguyen et al (2022) study (29). The selection of the batch is random and the fact that we performed the analyses on a single batch eliminates batch effects that could obscure variable associations. Genotypes and fitness data used were the same ones obtained in the previous study.

Yeast growth and single-cell RNA-sequencing protocol

To prepare strains for scRNA sequencing, we unfroze the batch of segregants and inoculated approximately 5*10^6 cells in YPD (1% Yeast Extract, 2% Peptone, 2% Dextrose) to saturation. The next day, about 10^7 cells were passaged to 5 mL of fresh YPD and grew for 4 hours to bring cells to log-phase. We then pelleted 100 ul of cells and resuspended them in spheroplasting solution (5 mg/mL zymolyase 20T, 10 mM DTT, 1 M Sorbitol, 100 mM Sodium Phosphate pH 7.4) at a concentration of 10^7 cells/mL. The cells were incubated at 37 degrees Celcius for approximately 10 minutes at which point spheroplasting was verified by mixing a small aliquot of cells with detergent to observe lysis. The cells at this point were quantified using a hemocytometer and prepared using the standard 10x Genomics Gel Beads-in-emulsion (GEM) protocol. We used the Chromium Next GEM Single-cell 3' Reagent Kit to prepare the sequencing libraries and sequenced on a NextSeq 500 high-output flow cell.

We note that the cells analyzed here were grown in bulk and assayed for their transcriptome in log-phase. Our fitness data was obtained from competitive bulk fitness assays which include several whole growth cycles over multiple days and thus capture lag phase, exponential growth, and saturation. Nevertheless, previous experiments had shown that fitness was mostly determined by exponential growth which suggests that our analysis is adequate even if the cells were prepared for sequencing at a single time point.

Single-cell RNA-sequencing data parsing

From the scRNA-seq reads, we obtained gene expression levels and allele counts using the pipeline count from CellRanger version 3.1.0 (55). For each of the ancestral strains, i.e RM11-1a and BY4741, the pipeline mapped the scRNA-seq reads to the reference genome, filtered the barcodes by comparing the UMI count per barcode distribution to a background model of empty gel-bead in-emulsion, and counted the number of UMI per gene per barcode. The barcode filtering retained 18,233 barcodes. For each barcode, we then counted the number of RM and BY alleles at each polymorphic site by parsing the RM and BY bam files using a Python script (https://github.com/arnaud00013/sc-eQTL/tree/main/II_scRNA-seq_genotyping). This script only keeps reads that mapped at the same loci on both reference genomes to increase the level of confidence of the mapping.

Correction and imputation of single-cell genotypes with a Hidden Markov Model

Because there are only two possible alleles at each polymorphic sites of the F2 RM/BY segregants, their genotype can be recapitulated by a quantitative variable measuring the proportion of reads from one of the parental strains, which is RM in our dataset. The raw allele count data provides a first estimate of this RM allele frequency at each polymorphic site. However, due to the low mean depth of coverage of scRNA-seq data (0.2x), the absence of reads in some polymorphic sites and the biases introduced during sequencing like index hopping/swapping, we expect that the raw data can be imputed and corrected for errors and uncertainty in the observed alleles. Therefore, we applied a Hidden Markov Model (HMM) on the observed allele count. Such model can infer accurate genotype data at sequencing depths as low as 0.1x (29,31,56). Nguyen Ba and collaborators (2022) designed an HMM to infer the segregants genotypes from bulk DNA sequencing by accounting for sequencing error rate, recombination rate and index swapping rate (29). Because scRNA-seq uses the reverse transcriptase, which has a higher error rate, and because it is a pooled assay with higher chances of index swapping, we expected the HMM parameter to differ for the single cell data. Therefore, we adapted the HMM to scRNA-seq data by measuring its parameters in our dataset (**Figure S1**). The scripts are available on GitHub (https://github.com/arnaud00013/sc-eQTL/tree/main/II_scRNA-seq_genotyping).

Assigning single cells to the reference panel strains

To evaluate the level of relatedness between the reference panel strains and the imputed single cell genotypes, we used the expected distance to identify the strain that best relate to each single cell:

$$Expected\ distance(g_c, g_s) = \sum_{i=1}^{41594} g_c + g_s - 2g_c g_s \qquad \text{(Eq.1)}$$

where g_c is the cell genotype and g_s is the strain genotype. Next, we assigned the single cell to its best match in the studied batch of 4,489 trains only if this match is better than the best match in randomly generated batches of the same size (**Figure S2**). This procedure is implemented and available at https://github.com/arnaud00013/sc-eQTL/tree/main/III_Genotype_analysis).

Partitioning the phenotypic variance into genetic and transcriptomic components

To analyze the yeast GPM at a broad scale and to evaluate the association between selection and the transcriptome, we estimated the contribution of genetic and transcriptomic variations to phenotypic variation from scRNA-seq data. More precisely, we performed a Genome-wide Complex Trait Analysis (GCTA) by fitting a linear mixed model to the data using the restricted maximum-likelihood (REML) method (57):

$$y = X\beta + W_g u_g + \varepsilon_g \qquad \text{(Eq.2)}$$

$$y = X\beta + W_e u_e + \varepsilon_e \qquad \text{(Eq.3)}$$

$$y = X\beta + W_g u_g + W_e u_e + \varepsilon \quad \text{(Eq.4)}$$

where y is the fitness vector for the n segregants with a significant association to batch 1 cells, X is the nxk matrix of k fixed effects, β is the vector of k coefficients of the fixed effects, W_g is the nxp genotype matrix, u_g is the vector of p SNP effects, W_e is the nxm expression matrix, u_e is the vector of m gene expression effects and ε is the error term. To decrease the effect of gene dropouts and noise in the gene expression profile of each cell, we fitted an imputation model called DISCERN to the expression data (33). DISCERN is an autoencoder that learns how to reconstruct the expression profile of high-coverage cells after embedding it in a lower dimension. These operations eliminate the gene dropouts and denoises the expression profile. We also controlled for the effect of the cell cycle on gene expression by using the consensus F2 segregants expression and genotypes from the associated cell data. The consensus data has been obtained using the cell with the maximum read count. Because the dataset does not include fixed effects, we set the fixed effect to a vector of ones such that its coefficients represent the mean fitness while the genotype and expression data are the random effects that explain the fitness variance along with the error terms. The REML solution assumes that the data follow a Gaussian distribution, so the data are standardized before fitting the model. We also divided the standardized expression counts by the cell sum of expression counts to control for molecule count biases across cells. The cell fitness is based on the fitness of the closest segregant in batch 1 as measured by the expected distance. Because this model is linear and additive, it can be compared to the estimates of narrow-sense heritability obtained by Nguyen Ba and collaborators (2022) (29). The difference between the variance explained in equation 4 and equations 2 or 3 allow to infer the variance explained only by the genotype or the expression component of the model. The code for the variance partitioning is available on GitHub (https://github.com/arnaud00013/sc-eQTL/tree/main/IV_variance_partitioning).

Estimating the expression heritability from scRNA-seq

To obtain this estimate from scRNA-seq data, we considered the fact that GCTA-REML only takes a vector as a response variable while the gene expression matrix is multi-dimensional. To solve this, we orthogonalized the gene expression matrix using principal component analysis (PCA), and used its first PC, which explains 99.6% of the variance, as a response variable of the model. In case 99% of the expression matrix is explained by more than one PC, it is possible to use a weighted sum to obtain the expression heritability. Indeed, if the expression PCs recapitulate the total expression variance and are orthogonal or independent to each other, then the sum of the PCs variance explained by genotype should be the expression heritability. If "k” is the number of PCs explaining at least 99% of expression variance:

$$Expression\ heritability = \sum_{i=1}^{k} PC_i\ eigen\ value * \text{"}PC_i \sim genotype\text{"}\ model\ R^2 \quad \text{(Eq.5)}$$

QTL mapping

To identify the loci that influence cell fitness, we performed a linear regression on the consensus genotypes of the strains from the scRNA-seq data and the strain fitness. We decided to use the consensus genotypes of the strains as they relate better to the bulk segregant genomes. To build the consensus genotypes, we defined cells from the same lineage as the ones that shared the same closest F2 segregant in batch 1. Next, we used the median to obtain cells' consensus genotypes as it is less sensitive to outliers and because it yields the best relatedness to the batch 1 reference genotypes (median $R^2 = 87.0\%$; μ=79.5%; σ=18.2; **Figure S3**). We selected the QTL in the linear models using cross-validation on the scRNA-seq data. This analysis consists in dividing the dataset into 10 random partitions of similar sample sizes and running a cross-validated stepwise forward linear regression on each partition. For each partition, the model starts with no QTL and a linear model "Fitness ~ Genotype" is fitted using the genotype data at each polymorphic site, where the correlation coefficient represents the effect size of the SNP. Then, the forward search starts and at each iteration, a new locus with the minimum linear model residual sum of squares (RSS) is added to the QTL model, which is updated with new effect sizes after the addition of a new SNP. Because the order of addition of QTL matters in the forward search and because some QTL are linked or collinear, the model can be refined by exploring different QTL around the local optima. These steps are repeated until the model RSS cannot be improved anymore or until the number of QTL reaches an arbitrary maximum far from the cross-validated number of QTL. After the forward search is completed in each partition, the algorithm calculates the optimal λ values that minimizes the objective function F_o:

$$F_o(\beta) = RSS(\beta) + Lasso\ penalty(\beta)$$

$$\|Y - X\beta\|_2^2 + \lambda\|\beta\|_0 \text{ (Eq.6)}$$

where β is the vector of SNP effect sizes in the QTL model, $\|Y - X\beta\|_2^2$ is the RSS of the linear QTL model, λ defines the penalty for adding a new SNP to the model and $\|\beta\|_0$ is the number of SNPs in the QTL model. This objective function has the property to add sparsity in the QTL model and thus avoid overestimating the number of QTL while being consistent (29). The optimal λ has a minimum of log(n) which corresponds to the Bayesian Information Criterion (BIC), which is known to yield correct models asymptotically (58). This allows to consider the possibility that a sparser model than the one found using the BIC could yield better predictive power on a test set while avoiding overfitting. The optimal λ values found in all the partitions are then averaged and the resulting mean λ is used to solve the objective function in the full dataset, which yields the optimal QTL model. The cross-validation assumes that the partitions are independent, such that the variance explained by the model and the number of relevant QTL are unbiased estimates.

Highlighting hotspots of gene regulation through eQTL mapping

To identify the loci regulating gene expression regulation, we adapted the QTL mapping framework using expression as the predicted phenotype. Because this approach had to be repeated for each of the 6,240 genes, we needed to modify it so that the execution time was convenient. To do so, the parameter λ was not estimated using cross-validation but rather from the Bayesian Inference Criterion (BIC), i.e. $\lambda = \log(n)$ where n is the number of cells. We found that the BIC was often selected by the cross-validation procedure when tested on a few genes and thus we do not believe that this approach will significantly change our results.

To acknowledge the uncertainty around the exact position of eQTL due to linkage disequilibrium, we define eQTL hotspots as 25 kb genomic windows that were repeatedly identified in the eQTL mapping procedure. Finally, we defined hotspots of gene regulation as genomic windows in the fourth quantile of the distribution of the number of regulated genes per window and the fourth quantile of the number of eQTL per window. The code for the single cell eQTL mapping is available on GitHub (https://github.com/arnaud00013/sc-eQTL/tree/main/V_sc_eQTL_mapping).

Functional enrichment analysis by gene ontology annotation

To highlight gene functions enriched at different levels of expression or expression heritability, we performed the panther database binomial test for statistical overrepresentation of gene ontology biological processes (39,59). A low level was defined as within the 25% bottom part of the distribution (<Q1) while a high level was defined as within the top 25% part of the distribution (>Q3). The *p*-values were corrected for multiple testing using the false discovery rate correction (FDR).

Matching QTL to eQTL

To evaluate the contribution of gene expression regulation to fitness variation, we created a model to match QTL and eQTL based on the similarity of loci and the similarity of predicted effect on gene expression. More precisely, for each of the 6,088 genes for which we could detect eQTL, we performed a new eQTL model by correlating the expression level of the gene to the genetic variation at QTL positions. This allowed us to measure the predicted effect of the QTL on gene expression. We then calculated the distance between the QTL and the real eQTL of the gene based on recombination distance within each chromosome, which decreases exponentially with genetic distance, and the difference in the predicted effect on the gene expression using the formulation developed by Nguyen Ba et al (2022) (29). Next, we used the same Needleman-Wunsch algorithm to find the most likely set of pairing between QTL and eQTL, where an unmatched QTL is also possible but penalized. Finally, we determined the proportion of genes for which gene expression regulation is associated with higher fitness. To do so, for each gene, we performed a permutation test by comparing the average rank of the matched QTL of the gene to the average rank of 999 random subsets of unmatched QTL of the same size. The p-value is the proportion of random subsets of unmatched QTL with a higher average QTL rank than the set of matched QTL.

Comparing cis- and trans-eQTL contribution to expression variation

We used the definition of local eQTL in Albert et al. (2018) to define cis-eQTL, i.e. any eQTL between 1,000 bp upstream of the gene and 200 bp downstream of the gene. Thus, we defined trans-eQTL as the eQTL that do not follow this criterion. For each gene, we then performed variance partitioning using the GCTA:

$$y = X\beta + W_{g\_cis}u_{g\_cis} + \varepsilon_{cis} \quad \text{(Eq.7)}$$

$$y = X\beta + W_{g\_trans}u_{g\_trans} + \varepsilon_{trans} \quad \text{(Eq.8)}$$

$$y = X\beta + W_{g\_cis}u_{g\_cis} + W_{g\_trans}u_{g\_trans} + \varepsilon \qquad \text{(Eq.9)}$$

where y is the vector of expression level of the gene across the n cells, X is the nxk matrix of k fixed effects, β is the vector of k coefficients of the fixed effects, W_{g_cis} is the nxp cis-eQTL genotype matrix, u_{g_cis} is the vector of p cis-eQTL effects on expression, W_{g_trans} is the nxm trans-eQTL expression matrix, u_e is the vector of m trans-eQTL effects on expression and ε represent the error terms. Because the dataset does not include fixed effects, we set the fixed effect to a vector of ones such that its coefficients represent the mean expression level while the cis-eQTL and trans-eQTL genotypes are the random effects that explain the expression variance along with the error terms. We can infer the variance explained by the cis-eQTL by the difference in variance explained between the models in equations 9 and 8. Likewise, the difference of variance explained by the models in equations 9 and 7 can help us estimate the variance explained by the trans-eQTL. Finally, we estimate the effect sizes using the absolute value of the correlation coefficients of each loci and compare the mean between the cis- and trans-eQTL from the same gene (paired data) with a Wilcoxon signed rank test.

DATA AVAILABILITY

The code used for this study is available and explained at https://github.com/arnaud00013/sc-eQTL and the original single-cell reads from the pooled segregants scRNA-seq assay have been uploaded in the NCBI BioProject database with the accession number PRJNA1022775. The single-cell barcodes expression data are also available at https://github.com/arnaud00013/sc-eQTL as an archive file named Matrix_gene_expression_barcodes_1_to_9000.csv.tar.gz or Matrix_gene_expression_barcodes_9001_to_18233.csv.tar.gz.

ACKNOWLEDGEMENTS

ANNB acknowledges support from the Natural Science and Engineering Research Council of Canada (NSERC RGPIN-2021–02716 and DGECR-2021–00117) and AN acknowledges support from the Natural Science and Engineering Research Council (NSERC CGS-D App Id: 569340-2022, UTF Fellowship from the University of Toronto and Rustom H. Dasture Graduate Scholarship in Cell and Systems Biology). The computations in this paper were enabled by resources provided by Compute Canada. We also thank the Centre for the Analysis of Genome Evolution and Function (CAGEF) and TCAG at SickKids for sequencing support.

REFERENCES

1. Bartoli C, Roux F. Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. Front Plant Sci [Internet]. 2017 May 23;8. Available from: ://WOS:000402030200001

2. Ferreira MA, Gamazon ER, Al-Ejeh F, Aittomaki K, Andrulis IL, Anton-Culver H, et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat Commun [Internet]. 2019 Apr 15;10. Available from: ://WOS:000464494100010

3. Aguet F, Alasoo K, Li YI, Battle A, Im HK, Montgomery SB, et al. Molecular quantitative trait loci. Nat Rev Methods Primer. 2023 Jan 25;3(1):1–22.

4. Tarantino LM, Eisener-Dorman AF. Forward Genetic Approaches to Understanding Complex Behaviors. Curr Top Behav Neurosci. 2012;12:25–58.

5. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int J Mol Sci. 2017 Jul 29;18(8):1652.

6. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019 Apr;51(4):592–9.

7. Transcriptome: Connecting the Genome to Gene Function | Learn Science at Scitable [Internet]. [cited 2023 Aug 31]. Available from: https://www.nature.com/scitable/topicpage/transcriptome-connecting-the-genome-to-gene-function-605/

8. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome Med. 2022 Jun 27;14(1):68.

9. King MC, Wilson AC. Evolution at Two Levels in Humans and Chimpanzees. Science. 1975 Apr 11;188(4184):107–16.

10. Jacob F. Evolution and Tinkering. Science. 1977 Jun 10;196(4295):1161–6.

11. Hoekstra HE, Coyne JA. The Locus of Evolution: Evo Devo and the Genetics of Adaptation. Evolution. 2007;61(5):995–1016.

12. Kratochwil CF, Meyer A. Evolution: Tinkering within Gene Regulatory Landscapes. Curr Biol. 2015 Mar 30;25(7):R285–8.

13. Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, et al. The core meiotic transcriptome in budding yeasts. Nat Genet. 2000 Dec;26(4):415–23.

14. Cavalieri D, Townsend JP, Hartl DL. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc Natl Acad Sci U S A. 2000 Oct 24;97(22):12369–74.

15. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012 Sep 7;337(6099):1190–5.

16. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012 Sep;22(9):1748–59.

17. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet. 2010 Apr 1;6(4):e1000888.

18. Dutta D, He Y, Saha A, Arvanitis M, Battle A, Chatterjee N. Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood. Nat Commun. 2022 Jul 26;13(1):4323.

19. Wang L, Babushkin N, Liu Z, Liu X. Trans-eQTL mapping in gene sets identifies network effects of genetic variants. Cell Genomics. 2024 Apr 1;4(4):100538.

20. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science [Internet]. 2002 Apr 26 [cited 2022 Apr 29]; Available from: https://www.science.org/doi/pdf/10.1126/science.1069516

21. Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L. Genetics of trans-regulatory variation in gene expression. Wittkopp PJ, editor. eLife. 2018 Jul 17;7:e35471.

22. Schwarz T, Boltz T, Hou K, Bot M, Duan C, Loohuis LO, et al. Powerful eQTL mapping through low-coverage RNA sequencing. Hum Genet Genomics Adv. 2022 Jul 14;3(3):100103.

23. Fan Y, Zhu H, Song Y, Peng Q, Zhou X. Efficient and effective control of confounding in eQTL mapping studies through joint differential expression and Mendelian randomization analyses. Bioinformatics. 2020 Aug 13;37(3):296–302.

24. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. Plos Comput Biol [Internet]. 2012 Dec;8(12). Available from: ://WOS:000312901500028

25. Lorenz K, Cohen BA. Small- and Large-Effect Quantitative Trait Locus Interactions Underlie Variation in Yeast Sporulation Efficiency. Genetics. 2012 Nov;192(3):1123-+.

26. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug- induced reprogramming as a mode of cancer drug resistance. Nature. 2017 Jun 15;546(7658):431-+.

27. Bloom JS, Ehrenreich IM, Loo WT, Lite TLV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. 2013 Feb;494(7436):234–7.

28. Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, et al. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature. 2010 Apr;464(7291):1039–42.

29. Nguyen Ba AN, Lawrence KR, Rego-Costa A, Gopalakrishnan S, Temko D, Michor F, et al. Barcoded Bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. Verstrepen KJ, editor. eLife. 2022 février;11:e73983.

30. Vermeersch L, Jariani A, Helsen J, Heineike BM, Verstrepen KJ. Single-Cell RNA Sequencing in YeastYeast Using the 10× Genomics Chromium Device. In: Devaux F, editor. Yeast Functional

Genomics: Methods and Protocols [Internet]. New York, NY: Springer US; 2022 [cited 2024 Jun 16]. p. 3–20. Available from: https://doi.org/10.1007/978-1-0716-2257-5_1

31. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med [Internet]. 2018 Aug 7;50. Available from: ://WOS:000441266700002

32. Jariani A, Vermeersch L, Cerulus B, Perez-Samper G, Voordeckers K, Van Brussel T, et al. A new protocol for single-cell RNA-seq reveals stochastic gene expression during lag phase in budding yeast. Rokas A, Barkai N, editors. eLife. 2020 May 18;9:e55320.

33. Hausmann F, Ergen C, Khatri R, Marouf M, Hänzelmann S, Gagliani N, et al. DISCERN: deep single-cell expression reconstruction for improved cell clustering and cell subtype and state detection. Genome Biol. 2023 Sep 20;24(1):212.

34. Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, et al. A High-Definition View of Functional Genetic Variation from Natural Yeast Genomes. Mol Biol Evol. 2014 Apr 1;31(4):872–88.

35. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017/10/19 ed. 2017 Nov 2;551(7678):45–50.

36. Johnson MS, Gopalakrishnan S, Goyal J, Dillingham ME, Bakerlee CW, Humphrey PT, et al. Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. Verstrepen KJ, Wittkopp PJ, Verstrepen KJ, Hodgins-Davis A, editors. eLife. 2021 Jan 19;10:e63910.

37. Sun XM, Bowman A, Priestman M, Bertaux F, Martinez-Segura A, Tang W, et al. Size-Dependent Increase in RNA Polymerase II Initiation Rates Mediates Gene Expression Scaling with Cell Size. Curr Biol. 2020 Apr 6;30(7):1217-1230.e7.

38. Marguerat S, Bähler J. Coordinating genome expression with cell size. Trends Genet. 2012 Nov 1;28(11):560–5.

39. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019 Mar;14(3):703–21.

40. Bähler J. Cell-cycle control of gene expression in budding and fission yeast. Annu Rev Genet. 2005;39:69–94.

41. Yu L, Castillo LP, Mnaimneh S, Hughes TR, Brown GW. A Survey of Essential Gene Function in the Yeast Cell Division Cycle. Mol Biol Cell. 2006 Nov;17(11):4736–47.

42. Roos J, Luz JM, Centoducati S, Sternglanz R, Lennarz WJ. ENP1, an essential gene encoding a nuclear protein that is highly conserved from yeast to humans. Gene. 1997 Jan 31;185(1):137–46.

43. Chen W, Bucaria J, Band DA, Sutton A, Sternglanz R. Enp1, a yeast protein associated with U3 and U14 snoRNAs, is required for pre-rRNA processing and 40S subunit synthesis. Nucleic Acids Res. 2003 Jan 15;31(2):690–9.

44. Roy A, Dement AD, Cho KH, Kim JH. Assessing Glucose Uptake through the Yeast Hexose Transporter 1 (Hxt1). PLOS ONE. 2015 Mar 27;10(3):e0121985.

45. Lu L, Roberts GG, Oszust C, Hudson AP. The YJR127C/ZMS1 gene product is involved in glycerol-based respiratory growth of the yeast Saccharomyces cerevisiae. Curr Genet. 2005 Oct 1;48(4):235–46.

46. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature. 2005 Aug;436(7051):701–3.

47. Vande Zande P, Hill MS, Wittkopp PJ. Pleiotropic effects of trans-regulatory mutations on fitness and gene expression. Science. 2022 Jul;377(6601):105–9.

48. Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, et al. Cis and trans effects differentially contribute to the evolution of promoters and enhancers. Genome Biol. 2020 Aug 20;21(1):210.

49. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012 Jun 18;13(7):505–16.

50. Bloom JS, Boocock J, Treusch S, Sadhu MJ, Day L, Oates-Barker H, et al. Rare variants contribute disproportionately to quantitative trait variation in yeast. Landry CR, Barkai N, editors. eLife. 2019 Oct 24;8:e49212.

51. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans-eQTLs as putative drivers of known disease associations. Nat Genet. 2013 Oct;45(10):1238–43.

52. Yao C, Joehanes R, Johnson AD, Huan T, Liu C, Freedman JE, et al. Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. Am J Hum Genet. 2017 Apr 6;100(4):571–80.

53. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019 May 2;177(4):1022-1034.e6.

54. Brauer MJ, Christianson CM, Pai DA, Dunham MJ. Mapping novel traits by array-assisted bulk segregant analysis in Saccharomyces cerevisiae. Genetics. 2006 Jul;173(3):1813–6.

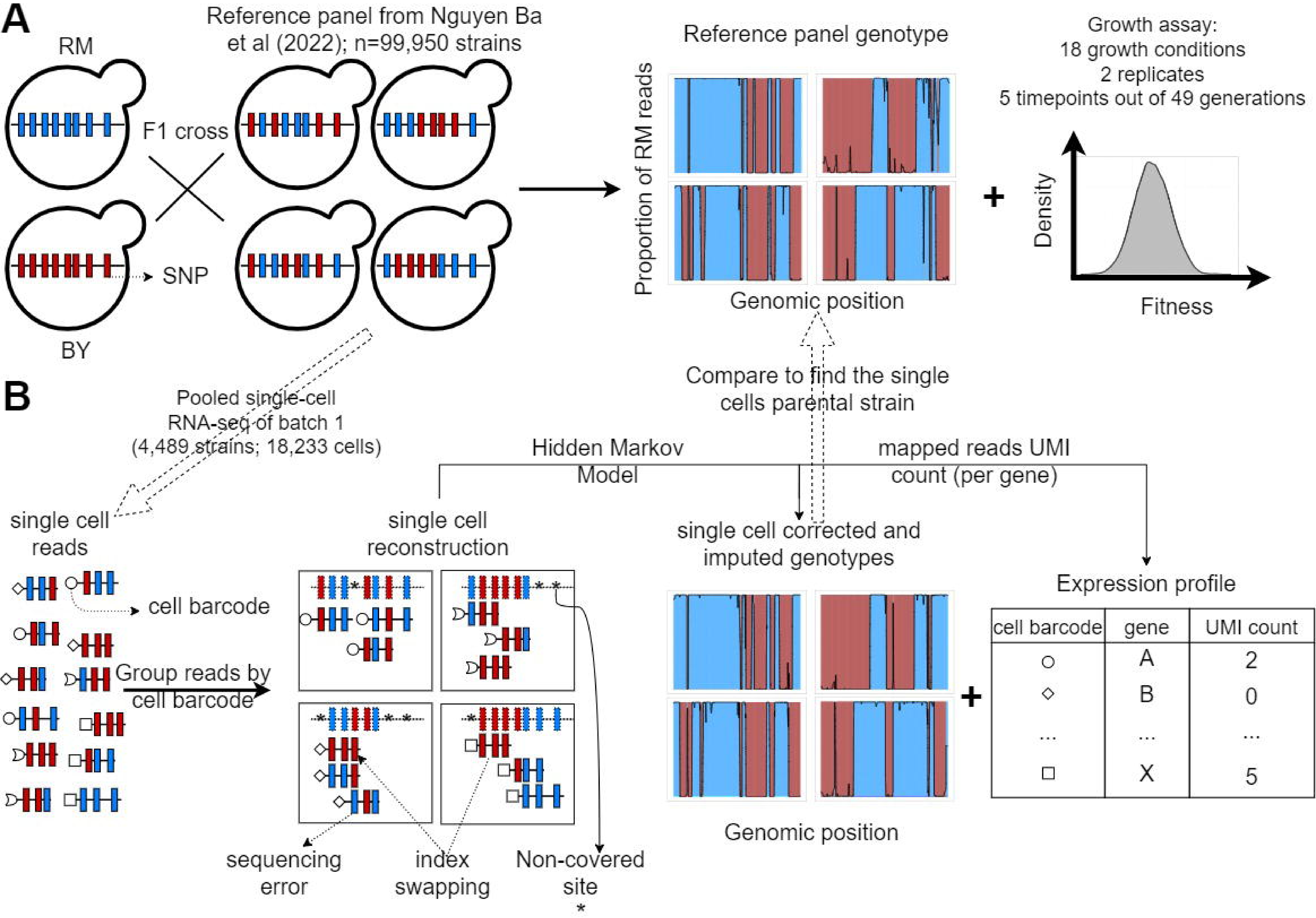
55. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017/01/17 ed. 2017 Jan 16;8:14049.

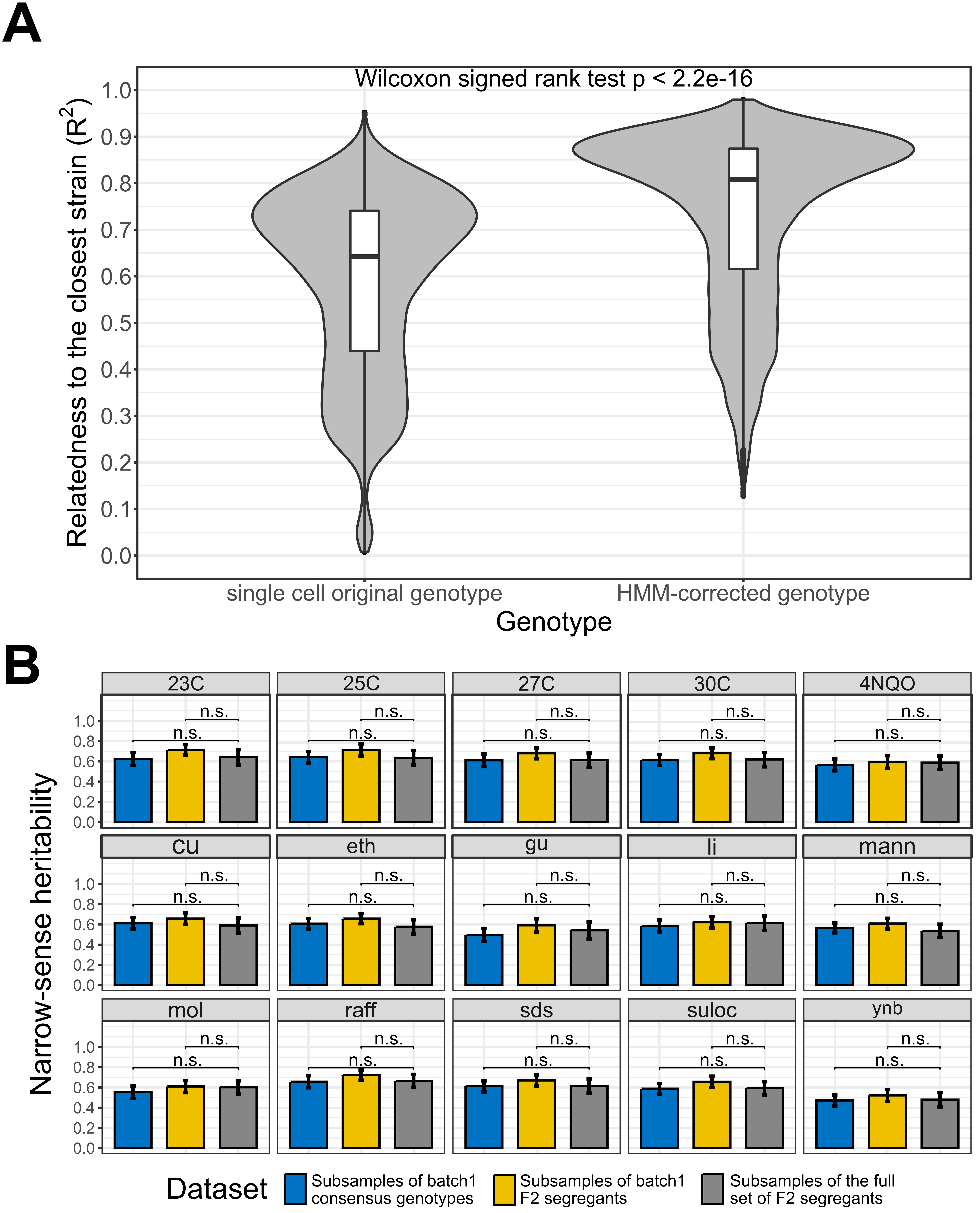
56. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019 Nov;20(11):631–56.

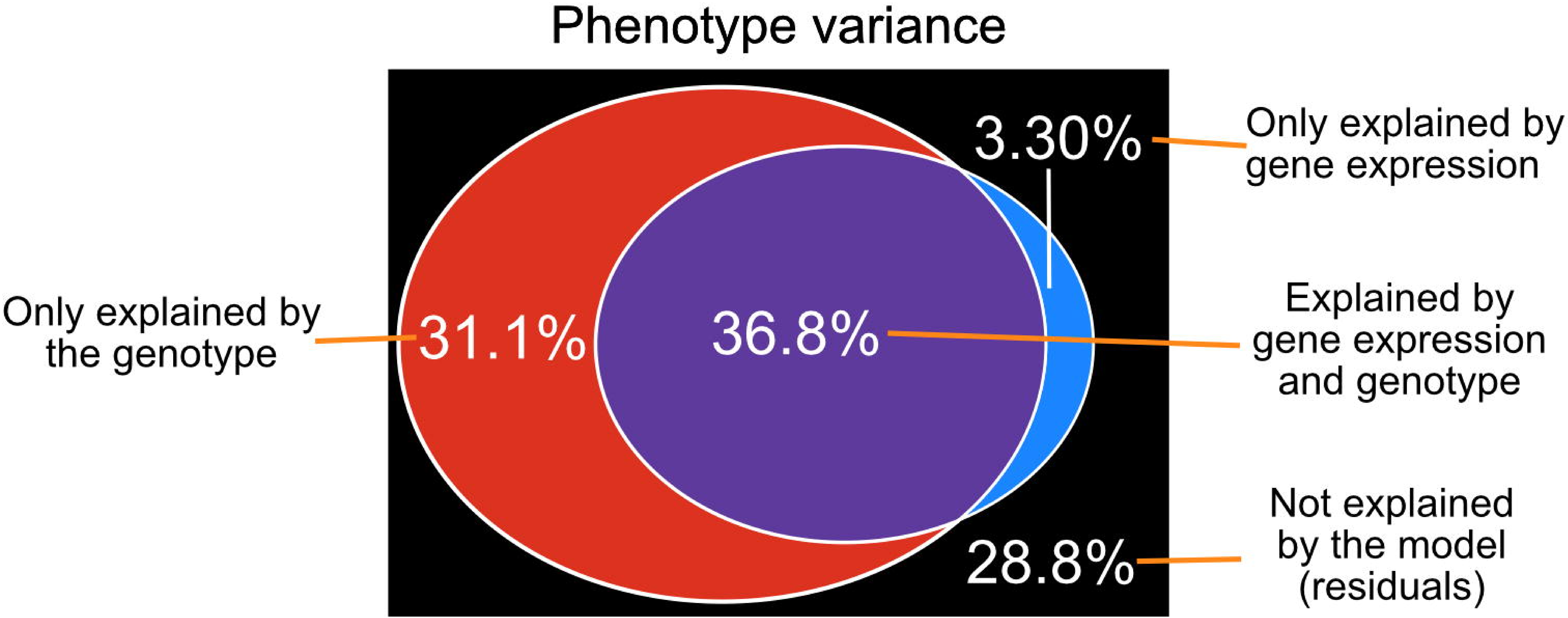
57. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2010/12/21 ed. 2011 Jan 7;88(1):76–82.

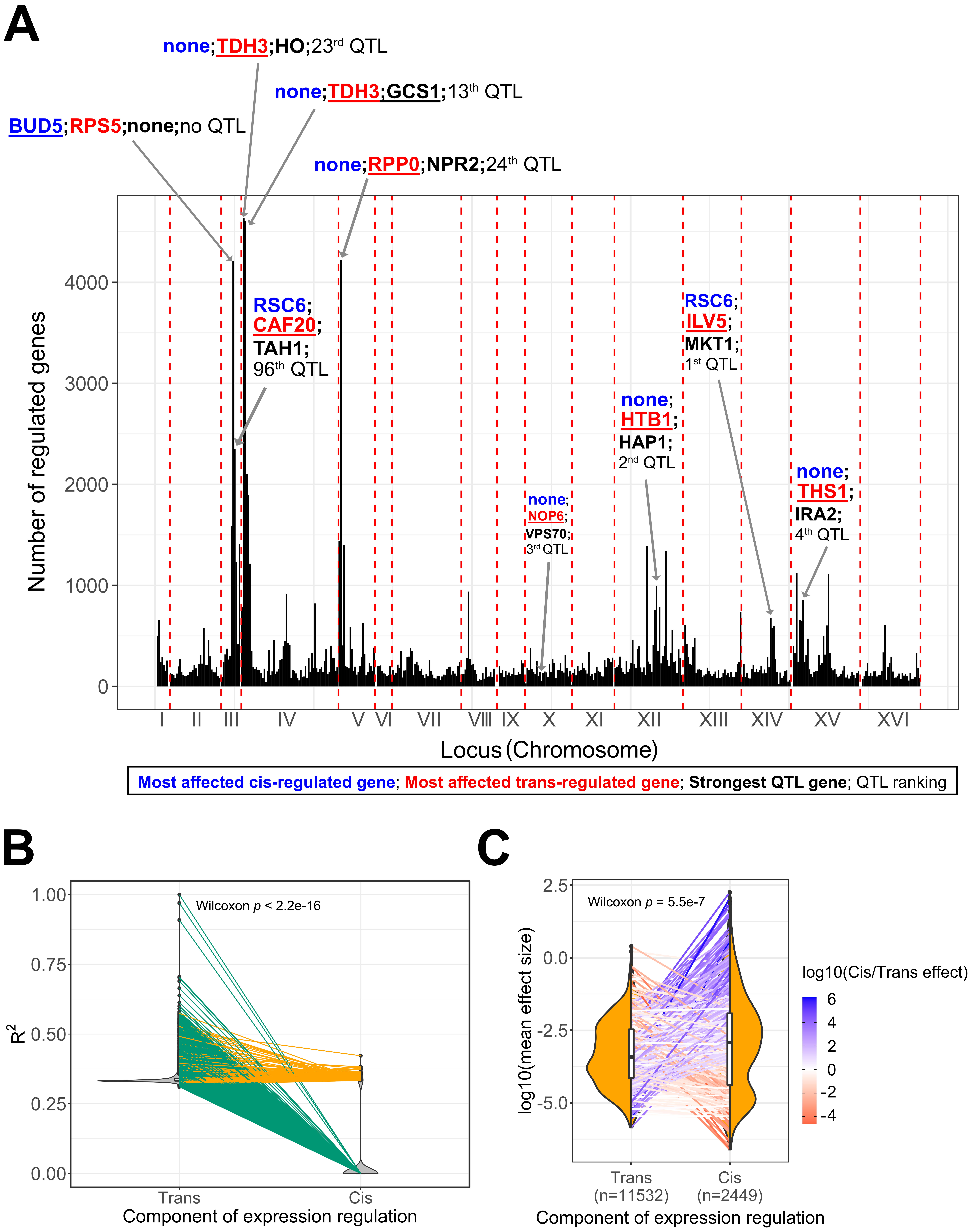
58. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis [Internet]. New York, NY: Springer; 2001 [cited 2023 Mar 29]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-1-4757-3462-1

59. PANTHER: Making genome-scale phylogenetics accessible to all [Internet]. [cited 2023 Sep 29]. Available from: https://onlinelibrary.wiley.com/doi/epdf/10.1002/pro.4218


A
RM
BY
F1 cross
SNP
Reference panel from Nguyen Ba et al (2022); n=99,950 strains
Reference panel genotype
Proportion of RM reads
Genomic position
+
Growth assay:
18 growth conditions
2 replicates
5 timepoints out of 49 generations
Density
Fitness
B
Pooled single-cell RNA-seq of batch 1 (4,489 strains; 18,233 cells)
Compare to find the single cells parental strain
Hidden Markov Model
mapped reads UMI count (per gene)
single cell reads
cell barcode
Group reads by cell barcode
single cell reconstruction
sequencing error
index swapping
Non-covered site
*
single cell corrected and imputed genotypes
Genomic position
+
Expression profile
cell barcode
gene
UMI count
○
A
2
◇
B
0
...
...
...
□
X
5


A
Wilcoxon signed rank test p < 2.2e-16
Relatedness to the closest strain (R2)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
single cell original genotype
HMM-corrected genotype
Genotype
B
Narrow-sense heritability
23C
25C
27C
30C
4NQO
cu
eth
gu
li
mann
mol
raff
sds
suloc
ynb
n.s.
1.0
0.8
0.6
0.4
0.2
0.0
Dataset
Subsamples of batch1 consensus genotypes
Subsamples of batch1 F2 segregants
Subsamples of the full set of F2 segregants


Phenotype variance
3.30%
Only explained by gene expression
Only explained by the genotype
31.1%
36.8%
Explained by gene expression and genotype
28.8%
Not explained by the model (residuals)


A
none;TDH3;HO;23rd QTL
none;TDH3;GCS1;13th QTL
BUD5;RPS5;none;no QTL
none;RPP0;NPR2;24th QTL
RSC6;
CAF20;
TAH1;
96th QTL
RSC6;
ILV5;
MKT1;
1st QTL
none;
HTB1;
HAP1;
2nd QTL
none;
THS1;
IRA2;
4th QTL
none;
NOP6;
VPS70;
3rd QTL
Number of regulated genes
4000
3000
2000
1000
0
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
Locus (Chromosome)
Most affected cis-regulated gene; Most affected trans-regulated gene; Strongest QTL gene; QTL ranking
B
Wilcoxon p < 2.2e-16
R2
1.00
0.75
0.50
0.25
0.00
Trans
Cis
Component of expression regulation
C
Wilcoxon p = 5.5e-7
log10(mean effect size)
2.5
0.0
-2.5
-5.0
log10(Cis/Trans effect)
6
4
2
0
-2
-4
Trans (n=11532)
Cis (n=2449)
Component of expression regulation